KM 7 (1982)
pag. 109-124

centraal bureau voor de statistiek

Hoofdafdeling Statistieken van Inkomen en Consumptie
Hoofdafdeling Statistische Methoden
Postbus 959, 2270 AZ VOORBURG

HET EFFECT VAN SELECTIEVE NON-RESPONS OP DE RESULTATEN VAN EEN REGRESSIEANALYSE

Huib van de Stadt
Tom Wansbeek

De meningen en standpunten in dit rapport zijn die van de auteurs en komen niet noodzakelijk overeen met die van het Centraal Bureau voor de Statistiek. De auteurs danken Wynand van de Ven en Jacques Thijssen voor hun medewerking bij de totstandkoming van dit rapport, en Albert Verbeek, Wouter Keller en de referees van KM voor hun uitgebreide en constructieve commentaar op een eerdere versie van dit rapport. De programmeerwerkzaamheden werden verricht door Eitel Homan op het Centrum voor Onderzoek van de Economie van de Publieke Sector van de Rijksuniversiteit te Leiden.

Proj. S9-79-298

Maart 1982

Herziene versie

HET EFFECT VAN SELECTIEVE NON-RESPONS OP DE RESULTATEN VAN EEN REGRESSIEANALYSE

Samenvatting

Wanneer een onderzoeker een regressievergelijking wil schatten op basis van een steekproef, zullen de kleinste-kwadraten-schatters inconsistent zijn als het wel of niet responderen bij de steekproef samenhangt met de storingsterm in de regressievergelijking. Een recent door Heckman (1979) ontwikkelde methode maakt het mogelijk voor een dergelijke samenhang te corrigeren en alsnog consistente schatters te verkrijgen. In dit rapport wordt de Heckman-methode uiteengezet en toegepast op een in het kader van het onderzoek Inkomenswaardering van het CBS dikwijls geschatte regressievergelijking.

1. Inleiding

Wanneer men een regressievergelijking wil schatten op basis van individuele gegevens die uit een steekproef afkomstig zijn, kunnen de resultaten worden beïnvloed door eventueel optredende non-respons. Wanneer het al of niet responderen onafhankelijk is van de storingsterm van de regressievergelijking (maar eventueel wel gecorreleerd met de onafhankelijke variabelen), dan is er geen groot probleem. Omdat aan alle assumpties van het regressiemodel is voldaan, zijn de kleinste-kwadratenschatters nog steeds consistent; alleen de nauwkeurigheid van de schatters kan suboptimaal zijn (zie bijvoorbeeld Ten Cate, 1981). Indien de non-respons echter selectief is en wél samenhangt met de storingsterm, treedt 'selection bias' op en zijn de schatters niet meer consistent.

Met behulp van een recent door Heckman (1979) ontwikkelde methode kan voor deze 'selection bias' gecorrigeerd worden ten einde alsnog consistente schatters te verkrijgen. Om de methode toe te passen is het echter noodzakelijk individuele informatie over de non-respondenten te hebben en deze zal doorgaans niet aanwezig zijn. In dit rapport zetten wij de methode van Heckman uiteen en laten zien hoe het probleem van ontbrekende informatie door middel van een benaderingsformule opgelost kan worden. De methode wordt toegepast op een regressievergelijking die dikwijls geschat wordt in het kader van het onderzoek Inkomenswaardering van het CBS.

De inhoud van dit rapport is als volgt. In paragraaf 2 en 3 wordt de gevolgde methode weergegeven, terwijl in paragraaf 4 de resultaten zijn vermeld. In paragraaf 5 staan enkele conclusies. De benaderingsformule is afgeleid in appendix A en de voor de analyse benodigde data zijn opgenomen in appendix B.

2. De methode

De relatie die zal worden onderzocht is de volgende (zie bijvoorbeeld Van Praag en Kapteyn, 1973, Van de Stadt, 1981, Kapteyn en Wansbeek, 1982):

$$\mu = \beta_0 + \beta_1 \log fs + \beta_2 \log y + \varepsilon. \qquad (1)$$

Hierin is fs de gezinsgrootte (family size), y het netto gezinsinkomen en μ

de logaritme van het inkomen dat met een 5 (bijna voldoende) wordt gewaardeerd.[1] De grootheid μ kan worden opgevat als een maat voor het behoeftenniveau van het huishouden; een hoge μ impliceert dat het gezin een hoog inkomen nodig heeft om een bepaald niveau van tevredenheid te bereiken. De storingsterm ε representeert de onverklaarde bijdrage aan μ.

Vergelijking (1) willen we schatten voor de gehele populatie, in dit geval de verzameling van alle Nederlandse huishoudens.[2] Hiertoe is een enquête gehouden onder een aselecte steekproef uit deze populatie. Gezien de aard van de vragen die moeten worden gesteld om μ te kunnen berekenen, is de respons op dit soort enquêtes (althans bij deze vragen) nogal laag ($\pm$ 60%) en loont het de moeite om na te gaan of de non-respons vertekenend werkt op de schattingsresultaten. Het responsgedrag van de gezinnen die in de steekproef gevallen zijn wordt beschreven door vergelijking (2)

$$a = \sum_{j=0}^{k} \gamma_j x_j + \delta. \qquad (2)$$

Hierin is a een latente continue variabele die de 'responsgeneigdheid' aangeeft. Als $a \leq 0$ zal het betreffende huishouden niet responderen, als $a > 0$ is er sprake van respons. De x_j's zijn een aantal variabelen die de respons zouden kunnen beïnvloeden (bijvoorbeeld inkomen, gemeentegrootte); $x_0 = 1$ is de constante term en δ is een storingsterm.

Om het model te kunnen schatten is het noodzakelijk een veronderstelling te maken over de verdeling van de storingstermen. We veronderstellen dat ε en δ een gezamenlijke tweedimensionale normale verdeling hebben met verwachtingen nul, varianties σ^2_ε en 1 en covariantie $\sigma_{\varepsilon\delta}$. Aangezien de schaling van de responsgeneigdheid a niet vastligt, kan de variantie van δ zonder verlies van algemeenheid gelijk worden gesteld aan 1.

---

1) Merk op dat het symbool μ wordt gebruikt om een waargenomen variabele aan te duiden, en dus niet (anders dan de andere Griekse symbolen) een modelparameter aangeeft. Deze notatie is de op dit onderzoeksgebied gebruikelijke.

2) De begrippen gezin, huishouden en individu worden zonder onderscheid gebruikt. Met (non-) respondenten bedoelen we huishoudens die (niet) aan de enquête deelnamen.

Essentieel bij de assumptie over de verdeling van ε en δ is dat de varianties en covarianties niet afhangen van fs, y of de x_j's. Deze veronderstelling is moeilijk te toetsen, maar lijkt wel plausibel.

In termen van het gezamenlijke model (1) en (2) betekent selectieve nonrespons dat $\sigma_{\varepsilon\delta}$ ongelijk aan nul is. We zullen nu trachten om door middel van de methode van Heckman een schattingsprocedure voor $\sigma_{\varepsilon\delta}$ te ontwikkelen en vervolgens te toetsen of de empirisch gevonden schatting significant van nul verschilt. De redenering van Heckman is als volgt:

Vergelijking (1) geldt voor de hele populatie en het zijn de populatieparameters die ons interesseren. We hebben echter alleen maar de respondenten en daarvoor geldt niet dat de verwachting van de storingsterm ε nul is. In plaats van (1) krijgen we

$$E(\mu|\text{individu respondeert}) = \beta_0 + \beta_1 \log fs + \beta_2 \log y + \\ + E(\varepsilon|\text{individu respondeert}), \qquad (3)$$

Voor de laatste term van (3) geldt:

$$E(\varepsilon|\text{individu respondeert}) = E(\varepsilon|a > 0) = \\ = E(\varepsilon|\delta > -\sum_{j=0}^{k} \gamma_j x_j) = \\ = \sigma_{\varepsilon\delta} \cdot \frac{n(Z)}{N(Z)}, \qquad (4)$$

met $Z = \sum_{j=0}^{k} \gamma_j x_j$, $n(.)$ de standaardnormale dichtheidsfunctie en $N(.)$ de standaardnormale verdelingsfunctie. Het laatste gelijkteken in (4) is een bekende eigenschap van de binormale verdeling, zie bijvoorbeeld Johnson en Kotz (1972, p. 112). De regressievergelijking voor de respondenten wordt nu

$$\mu = \beta_0 + \beta_1 \log fs + \beta_2 \log y + \sigma_{\varepsilon\delta} \frac{n(Z)}{N(Z)} + \zeta, \qquad (5)$$

waarbij ζ wél verwachting nul heeft. Indien nu de waarde van Z voor iedere respondent bekend zou zijn, dan kan door middel van (5) een waarde voor $\sigma_{\varepsilon\delta}$ worden geschat. Om Z voor iedere respondent te kunnen berekenen zijn eerst schattingen voor de γ_j's nodig. Hiertoe moet de responsvergelijking (2) voor de steekproef van respondenten plus non-respondenten worden geschat. In de volgende paragraaf wordt aangegeven hoe dit op grond van een populatiecovariantiematrix en de populatie frequenties van de variabelen x_j kan geschieden. Samenvattend luidt de schattingsprocedure:

a) Schat met behulp van een populatiecovariantiematrix vergelijking (2) voor de steekproef van respondenten plus non-respondenten. Dit levert consistente schattingen op voor de γ_j's.

b) Bereken voor alle respondenten de zgn. Heckman-term $Z = \sum_{j=0}^{k} \hat{\gamma}_j x_j$ en vervolgens $n(Z)/n(Z)$.

c) Schat (5) voor alle respondenten. Dit levert consistente schattingen op voor β_0, β_1, β_2 en $\sigma_{\varepsilon\delta}$. Als deze laatste schatting significant van nul verschilt is er sprake van invloed van selectieve non-respons op de regressievergelijking (1).

3. Het schatten van de responsvergelijking

De responsvergelijking (2) en de normaliteitsassumptie definieren een probit-model, zodat (2) eigenlijk door middel van probitanalyse geschat zou moeten worden. Hiervoor is echter individuele informatie over de waarden van de variabelen x_j voor de non-respondenten nodig. Omdat deze informatie in het algemeen niet aanwezig is, volgen we hier een andere methode, waarvoor geen individuele informatie nodig is, namelijk een eerste-orde benadering van probit. Hierbij doet zich de (althans voor dit doel) gelukkige omstandigheid voor dat de non-respond zich in de buurt van een half bevindt, zodat de benadering redelijk lijkt en mag worden verwacht dat de resultaten lijken op die van probit.

De gebruikte methode werkt als volgt. In plaats van (2) wordt een overeenkomstige regressievergelijking geschat waarin de afhankelijke variabele de waarde 1 of 0 (respons of non-respons) aangeeft. De resultaten van deze regressie worden vervolgens op een in appendix A beschreven wijze gebruikt om de probitcoëfficiënten te berekenen.

In het geval van regressie van een (0,1)-variabele is individuele informatie over non-respondenten niet nodig, maar kan volstaan worden met een covariantiematrix voor respondenten plus non-respondenten. Dit kan als volgt worden aange-

toond. Zij a^* een (0,1)-variabele, met $a^*=1$ voor een respondent en $a^*=0$ voor een non-respondent, dan wordt uitgegaan van de volgende regressie (in matrix-notatie):

$$a^* = X\alpha + \zeta, \tag{6}$$

met a^* I×1, X I×k, α k×1, en ζ I×1 een storingsterm; de eerste I_1 elementen (cq. rijen) van a^*, X en ζ hebben betrekking op de respondenten, de overige op de non-respondenten. Er geldt

$$a^* = \begin{pmatrix} \iota \\ 0 \end{pmatrix} \begin{matrix} I_1 \\ I-I_1 \end{matrix}, \tag{7}$$

met ι een I_1-vector van enen. Splits X:

$$X = \begin{pmatrix} X_1 \\ X_0 \end{pmatrix} \begin{matrix} I_1 \\ I-I_1 \end{matrix}, \tag{8}$$

dan geldt voor de kleinste-kwadratenschatter $\hat{\alpha}$ van α:

$$\hat{\alpha} = (X'X)^{-1}X'a^* = (X'X)^{-1}(X_1', X_0')\begin{pmatrix} \iota \\ 0 \end{pmatrix} = (X'X)^{-1}X_1'\iota . \tag{9}$$

Deze schatter bestaat dus uit twee elementen, namelijk $X_1'\iota$ en X'X. Daarvan heeft $X_1'\iota$ alleen betrekking op de respondenten en is dus bekend. X'X is echter niet bekend, want dat is I maal de matrix van 2^e momenten voor respondenten plus non-respondenten. Aangezien de steekproef (van respondenten plus non-respondenten) aselect getrokken was uit de gehele populatie mag echter verondersteld worden dat de momentenmatrix voor respondenten plus non-respondenten niet veel zal verschillen van de momentenmatrix voor de gehele populatie. De procedure die we gevolgd hebben is dat X'X berekend wordt uit een covariantiematrix afkomstig van een ander, groot onderzoek, dat veel minder door non-respons wordt geplaagd. Hiervoor kozen we het Woningbehoeftenonderzoek (WBO) van het CBS. We nemen dus aan dat de covariantiematrix uit het WBO een goede benadering is van de covariantiematrix voor respondenten plus non-respondenten van het inkomenswaarderingsonderzoek.

De keuze voor het WBO bracht enkele nadelen met zich mee. Het onderzoek vertoont niet veel, maar toch nog altijd wel enige non-respons, het is in een ander jaar gehouden en de gebruikte vraagstellingen verschillen enigszins. Op deze technische problemen wordt hier verder niet ingegaan. We verwachten echter dat de op deze manier geschatte $\hat{\alpha}$ een goede benadering is van de werkelijke α.

Zoals gezegd wordt in appendix A aangetoond hoe $\hat{\alpha}$ wordt gebruikt om schattingen voor γ in (2) (de probitvergelijking) af te leiden.

4. De resultaten

Vergelijking (6) is geschat met gebruikmaking van de covariantiematrix uit het WBO en de gemiddelde waarde van de achtergrondvariabelen uit het Inkomenswaarderingsonderzoek. Als achtergrondvariabelen zijn de variabelen die waarschijnlijk invloed hadden op de respons en die in beide onderzoeken voorkwamen gekozen. De regressieresultaten staan vermeld in tabel 1.

Tabel 1. Resultaten regressie responsvergelijking (6)

Variabele	$\hat{\alpha}_j$	standaardafwijking van $\hat{\alpha}_j$
1) Hoofd huishouden = vrouw	-0,00	0,04
2) Hoofd huishouden = gehuwd	-0,00	0,04
3) Grootte huishouden = 1 persoon	-0,09	0,04
4) Grootte huishouden ≥ 5 personen	-0,07	0,03
5) Inwonend huishouden	-0,31	0,06
6) Leeftijd hoofd huishouden < 30	-0,05	0,06
7) Leeftijd hoofd huishouden ≥ 65	-0,18	0,03
8) Inkomen beneden 29^{e} percentiel	0,19	0,03
9) Inkomen boven 72^{e} percentiel	-0,05	0,02
10) Gemeentegrootte > 50 000	-0,07	0,02
11) Hoofd huishouden = zelfstandige	-0,21	0,04
12) Hoofd huishouden = niet-werkzaam	-0,18	0,03
Constante term	0,59	0,05
Aantal respondenten	1 008	
Aantal respondenten plus non-respondenten	2 332	
R^2	0,08	

Alle variabelen zijn, om later te vermelden redenen, dummyvariabelen, d.w.z. zij hebben de waarde 1 indien de desbetreffende eigenschap wel van toepassing is en de waarde 0 indien zij niet van toepassing is. De regressiecoëfficiënten α_j kunnen worden geïnterpreteerd als het effect van de bijbehorende variabele op de respons, onder constanthouding van de overige variabelen. De resultaten van tabel 1 zijn in overeenstemming met de verwachtingen. Alleenstaanden, grote gezinnen, onwonende huishoudens, 65^+-ers, inwoners van grote steden, zelfstandigen en niet-werkzamen responderen significant slechter. Met name het effect van de variabelen 'inwonend huishouden' en 'zelfstandige' is opmerkelijk groot (respectievelijk 31 en 21% lagere respons).

Uit tabel 1 blijkt duidelijk dat een aantal variabelen significant samenhangen met de respons. Wat is nu het effect van deze selecte respons op de regressievergelijking (1)? Om deze vraag te beantwoorden berekenen we eerst uit de $\hat{\alpha}_j$ schattingen voor de γ_j (in vergelijking (2)), en vervolgens voor iedere respondent de 'Heckman-term' (4) met behulp van de geschatte γ_j. Vervolgens is de vergelijking (1) voor alle respondenten zowel met als zonder deze Heckman-term geschat. De resultaten staan in tabel 2. Met betrekking tot de standaarddeviaties die in de laatste kolom van tabel 2 gegeven worden is enige voorzichtigheid geboden. Deze standaarddeviaties zijn berekend onder de veronderstelling dat bij de berekening van de Heckman-term gebruik is gemaakt van de 'ware' γ_j. In werkelijkheid zijn hiervoor schattingen gebruikt die zelf ook onnauwkeurigheid vertonen, wat er toe leidt dat de standaarddeviaties van de laatste kolom van tabel 2 niet juist zijn. In zijn artikel claimt Heckman dat de ware standaarddeviaties altijd groter zijn, waardoor het nog steeds mogelijk zou

Tabel 2. Regressieresultaten met en zonder Heckman-term

	Zonder Heckman-term		Met Heckman-term	
	Parameter	Standaarddev.	Parameter	Standaarddev.
β_0 (constante term)	3,060	0,182	3,046	0,182
β_1 (coëff.gezinsgrootte)	0,091	0,015	0,083	0,015
β_2 (coëff. inkomen)	0,673	0,018	0,680	0,019
$\sigma_{\varepsilon\delta}$			-0,054	0,031
$\overline{R^2}$	0,653		0,653	
Aantal waarnemingen	1 008		1 008	

zijn de nulhypothese '$\sigma_{\varepsilon\delta} = 0$' niet te verwerpen. Onlangs (Greene, 1981) is echter aangetoond dat dit niet juist is en dat het ook mogelijk is dat de ware standaarddeviaties kleiner zijn dan de berekende. Over de significantie van onze conclusies met betrekking tot $\sigma_{\varepsilon\delta}$ moet dus enig voorbehoud worden gemaakt. Hiermee rekening houdende zien we dat de nulhypothese '$\sigma_{\varepsilon\delta} = 0$' (net) niet kan worden verworpen; dus dat de selecte non-respons geen significante invloed heeft op de regressie. De nieuwe schattingen voor β_0, β_1 en β_2 liggen allemaal binnen één (juiste) standaarddeviatie van de oude schattingen. De geschatte waarde van $\sigma_{\varepsilon\delta}$ is negatief wat inhoudt dat een relatief hoge responsgeneigdheid ($\delta > 0$) samengaat, met een relatief lage μ ($\varepsilon < 0$, een μ lager dan op grond van fs en y verwacht had mogen worden) en dus een relatief hoge tevredenheid met het inkomen. Tenslotte is het interessant op te merken dat de geschatte variantie van ε 0,048 bedraagt, zodat de correlatiecoëfficiënt van ε en δ -0,25 is.

5. Conclusies

In dit rapport is nagegaan wat de invloed van een selecte non-respons is op de resultaten van onderzoek door middel van regressierekening. Het model van Heckman (1979) is toegepast op een relatie die in het Inkomenswaarderingsonderzoek van het CBS wordt gemeten, het verband tussen het behoeftenniveau van een huishouden en de gezinsgrootte en het netto inkomen. De resultaten kunnen als volgt worden samengevat:

a) Verscheidene variabelen hangen significant samen met het al of niet responderen. Alleenstaanden, grote gezinnen, inwonende huishoudens, 65^+-ers, inwoners van grote steden, zelfstandigen en niet-werkzamen responderen significant slechter.
b) Deze selecte non-respons heeft een net niet significante invloed op de relatie die bij het Inkomenswaarderingsonderzoek wordt onderzocht, in die zin dat personen die relatief tevreden zijn met hun inkomen relatief vaker responderen.
c) De verandering in ieder van de coëfficiënten van de regressie is gering.

De hypothese dat tevreden personen meer geneigd zijn aan enquêtes mee te doen vindt ook ondersteuning buiten het onderhavige onderzoek. Uit resultaten van het Werknemersbudgetonderzoek 1974/1975 blijkt dat de gemiddelde inkomenswaardering in deze steekproef 0,75 bedroeg (Van Praag e.a., 1982), terwijl in de overige onderzoeken altijd een waarde tussen 0,58 en 0,69 is gevonden. Ook

indien er rekening mee wordt gehouden dat werknemers gemiddeld tevredener zijn dan de gehele bevolking, kan hieruit toch worden geconcludeerd dat de hoge non-respons bij het Budgetonderzoek er toe leidt dat personen die tevreden zijn met hun inkomen in de steekproef oververtegenwoordigd zijn.

Hieronder volgen tenslotte nog een tweetal slotopmerkingen:

a) Bij het onderhavige onderzoek konden alleen dummyvariabelen worden gebruikt omdat er geen covariantiematrix met continue variabelen beschikbaar was. Het lijkt wenselijk dat een populatiecovariantiematrix met continue variabelen ieder jaar wordt gepubliceerd, bijvoorbeeld om de effecten van selectieve non-respons na te gaan. Ten behoeve van andere onderzoekers is in appendix B de, met behulp van tabellen berekende, correlatiematrix uit het WBO opgenomen.
b) Het onderzoek heeft aangetoond dat niet alleen klassieke 'harde' variabelen, zoals urbanisatiegraad, samenhangen met de respons, maar ook 'zachte' attitude-variabelen als de relatieve tevredenheid met het inkomen. Indien ook maar enigszins mogelijk, zou bij de analyse van non-respons ook met dergelijke variabelen rekening moeten worden gehouden.

Appendix A. De benadering van probit

In deze appendix wordt aangegeven hoe de probit-coëfficiënten γ_j kunnen worden verkregen uit de regressiecoëfficiënten α_j. De elementen van X geven we aan met x_{ij}, en

$$x_{\cdot j} \equiv \frac{1}{I} \sum_{i=1}^{I} x_{ij} \tag{A.1}$$

$$\theta \equiv \sum_{j=0}^{k} \gamma_j x_{\cdot j} \; . \tag{A.2}$$

Er geldt nu, onder de probit-specificatie,

$$E(a_i^* | X) = P(a_i^* = 1 | X) = N\Big(\sum_{j=0}^{k} \gamma_j x_{ij}\Big) \approx N(\theta) + \sum_{j=0}^{k} (x_{ij} - x_{\cdot j}) n(\theta) \gamma_j \equiv$$

$$\equiv \alpha_0 + \sum_{j=1}^{k} \alpha_j x_{ij} \tag{A.3}$$

waarbij

$$\alpha_0 = N(\theta) - \sum_{j=1}^{k} x_{\cdot j} n(\theta) \gamma_j \tag{A.4}$$

$$\alpha_j = n(\theta) \gamma_j \qquad j=1,\ldots,k \; . \tag{A.5}$$

Schattingen voor de α_j's kunnen worden verkregen uit de regressie (6). Hiermee kunnen, met behulp van (A.4) en (A.5), schattingen worden verkregen voor de γ_j's. Door (A.5) in te vullen in (A.4), gaat (A.4), na herschrijven, over in

$$N(\theta) = \sum_{j=0}^{k} x_{\cdot j} \alpha_j = \bar{a}^* \tag{A.6}$$

met $\bar{a}^*$ het gemiddelde van de a_i^*'s (oftewel 1 008/2 332), zodat

$$n(\theta) = n(N^{-1}(\vec{a}^*)). \tag{A.7}$$

Hieruit volgt $n(\theta)$, en daarmee (zie (A.5)) $\gamma_1,\ldots,\gamma_k$. Tenslotte kan γ_0 worden verkregen door (A.6) te herschrijven als

$$N(\gamma_0 + \sum_{j=1}^{k} \gamma_j x_{\cdot j}) = \vec{a}^*, \tag{A.8}$$

en in deze formule de berekende waarden voor $\gamma_1,\ldots,\gamma_k$ in te vullen.

Appendix B. Data

De gemiddelden en de correlaties zijn vermeld in tabel 3 en 4. Deze gegevens zijn voldoende om de responsvergelijking (2) te schatten en kunnen dus ook bij andere onderzoeken waar sprake is van selectieve non-respons worden gebruikt.

Tabel 3. Gemiddelde frequenties van enkele achtergrondvariabelen in het Inkomenswaarderingsonderzoek 1980 en het Woningbehoeftenonderzoek 1977.

Variabele	Symbool	Gemiddelde IW 1980	Gemiddelde WBO 1977
Hoofd huishouden = vrouw	x_1	0,157	0,185
Hoofd huishouden = gehuwd	x_2	0,772	0,732
Grootte huishouden = 1 persoon	x_3	0,132	0,179
Grootte huishouden ≥ 5 personen	x_4	0,140	0,143
Inwonend huishouden	x_5	0,008	0,031
Leeftijd hoofd huishouden < 30	x_6	0,184	0,183
Leeftijd hoofd huishouden ≥ 65	x_7	0,137	0,209
Inkomen beneden 29^e percentiel	x_8	0,292	0,291
Inkomen boven 72^e percentiel	x_9	0,284	0,283
Gemeentegrootte > 50 000	x_{10}	0,398	0,444
Hoofd huishouden = zelfstandige	x_{11}	0,071	0,095
Hoofd huishouden = niet werkzaam	x_{12}	0,278	0,360

Tabel 4. Correlaties tussen enkele achtergrondvariabelen in het Woningbehoeftenonderzoek 1977

Variabele	Correlatie met										
	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}
x_1	1.000										
x_2	-0,768	1.000									
x_3	0,599	-0,755	1,000								
x_4	-0,157	0,203	-0,190	1,000							
x_5	0,130	-0,225	0,244	-0,055	1,000						
x_6	-0,043	-0,038	0,026	-0,159	0,210	1,000					
x_7	0,276	-0,272	0,298	-0,189	-0,027	-0,243	1,000				
x_8	0,432	-0,487	0,395	-0,187	0,132	-0,080	0,478	1,000			
x_9	-0,230	0,264	-0,215	0,131	-0,068	0,035	-0,234	-0,402	1,000		
x_{10}	0,104	-0,144	0,132	-0,100	0,061	0,032	0,039	0,052	-0,026	1,000	
x_{11}	-0,115	0,105	-0,095	0,125	-0,039	-0,064	-0,119	-0,140	-0,063	-0,094	1,000
x_{12}	0,384	-0,387	0,309	-0,189	0,084	-0,181	0,627	0,568	-0,315	0,091	-0,243

Referenties

Cate, A. ten, Lineaire regressieanalyse met steekproefgegevens. Interne CBS-nota (Centraal Bureau voor de Statistiek, Voorburg).

Greene, W.H., 1981, Sample selection bias as a specification error: Comment. Econometrica 49, pp. 795-798.

Heckman, J.J., 1979, Sample selection bias as a specification error. Econometrica 47, pp. 153-161.

Johnson, N. en S. Kotz, 1972, Distributions in Statistics: Continuous Multivariate Distributions (Wiley, New York).

Kapteyn, A. en T.J. Wansbeek, 1982, The individual welfare function: Measurement, explanation and policy applications. Statistical Studies 32 (Staatsuitgeverij, 's-Gravenhage).

Praag, B.M.S. van en A. Kapteyn, 1973, Further evidence on the individual welfare function of income: an empirical investigation in The Netherlands, European Economic Review 4, pp. 33-62.

Praag, B.M.S. van, J.S. Spit en H. van de Stadt, 1981, A comparison between the foodratio poverty-line and the Leyden poverty-line. Verschijnt in The Review of Economics and Statistics.

Stadt, H. van de, 1981, De waardering van inkomen in 1978. Statistisch Magazine 1, pp. 87-97.