

Ed P. van Duin en Aldi J.M. Hagenaars**

Samenvatting

Indien bij empirisch onderzoek gebruik wordt gemaakt van een steekproef is het van groot belang dat deze representatief is voor de bestudeerde populatie. In dit rapport wordt onderzocht hoe het gesteld is met de representativiteit van de steekproef die resulteerde uit de zgn. GPD-enquête, een enquête via een aantal regionale dagbladen. Voornaamste kenmerk van de steekproef is dat er geen respondenten in voorkomen uit grote(re) steden. Nagegaan wordt of dit gevolgen heeft voor de uitkomsten van op basis van de steekproef geschatte relaties, en of er gecorrigeerd kan worden voor genoemde onderdekking. Een en ander wordt geïllustreerd aan de hand van een tweetal relaties, t.w. een inkomenswaardings-vergelijking en een inkomens-vergelijking. Een aantal correctiemethoden wordt onderzocht op bruikbaarheid: 1) de methode van Heckman, 2) herweging naar een aantal achtergrondvariabelen, en 3) gewogen regressie. De uitkomsten op basis van de GPD-steekproef worden vergeleken met de uitkomsten op basis van een aantal andere steekproeven (ijkpunten).

Inleiding

Een enquête is een veel gebruikte methode om informatie te verkrijgen over een te onderzoeken populatie. Bij een onderzoek in de vorm van een enquête is het om praktische redenen onmogelijk de totale populatie mee te laten doen. De uitvoering, het verkrijgen en verwerken van het data-materiaal, worden een te kostbare en tijdrovende zaak. Men is daarom aangewezen op een steekproefonderzoek. Op grond van gegevens omtrent een gedeelte van de populatie worden dan uitspraken gedaan over de totale populatie. Van essentieel belang is dat de steekproef representatief is voor de doelgroep, dat wil zeggen het deel van de

* Het onderzoek waarover in dit rapport verslag wordt gedaan is onderdeel van het Inkomenswaardings Project. De auteurs danken Eitel Homan, Bernard van Praag, Klaas de Vos en Hans van Weeren voor hun commentaar op een eerdere versie van dit rapport. De eindverantwoordelijkheid voor de inhoud berust bij de auteurs.

** Erasmus Universiteit Rotterdam, Postbus 1738, 3000 DR Rotterdam, tel. 010-4081438.

populatie dat in de steekproef is opgenomen moet een juiste afspiegeling van de totale populatie zijn.

Op zaterdag 10 september 1983 hebben 10 regionale dagbladen (de Arnhemse Courant, het Brabants Nieuwsblad, de Gooi- en Eemlander, het Haarlems Dagblad, het Limburgs Dagblad, de Nieuwe Noord-Hollandse Courant, de Provinciale Zeeuwse Courant, Tubantia, de Typhoon en het Utrechts Nieuwsblad) alsmede de daarmee geassocieerde plaatselijke uitgaven (zoals de IJmuidense Courant, de Graafschapbode etc.), met in totaal een oplage van meer dan 700.000, in hun editie een grote vragenlijst afgedrukt, die lezers na invulling ongefrankeerd en anoniem naar het blad konden terugzenden. Deze onderneming is uitgevoerd door de Gemeenschappelijke Pers Dienst (voortaan aangeduid met GPD), waartoe al deze kranten behoren, in samenwerking met prof.dr B.M.S. van Praag en mw. dr A.J.M. Hagenaars, destijds beiden verbonden aan het Centrum voor Onderzoek van de Economie van de Publieke Sector (COEPS) van de Rijksuniversiteit Leiden.

De enquête bevatte een groot aantal vragen en maar liefst 24.000 ingevulde formulieren werden teruggestuurd. Deze enquête is uniek in zijn soort, niet alleen door de combinatie van diepgang en respons maar ook door de methode waarop de enquête gehouden is. Een enquête gehouden volgens de dagblad-methode, zoals de GPD-enquête, heeft als belangrijkste nadeel dat de steekproef niet te sturen is. Met betrekking tot de representativiteit spelen twee aspecten een rol, de zogenaamde dekkingsfout en de non-respons. De dekkingsfout ontstaat doordat het steekproefkader beperkt blijft tot lezers van de krant, waarin de enquête is opgenomen. Mensen die geen krant lezen en mensen die in gebieden wonen waar deze krant niet bezorgd wordt zijn bij voorbaat van deelname aan de enquête uitgesloten. Deze personen behoren echter wel tot de doelgroep. Niet alle elementen van de doelgroep zijn dus bevat in het steekproefkader. Dit probleem wordt wel aangeduid met onderdekking. Het maakt de enquête wellicht niet representatief voor de totale populatie.

De totale non-respons kan bij een dagblad-enquête hoger uitvallen dan bij een enquête gehouden op de gebruikelijke methode van ondervragen door een enquêteur/trice. Door de persoonlijke benadering van een enquêteur/trice is men eerder geneigd aan de enquête mee te doen.

Bovendien is door een ronde van aanvullende trekkingen de eventuele selectieve non-respons van bepaalde huishoudens bij dit soort enquêtes te verlagen en dus de representativiteit te verhogen. Bij een dagblad-enquête moet slechts afge wacht worden wie er zal reageren en kan geen aanvullende ronde gehouden worden. Dit kan de representativiteit nadelig beïnvloeden.

Ten aanzien van de onderdekking ligt de nadruk in dit rapport vooral op de

eventuele fout veroorzaakt door het niet-verspreidingsgebied van de GPD-bladen. De niet-kranten-lezers zijn wellicht minder belangrijk omdat vele gezinnen de krant delen met burens of kennissen, zo blijkt uit een onderzoek onder ex-abonnees [stops-onderzoek 1983], evenals uit een aantal bij de enquête gevoegde brieven. Wel zal bij de beoordeling van de uiteindelijke representativiteit dit aspect uiteraard mede aan de orde komen. Wat betreft de onderdekking door het beperkte verspreidingsgebied van de GPD-bladen hebben wij ons geconcentreerd op het afwezig zijn van grote steden. Wij willen onderzoeken in hoeverre bepaalde economische relaties worden verstoord door onderdekking met betrekking tot de grote steden. Voor dit onderzoek is informatie nodig over mensen die in het verspreidingsgebied van de GPD-bladen wonen als ook over mensen die er niet wonen. Hiervoor moet dus een ander dan het GPD-bestand gebruikt worden. Dit 'ijk'-bestand moet aan verschillende eisen voldoen.

Het bestand moet:

- groot zijn (veel gezinnen bevatten)
- recent zijn
- achtergrondvariabelen bevatten die ook in het GPD-bestand voorkomen
- de woonplaats van de respondent moet bekend zijn.

Geen enkel bij het COEPS beschikbaar bestand voldoet aan al deze eisen. In aanmerking komen de data-bestanden afkomstig van het PANEL-onderzoek uit 1983, dat wel recent is maar weinig waarnemingen bevat, en het Consumenten Conjunctuur Onderzoek (CCO) uit 1979: groot maar niet recent. Beide steekproeven zijn getrokken door het Centraal Bureau voor de Statistiek (CBS). Met behulp van het laatste bestand passen we een door Heckman (1979) ontwikkelde methode toe om te corrigeren voor onderdekking.

De non-respons is een probleem voor elke methode van enquêteren. Door non-respons is het mogelijk dat bepaalde groepen van mensen met dezelfde karakteristieken onder- of oververtegenwoordigd zijn in de steekproef. Een veel gebruikte methode om dit probleem op te lossen is herweging. De marginale frequentieverdeling in de steekproef van bepaalde gekozen achtergrondvariabelen kan, door aan iedere respondent een gewicht toe te kennen, gelijk gemaakt worden aan de marginale frequentieverdeling in de totale populatie. Voor deze procedure is wel vereist dat van elk der ondervertegenwoordigde groepen toch een aantal respondenten in de steekproef is opgenomen. Bij de gebruikelijke methode van enquêteren kan, als aan deze eis niet is voldaan, een volgende ronde van aanvullende trekkingen verricht worden. Bij de dagblad-methode kan dit niet. We verwachten echter niet dat dit in het GPD-bestand het geval is gezien het grote aantal respondenten.

De inhoud van dit rapport is als volgt. Hoofdstuk 1 gaat in op de representativiteit. De fouten die kunnen ontstaan bij onderzoek gebaseerd op steekproeven en het verschil van de dagblad-methode met de gebruikelijke methode van steekproef trekken worden hier besproken. In hoofdstuk 2 wordt de invloed van de onderdekking van het steekproef-ontwerp van de GPD-enquête op regressievergelijkingen onderzocht. De methode van Heckman wordt uiteengezet en getoetst op het CCO-bestand. Met het hierbij verkregen resultaat wordt de mogelijke toepassing van de methode op het GPD-bestand bepaald. De RAS-methode, een procedure om de gewichten per respondent te berekenen, wordt in hoofdstuk 3 behandeld. Een aantal verschillende data-bestanden wordt door middel van frequentie-berekeningen met elkaar vergeleken, zowel ongewogen als gewogen. Onderzocht wordt ook of de gewichten het probleem van onderdekking van de grote steden kunnen oplossen. In hoofdstuk 4 wordt een samenvatting van de belangrijkste resultaten gegeven.

1. Representativiteit

1.1 Inleiding

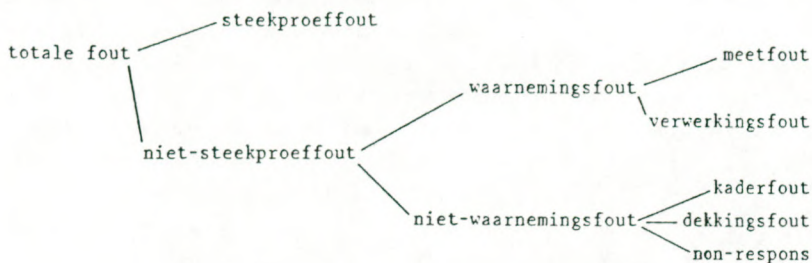
De enquête van de GPD is volgens de zogenaamde dagblad-methode gehouden. Deze methode is voor een uitgebreid onderzoek onder de Nederlandse bevolking nog nooit eerder toegepast. De informatie over de Nederlandse bevolking die tegenwoordig beschikbaar is is voornamelijk afkomstig van enquêtes die gehouden zijn door inschakeling van enquêteurs/trices.

Het enquêteren via een dagblad verschilt op een aantal punten nogal met de gebruikelijke methode van persoonlijke benadering toegepast door de opiniebureau's en het CBS. We willen nu onderzoeken of het data-bestand dat met de dagblad-methode verkregen is representatief is voor de populatie.

Bij het typeren van de mogelijke fouten maken we gebruik van een indeling zoals gegeven door Kersten en Bethlehem (1982). Het schema staat in figuur 1. Indien hierbij over 'fout' gesproken wordt is er niet altijd sprake van een foutieve handeling, hiermee wordt ook 'onzekerheid' bedoeld.

De totale fout is opgesplitst in de steekproeffout en de niet-steekproeffout. De steekproeffout ontstaat doordat men met gegevens van een deel van de populatie werkt en niet met de gegevens van de totale populatie. Dit soort fout is dus onafhankelijk van de steekproef-methode. Het deel van de fout dat wel van de methode afhankelijk is wordt omvat door de niet-steekproeffout.

We gaan nu de dagblad-methode vergelijken met de gebruikelijke methoden door de mogelijke niet-steekproeffouten te analyseren.



Tabel 1. Indeling van de totale fout in steekproefonderzoek.

[Kersten en Bethlehem (1982)]

1.2 Fout-analyse van de dagblad-methode

Wat de waarnemingsfouten betreft is het verschil in de beide methoden niet wezenlijk. De problemen die zich voor doen bij meetfouten hebben meer te maken met het verschil tussen mondelinge en schriftelijke ondervraging (het CBS werkt met beide soorten). Wel geldt dat een respondent van een dagblad-enquête geen mogelijkheid heeft over de gestelde vragen informatie in te winnen waardoor vragen die niet geheel duidelijk zijn onzorgvuldiger zullen worden ingevuld. Daartegenover staat dat door de enquêteur/trice niet 'gestuurd' kan worden. Ook zal bij de respondent de neiging om feiten te verdraaien om acceptabeler over te komen minder sterk zijn.

De essentiële verschillen bevinden zich in de niet-waarnemingsfouten. Bij het trekken van een steekproef zijn het steekproefkader en de doelpopulatie van belang. Het steekproefkader is de verzameling van alle elementen die in de steekproef kunnen worden opgenomen. Zo wordt bij het CBS veelal gewerkt met een lijst van gemeenten en adressen. Het steekproefkader bij een dagbladenquête is de groep van de op deze krant geabonneerde personen. Kaderfouten worden veroorzaakt door het opnemen van foutieve gegevens of het meerdere malen (ten onrechte) opgenomen zijn van dezelfde elementen. Zo kan het bijvoorbeeld voorkomen dat een persoon meerdere malen in het administratie-systeem van een gemeente (het steekproefkader) voorkomt, zodat de trekingskans van deze persoon vergroot wordt. Kaderfouten zijn in het algemeen zeer moeilijk te achterhalen. We zullen hier verder niet op ingaan.

Dekkingsfouten treden op indien het steekproefkader niet samenvalt met de doelpopulatie. Bevat het steekproefkader meer elementen dan de doelpopulatie

dan spreken we van overdekking, indien het steekproefkader minder elementen bevat dan de doelpopulatie van onderdekking. De doelpopulatie is de totale Nederlandse bevolking. Het CBS kan het steekproefkader zelf kiezen. Bij de dagblad-methode is het steekproefkader beperkt tot lezers van de betrokken bladen. Mensen die geen krant lezen en mensen die in plaatsen wonen die niet in het verspreidingsgebied van de krant liggen zijn bij voorbaat van deelname uitgesloten. Bij de GPD-enquête zijn personen uit een aantal grote steden zoals Amsterdam, Rotterdam en Den Haag niet in het steekproefkader opgenomen. Deze onderdekking kan de representativiteit van de dagblad-enquête in negatieve zin beïnvloeden.

Met non-respons worden alle elementen bedoeld die in de steekproef zijn opgenomen maar waarvan geen gegevens zijn verkregen. De non-respons kan bij een dagblad-enquête hoger uitvallen dan bij enquêtes gehouden door inschakeling van enquêteurs/trices. Men is minder snel geneigd een enquêteur/trice weg te sturen dan de vragenlijst in een krant aan de kant te schuiven. Ook bestaat bij de dagblad-methode niet de mogelijkheid om met een volgende ronde van aanvullende trekkingen de eventuele selectieve non-respons te verlagen, men moet maar afwachten wie er zal reageren. Het probleem van non-respons kan de representativiteit van de dagblad-enquête ongunstig beïnvloeden. De methode van enquêteren via een krant is echter veel goedkoper dan de methode met een corps van enquêteurs/trices. De steekproefgrootte kan daarom veel groter genomen worden. Het responspercentage is bij de GPD-enquête (3%) laag vergeleken met de conventionele enquête, de steekproefgrootte was echter groter dan gebruikelijk. Het uiteindelijk verkregen aantal (voor dit onderzoek gebruikte) waarnemingen van 12.000 is dan ook zeer ruim ten opzichte van de circa 1.000 tot 2.500 waarnemingen, die bij opiniebureaus en het CBS veelal verkregen aantallen.

De belangrijkste verschillen tussen de dagblad-methode en de conventionele methoden die bij de fout-analyse naar voren kwamen zijn de onderdekking van niet-krantenlezers en de personen buiten het verspreidingsgebied van de krant en in mindere mate de non-respons. Beide fouten kunnen veroorzaken dat bepaalde groepen mensen met dezelfde karakteristieken in de steekproef onder- of oververtegenwoordigd zijn ten opzichte van de doelpopulatie. De representativiteit kan met herweging worden verbeterd. Een vereiste hiervoor is dat van elk der ondervertegenwoordigde groepen toch een voldoende aantal respondenten in de steekproef aanwezig zijn: volledige afwezigheid kan met herweging niet worden gecorrigeerd. Bij de GPD-enquête zal, indien aan deze eis niet voldaan wordt, deze afwezigheid voornamelijk door de onderdekking veroorzaakt worden, daar door de grootte van de steekproef de kans dat door non-respons een groep

totaal niet vertegenwoordigd is zeer klein wordt. Het onderzoek in dit rapport richt zich daarbij voornamelijk op de fout veroorzaakt door het beperkte verspreidingsgebied van de GPD. Over de niet-krantenlezers is nagenoeg geen bruikbare informatie beschikbaar. Uit het stops-onderzoek van de GPD onder ex-abonnees (1983), evenals uit een aantal bij de enquête gevoegde brieven komt wel naar voren dat veel mensen de krant delen met kennissen of burens, waardoor de enquête hen toch bereikt kan hebben.

Problemen kunnen zich voordoen indien het GPD-bestand gebruikt gaat worden bij het schatten van verbanden tussen verschillende grootheden voor de Nederlandse bevolking.

Voor dit onderzoek zijn twee relaties gekozen:

- de in veel economisch onderzoek relevante verklaring van inkomen door een aantal verklarende variabelen waaronder opleiding, beroep en leeftijd
- daarnaast een belangrijke relatie die in de onderzoeken van het Inkomenswaardings Project van het COEPS centraal staat, namelijk het verband tussen inkomenswaardering en gezinsinkomen en -grootte.

2. De invloed van onderdekking van steden op regressie-analyse

2.1 Inleiding

De belangrijkste woongebieden die zich niet in het verspreidingsgebied van de GPD-bladen bevinden zijn een aantal steden met de hoogste urbanisatiegraad (meer dan 100.000 inwoners), waaronder Den Haag, Rotterdam en Amsterdam. Om enig inzicht te verkrijgen in het probleem van de uitsluiting van stedelingen wordt allereerst gekeken naar de te onderzoeken regressie-vergelijking met invoering van een indicator-variabele voor wonen in een grote stad. Op deze manier kan er getoetst worden op significante verschillen in de parameters van de regressie-vergelijking voor stedelingen en niet stedelingen.

Indien het wel of niet responderen bij een steekproef samenhangt met de storingsterm in de regressie-vergelijking zullen de kleinste kwadraten schatters onzuiver zijn. Er bestaat een methode om voor een dergelijke samenhang te corrigeren en alsnog zuivere schatters te krijgen. Deze methode is ontwikkeld door Heckman (1979). De methode komt er op neer de onzuiverheid terug te brengen tot een specificatiefout in de regressievergelijking. Dan wordt gecorrigeerd door de weggelaten variabele (de correctie-term) per respondent te berekenen en aan het model toe te voegen. De methode van Heckman wordt algemeen gebruikt voor het corrigeren van selectieve non-respons, wanneer bepaalde

groepen die wel in het steekproefkader zijn opgenomen besluiten niet aan de enquête mee te doen. Wij gaan in § 2.3 deze methode gebruiken om te corrigeren voor onderdekking. De selectie wordt dan opgelegd door de steekproefmethode. Voor het onderzoek is informatie nodig over zowel de mensen die in de stad (meer dan 100.000 inwoners) wonen als de mensen die niet in de stad wonen. Het GPD-bestand bevat deze informatie slechts voor een gering aantal steden. We gebruiken daarom gegevens afkomstig van het Consumenten Conjunctuur Onderzoek, een enquête gehouden onder de Nederlandse bevolking door het CBS in 1979 (CCO'79). De CCO-enquête werd gehouden op basis van een tweetrapssteekproef met als primaire eenheden de gemeenten en als secundaire de adressen. Gemeenten met meer dan 18.277 inwoners werden automatisch bij het onderzoek betrokken. Door aselechte trekkingen uit het Register van Volkstelling werden hierbij een aantal kleinere gemeenten gevoegd. Uit deze gemeenten werd een aantal gezinnen benaderd om mee te doen, waarbij het aantal uit één gemeente afhankelijk was van de gemeentegrootte. Voor de bij dit onderzoek benodigde gegevens zijn in het bestand 2.050 bruikbare waarnemingen aanwezig waarvan 548 stedelingen en 1.502 niet-stedelingen. Met deelbestanden van dit CCO-bestand zullen we de situatie van onderdekking van de GPD-steekproef nabootsen. Uit de vergelijking van de resultaten voor de gesimuleerde onderdekte steekproef en de resultaten voor het complete bestand hopen we conclusies te trekken met betrekking tot de waarde van de regressies op het GPD-bestand.

2.2 De regressie-vergelijking voor steden/niet steden

De te onderzoeken relatie is als volgt algemeen weer te geven:

$$(2.2.1) \quad y = X\beta + e$$

waarbij y een vector van de te verklaren variabele van dimensie N ; X een $N \times M$ matrix van verklarende variabelen; β de te schatten coëfficiëntenvector van dimensie M en e een vector van storingstermen van dimensie N is.

We definiëren de indicator-variabele als volgt:

$$D_i = \begin{cases} 1 & \text{indien de respondent } i \text{ in de stad woont} \\ 0 & \text{indien de respondent } i \text{ niet in de stad woont,} \\ & i=1, \dots, N. \end{cases}$$

De $(n \times m)$ matrix DX bevat als element DX_{ij} de waarde van X_{ij} indien $D_i=1$, en 0 indien $D_i=0$. Met de kleinste kwadraten methode op de vergelijking

$$y = X\beta + DX\alpha + e$$

$$E(e) = 0$$

$$E(ee') = \sigma^2 I_N$$

vinden we schattingen voor α , β en σ^2 .

De parameter-vector α geeft het verschil aan tussen wel en niet stadbewoners. Indien $\alpha=0$ dan zou dat betekenen dat de vergelijking (2.2.1) voor stedelingen niet verschilt met die voor de niet-stedelingen. Hieruit is dan de conclusie te trekken dat een schatting voor de coëfficiënten β met behulp van een databestand zonder mensen uit de grote steden een betrouwbaar beeld geeft van de samenhang in de totale populatie.

We hebben de hypothese $\alpha=0$ getoetst voor twee regressievergelijkingen. De eerste vergelijking betreft de parameter μ van de inkomenswaardingsvraag. Deze variabele kan globaal worden geïnterpreteerd als de logaritme van het bedrag dat mensen tussen "voldoende" en "onvoldoende" beschouwen. De verklarende variabelen zijn gezinsgrootte (fs) en gezinsinkomen (y_{gez}).

De resultaten van tabel 2 laten zien dat de coëfficiënten $\alpha_0, \alpha_1, \alpha_2$ elk afzonderlijk niet significant van nul verschillen. De gecombineerde nulhypothese ' $\alpha_0=\alpha_1=\alpha_2=0$ ' moet evenwel verworpen worden. Bij deze regressie bestaat er dus wel een verschil tussen steden en niet-steden. Dit verschil is echter maar zeer klein, en geen der afzonderlijke coëfficiënten verschilt significant van nul.

Tabel 2. Resultaten inkomenswaardings-vergelijking met een indicator voor wel/niet in de stad op het CCO'79 bestand

afhankelijke variabele: μ			
variabele		parameter	std afw.
constante	β_0	3.394	0.125
$\ln(fs)$	β_1	0.089	0.010
$\ln(y_{gez})$	β_2	0.641	0.013
D	α_0	-0.222	0.225
$D \times \ln(fs)$	α_1	0.011	0.019
$D \times \ln(y_{gez})$	α_2	0.024	0.023
$N = 2050$			
$\sigma^2 = 0.035$			
$R^2 = 0.723$			

toets $H_0 : \alpha_0=\alpha_1=\alpha_2=0$

$$F_0 = 4.306 > F_{0.05;3,\infty} = 2.60$$

Tabel 3. Resultaten inkomens-vergelijking van de hoofdkostwinner met een indicator voor wel/niet in de stad op het CCO'79 bestand

afhankelijk variabele: $\ln(\text{inkomen hkw})$			
variabele	parameter		std afw.
constante	β_0	0.967	0.927
<u>opleiding</u>			
- lbo lts huishoudschool	β_1	0.082	0.024
- ulo mulo mavo	β_2	0.260	0.030
- havo hbs lyceum	β_3	0.306	0.041
- mbo	β_4	0.217	0.026
- hbo	β_5	0.425	0.029
- universiteit	β_6	0.620	0.043
$\ln(\text{leeftijd})$	β_7	4.513	0.504
$\ln(\text{leeftijd})^2$	β_8	-0.577	0.068
<u>werkzaamheid</u>			
- geen werk	β_9	-0.199	0.031
- geschoold arbeider	β_{10}	-0.007	0.032
- employee	β_{11}	0.093	0.028
- zelfstandig/directeur			
/vrij beroep	β_{12}	0.025	0.037
- landbouwer/tuinder	β_{13}	-0.059	0.044
man	β_{14}	0.257	0.023
D	α_0	-2.343	1.593
<u>D × opleiding</u>			
- lbo lts huishoudschool	α_1	0.059	0.045
- ulo mulo mavo	α_2	-0.021	0.055
- havo hbs lyceum	α_3	-0.104	0.067
- mbo	α_4	0.029	0.050
- hbo	α_5	-0.116	0.053
- universiteit	α_6	-0.035	0.083
$\ln(\text{leeftijd})$	α_7	1.032	0.863
$\ln(\text{leeftijd})^2$	α_8	-0.105	0.116
<u>D × werkzaamheid</u>			
- geen werk	α_9	-0.098	0.062
- geschoold arbeider	α_{10}	0.040	0.073
- employee	α_{11}	0.019	0.060
- zelfstandig/directeur			
/vrij beroep	α_{12}	-0.166	0.082
- landbouwer/tuinder	α_{13}	0.013	0.201
D × man	α_{14}	-0.055	0.037

N = 2050

 $\sigma^2 = 0.071$ $R^2 = 0.556$ toets $H_0 : \alpha_0 = \dots = \alpha_{14} = 0$ $F_0 = 5.502 > F_{0.05; 15, \infty} = 1.67$

Alle variabelen die opleiding, werkzaamheid en geslacht weergeven zijn dummy-variabelen; de referentiegroep bestaat uit vrouwelijke hoofdkostwinners met lagere schoolopleiding, die werkzaam zijn als ongeschoold handarbeider.

De tweede vergelijking betreft de variabele "inkomen van de hoofdkostwinner". De verklarende variabelen zijn opleiding, beroep, $\ln(\text{leeftijd})$ en $[\ln(\text{leeftijd})]^2$ en geslacht van de hoofdkostwinner. De resultaten van deze regressie-vergelijking, met kruisprodukten van alle verklarende variabelen en de indicator D, staan beschreven in tabel 3. Deze tabel geeft eenzelfde beeld. Ook hier zijn de coëfficiënten α afzonderlijk nagenoeg alle niet significant van nul verschillend met uitzondering van α_5 'hbo-ers' en α_{12} de 'zelfstandigen'. De nulhypothese ' $\alpha_0 = \alpha_1 = \dots = \alpha_{14} = 0$ ' moet ook hier verworpen worden.

Er blijken bij regressie-vergelijkingen in de coëfficiënten kleine verschillen te bestaan indien toegepast op de subgroepen van stedelingen respectievelijk niet-stedelingen. Hieruit mag echter niet de conclusie getrokken worden dat er dan ook significante verschillen zullen bestaan tussen niet-stedelingen en de totale populatie. Bij de inkomenswaarderings-vergelijking blijkt geen der coëfficiënten afzonderlijk significant van nul verschillend; bij de inkomens-vergelijking is dat het geval voor twee coëfficiënten.

2.3 De methode van Heckman bij onderdekking van steden

We zullen nu onderzoeken in hoeverre de coëfficiënten van de relatie (2.2.1) gecorrigeerd moeten en kunnen worden indien slechts gegevens over één subgroep aanwezig zijn, te weten niet-stedelingen. De relatie (2.2.1) is per individu weer te geven als:

$$(2.3.1) \quad y_i = x_i' \beta + e_i$$

waarbij x_i' de i -de rijvector uit de matrix X is.

Het wel of niet meedoen aan de enquête wordt weergegeven in de selectievergelijking - 'sample selection rule' -:

$$(2.3.2) \quad r_i = z_i' \gamma + u_i$$

waarbij $r_i > 0$ indien de betreffende persoon i in de steekproef zit
 $r_i \leq 0$ indien de betreffende persoon i niet in de steekproef zit
 z_i een $(k \times 1)$ vector van relevante achtergrond-variabelen van individu i
 γ_i de bijbehorende $(k \times 1)$ coëfficiëntenvector
en u_i de storingsterm van individu i .

We nemen de relatie (2.3.1) alleen waar voor de respondenten, dus observaties i waarbij $r_i > 0$.

Relatie (2.3.1) is te schrijven als

$$\begin{aligned}
 E(y_i | x_i, \text{sample selection rule}) &= x_i' \beta + E(e_i | r_i > 0) \\
 &= x_i' \beta + E(e_i | u_i > -z_i' \gamma).
 \end{aligned}$$

Indien de conditionele verwachting van e gelijk aan nul is hebben we niet te maken met selectieve non-respons en is de schatting van β met de kleinste kwadraten methode zuiver. In het andere geval dient men de verwachting verder uit te schrijven. Onder de veronderstelling dat de simultane dichtheid van e en u normaal is met verwachting nul, varianties σ_e^2 en σ_u^2 respectievelijk, en covariantie σ_{eu} , geldt [Johnson en Kotz (1972)]:

$$(2.3.3) \quad E(e_i | u_i > -z_i' \gamma) = \frac{\sigma_{eu}}{\sigma_u} \lambda_i$$

$$\text{met } \lambda_i = \frac{\phi(v)}{1 - \Phi(v)} \quad v = \frac{-z_i' \gamma}{\sigma_u}$$

waarbij ϕ en Φ respectievelijk de normale dichtheid- en verdelingsfunctie weergeven.

Als we deze term in de regressie-vergelijking opnemen kunnen we afleiden dat de verwachting van de nieuwe storingsterm $v_i = e_i - E(e_i | u_i > -z_i' \gamma)$ wel gelijk aan nul is. Een toepassing van deze methode voor correctie voor selectieve non-respons wordt beschreven in bijvoorbeeld Van de Stadt en Wansbeek (1982).

Gebruik makend van deze resultaten kunnen we een zuivere schatting voor β berekenen met de volgende procedure:

- (1) met de selectievergelijking berekenen we een schatting voor γ met probit-analyse
- (2) voor elke respondent berekenen we de correctie-term λ_i
- (3) vervolgens berekenen we de kleinste kwadraten schatting voor β van de regressie vergelijking

$$(2.3.4) \quad y_i = x_i' \beta + \frac{\sigma_{eu}}{\sigma_u} \lambda_i + v_i \quad E(v_i) = 0.$$

2.4 Het schatten van de selectievergelijking

In de selectievergelijking (2.3.2) is r_i een continue variabele die de responsgeneigdheid van de persoon weergeeft. Deze variabele wordt echter niet waargenomen. Bovendien wordt het al dan niet responderen van een individu uit-

sluitend afhankelijk verondersteld van het al dan niet wonen in een stad. We definiëren daarom een dummy-variabele als volgt:

$$s_i = 1 \text{ als } r_i > 0, \text{ d.w.z. individu } i \text{ respondeert, dus woont} \\ \text{niet in de stad} \\ 0 \text{ als } r_i \leq 0, \text{ ofwel individu } i \text{ respondeert niet, dus} \\ \text{woont in een stad van meer dan 100.000 inwoners.}$$

Voor het schatten van y uit de selectievergelijking is informatie nodig over de waarden van de achtergrondvariabelen z_i van stedelingen en van niet-stedelingen. We hebben het effect van de correctieterm geschat door het selectieproces te simuleren in een "compleet" bestand, namelijk het CCO-bestand, waarin zowel stedelingen als niet-stedelingen aanwezig zijn. Dit bestand is in deze twee deelgroepen gesplitst, en op deze twee groepen is vergelijking (2.3.2) geschat.

Tabel 4. Resultaten selectievergelijking¹

variabele	parameter	std.afw.
gepensioneerd	γ_1 -0.095	0.099
in opleiding	γ_2 -0.889	0.239
werkloos	γ_3 0.034	0.089
employee	γ_4 -0.130	0.093
zelfstandig/directeur/ vrij beroep	γ_5 0.038	0.158
landbouwer/tuinder	γ_6 0.855	0.558
lbo lts huishoudschool	γ_7 0.122	0.083
mavo have lyceum etc.	γ_8 -0.038	0.104
mbo	γ_9 0.247	0.109
hbo	γ_{10} 0.045	0.111
universiteit	γ_{11} 0.012	0.185
geen kinderen	γ_{12} 0.071	0.072
meer dan een kind	γ_{13} 0.407	0.080
hoofdkostwinner man	γ_{14} 0.485	0.066

$R^2 = 0.187$

$N = 2050$

steden	goed voorspeld: 12% (66)
-----	fout voorspeld: 88% (482)
niet-steden	goed voorspeld: 96% (1438)
-----	fout voorspeld: 4% (64)

¹ In eerste instantie is de variabele $s_i = 0$ gedefinieerd als: wonend in Amsterdam, Rotterdam of Den Haag. Door het relatief kleine aantal respondenten uit deze steden t.o.v. de totale steekproef (245 van de 2050) is de op deze variabele geschatte selectievergelijking nog slechter in staat de beide groepen te onderscheiden: geen van de respondenten werd in de grote stad voorspeld.

Als achtergrondvariabelen zijn karakteristieken gekozen die van invloed zijn op het al dan niet wonen in de stad. De selectievergelijking zoals hier gedefinieerd is kan, nadat een schatter van γ berekend is, geïnterpreteerd worden als een voorspellingsfunctie of mensen al dan niet in de stad wonen. We berekenen dan de waarden van r_i door γ en de waarden van de achtergrondvariabelen z_i in vergelijking (2.3.2) te substitueren. Indien $\hat{r}_i \leq 0$ dan is de 'voorspelling' dat persoon i in de stad woont. Als criterium voor de kwaliteit van de selectievergelijking is genomen het aantal goed/verkeerd ingedeelde personen. Na verschillende combinaties van verklarende variabelen geprobeerd te hebben is de relatie waarvan de resultaten in tabel 4 staan tenslotte gekozen. Daarbij zijn als verklarende variabelen opgenomen: werkzaamheid en beroep, opleiding, geslacht van de hoofdkostwinner en gezinssamenstelling.

Alle variabelen zijn dummy-variabelen en hebben de waarde 1 indien de eigenschap van toepassing is en 0 anders.

Slechts enkele coëfficiënten, γ_2 , γ_9 , γ_{13} en γ_{14} zijn significant verschillend van nul. Hiervan is γ_2 negatief hetgeen betekent dat meer hoofdkostwinners die nog met hun opleiding bezig zijn zich in de stad bevinden. Mensen die als hoogste een 'mbo'-opleiding hebben (γ_9) bevinden zich meer in de niet-steden evenals gezinnen met meerdere kinderen (γ_{13}) en de gezinnen waar de hoofdkostwinner een man is (γ_{14}).

Van de huishoudens die in de stad 'voorspeld' zijn (66 resp. 64) woont iets meer dan de helft ook daadwerkelijk in de stad, dus ook de 'voorspellende' waarde van de selectievergelijking is niet erg groot. Het een en ander kan in verband gebracht worden met het in paragraaf 2.1 gevonden minimaal verschil tussen de steden en niet steden. Hierdoor is het onmogelijk een scherpe scheidingslijn tussen de beide groepen te krijgen.

(Merk op dat ook de coëfficiënten voor HBO'ers en 'zelfstandigen' in de selectievergelijking niet significant van nul verschillen).

2.5 Resultaten met de Heckman-correctieterm: simulaties

Ter controle van de bruikbaarheid van de methode van Heckman voor het corrigeren van de uitsluiting van personen uit de grote steden zijn de berekeningen uitgevoerd op het deelbestand "niet-stedelingen" van het CCO'79. Met behulp van $\hat{\gamma}$ is voor de personen die niet in de stad wonen de Heckman-term berekend.

$$(2.5.1) \quad \lambda_i = \frac{\phi(-z_i' \hat{\gamma})}{1 - \phi(-z_i' \hat{\gamma})} \quad (\sigma_u = 1, \text{ een normalisatierestrictie})$$

Tabel 5. Resultaten inkomenswaarderings-vergelijking
(methode van Heckman op het CCO'79)

variabele		afhankelijke variabele: mu					
		totaal CCO'79		alleen niet steden CCO'79			
		parm.	std.afw.	vóór correctie	na correctie	parm.	std.afw.
constante	β_0	3.334	0.104	3.394	0.122	3.356	0.129
ln(fs)	β_1	0.088	0.008	0.089	0.009	0.097	0.013
ln(ygez)	β_2	0.648	0.010	0.641	0.012	0.642	0.012
Heckman-term	σ_{eu}^2	-	-	-	-	0.038	0.042
R^2		0.7208		0.7082		0.7084	
N		2050		1502		1502	

Tabel 6. Resultaten inkomens-vergelijking van de hoofdkostwinner
(methode van Heckman op het CCO'79)

variabele		afhankelijke variabele: ln(inkomen hkw)					
		totaal CCO'79		alleen niet steden CCO'79			
		parm.	std.afw.	vóór correctie	na correctie	parm.	std.afw.
constante	β_0	-0.165	0.761	0.967	0.919	3.899	1.046
lbo lts huishoud	β_1	0.091	0.020	0.082	0.024	0.057	0.023
ulo mulo mavo	β_2	0.244	0.026	0.260	0.030	0.278	0.030
havo hbs lyceum	β_3	0.224	0.032	0.217	0.026	0.331	0.041
mbo	β_4	0.223	0.023	0.217	0.026	0.174	0.027
hbo	β_5	0.377	0.025	0.425	0.029	0.427	0.029
universiteit	β_6	0.606	0.037	0.620	0.043	0.612	0.043
ln(leeftijd)	β_7	5.041	0.413	4.513	0.504	3.092	0.554
ln(leeftijd) ²	β_8	-0.636	0.056	-0.577	0.068	-0.387	0.075
geen werk	β_9	-0.237	0.026	-0.199	0.031	-0.190	0.030
geschoold arbeider	β_{10}	0.007	0.029	-0.007	0.032	-0.009	0.031
employee	β_{11}	0.107	0.025	0.093	0.028	0.117	0.028
zelfst./direc.							
/vrij beroep	β_{12}	0.023	0.033	0.025	0.037	0.035	0.037
landbouw./tuinder	β_{13}	-0.072	0.043	-0.059	0.044	-0.172	0.048
man	β_{14}	0.243	0.018	0.257	0.023	0.141	0.030
Heckman-term	σ_{eu}	-	-	-	-	-0.465	0.082
R^2		0.5374		0.5237		0.5338	
N		2050		1502		1502	

Deze term wordt vervolgens aan het model toegevoegd en de regressie-vergelijking kan met de kleinste kwadraten methode geschat worden. Na toevoeging van de correctieterm bij het model voor dit deelbestand moeten de regressie-coëfficiënten β , indien de methode goed werkt, meer lijken op de coëfficiënten van het originele model (zonder Heckman-term) van het totale bestand, dan vóór correctie.

De resultaten van de twee onderzochte regressievergelijkingen voor het totale bestand en voor het deelbestand 'alleen niet-steden' zonder en met correctie-term staan in tabel 5 en 6.

De tabellen laten enige interessante resultaten zien:

Uit tabel 5 blijkt dat de coëfficiënten van de inkomenswaarderingsvergelijking van het totale bestand en het sub-bestand nagenoeg hetzelfde zijn (kolom 3 en 5). De coëfficiënt van de Heckman-term σ_{eu} is dan ook niet significant van nul verschillend, hetgeen inhoudt dat de onderdekking van de steden op deze vergelijking geen invloed heeft. Het toevoegen van de Heckman-term beïnvloedt echter wel de andere coëfficiënten; de coëfficiënt β_1 wordt slechter, in die zin dat deze minder op de coëfficiënt van de totale vergelijking lijkt dan β_1 in de niet-gecorrigeerde vergelijking. β_2 verandert nagenoeg niet en β_0 wordt enigszins in de goede richting gecorrigeerd.

Bij de inkomensvergelijking van de hoofdkostwinner in tabel 6 is de Heckman-term wel significant. De overige coëfficiënten - met uitzondering van de coëfficiënt voor universiteit - verschillen echter meer met die van het totale bestand dan de coëfficiënten van de regressie op het subbestand zonder Heckman-term. Merk overigens op dat β_{10} voor het totale bestand een ander teken heeft dan voor het deelbestand niet-steden, zowel met als zonder Heckman-term. β_{10} is echter in alle versies niet significant van nul verschillend.

Hoewel de Heckman-term een significante invloed aangeeft is de gecorrigeerde vergelijking eerder een verslechtering dan een verbetering. De methode van Heckman corrigeert derhalve niet op de juiste wijze voor selectiviteit in onze steekproef.

De tentatieve verklaring hiervoor is dat de sample selection rule een slecht onderscheidend vermogen heeft om stedelingen en niet-stedelingen te groeperen. Indien in dat geval de methode van Heckman toch wordt toegepast, krijgen we in plaats van een correctie een verslechtering van de resultaten. We concluderen dan ook dat bij de methode van Heckman een significante σ_{eu} met enig voorbehoud bekeken moet worden: de bijbehorende sample selection rule moet wel 'goed' geschat kunnen worden.

Voor de volledigheid worden in tabel 7 en 8 nog de resultaten van de beide vergelijkingen met en zonder correctie-term gegeven voor het GPD-bestand.

De correctie-term wordt in dit geval berekend door de eigenschappen z_1 van de GPD-respondenten in vergelijking (2.5.1) in te vullen waarbij voor γ de schatting uit het CCO-bestand genomen wordt. De regressies zijn gedaan op basis van

11.380 waarnemingen, dat is de netto steekproef die van de 24.000 GPD waarnemingen beschikbaar is gekomen.

Tabel 7. Resultaten inkomenswaarderings-vergelijking
(methode van Heckman op het GPD'83)

afhankelijke variabele: mu		zonder Heckman		met Heckman	
variabele		parm.	std.afw.	parm.	std.afw.
constante	β_0	2.930	0.049	2.873	0.050
ln(fs)	β_1	0.037	0.004	0.049	0.004
ln(ygez)	β_2	0.695	0.005	0.696	0.005
Heckman-term	σ_{eu}^2	-	-	0.062	0.011
R ²		0.6954		0.6962	
N		11380		11380	

Tabel 8. Resultaten inkomens-vergelijking van de hoofdkostwinner
(methode van Heckman op het GPD'83)

afhankelijke variabele: ln(inkomen hkw)		zonder Heckman		met Heckman	
variabele		parm.	std.afw.	parm.	std.afw.
constante	β_0	2.080	0.373	2.524	0.379
lbo lts huishoud	β_1	0.060	0.010	0.050	0.011
ulo mulo mavo	β_2	0.132	0.010	0.131	0.010
havo hbs lyceum	β_3	0.236	0.011	0.236	0.011
mbo	β_4	0.157	0.011	0.140	0.011
hbo	β_5	0.293	0.010	0.286	0.010
universiteit	β_6	0.512	0.013	0.506	0.013
ln(leeftijd)	β_7	3.671	0.204	3.473	0.206
ln(leeftijd) ²	β_8	-0.428	0.028	-0.402	0.028
geen werk	β_9	-0.095	0.020	-0.072	0.020
geschoold arbeider	β_{10}	0.120	0.020	0.122	0.020
employee	β_{11}	0.202	0.019	0.212	0.019
zelfst./direc.					
/vrij beroep	β_{12}	0.295	0.022	0.295	0.022
landbouw./tuinder	β_{13}	0.050	0.051	0.018	0.051
man	β_{14}	0.238	0.009	0.202	0.010
Heckman-term	σ_{eu}^2	-	-	-0.123	0.019
R ²		0.4792		0.4812	
N		11360		11360	

Opvallend is dat σ_{eu} bij de inkomenswaarderings-vergelijking van het GPD-be-stand in tegenstelling tot het CCO nu wel significant van nul verschillend is.

Een verdere vergelijking met tabel 5 laat zien dat de overige coëfficiënten op dezelfde manier op de Heckman-term reageren: β_0 wordt verlaagd, β_1 wordt verhoogd en β_2 blijft nagenoeg onveranderd.

Bij de inkomensvergelijking van de hoofdkostwinner is, evenals bij het CCO in tabel 6, ook bij het GPD de coëfficiënt van de Heckman-term significant van nul verschillend. De andere coëfficiënten in tabel 8 veranderen maar weinig door het toevoegen van de term, alleen de constante wordt sterk verhoogd.

We concluderen dat de methode van Heckman niet zonder meer kan worden toegepast bij het bepalen van de invloed van de uitsluiting van respondenten uit steden met meer dan 100.000 inwoners. Daar de selectieregel niet goed geschat kan worden heeft een significant van nul verschillende coëfficiënt van de Heckman-term niet de gewenste betekenis: uit de simulatie op het CCO-bestand blijkt de correctie geen verbetering van de schattingen op te leveren.

Het niet kunnen onderscheiden van stedelingen en niet-stedelingen is wel een indicatie dat de invloed van de afwezigheid van stedelingen niet groot kan zijn, hetgeen overeenkomt met de conclusie van § 2.1. Of dit echter een specifiek probleem van het CCO bestand is, dan wel een algemeen verschijnsel, kunnen wij met de ons ter beschikking staande gegevens niet analyseren.

3. Herwegen met RAS-gewichten

3.1 De RAS-methode

De respondenten van de GPD-enquête worden herwogen zodanig dat de relatieve frequenties van bepaalde gekozen achtergrondvariabelen overeenkomen met de relatieve frequenties in de landelijke bevolking.

Een vereiste voor deze herwegingsprocedure is dat van elke 'cel', gecreëerd door de kenmerken, een aantal respondenten in de steekproef opgenomen is, indien die cel in de totale populatie niet leeg is. Ongeschoolde arbeiders met een universitaire opleiding zullen bijvoorbeeld niet in de populatie te vinden zijn, dus ook niet in de steekproef.

Daar bij een dagblad-enquête maar moet worden afgewacht wie er zal reageren is niet a priori zeker dat aan de voorwaarde van celvulling voldaan is. Door de aanzienlijke steekproef-grootte van 12.000 waarnemingen en het resultaat uit het vorige hoofdstuk, dat het uitsluiten van stedelingen alleen minimale verschillen laat zien, zal de kans dat bepaalde cellen niet voldoende zijn gevuld

echter zeer klein zijn. Het gebruik van de herwegingsprocedure is hierdoor grotendeels gerechtvaardigd.

Bij de GPD-enquête zijn drie kenmerken voor de herwegingsprocedure gebruikt:

leeftijd : onderverdeeld in 10 categorieën
 opleiding : onderverdeeld in 7 categorieën
 politieke keuze in 1982 : onderverdeeld in 14 categorieën.

Elke respondent van wie deze kenmerken bekend zijn is zo in te delen in een bepaalde cel. Een voorbeeld van een cel is personen tussen de 30 en 35 jaar met een mbo-opleiding die in 1982 op het CDA gestemd hebben.

Uit de enquête is de simultane frequentie voor elke combinatie van karakteristieken te berekenen en deze worden aangeduid met F_{ijk} , $i=1, \dots, 10$; $j=1, \dots, 7$; $k=1, \dots, 14$.

Er wordt nu gezocht naar een herwegingsfactor $w(i,j,k)$ om gecorrigeerde simultane steekproef-frequenties te krijgen die gelijk zijn aan de simultane frequenties in de totale populaties: P_{ijk} .

We veronderstellen dat een cel X_t met $X_t = (t; \text{leeftijd}(t) = i, \text{opleiding}(t) = j, \text{politiek}(t) = k)$ homogeen verdeeld is, hetgeen erop neerkomt dat elke persoon uit X_t hetzelfde gewicht krijgt.

De herwegingsfactoren zouden gemakkelijk te berekenen zijn uit de vergelijking:

$$P_{ijk} = w(i,j,k)F_{ijk}.$$

We hebben echter niet de beschikking over de simultane populatiefrequenties P_{ijk} , maar weten alleen de marginale frequenties in de populatie $P_{i..}$; $P_{.j.}$ en $P_{..k}$, met

$$P_{i..} = \sum_{j=1}^7 \sum_{k=1}^{14} P_{ijk}$$

$$(3.1.1) \quad P_{.j.} = \sum_{i=1}^{10} \sum_{k=1}^{14} P_{ijk}$$

$$P_{..k} = \sum_{i=1}^{10} \sum_{j=1}^7 P_{ijk}$$

Gebruik makend van stellingen uit de Informatie-theorie [Ireland en Kullback (1968)] geldt dat de oplossing van

$$(3.1.2) \quad \min_P I(F;P) = \min_{P_{ijk}} \sum_{ijk} P_{ijk} \frac{\ln(P_{ijk})}{\ln(F_{ijk})}$$

onder restrictie van (3.1.1)

resulteert in de volgende benadering voor P_{ijk} :

$$(3.1.3) \quad P_{ijk}^* = w_1(i) w_2(j) w_3(k) F_{ijk}$$

waarbij P_{ijk}^* , w_1 , w_2 en w_3 met een iteratief proces benaderd worden. Dit proces staat bekend onder de naam RAS-methode [Stone en Brown (1964)].

Het gewicht per individu blijkt uitgesplitst te kunnen worden naar de drie karakteristieken:

$$(3.1.4) \quad w(i,j,k) = w_1(i) w_2(j) w_3(k).(*)$$

$w_1(i) = 1$ indien de personen die bij karakteristiek 1 (leeftijd) in klasse i behoren een normaal responsgedrag vertonen,

$w_1(i) > 1$ indien deze personen een lagere respons dan gemiddeld hebben,

$w_1(i) < 1$ indien deze personen een hogere respons dan gemiddeld hebben.

3.2 Herwegen van de GPD-enquête

Ter illustratie zijn in tabel 9 de marginale frequenties van de drie karakteristieken leeftijd, opleiding en politieke voorkeur in 1982 opgenomen, zoals gevonden in de steekproef en voor de populatie (de totale Nederlandse bevol-

$$(*) \quad \text{N.B.} \quad q(t) = \frac{F_{ijk}}{P_{ijk}} = [w(i,j,k)]^{-1} \quad \text{voor } t \in X_t$$

kan geïnterpreteerd worden als een conditionele responskans, d.w.z. de kans dat een individu met waarden i,j,k voor de respectievelijke kenmerken, in de steekproef aanwezig is. In dat geval moeten alle waarden van $q(t)$ met eenzelfde factor vermenigvuldigd worden om $0 \leq q(t) \leq 1$ en $\sum q(t) = 1$ te krijgen. Voor de gewichten is dit niet noodzakelijk omdat daar alleen de verhouding van belang is. De methode veronderstelt dat de verschillende kenmerken onderling onafhankelijk zijn. Indien dit in werkelijkheid niet het geval is, zijn herwogen deelbestanden niet zonder meer representatief voor de populatiedeelbestanden.

Tabel 9. Marginale frequentie in de steekproef van de GPD en in de populatie

1. leeftijd

	steekproef GPD'83	populatie
13 - 24	0.040	0.068
25 - 29	0.147	0.124
30 - 34	0.193	0.134
35 - 39	0.176	0.114
40 - 44	0.112	0.086
45 - 49	0.083	0.077
50 - 54	0.079	0.077
55 - 59	0.070	0.073
60 - 64	0.053	0.069
65 +	0.047	0.180

2. opleiding

	steekproef GPD'83	populatie
lagere school	0.086	0.225
lbo lts huishoud	0.154	0.273
ulo mulo mavo	0.189	0.127
havo hbs lyceum	0.120	0.132
mbo	0.138	0.058
hbo	0.243	0.128
universiteit	0.071	0.057

3. politieke voorkeur in 1982

	steekproef GPD'83	populatie
pvda	0.272	0.245
cda	0.223	0.236
vvd	0.278	0.186
d'66	0.079	0.035
cpn	0.016	0.014
psp	0.033	0.018
sgp	0.008	0.015
ppr	0.023	0.013
gpv	0.004	0.007
rpf	0.010	0.012
cp	0.004	0.007
evp	0.010	0.005
overig	0.009	0.012
geen	0.031	0.195

king). De marginale frequenties voor de populatie zijn wat de politieke voorkeur betreft afkomstig van de uitslag van de tweede kamer verkiezing in 1982. De gegevens van de karakteristieken leeftijd en opleiding zijn verstrekt door het CEBUCO. Deze instantie heeft o.a. de oplage specificatie en verspreidingscijfers van de dagbladen onder haar beheer. De landelijke gegevens zijn door het CEBUCO berekend uit een - herwogen - steekproef van het NIPO uit 1981 met een omvang van 30.870 waarnemingen.

In de GPD steekproef vindt men een oververtegenwoordiging van de hoogste opleidingen en de politiek meer bewuste personen. Een ondervertegenwoordiging is te zien in de laagste en hoogste leeftijdsgroepen, de mensen met een lagere opleiding en de niet-stemmers. Na herweging zijn de marginale frequenties in de steekproef gelijk aan de marginale frequenties in de totale Nederlandse bevolking.

De herwegingskarakteristieken kunnen zelf gekozen worden. Deze keuze zal sterk afhankelijk zijn van het onderzoek dat met het data-bestand gedaan wordt. De procedure garandeert namelijk niet dat de steekproef na herweging in alle opzichten lijkt op de populatie, maar wel voorzover samenhangend met de gekozen herwegingskarakteristieken.

De respondenten die in de simultaan ondervertegenwoordigde klasse zitten, bijvoorbeeld 65-plussers met een lagere school opleiding die niet gestemd hebben, zullen in de GPD-enquête een groot gewicht meekrijgen.

3.3 Gewichten in frequentieberekeningen

Om te onderzoeken of herweging voldoende is om de GPD-enquête representatief te maken hebben we deze steekproef vergeleken met andere steekproeven.

De relatieve frequenties [in percentages] zijn voor diverse grootheden berekend, zowel ongewogen als gewogen, uit het GPD-, CCO- en PANEL-bestand. Het PANEL-bestand is afkomstig van de vierde jaarlijks gehouden panel-enquête van het CBS (maart 1983). Bij de start van dit onderzoek in 1980 zijn ongeveer 2.000 willekeurig gekozen Nederlandse huishoudens benaderd om mee te doen aan deze enquête. Elk opvolgend jaar zijn dezelfde huishoudens door een enquêtrice bezocht. Door uitvallers (sterfgevallen en verhuizingen) en afvallers (niet langer geïnteresseerd) liep het aantal deelnemers sterk terug. Het PANEL-bestand telt nu 466 voor ons onderzoek bruikbare waarnemingen waarvan 95 mensen uit de grote steden en 371 zogenoemde niet-stedelingen.

Bij het GPD- en het PANEL-bestand zijn de gewichten berekend met de RAS-methode, zoals beschreven in § 3.1, op de karakteristieken leeftijd, opleiding en

politieke voorkeur. De politieke voorkeur is bij het PANEL alleen in minder categorieën ingedeeld (10). Het CCO-bestand is destijds ('79) eveneens met behulp van de RAS-methode gewogen; daarbij zijn echter andere karakteristieken gebruikt die toen door het CBS ter beschikking zijn gesteld, namelijk leeftijd, werkzaamheid en gezinssamenstelling. De gewichten die toen berekend zijn hebben we hier ook gebruikt.

Om ruimte te besparen hebben we de resulterende frequentietabellen niet afgedrukt. We geven hier enkel de conclusies.

Leeftijd

De leeftijd geeft na herweging voor alle steekproeven nagenoeg dezelfde frequentieverdeling van de Nederlandse bevolking.

Opleiding

Op het 'havo'- en 'mbo'-niveau blijven na herweging grote verschillen bestaan tussen het CCO en de andere bestanden.

Politieke voorkeur

De politieke voorkeur geeft na herweging voor beide steekproeven nagenoeg dezelfde frequentieverdeling. Aangezien politieke voorkeur een herwegingskarakteristiek is, is deze verdeling gelijk aan de feitelijke frequentieverdeling.

Beroep

Het beroep van de hoofdkostwinner is de eerste uit de rij die niet als herwegingskarakteristiek gebruikt is. De frequentie van arbeiders als groep is bij alle steekproeven nagenoeg even groot. De uitsplitsing in geschoold en ongeschoold (die niet gemaakt is bij het PANEL) toont echter wat verschillen. Opvallend is dat na herweging de GPD goed in de richting van de landelijke resultaten van het CEBUCO komt, terwijl er bij het CCO bijna niets verandert.

De landbouwers blijven in het GPD-bestand ondervertegenwoordigd. De totale groep van niet-werkenden is bij alle steekproeven even groot en ook hier zitten de grote verschillen in de uitsplitsing. Het CEBUCO maakt deze uitsplitsing niet, waardoor helaas de resultaten van het GPD hiermee niet verder vergeleken kunnen worden. De grote verschillen bevinden zich in de categorieën huishouding, werkloos en opleiding. De gepensioneerden in het GPD-bestand zijn door de herweging goed opgetrokken. Verder geeft de herweging van werklozen bij het PANEL een verlaging van de relatieve frequentie terwijl in het GPD de relatieve frequentie wordt verhoogd.

Geslacht

Er bestaat een duidelijke oververtegenwoordiging van mannen in het GPD-bestand. De weging werkt wel verhogend voor de frequentie van de vrouwelijke hoofdkostwinners, de verhoging is echter niet groot genoeg.

Huishoudensamenstelling

De relatieve frequenties van het aantal kinderen van volledige en een-ouder gezinnen komen in de steekproeven aardig overeen.

Van de een-persoons-huishoudens en de twee-persoons zonder kinderen zijn de relatieve frequenties over de drie steekproeven alle verschillend.

In 'het statistisch zakboek 1983' van het CBS is niet de indeling gemaakt die hier gebruikt is, maar alleen een indeling die afhankelijk is van het aantal personen. Dit geeft weer een ander beeld van de frequentie-verdeling.

Inkomen

De inkomens in de tabel zijn de maximale bedragen van de tien-procents groepen. Zo zal volgens het GPD 10% van de Nederlandse bevolking minder dan 20,200 gulden verdienen. Bij het vergelijken van de kolommen moet overwogen worden dat het tijdsverschil op het inkomen een grote invloed heeft (nivellerings, prijscompensaties etc.).

Tabel 10. De decielgrenzen van de gewogen verdeling van het netto-inkomen van de huishoudens

(in guldens)	GPD'83	PANEL'83	CCO'79	CBS'83 (1979)
1e groep	20,200	16,000	14,000	15,000
2e groep	23,720	19,330	17,370	19,700
3e groep	26,540	23,840	20,380	23,300
4e groep	29,300	27,340	22,800	26,600
5e groep	31,980	29,820	25,200	29,900
6e groep	35,180	32,910	28,270	33,700
7e groep	39,000	36,260	32,100	38,200
8e groep	43,500	43,900	36,250	44,300
9e groep	50,500	50,220	43,000	54,600

Te zien is dat het GPD in de laagste groepen hogere inkomens vindt dan de andere steekproeven. Het CCO legt de grenzen beduidend lager in die klassen, maar deze steekproef is 4 jaar eerder gehouden. Het PANEL laat ondanks de 4 jaar die er tussen zitten dezelfde getallen zien als het CBS in hun statistisch handboek.

Het vergelijkingsmateriaal is afkomstig van enquêtes die net als het GPD-bestand steekproeven zijn van de Nederlandse bevolking. Enige verschillen houden we dus altijd. Over het algemeen blijkt uit de relatieve frequenties dat de GPD-enquête zeker niet minder representatief is dan de andere enquêtes.

Voorzichtigheid is geboden bij het conclusies trekken uit het materiaal over enkele subgroepen ten opzichte van de totale populatie, met name bij de groep 'landbouwers en tuinders'. Door de veel grotere aantallen respondenten van de GPD-enquête zijn de ondervertegenwoordigde groepen zoals de vrouwelijke hoofdkostwinners in absolute aantallen toch nog groter dan in de andere steekproeven (GPD: 898; CCO: 313; PANEL: 84). Het onderzoeken van eigenschappen van deze groep zal dan ook zonder meer met het GPD-bestand gedaan kunnen worden.

We hebben als wegingskarakteristieken leeftijd, opleiding en politieke voorkeur gekozen. Indien men andere specifieke aspecten van de populatie wil onderzoeken kunnen andere achtergrondvariabelen voor herweging gebruikt worden. Deze variabelen moeten goed overwogen gekozen worden. Het tijdsverschil, het enigszins anders gedefinieerd zijn van de variabelen (telt bijvoorbeeld bij opleiding alleen een behaald diploma of wordt een (nog) niet afgemaakte studie als hoogste geteld), een afwijkende klassenindeling van de variabelen en wie die indeling maakt (bepaalt een enquêtrice of iemand een geschoold danwel een ongeschoold arbeider is of doet de respondent dat zelf); dit alles moet afgewogen worden bij het kiezen van de karakteristieken (en het 'ijk'-bestand) voor het verkrijgen van 'goede' gewichten.

3.4 Corrigeren van de dekkingsfout van het steekproefontwerp ten aanzien van de steden bij regressie-analyse met gewichten: simulaties

Gebleken is dat met behulp van gewichten de marginale frequenties in de steekproef kunnen worden aangepast aan die van de populatie. We onderzoeken in hoeverre deze gewichten bij regressie-analyse nuttig kunnen zijn. De methode van gewogen regressie met RAS-gewichten wordt getoetst op het CCO-bestand. Gekeken wordt of het invoeren van de gewichten de fout in de coëfficiënten veroorzaakt door het uitsluiten van respondenten uit de stad corrigeert. We nemen daarvoor aan dat het volledige CCO-bestand de doelgroep is waar we de relatie voor willen bepalen.

Met het totale CCO-bestand wordt van de onderzochte regressie-vergelijking de coëfficiënten vector geschat met de (ongewogen) kleinste kwadraten methode.

Om de situatie van de GPD-enquête te simuleren nemen we van het CCO-bestand

weer het deelbestand bestaande uit mensen die in gemeenten wonen met minder dan 100.000 inwoners. Dit sub-bestand van niet-stedelingen wordt op de manier als beschreven in § 3.1 herwogen tot de doelgroep: het totale CCO-bestand. De karakteristieken voor het herwegen zijn leeftijd en opleiding in dezelfde categorieën opgesplitst als bij het herwegen van de GPD-enquête (zie tabel 9). De politieke voorkeur van de respondenten is in de CCO-enquête niet gevraagd. Voor het deelbestand is op de regressie-vergelijking zowel de ongewogen als de gewogen kleinste kwadraten methode toegepast. De resultaten staan in tabel 11 en 12.

Bij het berekenen van de gewichten bleek dat de relatieve frequenties van de niet-steden en die van het totale bestand niet veel van elkaar verschilden. De gewichten per cel variëren dan ook rond 1.0; met een minimum van 0.8 en een maximum van 1.8. De coëfficiënten in de inkomenswaarderingsrelatie van de niet-steden veranderen na herweging dan ook maar weinig. Het verschil van de constante term β_0 tussen de regressie op het totale bestand en de herwogen versie van het deelbestand is kleiner geworden dan het verschil tussen het totale bestand en de ongewogen versie. Het herwegen werkt bij β_1 [$\ln(fs)$] in de verkeerde richting terwijl het β_2 zeer goed corrigeert.

Ook in tabel 12 verschillen de coëfficiënten van de ongewogen en de gewogen versie niet veel. Het herwegen geeft als resultaat dat, met uitzondering van β_4 'mbo' en β_{12} 'zelfstandigen etc', de coëfficiënten van de regressie op het deelbestand na het herwegen meer overeen komen met de regressie op het totale bestand dan voor het herwegen. Opvallend is de goede correctie van β_7 en β_8 , die met leeftijd te maken hebben, een van de herwegingskarakteristieken, en β_{14} 'man'.

Tabel 11. Resultaten inkomenswaarderings-vergelijking CCO'79
Ongewogen en gewogen regressie op het deelbestand niet-steden

afhankelijke variabele: mu		totaal CCO'79 ongewogen		alleen niet steden CCO'79 ongewogen gewogen			
variabele		parm. std.afw.		parm. std.afw.		parm. std.afw.	
constante	β_0	3.334	0.104	3.394	0.122	3.319	0.122
$\ln(fs)$	β_1	0.088	0.008	0.089	0.009	0.093	0.010
$\ln(y_{gez})$	β_2	0.648	0.010	0.641	0.012	0.647	0.012
R^2		0.7208		0.7082		0.7154	
N		2050		1502		1502	

Tabel 12. Resultaten inkomensvergelijking van de hoofdkostwinner CCO'79
Ongewogen en gewogen regressie op het deelbestand niet-steden

afhankelijke variabele: ln(inkomen hkw)		totaal CCO'79 ongewogen		alleen niet steden CCO'79 ongewogen gewogen			
variabele		parm.	std.afw.	parm.	std.afw.	parm.	std.afw.
constante	β_0	-0.165	0.761	0.967	0.919	0.309	0.890
lbo lts huishoud	β_1	0.091	0.020	0.082	0.024	0.083	0.024
ulo mulo mavo	β_2	0.244	0.026	0.260	0.030	0.255	0.030
havo hbs lyceum	β_3	0.224	0.032	0.306	0.041	0.294	0.036
mbo	β_4	0.223	0.023	0.217	0.026	0.214	0.027
hbo	β_5	0.377	0.025	0.425	0.029	0.422	0.029
universiteit	β_6	0.606	0.037	0.620	0.043	0.615	0.043
ln(leeftijd)	β_7	5.041	0.413	4.513	0.504	4.847	0.485
ln(leeftijd) ²	β_8	-0.636	0.056	-0.577	0.068	-0.618	0.065
geen werk	β_9	-0.237	0.026	-0.199	0.031	-0.213	0.031
geschoold arbeider	β_{10}	0.007	0.029	-0.007	0.032	-0.001	0.033
employee	β_{11}	0.107	0.025	0.093	0.028	0.096	0.029
zelfst./direc.							
/vrij beroep	β_{12}	0.023	0.033	0.025	0.037	0.053	0.038
landbouw./tuinder	β_{13}	-0.072	0.043	-0.059	0.044	-0.061	0.046
man	β_{14}	0.243	0.018	0.257	0.023	0.247	0.023
R ²		0.5374		0.5237		0.5317	
N		2050		1502		1502	

Merk op dat bij β_{10} door alleen de niet-steden te nemen het teken omklapt. In alle regressies is β_{10} echter niet significant verschillend van nul.

3.5 Resultaten met gewogen regressie-analyse

De simulatie op het CCO-bestand laat zien dat gewogen regressie-analyse een verbetering te zien geeft, met name voor variabelen die als herwegingskarakteristiek zijn gebruikt. In deze paragraaf vergelijken we de resultaten van de regressie-vergelijkingen op het GPD-bestand met die van de overeenkomstige relaties van het PANEL- en het CCO-bestand.

Zowel de gewogen als de ongewogen kleinste kwadraten methode zijn op de relaties toegepast.

De verschillen in de coëfficiënten bij de ongewogen versie van de inkomenswaarderings-vergelijking toegepast op de diverse bestanden verdwijnen nagenoeg voor de constante en de coëfficiënt van 'ln(ygez)'. De ongelijkheid van β_1 [ln(fs)] blijft echter bestaan. Voor kinderloze huishoudens treden afwijkende frequenties op tussen de verschillende bestanden. Door gezinsgrootte als een der herwegingskarakteristieken te kiezen zou dit probleem opgelost kunnen worden.

Tabel 13. Resultaten inkomenswaarderings-vergelijking
(ongewogen en gewogen versies)

a. ongewogen

afhankelijke variabele: mu

variabele		GPD'83		PANEL'83		CCO'79	
		parm.	std.afw.	parm.	std.afw.	parm.	std.afw.
constante	β_0	2.930	0.049	3.056	0.233	3.334	0.104
ln(fs)	β_1	0.037	0.004	0.070	0.017	0.088	0.008
ln(y _{gez})	β_2	0.695	0.005	0.681	0.023	0.648	0.010
R ²		0.6954		0.7292		0.7208	
N		11380		466		2050	

b. gewogen

afhankelijke variabele: mu

variabele		GPD'83		PANEL'83		CCO'79	
		parm.	std.afw.	parm.	std.afw.	parm.	std.afw.
constante	β_0	3.115	0.048	3.135	0.231	3.292	0.102
ln(fs)	β_1	0.030	0.004	0.063	0.017	0.090	0.008
ln(y _{gez})	β_2	0.677	0.005	0.672	0.023	0.652	0.010
R ²		0.6835		0.7478		0.7354	
N		11380		466		2050	

Door het grote aantal categorieën in de inkomensvergelijking van de hoofdkostwinner (tabel 14) is het lastig een vergelijking te maken tussen de relaties van de diverse bestanden. We gaan daarom stapsgewijs de coëfficiënten langs.

- De grootste verschillen bevinden zich in de coëfficiënten β_0 'constante', β_7 'ln(leef tijd)' en β_8 'ln(leef tijd)²'. In eerste instantie lijken de coëfficiënten van het PANEL- en het CCO-bestand het meest op elkaar.

ln(inkomen) is waarschijnlijk een kwadratische functie van ln(leef tijd) want in alle versies van de relatie zijn de coëfficiënten β_7 en β_8 significant van nul verschillend. Bekijken we de drie coëfficiënten β_0 , β_7 en β_8 gezamenlijk bij het constant houden van de overige variabelen dan is het verschil tussen de relaties te onderzoeken met de vergelijking $x = \beta_0 + \beta_7 \ln(z) + \beta_8 \ln(z)^2$ waarbij z de leef tijd en x ln(inkomen).

Nu blijkt dat voor de relevante leeftijden van hoofdkostwinners (20-70 jaar) de x waarden verkregen met de coëfficiënten van het GPD en het CCO het minst verschillen (tabel 15). Het herwegen heeft als gevolg dat de verschillen in x vanaf 45 jaar kleiner worden.

Tabel 14. Resultaten inkomensvergelijking van de hoofdkostwinner
(ongewogen en gewogen versies)a. ongewogen

afhankelijke variabele: ln(inkomen hkw)

variabele		GPD'83		PANEL'83		CCO'79	
		parm.	std.afw.	parm.	std.afw.	parm.	std.afw.
constante	β_0	2.078	0.373	-0.922	2.211	-0.165	0.761
lbo lts huishoud	β_1	0.059	0.010	0.067	0.041	0.091	0.020
ulo mulo mavo	β_2	0.133	0.010	0.170	0.066	0.244	0.026
havo hbs lyceum	β_3	0.237	0.011	0.142	0.069	0.224	0.032
mbo	β_4	0.157	0.011	0.132	0.043	0.223	0.023
hbo	β_5	0.293	0.010	0.282	0.053	0.377	0.025
universiteit	β_6	0.512	0.013	0.430	0.069	0.606	0.037
ln(leeftijd)	β_7	3.668	0.204	5.482	1.184	5.041	0.413
ln(leeftijd) ²	β_8	-0.428	0.028	-0.688	0.158	-0.636	0.055
geen verk	β_9	-0.096	0.020	-0.216	0.048	-0.237	0.027
geschoold arbeider	β_{10}	0.120	0.020	-	-*	0.007	0.029
employee	β_{11}	0.202	0.019	0.156	0.041	0.107	0.025
zelfst./direc.							
/vrij beroep	β_{12}	0.296	0.022	0.119	0.067	0.023	0.033
landbouw./tuinder	β_{13}	0.050	0.051	0.261	0.091	-0.072	0.043
man	β_{14}	0.238	0.009	0.170	0.039	0.243	0.018
R ²		0.4792		0.4569		0.5374	
N		11380		466		2050	

b. gewogen

afhankelijke variabele: ln(inkomen hkw)

variabele		GPD'83		PANEL'83		CCO'79	
		parm.	std.afw.	parm.	std.afw.	parm.	std.afw.
constante	β_0	3.119	0.316	-0.529	1.910	-0.695	0.758
lbo lts huishoud	β_1	0.058	0.007	0.095	0.039	0.097	0.020
ulo mulo mavo	β_2	0.132	0.009	0.228	0.048	0.246	0.026
havo hbs lyceum	β_3	0.244	0.009	0.100	0.049	0.200	0.032
mbo	β_4	0.173	0.011	0.150	0.064	0.225	0.023
hbo	β_5	0.316	0.009	0.306	0.050	0.387	0.025
universiteit	β_6	0.530	0.012	0.498	0.064	0.606	0.039
ln(leeftijd)	β_7	3.143	0.172	5.343	1.033	5.304	0.410
ln(leeftijd) ²	β_8	-0.364	0.023	-0.677	0.139	-0.668	0.055
werkloos	β_9	-0.075	0.015	-0.226	0.047	-0.246	0.027
geschoold arbeider	β_{10}	0.124	0.015	-	-*	0.008	0.029
employee	β_{11}	0.188	0.015	0.124	0.041	0.107	0.025
zelfst./direc.							
/vrij beroep	β_{12}	0.245	0.019	0.198	0.086	0.024	0.036
landbouw./tuinder	β_{13}	0.084	0.057	0.155	0.125	-0.078	0.045
man	β_{14}	0.250	0.009	0.123	0.036	0.231	0.018
R ²		0.4565		0.4932		0.5539	
N		11380		466		2050	

* In de PANEL-enquête is bij de beroepsvraag geen uitsplitsing gemaakt in ongeschoold en geschoold arbeider. De referentiegroep hier is de totale groep van arbeiders.

Tabel 15. Vergelijking van log-inkomens als functie van leeftijd

$$x = \beta_0 + \beta_7 \ln(z) + \beta_8 \ln(z)^2$$

z	ongewogen			gewogen		
	gpd	panel	cco	gpd	panel	cco
20	9.23	9.32	9.23	9.27	9.40	9.20
25	9.45	9.60	9.47	9.46	9.65	9.46
30	9.60	9.76	9.62	9.60	9.81	9.62
35	9.71	9.87	9.72	9.69	9.91	9.72
40	9.78	9.94	9.78	9.76	9.97	9.78
45	9.84	9.98	9.81	9.81	10.00	9.82
50	9.88	9.99	9.82	9.84	10.01	9.83
55	9.90	10.00	9.82	9.87	10.01	9.83
60	9.92	9.99	9.81	9.89	10.00	9.82
65	9.93	9.97	9.80	9.90	9.98	9.81
70	9.94	9.95	9.77	9.90	9.95	9.78

- De totaal verschillende coëfficiënten β_1 ('lbo lts huishoudschool') worden door herwegen gelijk bij het PANEL en het CCO.

- In β_2 ('ulo mulo mavo'), β_3 ('havo hbs lyceum') en β_4 ('mbo') blijven de verschillen na herwegen bestaan.

- De gewichten zorgen ervoor dat β_5 ('hbo') in alle drie de bestanden verhoogd worden en dichter bij elkaar komen. Hetzelfde geldt voor β_6 ('universiteit') waar alleen de (hoge) waarde in het CCO constant blijft.

- β_9 ('geen werk') is bij het GPD-bestand beduidend lager t.o.v. de twee andere bestanden.

- In het PANEL-enquête is geen onderscheid gemaakt tussen ongeschoolde en geschoolde arbeiders. Tussen het GPD- en het CCO-bestand bestaat in de coëfficiënt β_{10} een groot verschil. Dit verschil is een gevolg van de afwijkende aantallen ongeschoolde arbeiders in de beide bestanden. De coëfficiënt β_{10} is in de relatie van het CCO overigens niet significant van nul verschillend.

- In de coëfficiënten β_{11} ('employee') en β_{12} ('zelfstandig/directeur/ vrij beroep') vertoont het CCO-bestand een verschillend beeld t.o.v. het GPD en het PANEL. Na herwegen blijft de coëfficiënt β_{11} in het CCO onveranderd, in het PANEL- en het GPD-bestand wordt β_{11} verlaagd en worden de verschillen kleiner, ook t.o.v. het CCO.

Het verschil in β_{12} tussen PANEL en GPD wordt door het wegen sterk verminderd.

- In de ongewogen versies is β_{13} ('landbouwer/tuinder') alleen bij het PANEL significant van nul verschillend. Na het herwegen is de coëfficiënt in alle drie de bestanden niet significant van nul verschillend. Merk overigens op dat de coëfficiënt alleen in de CCO-versie negatief is.

- Het PANEL verschilt in β_{14} ('man') nogal met de andere twee bestanden. Bij het herwegen wordt β_{14} bij het GPD verhoogd en bij het PANEL en het CCO-bestand verlaagd.

Uit de tabellen 13 en 14 is af te lezen dat de verschillen in de relaties van de diverse bestanden over het algemeen niet groot zijn en dat door het herwegen de verschillen kleiner worden. Uit de resultaten is te concluderen dat het herwegen met RAS-gewichten de voorkeur verdient boven ongewogen regressie. Er blijven echter nog altijd verschillen bestaan.

Mogelijk kunnen de resultaten met een andere keuze van de herwegingskarakteristieken nog meer verbeterd worden. Verschillen zullen er altijd blijven daar de bestanden steekproeven van de populatie zijn. Te zien is ook dat regressie-analyse op het PANEL- en het GPD-bestand in de relaties verschillen geven van dezelfde orde van grootte als de verschillen tussen PANEL en CCO of GPD en CCO. Bezwaren op grond van representativiteit van het GPD-bestand gelden dan ook in even grote mate voor deze andere bestanden.

4. Conclusies

In dit rapport is de representativiteit van een dagbladenquête onderzocht.

De resultaten uit de dagbladenquête van de Gemeenschappelijke Pers Dienst zijn in dit onderzoek vergeleken met resultaten afkomstig van berekeningen op twee steekproeven van het CBS die getrokken zijn voor het PANEL-onderzoek (1983) en het Consumenten Conjunctuur Onderzoek (1979). Verder is als vergelijkingsmateriaal gebruik gemaakt van gegevens uit het statistisch zakboek 1983 en gegevens afkomstig van CEBUCO (1981).

Het belangrijkste verschil tussen de gebruikelijke manier van enquêteren (door persoonlijke benadering van een enquêteur/trice) en de dagbladmethode (een vragenlijst in de krant) is het ongebruikelijke steekproefontwerp van de dagbladenquête. Men kan ondervertegenwoordigde groepen van bepaalde sociale klassen niet aanvullen. Bovendien worden door de manier van enquêteren de mensen die geen krant lezen en mensen die niet in het bezorgingsgebied van de krant wonen bij voorbaat van deelname uitgesloten, hetgeen met onderdekking wordt aangeduid.

De nadruk bij dit onderzoek ligt op de gevolgen van het door deze methode veroorzaakte uitsluiten van de mensen die niet in het verspreidingsgebied van de krant wonen.

Met de methode van Heckman is de invloed van deze onderdekking op twee relaties onderzocht, te weten de samenhang tussen inkomenswaardering en gezinsgrootte en -inkomen en het verband van het inkomen van de hoofdkostwinner met een aantal verklarende variabelen waaronder opleiding, beroep en leeftijd. Daar de selectievergelijking geen goed onderscheid mogelijk maakt tussen mensen die wel en mensen die niet in het bezorgingsgebied van de krant wonen heeft een correctie voor non-respons met de methode van Heckman geen zin.

Door het uitzonderlijke steekproefontwerp bij een dagbladenquête is het zeer waarschijnlijk dat de relatieve frequenties die in de steekproef gevonden worden niet overeen zullen komen met de relatieve frequenties in de populatie.

Met de RAS-methode is het mogelijk gewichten te berekenen zodanig dat met deze gewichten de marginale frequenties in de steekproef van een aantal gekozen karakteristieken gelijk worden aan de marginale frequenties in de populatie. Bij dit onderzoek zijn leeftijd, opleiding en politieke voorkeur gebruikt om de GPD-enquête te herwegen. De karakteristieken kunnen echter zelf gekozen worden afhankelijk van een aantal voorwaarden, waaronder betrouwbare 'ijk'gegevens, en het doel van het onderzoek waarvoor de steekproef gebruikt gaat worden.

Met de door ons berekende gewichten is wat betreft het bepalen van frequenties en gemiddelden de GPD-enquête zeker zo representatief als de andere enquêtes. Bij enkele subgroepen is door ondervertegenwoordiging voorzichtigheid geboden bij het conclusies trekken over deze groep ten opzichte van de totale steekproef c.q. populatie, met name bij de landbouwers en de groep van vrouwelijke hoofdkostwinners. Door andere herwegingskarakteristieken te kiezen zou dat probleem verholpen kunnen worden. De grootte van de GPD-steekproef heeft als consequentie dat ook de ondervertegenwoordigde subgroepen in absolute aantallen meestal beter vertegenwoordigd zijn dan in andere steekproeven. Daardoor kan onderzoek binnen deze groepen met het GPD-bestand zonder meer verricht worden, met het voorbehoud dat deelbestanden alleen dan representatief zijn voor de betreffende deelpopulaties als de gebruikte herwegingskarakteristieken onafhankelijk zijn.

De nadelige invloed van het uitsluiten van respondenten uit steden met meer dan 100.000 inwoners, de voornaamste groep binnen het niet-bezorgingsgebied van de GPD, op de relaties is door middel van gewogen regressie te verminderen. Over de representativiteit van de GPD-enquête bij het gebruik in regressie-analyse is het moeilijk om al definitieve conclusies te trekken. We hebben

als vergelijkingsmateriaal voor de twee regressievergelijkingen alleen de beschikking over de resultaten afkomstig van berekeningen op het PANEL- en het CCO-bestand.

Hoewel de herweging ervoor zorgt dat de schatters van de coëfficiënten van de relaties meer op elkaar gaan lijken blijven er tussen de drie steekproeven verschillen bestaan. De verschillen zijn tussen de drie bestanden onderling echter even groot. Het is niet mogelijk te bepalen welke van de drie bestanden de samenhang tussen de grootheden in de werkelijke populatie het beste beschrijft.

Concluderend kunnen we stellen dat uit dit onderzoek gebleken is dat de herwogen dagbladenquête van het GPD niet minder representatief is voor de totale Nederlandse bevolking dan andere steekproeven, via de traditionele methoden getrokken.

Literatuurlijst

- Centraal Bureau voor de Statistiek, Statistisch Zakboek 1983, Staatsuitgeverij, 's-Gravenhage.
- CEBUCO (1981), Oplage specificatie en verspreidingscijfers Dagbladen.
- Deming, W.E. en F.F. Stephan (1940), "On a least squares adjustment of a sampled frequency table when the expected marginal totals are known", Annals of Mathematical Statistics II, pp.427-444.
- Heckman, J.J. (1979), "Sample selection bias as a specification error", Econometrica 47, pp.153-161.
- Ireland, C.T. en S. Kullback (1968), "Contingency tables with given marginals", Biometrika 55, pp.179-188.
- Johnson, N. en S. Kotz (1972), Distribution in Statistics: continuous multivariate distributions, Wiley, New York.
- Kersten, H.M.P. en J.G. Bethlehem (1982), "Foutenbronnen in steekproefonderzoek", Jaarboek van de Nederlandse Vereniging van Marktonderzoekers, pp.49-65.
- Leyden Income Evaluation Project (1984), Summary of Activities, Centrum voor Onderzoek van de Economie van de Publieke Sector.
- Maddala, G.S. (1983), Limited-dependent and qualitative variables in econometrics, Cambridge University Press.
- Montgomery, D.C. en E.A. Peck (1982), Introduction to linear regression analysis, Wiley, New York.
- Stadt, H. van de en T.J. Wansbeek (1982), "Het effect van selectieve non-respons op het resultaat van een regressie-analyse", Kwantitatieve Methoden, vol.3, pp.109-124.
- Stone, R. en A. Brown (1964), "A computable model of econometric growth", A programme for growth, vol.1, Londen.
- Theil, H. (1967), Economics and Information Theory, North Holland Publishing Company, Amsterdam.

Ontvangen: 24-09-1986

Geaccepteerd: 14-04-1988