Wegen als correctie voor non-respons

08

Jelke Bethlehem

Statistische Methoden (08005)

Centraal Bureau voor de Statistiek

Voorburg/Heerlen, 2008

Verklaring van tekens

.	= gegevens ontbreken
*	= voorlopig cijfer
x	= geheim
–	= nihil
–	= (indien voorkomend tussen twee getallen) tot en met
0 (0,0)	= het getal is kleiner dan de helft van de gekozen eenheid
niets (blank)	= een cijfer kan op logische gronden niet voorkomen
2005–2006	= 2005 tot en met 2006
2005/2006	= het gemiddelde over de jaren 2005 tot en met 2006
2005/'06	= oogstjaar, boekjaar, schooljaar enz., beginnend in 2005 en eindigend in 2006
2003/'04–2005/'06	= oogstjaar, boekjaar enz., 2003/'04 tot en met 2005/'06

In geval van afronding kan het voorkomen dat de som van de totalen afwijkt van het totaal.

Colofon

Uitgever
Centraal Bureau voor de Statistiek
Prinses Beatrixlaan 428
2273 XZ Voorburg

tweede helft van 2008:
Henri Faasdreef 312
2492 JP Den Haag

Prepress
Centraal Bureau voor de Statistiek - Facilitair bedrijf

Omslag
TelDesign, Rotterdam

Inlichtingen
Tel. (088) 570 70 70
Fax (070) 337 59 94
Via contactformulier: www.cbs.nl/infoservice

Bestellingen
E-mail: verkoop@cbs.nl
Fax (045) 570 62 68

Internet
www.cbs.nl

ISSN: 1876-0333

6016508005 X-37

Inhoudsopgave

1. Over wegen

1.1 Algemene beschrijving en leeswijzer

Bij het uitvoeren van een survey-onderzoek kunnen zich allerlei problemen voordoen. Een van de belangrijkste problemen is het optreden van *non-respons*. Indien in een enquête van respondenten de gewenste informatie niet wordt verkregen, terwijl ze toch tot de doelpopulatie van het onderzoek behoren en in de steekproef zijn getrokken (en dus de gegevens hadden moeten verstrekken), dan wordt dat *non-respons* genoemd.

Bij non-respons wordt een tweedeling gemaakt in unit-non-respons en item-non-respons. Bij *unit-non-respons* wordt van het betreffende element in de steekproef geen enkele informatie verkregen (de vragenlijst blijft leeg).

Bij *item-non-respons* blijven alleen enkele vragen onbeantwoord. Dat betreft meestal vragen over gevoelig liggende onderwerpen zoals inkomen, zwart geld, seksueel gedrag en crimineel verleden. Dit thema gaat vooral over de behandeling van unit-non-respons. De aanpak van item-non-respons komt hier niet aan de orde maar bij het thema “controle/correctie en imputatie”.

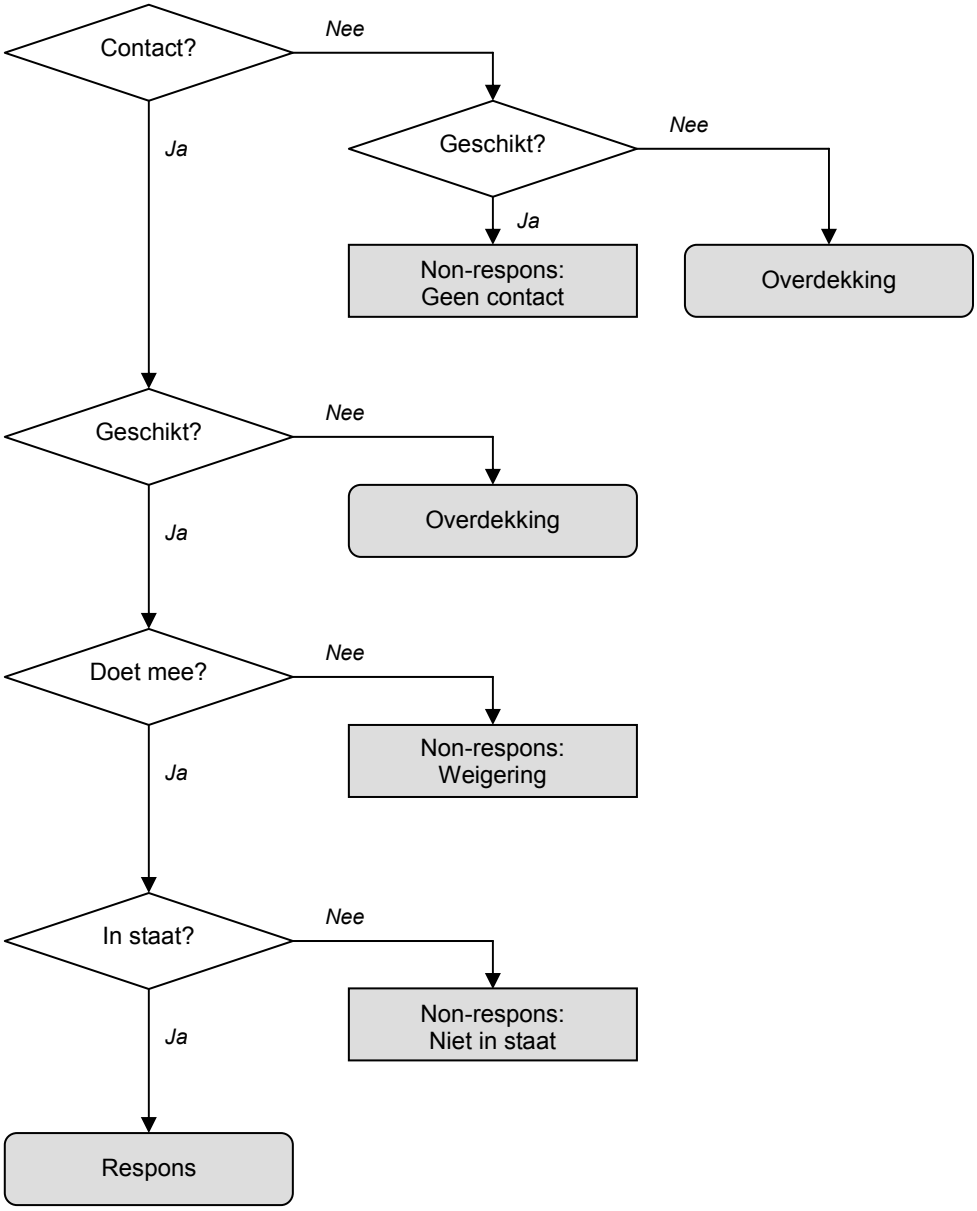
Non-respons leidt tot een geringer aantal waarnemingen dan was gepland, maar in principe hoeft dat niet tot onjuiste uitkomsten te leiden. Wel zullen de betrouwbaarheidsmarges van de schattingen wat groter zijn. De situatie is ernstiger wanneer de non-respons *selectief* is. Dit verschijnsel doet zich voor als, ten gevolge van non-respons, bepaalde groepen onder- of oververtegenwoordigd zijn in het onderzoek. Gedraagt een ondervertegenwoordigde groep zich duidelijk anders met betrekking tot de te onderzoeken variabelen dan de oververtegenwoordigde groep, dan leidt dit tot een *vertekening* in de uitkomsten. Anders gezegd: een schatting valt systematisch te hoog of te laag uit.

De praktijk heeft uitgewezen dat helaas non-respons vaak selectief is. Dat is ook gebleken bij een aantal onderzoeken van het CBS. Bij de Enquête Slachtoffers Misdrijven bleken mensen die 's avonds thuis bang zijn, minder bereid om mee te doen aan een vervolgonderzoek. Bij woningbehoeftenonderzoeken was de tevredenheid met de huidige woning groter onder weigeraars dan onder respondenten. En bij het Onderzoek Verplaatsingsgedrag werden mobielere mensen minder vaak thuis aangetroffen (en dat was juist het onderwerp van het onderzoek).

De non-respons kan verschillende oorzaken hebben. Het is goed de non-respons op basis van deze oorzaken in groepen te verdelen. Uit onderzoek van non-respondenten is gebleken dat de diverse groepen nogal kunnen verschillen. Elk type non-respons kan aanleiding geven tot een ander soort vertekening. Een goede indeling van de non-respons is dus onontbeerlijk. Dit geldt niet alleen voor de analyse van de non-respons, maar ook voor een goede verantwoording van het veldwerk is een duidelijke classificatie belangrijk. In figuur 1 wordt aangegeven hoe

de verschillende vormen van non-respons kunnen ontstaan bij de poging de medewerking van een geselecteerde persoon te krijgen.

Figuur 1.Mogelijke uitkomsten in het veldwerk



Eerst moeten we contact maken met de personen die zijn geselecteerd in de steekproef. Als dat niet lukt, dan zijn er twee mogelijkheden. Als die personen tot de doelpopulatie behoren, dan horen ze thuis in de steekproef en zouden we hun gegevens moeten hebben. Er is dan sprake van non-respons als gevolg van *geen contact*. Als de desbetreffende personen geen deel uitmaken van de doelpopulatie, dan horen ze ook niet thuis in de steekproef. Dit is een geval van *overdekking*. Er hoeft dan verder geen actie meer te worden ondernomen. Merk op dat in de praktijk meestal niet kan worden vastgesteld of het gaat om non-respons of overdekking. Dat maakt het bijvoorbeeld lastig (zo niet onmogelijk) het percentage respons te berekenen.

Is het wel gelukt om contact te maken met personen, dan kunnen we vaststellen of ze behoren tot de doelpopulatie. Is dat niet het geval, dan zijn we klaar. Ze horen dan niet thuis in de steekproef. Ze hoeven daarom geen vragenlijst in te vullen. We kunnen deze personen negeren als een geval van *overdekking*. Behoren personen wel tot de doelpopulatie, dan moeten we ze overhalen om mee te werken aan het onderzoek. Lukt dat niet dan is er sprake van non-respons als gevolg van *weigering*.

Ook al behoren de geselecteerde personen tot de doelpopulatie en willen ze meewerken aan het onderzoek, dan kunnen er toch nog omstandigheden zijn die het onmogelijk maken om mee te werken. Voorbeelden zijn ziekte of taalproblemen. Er is dan sprake van non-respons omdat ze *niet in staat* zijn om mee te werken.

Als geselecteerde personen behoren tot de doelpopulatie, het mogelijk is contact met ze te leggen, ze ook wel willen meewerken en er geen andere omstandigheden zijn die dat verhinderen, dan pas is er sprake van *respons*.

Er zijn voldoende aanwijzingen dat in veel onderzoek het optreden van non-respons tot vertekeningen in de schattingen leidt. Dat betekent dat het niet verantwoord is om zonder verdere correcties over te gaan tot het berekenen van schattingen en het publiceren van uitkomsten. Een veel toegepaste methode om de uitkomsten te corrigeren is het uitvoeren van een *weegprocedure*. Daarbij wordt aan elk geobserveerd element i een gewicht w_i toegekend. In de schattingsprocedures worden vervolgens deze gewichten meegenomen.

De effectiviteit van een wegingprocedure staat of valt met de beschikbaarheid van geschikte *hulpvariabelen*. Binnen de context van wegen zijn dat variabelen die in het onderzoek zijn gemeten, en waarvoor op het niveau van de populatie (of de volledige steekproef) de verdeling bekend is. Door de verdeling van een hulp-variabele in de respons te vergelijken met die van de populatie krijgen we een indruk in hoeverre de respons representatief is met betrekking tot de hulpvariabele. Constateren we wezenlijke verschillen tussen beide verdelingen, dan kunnen we concluderen dat non-respons, althans voor deze variabele, heeft geleid tot een selectieve respons.

We gebruiken deze hulpvariabelen voor het berekenen van de gewichten. Die gewichten zijn zodanig dat de gewogen verdeling van de hulpvariabele in de respons exact gelijk wordt aan die in de populatie. Dit bereiken we door onder-vertegenwoordigde groepen een hoger gewicht te geven en oververtegenwoordigde groepen een lager gewicht. We kunnen dan zeggen dan dat onze gewogen gegevens representatief zijn met betrekking tot deze variabelen.

Bij de opzet van een onderzoek wordt een *steekproefontwerp* gespecificeerd. Dit steekproefontwerp beschrijft de mogelijke steekproeven en de bijbehorende kansen. We beperken ons hier tot steekproeven die *zonder teruglegging* worden getrokken. Dit betekent dat een element maar hooguit één keer in een bepaalde steekproef kan voorkomen. Bij een steekproef zonder teruglegging kan voor elke element i de *insluitkans* π_i worden bepaald. Dat is de kans dat het desbetreffende element in de steekproef wordt getrokken. Verder introduceren we voor een geselecteerd element i

het *insluitgewicht* $d_i = 1 / \pi_i$ als het omgekeerde van de *insluitkans* π_i, voor $i = 1, 2, \ldots, n$. Hierin is n de omvang van de steekproef.

We gaan er in deze bijdrage vanuit dat we het populatiegemiddelde van een variabele Y willen schatten. Uiteraard is de theorie niet essentieel anders voor het schatten van populatietotalen.

Gegeven de insluitkansen kan altijd een zuivere schatter worden gedefinieerd voor het populatiegemiddelde. Die schatter heet de *Horvitz-Thompson-schatter*. De Horvitz-Thompson-schatter voor het populatiegemiddelde van de doelvariabele Y kunnen we schrijven als

$$\bar{y}_{HT} = \frac{1}{N}\sum_{i=1}^{n} d_i y_i \tag{1.1.1}$$

Merk op dat we de Horvitz-Thompson-schatter ook kunnen opvatten als een vorm van wegen. De gemeten waarden van de variabele Y worden immers gewogen met de insluitgewichten. Daarom wordt de correctie voor non-respons ook wel eens aangeduid als *herwegen*. Wij zullen deze non-responscorrectie aanduiden als *wegen*.

Na het uitvoeren van een weegprocedure (ter correctie voor non-respons) vervangen we de Horvitz-Thompson-schatter schatter door een nieuwe schatter

$$\bar{y}_W = \frac{1}{N}\sum_{i=1}^{n} w_i y_i. \tag{1.1.2}$$

waarin het gewicht w_i gelijk is aan

$$w_i = c_i \times d_i. \tag{1.1.3}$$

Hierin is c_i een *correctiegewicht* dat we verkregen hebben uit een nader te definiëren weegprocedure.

Voor alle hulpvariabelen die we gebruiken in de weegprocedure moet gelden dat de gewogen verdeling in de respons gelijk is aan de verdeling in de populatie. Als we schatter (1.1.2) zouden gebruiken om het gemiddelde van een hulpvariabele X te schatten, dan moet die schatting dus exact overeenkomen met het populatie-gemiddelde. In formule:

$$\bar{x}_W = \frac{1}{n}\sum_{i=1}^{n} w_i X_i = \bar{X}. \tag{1.1.4}$$

Als aan de voorwaarde (1.1.4) is voldaan, dan noemen we de gewogen steekproef *representatief* met betrekking tot de desbetreffende hulpvariabele.

Als het mogelijk is om de steekproef tegelijk representatief te maken met betrekking tot een flink aantal hulpvariabelen, en al die hulpvariabelen zijn allemaal sterk gecorreleerd met de doelvariabelen van het onderzoek, dan zal de (gewogen) steekproef ook (bij benadering) representatief zijn met betrekking tot de doelvariabelen. Daardoor zullen schattingen van de doelvariabele gebaseerd op de gewogen steekproef beter zijn dan schattingen die zijn gebaseerd op de ongewogen steekproef.

In de volgende hoofdstukken behandelen we een aantal methoden van wegen. De eenvoudigste en meest gebruikte methode is *post-stratificatie*. Deze methode komt aan bod in hoofdstuk 2.

Het is lang niet altijd mogelijk om post-stratificatie toe te passen. Er zijn twee weegmethoden die we kunnen gebruiken daar waar post-stratificatie niet kan. De eerste daarvan is *lineair wegen*. Deze methode is gebaseerd op de algemene regressieschatter. We tonen aan dat post-stratificatie een speciaal geval is van lineair wegen. We beschrijven lineair wegen in hoofdstuk 3. De tweede weegmethode is *multiplicatief wegen*. Daarbij worden gewichten uitgerekend via een iteratief proces. We behandelen deze methode in hoofdstuk 4.

Hoewel lineair wegen en multiplicatief wegen zo op het oog twee totaal verschillende methoden lijken te zijn, is het toch mogelijk een soort algemeen raamwerk voor wegen te maken waarvan beide genoemde technieken passen. Dit algemene raamwerk is ontwikkeld door Deville and Särndal (1992). Ze noemen dit algemene theoretische kader *calibratie*. Hierover is meer te vinden in hoofdstuk 5.

In deze bijdrage zijn belangrijke termen uit de methodologie cursief gedrukt. Daarmee verwijzen we naar de thesaurus van de Methodenreeks. Daarin worden deze termen uitgelegd.

De theorie van het verbeteren van schatters onder het optreden van non-respons wordt ook uitstekend behandeld in het boek van Särndal & Lundström (2005).

1.2 Afbakening en relatie met andere thema's

In dit thema beschrijven we de toepassing van weegtechnieken om te corrigeren voor de negatieve effecten van unit-non-respons. Er bestaan andere technieken voor de correctie van non-respons, zoals de na-enquête onder non-respondenten en de Methoden van de Centrale Vraag. Deze technieken worden hier niet behandeld.

Wegen van steekproeven vindt ook om andere reden plaats. Zo moet er ook worden gewogen als de steekproe met ongelijke kansen is getrokken. Deze vorm van wegen wordt besproken in het thema *Steekproeven* van de Methodenreeks.

We beschrijven het wegen van steekproeven hier uitsluitend binnen het kader van onderzoek onder personen. De theorie is echter ook net zo goed toepasbaar bij onderzoek van andere populaties, zoals bijvoorbeeld bedrijven.

Naast unit-non-respons is er ook nog item-non-respons. Hierbij wordt slechts een (klein) deel van de vragen niet beantwoord. Om te corrigeren voor item-non-respons maken we meestal gebruik van imputatie-technieken. Deze technieken worden hier ook niet behandeld.

Meer over non-respons is te vinden in het thema *Benaderingsstrategieën* van de Methodenreeks.

We laten in voorbeelden ook zien hoe de theorie van wegen is geïmplementeerd in het computerprogramma *Bascula*. Dit programma is hiervoor speciaal ontwikkeld.

De desbetreffende voorbeelden kunnen desgewenst worden overgeslagen. Ze zijn niet nodig voor het begrip van de stof.

Voor een beter begrip van de hier te behandelen stof verdient het aanbeveling enige kennis te hebben van de steekproeftheorie. Zie hiervoor het desbetreffende thema.

1.3 Plaats in het statistische proces

Een weegprocedure wordt uitgevoerd tijdens de bewerkingsfase van de gegevens. Eerst worden de gegevens verzameld, vervolgens worden de fouten eruit gehaald (gaafmaken), en daarna komt de weegprocedure aan de beurt. Uiteraard moet de weging plaatsvinden voordat de gegevens gaan worden getabelleerd en geanalyseerd. Immers, een analyse van ongewogen gegevens kan zeer goed tot onjuiste uitkomsten leiden.

1.4 Definities en notaties

De te onderzoeken doelpopulatie geven we aan met U. De doelpopulatie is eindig en bestaat uit N elementen. We geven de doelpopulatie aan met

$$U = \{1,\ 2,\ ...,\ N\} \tag{1.4.1}$$

Hierin is N de *omvang* van de doelpopulatie. De nummers 1, 2, …, N duiden de volgnummers van de elementen in de doelpopulatie aan.

Laat Y een *doelvariabele* van het onderzoek zijn. De waarden van Y noteren we met $Y_1, Y_2, \ldots, Y_N$.

Laat X een *hulpvariabele* zijn. We zullen hulpvariabelen gebruiken voor het opstellen van weegmodellen. De waarden van hulpvariabele X noteren we met $X_1, X_2, \ldots X_N$.

Op grond van de verzamelde steekproefgegevens moeten we uitspraken doen over de doelpopulatie. Concreet komt het erop neer dat we het gedrag, de structuur van de populatie vastleggen in enkele kengetallen. Dergelijke kengetallen noemen we populatiegrootheden.

In deze bijdrage gaan we er vanuit dat het doel van het onderzoek is het schatten van het populatiegemiddelde van de doelvariabele Y. Deze populatiegrootheid schrijven we als

$$\overline{Y} = \frac{1}{N}\sum_{k=1}^{N} Y_k = \frac{Y_1 + Y_2 + ... + Y_N}{N} \tag{1.4.2}$$

Er is nog een andere belangrijke populatiegrootheid die zeker moet worden genoemd. Dat is de populatievariantie. Deze zegt iets zegt over de mate waarin de waarden van de doelvariabele fluctueren. De *aangepaste populatievariantie* van een doelvariabele Y is gelijk aan

$$S^2 = \frac{1}{N-1}\sum_{k=1}^{N}(Y_k - \overline{Y})^2 \tag{1.4.3}$$

Schatten van de populatievariantie zal meestal geen doel op zich zijn. Deze grootheid is echter wel van groot belang bij het bepalen van de nauwkeurigheid van andere schattingen.

Uit de doelpopulatie trekken we een steekproef. Daarvoor moeten we eerst een steekproefontwerp kiezen. In het *steekproefontwerp* leggen we voor elke mogelijke steekproef vast hoe groot de kans op trekking ervan is.

We beperken ons hier tot *steekproeven zonder teruglegging*. Dat zijn steekproeven waarbij een element nooit meer dan één keer in dezelfde steekproef terecht kan komen.

Een *steekproef a* uit een populatie $U = \{1, 2, \ldots, N\}$ geven we aan met een reeks van indicatoren $a_1, a_2, \ldots, a_N$. Als element k in de steekproef wordt getrokken, dan krijgt a_k de waarde 1. Zit element k niet in de steekproef, dan krijgt a_k de waarde 0.

De *steekproefomvang n* van de steekproef is gelijk aan

$$n = \sum_{k=1}^{N} a_k \tag{1.4.4}$$

Een steekproefontwerp waarbij we elementen zonder teruglegging trekken, kunnen we karakteriseren door een reeks eerste orde en tweede orde insluitkansen.

De *eerste orde insluitkans* π_k van element k is gedefinieerd als

$$\pi_k = E(a_k) \tag{1.4.5}$$

voor $k = 1, 2, \ldots, N$. Hierin duidt E de verwachting aan van de stochastische variabele a_k. De eerste orde insluitkans π_k van element k is dus gelijk aan de kans dat element k in de steekproef wordt getrokken.

De *tweede orde insluitkans* π_{kl} van twee elementen k en l is gedefinieerd als

$$\pi_{kl} = E(a_k a_l) \tag{1.4.6}$$

voor $k = 1, 2, \ldots, N$ en $l = 1, 2, \ldots, N$. De tweede orde insluitkans π_{kl} van de elementen k en l is dus gelijk aan de kans dat beide elementen k en l tegelijk in de steekproef worden getrokken.

De *Horvitz-Thompson-schatter* voor het populatiegemiddelde van de doelvariabele Y is gedefinieerd als

$$\bar{y}_{HT} = \frac{1}{N} \sum_{k=1}^{N} a_k \frac{Y_k}{\pi_k} \tag{1.4.7}$$

Dit is een zuivere schatter voor het populatiegemiddelde van Y.

In het eenvoudige geval van een aselecte steekproef met gelijke kansen (en zonder teruglegging) reduceert de Horvitz-Thompson-schatter tot het gewone steekproef-gemiddelde. Hierbij is dan aangetoond dat in deze situatie het steekproefgemiddelde een zuivere schatter is voor het populatiegemiddelde.

Wat gebeurt er nu als er non-respons optreedt? We veronderstellen dat het *Stochastische Responsmodel* van toepassing is. Dan wordt aan elk element k in de populatie een vaste, maar onbekende kans ρ_k op responderen toegekend, gegeven dat het element in de steekproef zit. Het al of niet responderen van een element in de steekproef is dus het resultaat van een kansmechanisme.

Als een aselecte steekproef met gelijke kansen en zonder teruglegging wordt getrokken, en er treedt non-respons op, dan is het steekproefgemiddelde geen zuivere schatter meer. De verwachting van deze schatter is gelijk aan

$$E(\bar{y}_R) \approx \bar{Y}^* \equiv \frac{1}{N}\sum_{k=1}^{N}\frac{Y_k \rho_k}{\bar{\rho}}, \tag{1.4.8}$$

waarin

$$\bar{\rho} = \frac{1}{N}\sum_{k=1}^{N}\rho_k \tag{1.4.9}$$

het gemiddelde van alle responskansen is. De vertekening is bij benadering gelijk aan:

$$B(\bar{y}_R) = E(\bar{y}_R) - \bar{Y} \approx \bar{Y}^* - \bar{Y} = \frac{C_{\rho y}}{\bar{\rho}}, \tag{1.4.10}$$

waarin

$$C_{\rho Y} = \frac{1}{N}\sum_{k=1}^{N}(\rho_k - \bar{\rho})(Y_k - \bar{Y}). \tag{1.4.11}$$

Uit formules (1.4.10) en (1.4.11) volgt dat de vertekening van de schatter door twee factoren wordt bepaald. De teller van (1.4.10) valt te interpreteren als een covariantie tussen responsgedrag en doelvariabele. Naarmate de samenhang tussen die twee groter is, is de vertekening ook groter. De noemer van (1.4.10) is de gemiddelde responskans. Dus naarmate de responskansen kleiner zijn, wordt de vertekening ernstiger.

We onderscheiden drie verschillende mechanismen die tot non-respons kunnen leiden:

- *Missing Completely At Random* (MCAR). Non-respons treedt volledig willekeurig op. Er is geen enkel verband tussen het responsgedrag (R) en de doelvariabele (Y). Er is geen direct effect van Y op R, en ook geen indirect effect via een hulpvariabele (X). De non-respons is dan niet selectief. Schatters hebben dus geen vertekening.

- *Missing at Random* (MAR). Dit is de situatie waarin er geen direct effect is van de doelvariabele (Y) op het responsgedrag (R). Er is echter wel een direct effect van de hulpvariabele (X) op het responsgedrag (R). Als dit zich voordoet, dan is de respons selectief en schattingen hebben een vertekening. Het is echter wel

mogelijk om hiervoor te corrigeren, mits een correctietechniek wordt toegepast waarin de informatie over de hulpvariabele *X* wordt gebruikt.

- *Not Missing at Random* (NMAR). Dan is er een directe samenhang tussen het responsgedrag (*R*) en de doelvariabele (*Y*). Dit verband loopt niet (zoals bij MAR) via een hulpvariabele (*X*). De non-respons is selectief. Schattingen hebben een vertekening. Correctietechnieken helpen niet om die vertekening weg te werken.

2. Post-stratificatie

2.1 Korte beschrijving

Post-stratificatie is een eenvoudige, bekende en veel gebruikte weegtechniek. De populatie wordt in een aantal strata verdeeld. Binnen een stratum krijgen alle waargenomen elementen hetzelfde gewicht. Die gewichten wordt verkregen door de fractie populatie-elementen in dit stratum te delen door de fractie steekproef-elementen in het stratum.

Post-stratifcatie is effectief (vermindert de vertekening) als de strata homogeen zijn. Hiermee wordt bedoeld dat binnen een stratum de elementen in zoveel mogelijk aspecten op elkaar moeten lijken.

We geven een eenvoudig voorbeeld van de toepassing van post-stratificatie. We beschouwen daartoe de totale bevolking van Samplonië. Deze bestaat uit 1000 zielen. We hebben de beschikking over de hulpvariabele geslacht. Stel, we weten dat er 511 mannen zijn in Samplonië en 489 vrouwen. Uit deze populatie trekken we een steekproef. De uiteindelijke respons bestaat uit 100 elementen. De resultaten staan in tabel 1.

Tabel 1. Het wegen van een steekproef met de hulpvariabele geslacht

Respons		
	Aantal	Perc
Man	48	48 %
Vrouw	52	52 %
Totaal	100	100 %

Populatie		
	Aantal	Perc
Man	511	51 %
Vrouw	489	49 %
Totaal	1000	100 %

Gewicht	
Man	1,065
Vrouw	0,940

Uit de tabel blijkt dat de verhouding man/vrouw in de respons anders is dan in de populatie: de respons bestaat voor 48% uit mannen en in de populatie is dat 51%. We kunnen nu de respons representatief maken met betrekking tot de variabele geslacht door de mannen een gewicht

0,511 / 0,480= 1,065

te geven en de vrouwen een gewicht

0,489 / 0,520 = 0,940.

Dat de mannen een gewicht groter dan 1 krijgen, is niet verwonderlijk; ze zijn ondervertegenwoordigd in de respons. In feite telt nu elke man in de respons mee voor 1,065 man. Vrouwen zijn oververtegenwoordigd en krijgen een gewicht kleiner dan 1. Elke vrouw in de respons telt mee voor 0,940 vrouw.

2.2 Toepasbaarheid

We kunnen post-stratificatie toepassen om een mogelijke vertekening ten gevolge van non-respons te verminderen. Daarvoor zijn hulpvariabelen nodig. Die moeten zijn gemeten in het onderzoek en ook moet de verdeling in de populatie (of in de bruto-steekproef) bekend zijn.

Wordt meer dan één hulpvariabele gebruikt voor de weging, dan moeten de populatiefracties bekend zijn in alle cellen (strata) die ontstaan door het kruisen van de te gebruiken hulpvariabelen.

Om gewichten te kunnen berekenen moet er in elke cel minstens één waarneming beschikbaar zijn. Voor het berekenen van standaardfouten van schattingen moet elke cel minstens twee waarnemingen bevatten (maar bij voorkeur meer dan vijf).

Post-stratificatie is ook een zinvolle techniek als er geen non-respons is opgetreden. Indien de strata homogeen zijn, zal post-stratifcatie leiden tot nauwkeuriger schattingen.

2.3 Uitgebreide beschrijving

Voor het uitvoeren van post-stratificatie hebben we één of meer kwalitatieve hulpvariabelen nodig. Stel dat we een dergelijke hulpvariabele X hebben met L categorieën. Dan verdeelt hij de doelpopulatie U in L strata U_1, U_2, …, U_L. We geven het aantal populatie-elementen in stratum U_h aan met N_h, voor $h = 1, 2, ..., L$. Er geldt dus $N = N_1 + N_2 + ... + N_L$.

Uit deze populatie trekken we een enkelvoudige aselecte steekproef van omvang n. Laat n_h het aantal steekproefelementen zijn in stratum U_h (voor $h = 1, 2, ..., L$). Dan geldt $n = n_1 + n_2 + ... + n_L$. Merk op dat de waarden van de n_h het resultaat zijn van aselect selectieproces. Het zijn dus kansvariabelen.

We beschrijven eerst de situatie waarin er geen non-respons optreedt. Post-stratificatie kent aan alle responderende elementen binnen een stratum hetzelfde gewicht toe. In geval van een enkelvoudige aselecte steekproef zonder teruglegging is het correctiegewicht c_i voor een responderend element i in stratum U_h gelijk aan

$$c_i = \frac{N_h / N}{n_h / n} \tag{2.3.1}$$

Als we de insluitgewichten ($d_i = N / n$) en correctiegewichten (c_i) substitueren in schatter (1.1.2), dan krijgen we de *post-stratificatie-schatter*

$$\bar{y}_{PS} = \frac{1}{N} \sum_{h=1}^{L} N_h \bar{y}^{(h)} \tag{2.3.2}$$

Hierin is $\bar{y}^{(h)}$ het gemiddelde van de waarnemingen in stratum h.

De post-stratificatie-schatter is dus gelijk is aan de gewogen som van de waargenomen gemiddelden in de strata.

We geven een voorbeeld van post-stratificatie met twee variabelen. Stel we hebben bij het Samplonische onderzoek de beschikking over de hulpvariabelen geslacht en leeftijd (in drie categorieën jong, middelbaar en oud). Dan is er bij wegen naar de hulpvariabelen leeftijd en geslacht een stratum voor elke combinatie van leeftijd en geslacht en dat geeft hier 2 × 3 = 6 strata. Kennen we nu de verdeling van de populatie over de aldus gevormde strata dan kunnen we voor elk stratum een gewicht bepalen.

In tabel 2 staan de uitkomsten van een steekproef uit de bevolking van Samplonië. De gewichten zijn op dezelfde manier bepaald als in tabel 1. Maar nu hebben we bereikt dat de steekproef representatief is geworden met betrekking tot zowel leeftijd als geslacht. Sterker nog, de steekproef is ook representatief voor geslacht binnen elke leeftijdscategorie en, omgekeerd, voor leeftijd binnen elk geslacht.

Tabel 2. Wegen met twee hulpvariabelen

Steekproef n=100		
	Man	Vrouw
Jong	28	17
Middel	16	20
Oud	8	11

Populatie N=1000		
	Man	Vrouw
Jong	226	209
Middel	152	144
Oud	133	136

Gewicht		
	Man	Vrouw
Jong	0,807	1,229
Middel	0,950	0,720
Oud	1,663	1,236

Stel nu dat er non-respons optreedt in het onderzoek. De post-stratificatie-schatter krijgt dan de volgende vorm

$$\bar{y}_{PS,R} = \frac{1}{N}\sum_{h=1}^{L} N_h \bar{y}_R^{(h)}, \tag{2.3.3}$$

Hierin is $\bar{y}_R^{(h)}$ het gemiddelde van de responderende elementen in stratum h. De vertekening van deze schatter is gelijk aan

$$B(\bar{y}_{PS,R}) = \frac{1}{N}\sum_{h=1}^{L} N_h B(\bar{y}_R^{(h)}). \tag{2.3.4}$$

De vertekening is dus gelijk aan de gewogen som van de vertekeningen binnen de strata. Uitgaande van het Stochastische Responsmodel kunnen we de vertekening schrijven als

$$B(\bar{y}_{PS,R}) = \frac{1}{N}\sum_{h=1}^{L} N_h \frac{C_{\rho Y}^{(h)}}{\bar{\rho}^{(h)}}, \tag{2.3.5}$$

waarin $C_{\rho Y}^{(h)}$ de covariantie is tussen de waarden van Y en de waarden van ρ in stratum h.

Nadere bestudering van (2.3.5) geeft aanknopingspunten hoe stratificatie kan helpen bij het reduceren van de vertekening ten gevolge van non-respons. Het gaat erom de stratificatie zo te kiezen dat de covarianties tussen doelvariabele en responsgedrag binnen de strata klein zijn. Dat kan op de volgende manieren:

- Kies de strata zo dat ze homogeen zijn met betrekking tot de doelvariabele. De waarden variëren dan niet veel binnen de strata maar wel tussen de strata. De variatie in de waarden van de doelvariabele uit zich vooral in niveauverschillen tussen de strata (verschillen in stratumgemiddelden), en juist niet in variatie binnen de strata.

- Kies de strata zo dat ze homogeen zijn met betrekking tot de responskansen. De kansen variëren dan niet veel binnen de strata maar wel tussen de strata.

De eerste regel wordt ook toegepast bij post-stratificatie onder volledige respons. Het is een bekende regel die in het algemeen ook resulteert in kleine varianties.

De tweede regel concentreert zich op de responskansen. Uiteraard is het niet mogelijk om individuele responskansen te schatten, maar als het bijvoorbeeld mogelijk is de strata te verdelen in sub-strata, en voor elk sub-stratum kunnen we de gemiddelde responskans schatten, dan biedt vergelijking van deze gemiddelden enige houvast over de variatie in de responskansen.

Uiteraard valt die stratificatie te prefereren waarbij beide regels opgaan. Maar het zal lang niet altijd eenvoudig zijn om dat te realiseren, omdat geschikte hulpvariabelen bepaald niet voor het oprapen liggen.

2.4 Voorbeeld

De hier behandelde weegmethoden zijn alle geïmplementeerd in het programma Bascula. We laten aan de hand van een voorbeeld zien hoe we met Bascula een post-stratificatie kunnen uitvoeren.

Uit de bevolking van Samplonië (N = 1000) hebben we een steekproef (met gelijke kansen en zonder teruglegging) getrokken van n = 100. Er zijn twee hulpvariabelen die we willen gebruiken voor wegen: geslacht (man, vrouw) en leeftijdsklasse (jong, middelbaar, oud). Tabel 3 bevat de verdeling in de respons en in de populatie van beide variabelen.

Tabel 3. Post-stratificatie met twee hulpvariabelen

Respons			
	Man	Vrouw	Totaal
Jong	28	17	45
Middel	16	20	36
Oud	8	11	19
Totaal	52	48	100

Populatie			
	Man	Vrouw	Totaal
Jong	226	209	435
Middel	152	144	296
Oud	133	136	269
Totaal	511	489	1000

Gewichten		
	Man	Vrouw
Jong	0,807	1,229
Middel	0,950	0,720
Oud	1,663	1,236

Er is geen sprake van een representatieve steekproef. Zo is het percentage oude mannen in de steekproef gelijk aan 8,0% terwijl het in de populatie gelijk is aan 13,3%. De steekproef bevat dus te weinig oude mannen.

Door toepassing van post-stratificatie krijgen alle oude mannen een gewicht gelijk aan (133/1000) / (8/100) = 1,663. Aangezien oude mannen ondervertegenwoordigd zijn, krijgen ze een gewicht groter dan 1.

De uiteindelijke gewichten krijgen we door de correctiegewichten in tabel 3 te vermeningvuldigen met de insluitgewichten. Aangezien een steekproef van 100 met gelijke kansen is getrokken uit een populatie van 1000, zijn alle insluitkansen gelijk aan $n / N = 100 / 1000 = 0{,}1$. Dus zijn alle insluitgewichten gelijk aan $N / n = 10$.

De gewogen schatting voor het aantal oude mannen in de populatie is gelijk aan

$$10 \times 1{,}663 \times 8 = 133$$

en dat is precies gelijk aan het aantal oude mannen in de populatie. Dit bevestigt nog eens dat toepassing van de gewichten in schattingen voor de gebruikte hulp-variabelen exact de populatiegrootheden oplevert.

We gaan er in ons voorbeeld vanuit dat de gegevens voor het onderzoek zijn verzameld met Blaise. De Blaise-vragenlijst is weergegeven in figuur 2.

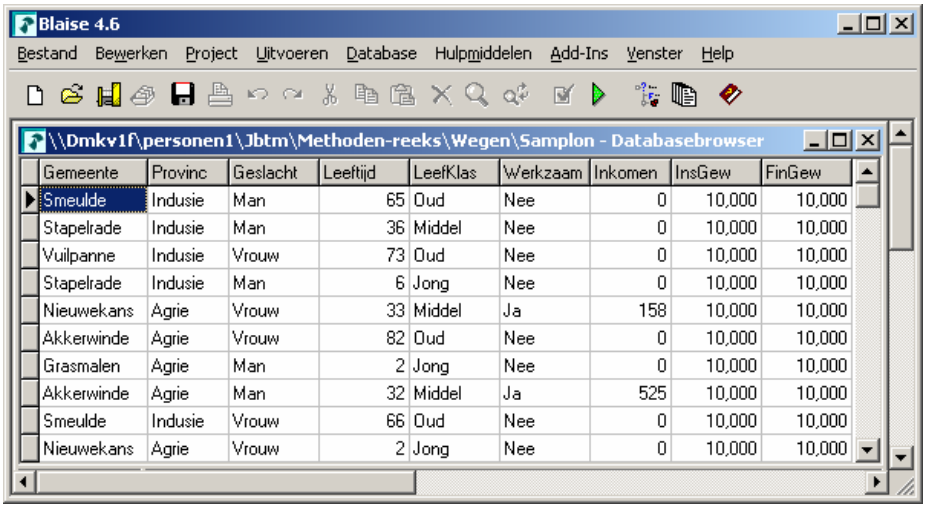
Figuur 2. De Blaise-vragenlijst

```
DATAMODEL Samplon "De Samplonische Survey"

FIELDS
   Gemeente  "In welke gemeente woont u?":
             (Akkerwinde, Grasmalen, Nieuwekans, Lommerdal,
              Smeulde, Stapelrade, Vuilpanne)
   Provinc   "In welke provincie woont u?": (Agrie, Indusie)
   Geslacht  "Wat is uw geslacht?": (Man, Vrouw)
   Leeftijd  "Wat is uw leeftijd?": 0..99
   LeefKlas  "Leeftijdsklasse": (Jong, Middel, Oud)
   Werkzaam  "Heeft u betaald werk?": (Ja, Nee)
   Inkomen   "Wat is uw maandelijks netto inkomen?": 0..6000
   InsGew    "Insluitgewicht": 0.000..1000.000
   FinGew    "Finaal gewicht": 0.000..1000.000

RULES
   Gemeente Provinc Geslacht Leeftijd
   IF Leeftijd <= 30 THEN
      LeefKlas:= Jong
   ELSEIF Leeftijd <= 55 THEN
      LeefKlas:= Middel
   ELSE
      LeefKlas:= Oud
   ENDIF
   Werkzaam Inkomen
   InsGew:= 10.000
   FinGew:=  1.000
ENDMODEL
```

De respondenten moesten antwoord geven op zes vragen: gemeente, provincie, geslacht, leeftijd, werkzaam en inkomen. De vraag *LeefKlas* (leeftijdsklasse) werd niet gesteld. Het antwoord werd afgeleid uit het antwoord op de vraag leeftijd.

Er zijn nog twee andere vragen in de vragenlijst opgenomen. Dat zijn *InsGew* (het insluitgewicht) en *FinGew* (het finale gewicht). Beide vragen werden niet gesteld. Hun antwoorden werden uitgerekend. De corresponderende variabelen worden daarom in het gegevensbestand opgenomen. Aangezien het insluitgewicht van iedere respondent gelijk was aan 10, is het antwoord op de vraag *InsGew* op 10 gezet. Het finale gewicht is voorlopig ook even op 10 gezet. Dit gewicht zal later worden aangepast in Bascula. In figuur 3 zijn de gegevens te zien van de eerste 10 personen in de steekproef.

Figuur 3. De eerste 10 personen in de steekproef

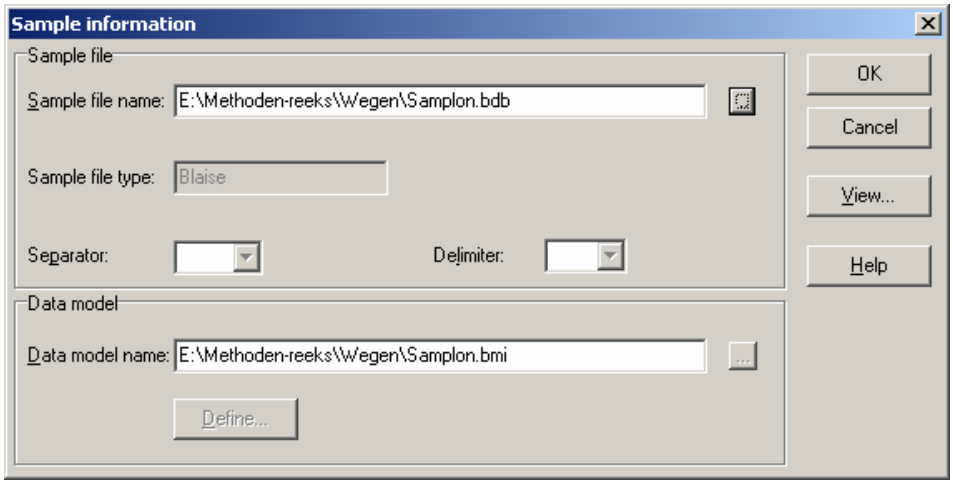


Gemeente	Provinc	Geslacht	Leeftijd	LeefKlas	Werkzaam	Inkomen	InsGew	FinGew
Smeulde	Indusie	Man	65	Oud	Nee	0	10,000	10,000
Stapelrade	Indusie	Man	36	Middel	Nee	0	10,000	10,000
Vuilpanne	Indusie	Vrouw	73	Oud	Nee	0	10,000	10,000
Stapelrade	Indusie	Man	6	Jong	Nee	0	10,000	10,000
Nieuwekans	Agrie	Vrouw	33	Middel	Ja	158	10,000	10,000
Akkerwinde	Agrie	Vrouw	82	Oud	Nee	0	10,000	10,000
Grasmalen	Agrie	Man	2	Jong	Nee	0	10,000	10,000
Akkerwinde	Agrie	Man	32	Middel	Ja	525	10,000	10,000
Smeulde	Indusie	Vrouw	66	Oud	Nee	0	10,000	10,000
Nieuwekans	Agrie	Vrouw	2	Jong	Nee	0	10,000	10,000

We gaan nu met Bascula gewichten berekenen door toepassing van post-stratificatie. We kunnen Bascula vanuit Blaise starten door in het menu *Hulpmiddelen* te kiezen voor de optie *Bascula.* Het is overigens ook mogelijk om Bascula buiten Blaise om te gebruiken. Het voordeel van toepassing vanuit Blaise is echter dat Bascula alle Blaise metadata vanuit Blaise kan overnemen.

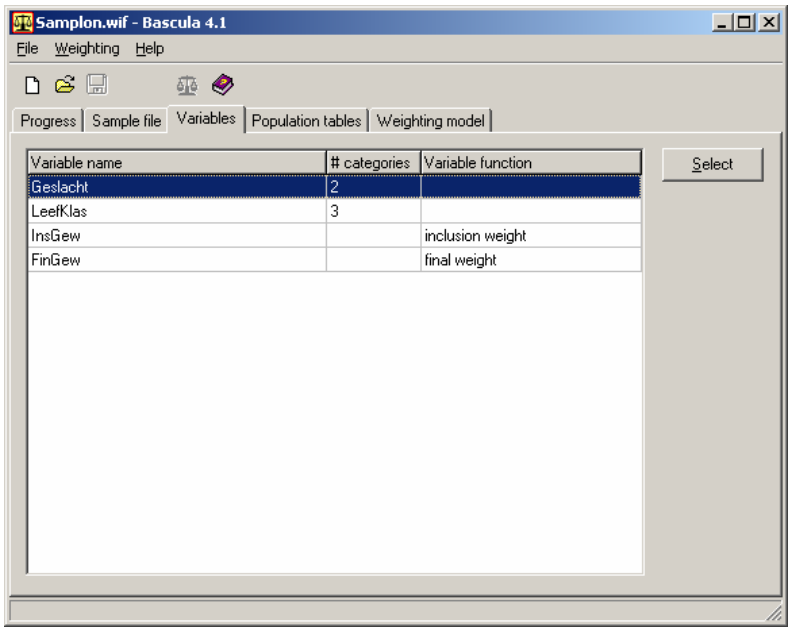
De eerste stap is het opgeven aan Bascula in welk bestand de steekproefgegevens zitten. In dit geval is het een Blaise-database met de naam *Samplon.dbd*. Zie ook figuur 4. Alle data en metadata komen hiermee voor Bascula beschikbaar.

Figuur 4. Specificatie van het steekproefbestand



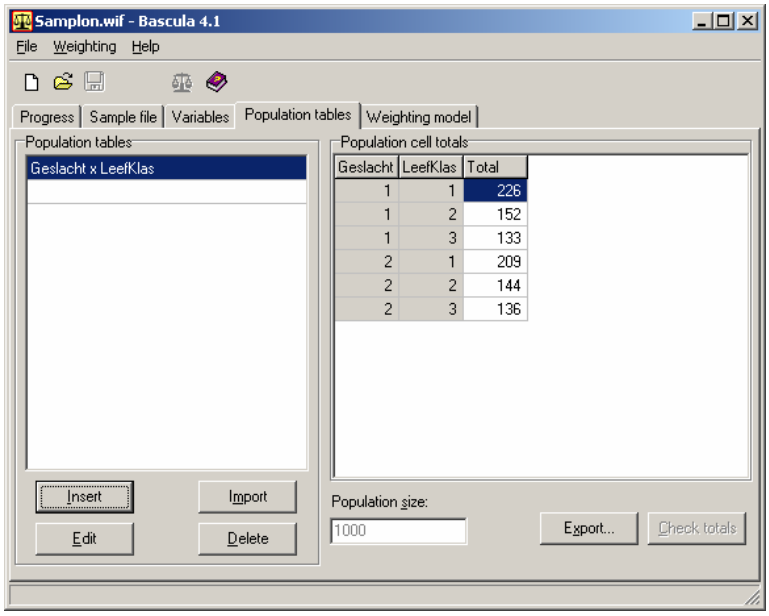
Nu moeten we aangeven welke variabelen kunnen worden gebruikt voor het weegmodel, welke variabele het insluitgewicht bevat, en welke variabele het finale gewicht gaat bevatten. Dat gebeurt op het tab-blad *Variables*, zie figuur 5.

Figuur 5. Het toekennen van functies aan de variabelen.



In dit voorbeeld voeren we een post-stratificatie uit, waarbij we de hulpvariabelen *Geslacht* een *LeefKlas* met elkaar kruisen. Aangezien *Geslacht* twee categorieën heeft en *LeefKlas* zes, zijn er 6 strata. Elk stratum krijgt een ander gewicht.

Figuur 6. Invoeren van de populatie- tabellen

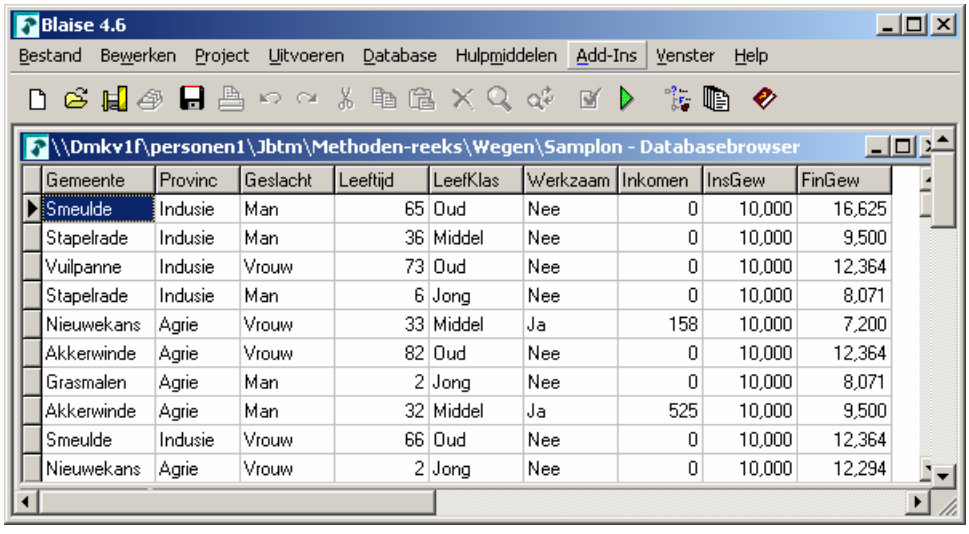


Bascula weet nu dat de variabele *InsGew* het insluitgewicht bevat en dat het uiteindelijke gewicht moet worden opgeslagen in de variabele *FinGew*.

Nu heeft Bascula alle benodigde informatie voor het uitvoeren van een weging. Aan gezien er voldoende waarnemingen per cel zijn (minstens 8), is het inderdaad mogelijk om een post-stratificatie uit te voeren. Kiezen voor lineair of multiplicatief wegen is in deze situatie niet mogelijk.

Na het berekenen van de gewichten sluiten we Bascula af en keren we weer terug naar Blaise. Vanuit Blaise kunnen we een kijkje nemen in het bestand met steek-proefgegevens. De eerste 10 records van het bestand zien er uit zoals weergegeven in figuur 7.

Figuur 7. De eerste 10 records na het berekenen van de gewichten.



Gemeente	Provinc	Geslacht	Leeftijd	LeefKlas	Werkzaam	Inkomen	InsGew	FinGew
Smeulde	Indusie	Man	65	Oud	Nee	0	10,000	16,625
Stapelrade	Indusie	Man	36	Middel	Nee	0	10,000	9,500
Vuilpanne	Indusie	Vrouw	73	Oud	Nee	0	10,000	12,364
Stapelrade	Indusie	Man	6	Jong	Nee	0	10,000	8,071
Nieuwekans	Agrie	Vrouw	33	Middel	Ja	158	10,000	7,200
Akkerwinde	Agrie	Vrouw	82	Oud	Nee	0	10,000	12,364
Grasmalen	Agrie	Man	2	Jong	Nee	0	10,000	8,071
Akkerwinde	Agrie	Man	32	Middel	Ja	525	10,000	9,500
Smeulde	Indusie	Vrouw	66	Oud	Nee	0	10,000	12,364
Nieuwekans	Agrie	Vrouw	2	Jong	Nee	0	10,000	12,294

Ten opzichte van figuur 3 zijn alleen de waarden van de variabele *FinGew* veranderd. Deze variabele bevat nu de gewichten die met post-stratificatie zijn berekend. Merk op dat het eerste record gegevens over een oude man bevat. De waarde van *FinGew* (16,625) is verkregen door het correctiegewicht (1,663, zie ook tabel 3) te vermenigvuldigen met het insluitgewicht in *InsGew* (10,000).

2.5 Kwaliteitsindicatoren

Een goede weegprocedure verbetert de kwaliteit van de schattingen ook al als er geen non-respons zou optreden. Wanneer de strata homogeen zijn met betrekking tot de doelvariabelen van het onderzoek, dan zal de variantie van de schattingen aanzienlijk kleiner worden. Zonder post-stratificatie is de variantie in feite opgebouwd uit een bijdrage veroorzaakt door variatie binnen de strata en een bijdrage die afkomstig is van de variatie tussen de strata. Door toepassing van post-stratificatie vervalt de bijdrage van de variatie tussen de strata. Alleen de variatie binnen de strata blijft over. Die is klein als de strata homogeen zijn.

Als er non-respons optreedt, zal een goede weegprocedure de vertekening van schattingen verwijderen of aanzienlijk verkleinen. Helaas kunnen we nooit in de praktijk nagaan of een weging de vertekening volledig elimineert. Daarvoor zouden we de populatiewaarden van de te schatten grootheden moeten kennen. Die zijn echter niet bekend. Anders zou het onderzoek niet nodig zijn geweest.

Wel is het zo dat naarmate we meer relevante hulpvariabelen gebruiken in de weging, een schatting steeds meer verschuift in een bepaalde richting. Dit geeft echter nog steeds niet de garantie dat een vertekening volledig is verdwenen.

In uitzonderlijke situaties is het mogelijk om de effectiviteit van een weegprocedure te controleren. Dat is het geval als we onderzoek hebben gedaan in een populatie waarvan we een aantal doelvariabelen van te voren al kennen, bijvoorbeeld uit een andere bron.

3. Lineair wegen

3.1 Korte beschrijving

Het is belangrijk om zoveel mogelijk hulpvariabelen te gebruiken in een weeg-procedure, want dan wordt de steekproef in zoveel mogelijk opzichten representatief gemaakt. Dit is een zinvol principe, maar vaak maken praktische problemen de zaak ingewikkeld.

Een eerste probleem is de vulling van de strata. Als er veel hulpvariabelen worden gebruikt voor wegen, zullen er ook veel strata zijn. Uiteraard kan slechts een schatting voor een stratum worden gemaakt als er waarnemingen in dat stratum beschikbaar zijn. Met veel strata en een niet heel grote steekproefomvang is het niet denkbeeldig dat er lege strata zullen zijn. Wordt een schatting gemaakt op basis van alleen de gevulde strata, dan zegt de uitkomst alleen maar iets over de populatie exclusief de lege strata. Er wordt dan dus een verkeerde conclusie getrokken.

Er zijn verschillende manieren om het probleem van de lege strata aan te pakken. Een (wellicht dure) methode is het vergroten van de steekproefomvang. Ook zou we kunnen overwegen om minder hulpvariabelen te gebruiken, wat tot minder strata leidt. Een derde mogelijkheid is het samenvoegen van een leeg stratum met een daarop lijkend gevuld stratum.

Het is niet voldoende om te zorgen dat alle strata gevuld zijn. De strata moeten ook voldoende gevuld zijn. Met maar één waarneming per stratum kunnen we wel een schatting uitrekenen voor het populatiegemiddelde, maar voor het schatten van de variantie hebben we minimaal twee waarnemingen per stratum nodig. Bovendien is het zo dat die variantieschatting erg instabiel is met een gering aantal waarnemingen per stratum. Aanbevolen wordt om toch minstens wel 5 waarnemingen per stratum te hebben.

Bij de toepassing van post-stratificatie kunnen zich ook nog andere problemen voordoen. Die hebben meer te maken met de beschikbare informatie over de populatie. Tabel 4 schetst zo'n probleem.

Tabel 4. Ontbrekende populatie-informatie

Steekproef			
	Man	Vrouw	Totaal
Jong	28	17	45
Middel	16	20	36
Oud	8	11	19
Totaal	48	52	100

Populatie			
	Man	Vrouw	Totaal
Jong	?	?	435
Middel	?	?	296
Oud	?	?	269
Oud	511	489	1000

We zouden beide hulpvariabelen leeftijd en geslacht graag willen gebruiken, maar helaas is voor beide hulpvariabelen niet de volledige informatie beschikbaar: alleen

de marginale verdelingen zijn bekend en niet de verdeling over de cellen. We kunnen geen gewichten berekenen, omdat de populatie-aantallen in de strata niet bekend zijn. Dit probleem zouden we kunnen oplossen door maar één van de twee hulpvariabelen te gebruiken: weeg òf naar leeftijd òf naar geslacht, maar niet naar allebei. Dan maken we niet gebruik van alle aanwezige informatie, zodat de correctie voor non-respons wel eens minder effectief zou kunnen zijn.

Een betere oplossing voor de hierboven genoemde problemen wordt geboden door *lineair wegen* en *multiplicatief wegen*. Dit hoofdstuk is gewijd aan lineair wegen en het volgende aan multiplicatief wegen.

3.2 Toepasbaarheid

Lineair wegen kunnen we toepassen in alle situaties waarin ook post-stratificatie mogelijk is. Beide technieken zijn dan identiek aan elkaar. Lineair wegen kan echter in veel meer situaties worden gebruikt. We kunnen overwegen om lineair wegen toe te passen als

- Post-stratificatie niet mogelijk is als gevolg van lege cellen (strata) of als er cellen zijn met maar heel weinig waarnemingen (zeg minder dan 5).
- Voor de populatie niet de frequentieverdeling van de kruising van alle hulpvariabelen bekend is (maar bijvoorbeeld wel alle marginale frequentieverdelingen).
- Er ook kwantitatieve (continue) hulpvariabelen voor wegen moeten worden gebruikt.

We voeren lineair wegen uit om een mogelijke vertekening ten gevolge van non-respons te verwijderen of te verminderen. Voor alle hulpvariabelen (of kruisingen van hulpvariabelen) in het weegmodel moeten de overeenkomstige frequentieverdelingen in de populatie beschikbaar zijn.

Lineair wegen is ook een zinvolle techniek als er geen non-respons is opgetreden. Indien de strata homogeen zijn, zal post-stratifcatie leiden tot nauwkeuriger schattingen.

3.3 Uitgebreide beschrijving

3.3.1 De algemene regressieschatter

Om lineair wegen te kunnen uitleggen, introduceren we eerst de *algemene regressieschatter*. We laten zien dat toepassing van deze algemene regressieschatter neer komt op een vorm van wegen. Die vorm van wegen noemen we lineair wegen.

We beginnen met de situatie waarin er geen non-respons optreedt. We veronderstellen dat we p hulpvariabelen tot onze beschikking hebben. De vector van waarden van de hulpvariabelen voor element k geven we aan met $X_k = (X_{k1}, X_{k2}, \ldots, X_{kp})'$. Hierin betekent het symbool ' dat de vector of matrix wordt getransponeerd. De vector van populatiegemiddelden van de hulpvariabelen wordt genoteerd met $\overline{X}$.

Als de hulpvariabelen gecorreleerd zijn met de doelvariabele, dan is het mogelijk een vector $B = (B_1, B_2,..., B_p)'$ van regressiecoëfficiënten te vinden zodanig dat de residuen $E_k = Y_k - X_k'B$ minder variëren dan de waarden van de doelvariabele zelf. De waarde van B kan worden gevonden door toepassing van de methode der kleinste kwadraten, en is gelijk aan

$$B = \left(\sum_{k=1}^{N} X_k X_k^{'} \right)^{-1} \left(\sum_{k=1}^{N} X_k Y_k \right). \tag{3.3.1}$$

In geval van volledige respons kunnen we deze vector van coëfficiënten schatten met

$$b = \left(\sum_{i=1}^{n} x_i x_i^{'} \right)^{-1} \left(\sum_{i=1}^{n} x_i y_i \right), \tag{3.3.2}$$

waarin y_i en x_i de steekproefwaarden zijn van de doelvariabele Y en de vector van hulpvariabelen X. Ter wille van de duidelijkheid gaan we er hier vanuit dat de gegevens zijn verkregen via een enkelvoudige aselecte steekproef zonder teruglegging. Het is echter goed mogelijk de theorie uit te breiden voor willekeurige steekproefontwerpen.

Het is niet zo dat b een zuivere schatter is voor B. Wel kan worden aangetoond dat de vertekening kleiner wordt bij een toenemende steekproefomvang. Voor grotere steekproeven is de vertekening verwaarloosbaar klein. Met andere woorden: b is een *asymptotisch zuivere schatter* voor B.

De *algemene regressieschatter* is nu gedefinieerd als

$$\bar{y}_{AR} = \bar{y} + (\bar{X} - \bar{x})'b\,. \tag{3.3.3}$$

Hierin is $\bar{x}$ de vector van steekproefgemiddelden van de hulpvariabelen. Omdat b een asymptotisch zuivere schatter is voor B, is de algemene regressieschatter een asymptotisch zuivere schatter voor het populatiegemiddelde van de doelvariabele. Zie voor meer details hierover Bethlehem (1988).

Als er een vector c van vaste getallen bestaat zodanig dat $Xc = \iota$, waarin $\iota = (1, 1, \ldots, 1)'$ een vector van énen is, dan kunnen we de algemene regressieschatter ook schrijven als

$$\bar{y}_{AR} = \bar{X}'b. \tag{3.3.4}$$

Aan deze voorwaarde is voldaan als het onderliggende regressiemodel een constante term bevat. De waarde X_{k1} van de eerste variabele in de vector $X_k = (X_{k1}, X_{k2}, \ldots, X_{kp})'$ heeft dan altijd de waarde 1. Kiezen we ($c = (1, 0, 0, \ldots, 0)'$), dan levert Xc een vector van enen op. Aan de voorwaarde is ook voldaan in het geval van post-stratificatie. De vector $X_k = (X_{k1}, X_{k2}, \ldots, X_{kp})'$ bevat dan altijd één keer een 1 en de overige waarden zijn 0. Met de keuze $c = (1, 1, 1, \ldots, 1)'$ levert Xc dan een vector van enen op.

Tot nu toe hebben we de algemene regressieschatter beschreven voor de situatie waarin van elk getrokken element de gewenste informatie kan worden verkregen. Wat gebeurt er nu als er non-respons optreedt? In het geval van MCAR (*Missing Completely at Random*) is de non-respons volledig willekeurig. Er is dan geen probleem. Er is slechts sprake van een kleinere steekproef. De hiervoor beschreven theorie kan worden toegepast. Is de non-respons MAR (*Missing at Random*) dan kan een vertekening worden verminderd door gebruik te maken van geschikte hulpvariabelen. In deze situatie kunnen we de volgende aangepaste algemene regressieschatter definiëren:

$$\bar{y}_{AR,R} = \bar{y}_R + (\bar{X} - \bar{x}_R)' b_R . \tag{3.3.5}$$

Hierin geeft het subscript R aan dat het gaat om grootheden gaat die we hebben berekend op basis van de beschikbare respons. Dus

$$\bar{y}_R = \frac{1}{n_R} \sum_{k=1}^{N} a_k R_k Y_k , \tag{3.3.6}$$

met R_k de responsindicator voor element k (R_k, = 1 in geval van respons, en anders R_k, = 0), $\bar{x}_R$ de vector van responsgemiddelden voor de p hulpvariabelen, en b_R gelijk aan

$$b_R = \left(\sum_{k=1}^{N} a_k R_k X_k X_k' \right)^{-1} \left(\sum_{k=1}^{k} a_k R_k X_k Y_k \right) \tag{3.3.7}$$

Bethlehem (1988) toont aan dat de vertekening van de aangepaste algemene regressieschatter onder het Stochastische Responsmodel (met $E(R_k) = P(R_k=1) = \rho_k$) gelijk is aan

$$B(\bar{y}_{AR,R}) = \bar{X}' B_R - \bar{Y}, \tag{3.3.8}$$

waarin B_R is gedefinieerd als

$$B_R = \left(\sum_{k=1}^{N} \rho_k X_k X_k' \right)^{-1} \left(\sum_{k=1}^{N} \rho_k X_k Y_k \right). \tag{3.3.9}$$

Uit formule (3.3.8) kunnen we afleiden dat de vertekening verdwijnt als $B_R = B$. Dus de algemene regressieschatter zal een zuivere schatter zijn als de non-respons de schattingen voor de regressiecoëfficiënten niet aantast.

De vector B_R kan worden geschreven als

$$B_R = B + \left(\frac{1}{N} \sum_{k=1}^{N} \frac{\rho_k X_k X_k'}{\bar{\rho}} \right) \bar{E}_R, \tag{3.3.10}$$

waarin

$$\bar{E}_R = \frac{1}{N} \sum_{k=1}^{N} \frac{\rho_k E_k}{\bar{\rho}} \tag{3.3.11}$$

Hieruit kunnen we de conclusie trekken dat de vertekening van de aangepaste algemene regressieschatter vermindert als grootheid (3.3.11) klein is. Dit is in twee situaties het geval:

- Er kan een goed passend regressiemodel worden gevonden. In dit geval liggen de residuen dicht bij 0, en is dus ook (3.3.11) klein.

- Er is weinig of geen correlatie tussen de residuen en de responskansen. Dan ligt de waarde van de covariantie tussen responskansen en residuen dicht bij 0, en is dus ook (3.3.11) klein.

We kunnen aantonen dat, in het geval van de non-respons van het type MAR is, en wanneer de relevante hulpvariabelen opgenomen zijn in het regressiemodel, grootheid (3.3.11) inderdaad 0 wordt.

De theorie in deze paragraaf toont aan dat de algemene regressieschatter de potentie heeft om een mogelijke vertekening ten gevolge van non-respons te verminderen. Daarom vormt de schatter de basis voor een aantal correctietechnieken. Hieronder gaan we nader in op het gebruik van kwalitatieve hulpvariabelen.

3.3.2 Lineair wegen met kwalitatieve variabelen

De techniek van lineair wegen is gebaseerd op de theorie van de algemene regressieschatter. We zullen laten zien dat de regressieschatter eigenlijk niets anders doet dan gewichten toekennen aan de waargenomen elementen. Vervolgens zullen we de theorie in deze deelparagraaf uitwerken voor kwalitatieve hulpvariabelen.

Veronderstel dat er p hulpvariabelen beschikbaar zijn. Voor het moment nemen we nog even aan dat dit kwantitatieve variabelen zijn. De vector van waarden van deze variabelen voor element k geven we aan met $X_k = (X_{k1}, X_{k2}, ..., X_{kp})'$. De vector van populatiegemiddelden van de hulpvariabelen noteren we met $\overline{X}$.

In deelparagraaf 3.3.1 hebben we laten zien dat we de algemene regressieschatter kunnen schrijven als $\overline{y}_{AR} = \overline{X}'b$, waarin b de vector van geschatte regressie-coëfficiënten is. Dit kan als er een vector c van vaste getallen bestaat zodanig dat $Xc = \iota$, waarin ι een vector van énen is. Bethlehem en Keller (1987) laten zien dat deze schatter kan worden herschreven in de vorm

$$\overline{y}_W = \frac{1}{n}\sum_{i=1}^{n} w_i y_i \tag{3.3.12}$$

waarbij het gewicht w_i gelijk is aan $w_i = v' x_i$. Hierin is

$$v = n\left(\sum_{i=1}^{n} x_i x_i'\right)^{-1} \overline{X}. \tag{3.3.13}$$

een vector van *gewichtscoëfficiënten*. Merk op dat het gewicht volledig wordt bepaald door de individuele waarden van de hulpvariabelen voor de geobserveerde

elementen en door de populatiegemiddelden daarvan. De doelvariabele speelt geen rol.

We zullen nu laten zien dat post-stratificatie een speciaal geval is van lineair wegen. Daartoe vervangen we elke kwalitatieve hulpvariabele door een reeks dummy-variabelen. Stel dat de hulpvariabelen worden gebruikt voor een post-stratificatie met L strata. Dan worden die hulpvariabelen vervangen door L dummy-variabelen, die we hier aangeven met $X_1, X_2, \ldots, X_L$. Voor een waarneming in stratum h geven we de bijbehorende dummy-variabele X_h de waarde 1, en de overige dummy-variabelen krijgen de waarde 0. Als gevolg hiervan kunnen we de vector van de populatiegemiddelden van de dummy-variabelen schrijven als

$$\bar{X} = \left(\frac{N_1}{N}, \frac{N_2}{N}, \ \ldots, \ \frac{N_L}{N} \right)', \tag{3.3.14}$$

en de vector v van gewichtscoëfficiënten is gelijk aan

$$v = \frac{n}{N} \left(\frac{N_1}{n_1}, \frac{N_2}{n_2}, \ldots, \frac{N_L}{n_L} \right)' \tag{3.3.15}$$

We illustreren dit nog eens aan de hand van tabel 2, waarbij gewogen wordt met de hulpvariabelen geslacht (2 categorieën) en leeftijd (3 categorieën). Voor elke cel in de tabel, verkregen door de twee variabelen met elkaar te kruisen, wordt een dummy-variabele ingevoerd. Zit een waarneming in de cel, dan krijgt de corresponderende dummy-variabele de waarde 1, en alle andere dummy-variabelen krijgen de waarde 0. Er zijn in totaal 2 × 3 = 6 cellen, en dus ook 6 dummy-variabelen. De mogelijke waarden van de dummy-variabelen staan in tabel 5. De tabel bevat ook de waarden van de populatiegemiddelden van de dummy-variabelen. Deze waarden komen overeen met de fracties elementen in de populatietabel.

Tabel 5. De waarden van de dummy-variabelen voor Geslacht × Leeftijd

Geslacht	Leeftijd	X_1	X_2	X_3	X_4	X_5	X_6
Man	Jong	1	0	0	0	0	0
Man	Middelbaar	0	1	0	0	0	0
Man	Oud	0	0	1	0	0	0
Vrouw	Jong	0	0	0	1	0	0
Vrouw	Middelbaar	0	0	0	0	1	0
Vrouw	Oud	0	0	0	0	0	1
Populatiegemiddelden		0,226	0,152	0,133	0,209	0,144	0,136
Gewichtscoëfficiënten		0,807	0,950	1,663	0,967	0,720	1,236

De gewichtscoëfficiënten staan in de onderste rij van tabel 5. Deze waarden gebruiken we voor de berekening van de gewichten die worden toegekend aan de waargenomen elementen. Zo wordt het gewicht voor een jonge man bijvoorbeeld 0,807. Merk op dat dit hetzelfde gewicht is als in tabel 2.

Het wegingschema in tabel 5 ontstaat door kruisen van de twee hulpvariabelen *Geslacht* en *Leeftijd*. Daarom wordt het genoteerd met *Geslacht* × *Leeftijd*.

We laten nu zien hoe met lineair wegen het probleem van de ontbrekende populatie-informatie kan worden aangepakt. Lineair wegen biedt de mogelijkheid om beide variabelen in het wegingschema op te nemen zonder de populatiefrequenties in de cellen van de tabel van *Geslacht* tegen *Leeftijd* te kennen. De truc is om een andere reeks dummy-variabelen te gebruiken. In plaats van gebruik te maken van de $2 \times 3 = 6$ dummy-variabelen corresponderend met de cellen in de tabel, worden twee reeksen dummy-variabelen gebruikt: een reeks van twee dummy-variabelen corresponderend met de twee categorieën van *Geslacht*, en een reeks van drie dummy-variabelen corresponderend met de categorieën van *Leeftijd*.

Verder wordt er een dummy-variabele toegevoegd die altijd de waarde 1 heeft. Deze dummy-variabele correspondeert met de de constante term in het regressiemodel. Die term is analogie met 'gewone' regressie-analyse toegevoegd. Dit is net per se noodzakelijk, maar daardoor wordt het wel duidelijker welke bijdrage de variabelen aan de gewichten leveren. Bovendien kan zo simpel worden aangetoond dat er een vector c van vaste getallen bestaat zodanig dat $Xc = \iota$. Met de keuze $c = (1, 0, 0, \ldots, 0)'$ is hieraan voldaan. De regressieschatter kan dan in de vorm (3.3.4) worden geschreven.

In totaal zijn er dan dus $2 + 3 + 1 = 6$ dummy-variabelen. In elke reeks heeft altijd één dummy-variabele de waarde 1, terwijl de overige waarden 0 zijn. De mogelijke waarden van de dummy-variabelen staan in tabel 6.

Tabel 6. De waarden van de dummy-variabelen voor Geslacht + Leeftijd

Geslacht	Leeftijd	X_1	X_2	X_3	X_4	X_5	X_6
Man	Jong	1	1	0	1	0	0
Man	Middelbaar	1	1	0	0	1	0
Man	Oud	1	1	0	0	0	1
Vrouw	Jong	1	0	1	1	0	0
Vrouw	Middelbaar	1	0	1	0	1	0
Vrouw	Oud	1	0	1	0	0	1
Populatiegemiddelden		1,000	0,511	0,489	0,435	0,296	0,269
Gewichtscoëfficiënten		1,068	-0,009	0,009	-0,099	-0,247	0,346

De eerste dummy-variabele stelt de constante term van het weegmodel voor. Deze variabele heeft altijd de waarde 1. De tweede en derde dummy-variabele corresponderen met de twee geslachten. De laatste drie dummy-variabelen hebben betrekking op de drie leeftijdscategorieën. De populatiegemiddelden van de dummy-variabelen zijn nu gelijk aan de fracties personen in de betreffende categorieën.

Merk op dat de populatiegemiddelden van de tweede en derde dummy-variabele tot 1 optellen. Hetzelfde is het geval voor de laatste drie dummy-variabelen. Merk ook op dat in dit wegingschema altijd drie dummy-variabelen de waarde 1 hebben.

Aangezien bij dit wegingschema geen informatie wordt gebruikt uit de tabel van *Geslacht* tegen *Leeftijd*, maar wel van de marginale totalen van *Geslacht* en *Leeftijd*, wordt dit schema aangegeven met *Geslacht* + *Leeftijd*.

We kunnen formule (3.3.13) gebruiken voor het berekenen van de gewichts-coëfficiënten. Vanwege de speciale structuur van het wegingschema kan de berekening echter niet worden uitgevoerd zonder het opleggen van extra condities. Er is sprake van multicollineariteit in het regressiemodel. Zo is in het bovenstaande voorbeeld de som van de twee dummy-variabelen voor geslacht altijd gelijk aan de som van de drie dummy-variabelen voor leeftijd. Dit wordt opgelost door voor elke hulpvariabele de conditie op te leggen dat de som van de bijbehorende gewichtscoëfficiënten gelijk moet zijn aan 0. De onderste regel van de tabel bevat de berekende coëfficiënten na opleggen van deze condities.

Het gewicht van een waargenomen element wordt nu verkregen door optelling van de relevante waarden uit de vector van gewichten. De eerste waarde correspondeert met de dummy-variabele X_1 die altijd gelijk is aan 1. Daarom is er altijd een bijdrage ter grootte van 1,068 aan het gewicht. De volgende twee waarden corresponderen met de twee geslachten man en vrouw. Voor mannen wordt een hoeveelheid 0,009 afgetrokken van het gewicht, en voor vrouwen komt er juist 0,009 bij. De laatste drie waarden corresponderen met de drie leeftijdscategorieën. Afhankelijk van de leeftijd wordt er een hoeveelheid opgeteld bij, dan wel afgetrokken van het gewicht. Zo wordt het gewicht van een oude man 1,068 – 0,009 + 0,346 = 1,405.

De gewichten die zijn verkregen voor het model *Geslacht* × *Leeftijd* zijn niet gelijk aan die van het model *Geslacht* + *Leeftijd*. Zie hiervoor tabel 7. Dat is niet erg verbazingwekkend. Het model *Geslacht* × *Leeftijd* gebruikt immers meer informatie dan het model *Geslacht* + *Leeftijd*

Tabel 7. Gewichten bij post-stratificatie en lineair wegen

Post-stratificatie		
	Man	Vrouw
Jong	0,807	0,967
Middel	0,950	0,720
Oud	1,663	1,236

Lineair wegen		
	Man	Vrouw
Jong	0,960	0,978
Middel	0,812	0,831
Oud	1,405	1,424

In de voorbeelden die we hiervoor gebruikten om lineair wegen te illustreren, zaten slechts twee hulpvariabelen. De weegmodellen hoeven echter niet beperkt te blijven tot twee variabelen. Naarmate we meer hulpvariabelen beschikbaar hebben, zijn er ook steeds meer weegmodellen mogelijk. Stel, bijvoorbeeld, eens dat er drie hulpvariabelen zijn: Geslacht, Leeftijd en Burgerlijke staat. De laatste variabele meet burgerlijke staat in vier categorieën. Indien de verdeling over de volledige kruising van Geslacht, Leeftijd en Burgerlijke staat bekend is, dan kunnen we het weegmodel *Geslacht* × *Leeftijd* × *Burgerlijk staat* gebruiken. Als alleen de verdelingen bekend zijn voor elk tweedimensionale tabel, dan kan het volgende model worden gehanteerd: *(Geslacht* × *Leeftijd)* + *(Geslacht* × *Burgerlijk staat)* +

(Leeftijd × Burgerlijk staat). In feite komt dit schema erop neer dat drie post-stratificaties tegelijkertijd worden uitgevoerd. En als alleen de marginale verdeling van elke variabele apart bekend is, dan kunnen we nog steeds het schema *Geslacht + Leeftijd + Burgerlijk staat* gebruiken.

Het lineair wegen zoals dat is beschreven in deze paragraaf is een speciale toepassing van de algemene regressieschatter. De kwalitatieve hulpvariabelen zijn hier omgezet in dummy-variabelen. Het is uiteraard ook mogelijk om alleen kwantitatieve hulpvariabelen te gebruiken. En zelfs kunnen kwantitatieve en kwalitatieve variabelen in één model worden gecombineerd. Meer informatie hierover is te vinden in Bethlehem en Keller (1987) en Bethlehem (2002).

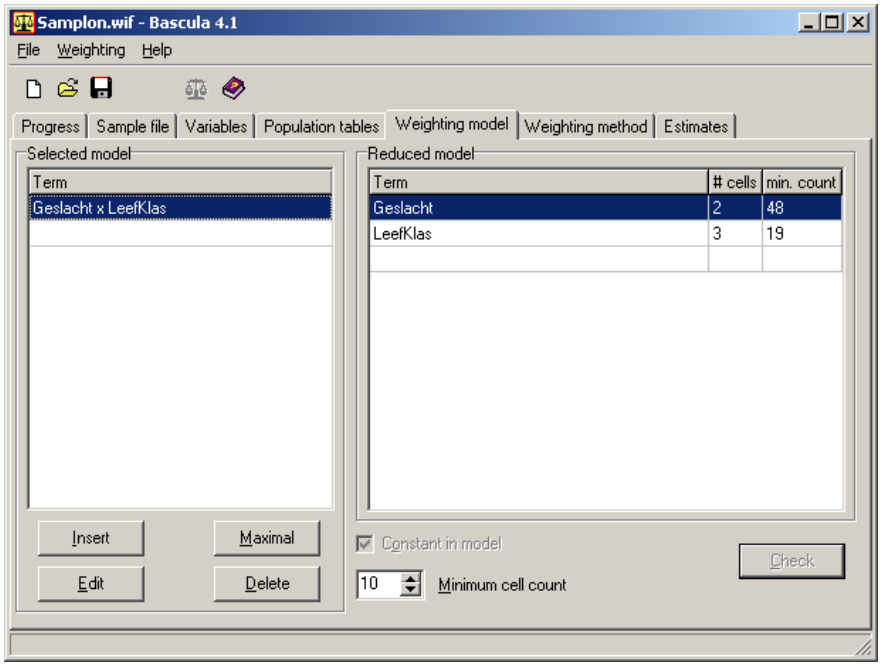
3.4 Voorbeeld

We kunnen het programma Bascula gebruiken voor het uitvoeren van een lineaire weging. We laten dit zien aan de hand van het zelfde voorbeeld dat we gebruikt hebben bij post-stratificatie.

Uit de bevolking van Samplonië (N = 1000) hebben we een steekproef (met gelijke kansen en zonder teruglegging) getrokken van n = 100. Er zijn twee hulpvariabelen die we willen gebruiken voor wegen: geslacht (man, vrouw) en leeftijdsklasse (jong, middelbaar, oud). Tabel 3 bevat de verdeling in de steekproef en in de populatie van beide variabelen.

De gang van zaken in Bascula is in grote lijnen hetzelfde als bij post-stratificatie. Eerst moeten we het steekproefbestand specificeren. Dat is hier de Blaise-database *Samplon.dbd*.

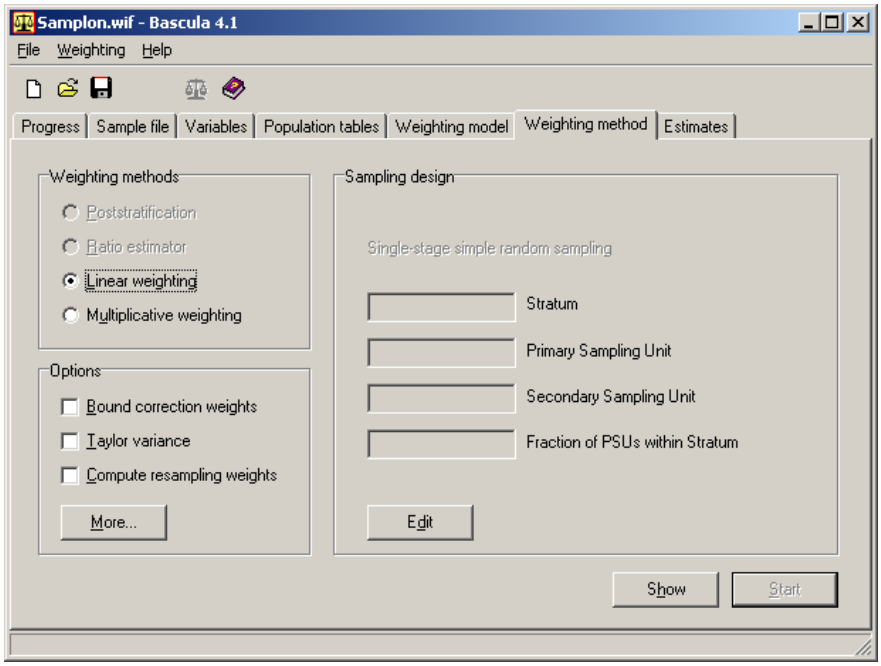
Figuur 8. Bepaling van het maximale weegmodel.



Vervolgens moeten we aangeven welke variabelen kunnen worden gebruikt voor het weegmodel, welke variabele het insluitgewicht bevat, en welke variabele het finale gewicht gaat bevatten

.

Figuur 9. Keuze van de weegtechniek

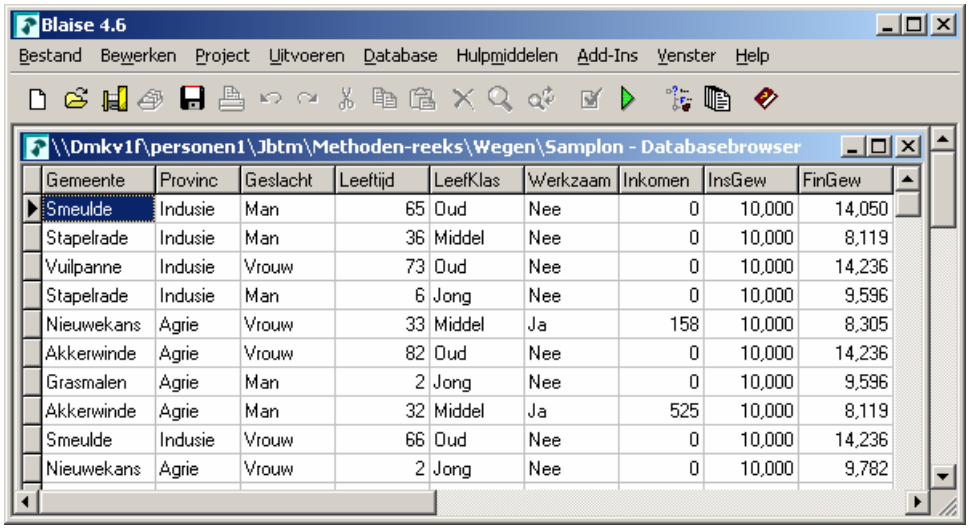


Daarna moeten we de beschikbare populatie-informatie specificeren. We hebben de gegevens voor de kruising van *Geslacht* en *Leeftijd*. Die moeten we invoeren.

We willen stabiele schattingen en eisen daarom minstens 10 waarnemingen in elk stratum. Deze eis kunnen we opgeven bij *Minimal cell count*, zie figuur 8. Door de deze eis valt post-stratificatie als weegtechniek af. Het stratum met de oude mannen bevat immers maar 8 waarnemingen. Bascula reduceert het maximaal mogelijke weegmodel tot *Geslacht* + *LeefKlas* (zie *Reduced model* in figuur 8).

Op het tab-blad *Weighting method* biedt Bascula de keuze uit twee weegtechnieken. Dat zijn lineair wegen en multiplicatief wegen. We kiezen voor lineair wegen. Zie figuur 9. De berekeningen kunnen nu worden gestart. Na afloop zijn de berekende gewichten opgeslagen in het steekproefbestand. In figuur 10.

Figuur 10. De berekende gewichten



Gemeente	Provinc	Geslacht	Leeftijd	LeefKlas	Werkzaam	Inkomen	InsGew	FinGew
Smeulde	Indusie	Man	65	Oud	Nee	0	10,000	14,050
Stapelrade	Indusie	Man	36	Middel	Nee	0	10,000	8,119
Vuilpanne	Indusie	Vrouw	73	Oud	Nee	0	10,000	14,236
Stapelrade	Indusie	Man	6	Jong	Nee	0	10,000	9,596
Nieuwekans	Agrie	Vrouw	33	Middel	Ja	158	10,000	8,305
Akkerwinde	Agrie	Vrouw	82	Oud	Nee	0	10,000	14,236
Grasmalen	Agrie	Man	2	Jong	Nee	0	10,000	9,596
Akkerwinde	Agrie	Man	32	Middel	Ja	525	10,000	8,119
Smeulde	Indusie	Vrouw	66	Oud	Nee	0	10,000	14,236
Nieuwekans	Agrie	Vrouw	2	Jong	Nee	0	10,000	9,782

3.5 Kwaliteitsindicatoren

Een weegprocedure is effectief als hij de vertekening van schattingen verwijdert of verkleint. Daarnaast kan hij ook nog de variantie van de schattingen verkleinen. Er valt nooit te bewijzen of lineair wegen een vertekening volledig elimineert. Daarvoor zouden de waarden van de te schatten grootheden bekend moeten zijn. Die zijn echter niet bekend. Anders zou het onderzoek niet nodig zijn geweest.

Wel is het zo dat naarmate we meer relevante hulpvariabelen worden gebruiken in de weging, een schatting meestal steeds meer verschuift in een bepaalde richting. Dit geeft echter nog steeds niet de garantie dat een vertekening volledig is verdwenen.

In uitzonderlijke situaties is het mogelijk om de effectiviteit van een weegprocedure te controleren. Dat is het geval als onderzoek is gedaan in een populatie waarvan we een aantal doelvariabelen van te voren al kennen, bijvoorbeeld uit een andere bron.

4. Multiplicatief wegen

4.1 Korte beschrijving

In hoofdstuk 3 hebben we lineair wegen beschreven als een weegtechniek die we kunnen toepassen als post-stratificatie niet mogelijk is. In die situaties kunnen we ook overwegen om multiplicatief wegen toe te passen.

In principe kunnen we overal multiplicatief wegen waar we ook lineair kunnen wegen. Er is echter één verschil. Multiplicatief wegen kan alleen met kwalitatieve hulpvariabelen. Bij lineair wegen mogen de hulpvariabelen kwalitatief of kwantitief zijn.

Aan multiplicatief wegen ligt een andere aanpak ten grondslag. Dat leidt ertoe dat gewichten worden berekend door het met elkaar *vermenigvuldigen* van een aantal relevante gewichtsfactoren. Bij lineair wegen ontstaan de gewichten door *optellen* van de relevante gewichtscoëfficiënten.

4.2 Toepasbaarheid

We kunnen multiplicatief wegen toepassen in alle situaties waarin ook post-stratificatie mogelijk is. Beide technieken zijn dan identiek aan elkaar. We kunnen echter multiplicatief wegen in meer situaties gebruiken:

- Als post-stratificatie niet mogelijk is als gevolg van lege cellen (strata) of als er cellen zijn met maar heel weinig waarnemingen (zeg minder dan 5).
- Als voor de populatie niet de frequentieverdeling van de kruising van alle hulpvariabelen bekend is (maar bijvoorbeeld wel alle marginale frequentie-verdelingen).

We kunnen multiplicatief wegen toepassen om een mogelijke vertekening ten gevolge van non-respons te verminderen. Voor alle hulpvariabelen (of kruisingen van hulpvariabelen) in het weegmodel moeten de overeenkomstige frequentie-verdelingen in de populatie beschikbaar zijn.

4.3 Uitgebreide beschrijving

Multiplicatief wegen staat ook wel bekend onder de namen '*raking*', '*raking ratio estimation*' en '*iterative proportional fitting*'.

In het algemeen kunnen we multiplicatief wegen in dezelfde situaties toepassen als lineair wegen, zolang de hulpvariabelen maar kwalitatief zijn. Een voorbeeld is de situatie waarin niet de frequenties beschikbaar zijn van de kruising van alle hulpvariabelen, maar wel voor de kruisingen van kleinere groepjes hulpvariabelen. Bij multiplicatief wegen worden de gewichten uiteindelijk verkregen in een iteratief proces. Het gewicht bestaat dan uit het product van een aantal gewichtsfactoren,

waarbij elke gewichtsfactor de bijdrage is van de kruising van een aantal variabelen in het weegmodel.

Er bestaat geen algemeen theoretisch kader voor multiplicatief wegen. Het is een stapsgewijs rekenproces dat uiteindelijk tot bevredigende resultaten leidt. Het algemene schema van de berekening is als volgt:

1) Introduceer een gewichtsfactor voor elk stratum in elke term (kruising van variabelen). Zet de waarden van deze factoren in eerste instantie op 1.

2) Pas de gewichtsfactoren voor de eerste term zodanig aan dat de gewogen steekproef representatief wordt met betrekking tot de hulpvariabelen die deel uitmaken van die term.

3) Pas de gewichtsfactoren voor de volgende term in het model zodanig aan dat de gewogen steekproef representatief wordt met betrekking tot de hulpvariabelen die deel uitmaken van die term. In het algemeen zal deze actie de representativiteit voor de eerste term verstoren.

4) Herhaal dit aanpassingsproces voor alle termen in het model.

5) Herhaal stappen 2, 3 en 4 net zolang totdat de gewichtsfactoren (bijna) niet meer veranderen.

We illustreren dit schema aan de hand van een simpel voorbeeld. We hebben weer een enkelvoudige aselecte steekproef getrokken van omvang 100 uit een populatie van omvang 1000. We gebruiken de twee hulpvariabelen *Geslacht* (2 categorieën) en *Leeftijd* (3 categorieën). Veronderstel nu dat alleen de populatieverdelingen voor *Geslacht* en *Leeftijd* apart bekend zijn, en niet voor de kruising van Geslacht met Leeftijd. Tabel 8 bevat de startsituatie.

Tabel 8. Startsituatie voor multiplicatief wegen

	Man	Vrouw	Gewichts-factor	Gewogen som	Populatie-verdeling
Jong	0,280	0,170	1,000	0,450	0,435
Middelbaar	0,160	0,200	1,000	0,360	0,296
Oud	0,080	0,110	1,000	0,190	0,269
Gewichtsfactor	1,000	1,000			
Gewogen som	0,520	0,480		1,000	
Popul. verdeling	0,511	0,489			1,000

Het blok linksboven in de tabel bevat de gewogen relatieve frequenties in de steekproef voor elke combinatie van *Geslacht* en *Leeftijd*. Het zijn de Horvitz-Thompson-schattingen voor de overeenkomstige populatiefracties.

De rij en kolom met de naam *Gewichtsfactor* bevatten de initiële waarden van de gewichtsfactoren. De waarden in de rij en kolom met de naam *Gewogen som* krijgen we door eerst het gewicht te bepalen voor elke cel in de tabel (door de relevante rij- en kolomfactor met elkaar te vermenigvuldigen), en vervolgens de gewogen

celfracties bij elkaar op te tellen. Bijvoorbeeld, het gewicht voor een jonge man is 1.000 × 1,000 = 1,000 en het gewicht voor een jonge vrouw is ook 1,000 × 1,000 = 1,000. De gewogen som voor leeftijdscategorie Jong wordt daarmee gelijk aan 0,280 × 1,000 + 0,170 × 1,000 = 0,450.

Aangezien de startwaarden van alle factoren gelijk zijn aan 1, zijn de gewogen sommen hier nog gelijk aan de ongewogen sommen. De waarden in de rij en kolom *Populatieverdeling* bevatten de populatieverdelingen van *Geslacht* en *Leeftijd*.

Het iteratieproces moet ervoor zorgen dat de rij- en kolomfactoren zodanige waarden krijgen dat de gewogen sommen gelijk worden aan de populatiefracties. In de startsituatie van tabel 8 is dat nog niet het geval. Eerst worden nu de gewichtsfactoren voor de rijen aangepast. Het resultaat hiervan staat in tabel 9. Bijvoorbeeld, om de gewogen som voor de categorie Jong op 0,435 te krijgen wordt de desbetreffende rij-factor verlaagd van 1,000 naar 0,435 / 0,450 = 0, 967.

Tabel 9. Situatie na aanpassing voor de rijen

	Man	Vrouw	Gewichts-factor	Gewogen som	Populatie-verdeling
Jong	0,280	0,170	0,967	0,435	0,435
Middelbaar	0,160	0,200	0,822	0,296	0,296
Oud	0,080	0,110	1,416	0,269	0,269
Gewichtsfactor	1,000	1,000			
Gewogen som	0,515	0,485		1,000	
Popul. verdeling	0,511	0,489			1,000

De gewogen sommen voor de rijen zijn nu correct, maar er is nog steeds sprake van een afwijking tussen de gewogen sommen van de kolommen en de bijbehorende populatiefracties. Daarom is nu de volgende stap het aanpassen van de kolomfactoren. Dit is gedaan in tabel 10.

Tabel 10. Situatie na aanpassing voor de kolommen

	Man	Vrouw	Gewichts-factor	Gewogen som	Populatie-verdeling
Jong	0,280	0,170	0,967	0,434	0,435
Middelbaar	0,160	0,200	0,822	0,296	0,296
Oud	0,080	0,110	1,416	0,269	0,269
Gewichtsfactor	0,991	1,009			
Gewogen som	0,511	0,489		1,000	
Popul. verdeling	0,511	0,489			1,000

Merk op dat de aanpassing voor de kolommen de aanpassing voor de rijen heeft verstoord. De gewogen sommen voor de rijen komen niet meer overeen met de bijbehorende populatiefracties. De verschillen zijn echter veel kleiner dan in de startsituatie.

Het proces van beurtelings aanpassen rij- en kolomfactoren wordt nu net zolang herhaald tot de waarden niet meer veranderen. Een dergelijke eindsituatie wordt

meestal bereikt na slechts een handvol iteratiestappen. Die eindsituatie is weergegeven in tabel 11.

Tabel 11. Eindsituatie

	Man	Vrouw	Gewichts-factor	Gewogen som	Populatie-verdeling
Jong	0,230	0,150	0,969	0,435	0,435
Middelbaar	0,160	0,170	0,821	0,296	0,296
Oud	0,130	0,160	1,413	0,269	0,269
Gewichtsfactor	0,991	1,010			
Gewogen som	0,511	0,489		1,000	
Popul. verdeling	0,511	0,489			1,000

Het gewicht voor een waargenomen element wordt nu verkregen door vermenigvuldigen van de betrokken factoren. Tabel 12 bevat die berekening voor alle combinaties van *Geslacht* en *Leeftijd*. De eerste twee kolommen onder Gewicht bevatten de gewichtsfactoren voor de mannen (0,991) en de vrouwen (1,010). De laatste drie kolommen bevatten de gewichtsfactoren voor de jongeren (0,969), middelbaren (0,821) en ouderen (1,413).

Tabel 12. Berekening van de gewichten

Geslacht	Leeftijd	Gewicht					
Man	Jong	0,991		x 0,969			= 0,960
Man	Middel	0,991			x 0,821		= 0,814
Man	Oud	0,991				x 1,413	= 1,401
Vrouw	Jong		1,010	x 0,969			= 0,978
Vrouw	Middel		1,010		x 0,821		= 0,829
Vrouw	Oud		1,010			x 1,413	= 1,427
Factoren		0,991	1,010	0,969	0,821	0,923	

De waarden van de gewichten in tabel 12 wijken maar heel weinig af van de gewichten die worden verkregen met lineair wegen in een overeenkomstige situatie.

In situaties waarin het mogelijk is om zowel lineair als multiplicatief te wegen, kan het bij het maken van een keuze misschien helpen om enkele voor- en nadelen op een rij te zetten:

- Lineair wegen heeft het voordeel dat het gebaseerd is op een simpel lineair model. Daaruit wordt duidelijk in welke situaties lineair wegen wel en niet werkt. En ook kunnen we vaststellen wat het effect van lineair wegen is op de variantie van de schattingen.

- Toepassing van lineair wegen kan *negatieve gewichten* opleveren. Zulke gewichten zijn niet fout, maar simpelweg de consequentie van toepassing van een lineair model dat geen beperkingen oplegt en het teken van de gewichten. Gewoonlijk zijn negatieve gewichten een indicatie van een slecht passend model (grote residuen, weinig verklarende kracht). Dit zou bijvoorbeeld kunnen

worden veroorzaakt door het ontbreken van interactietermen in het model. Lastig is wel dat sommige statistische pakketten (bijvoorbeeld SPSS) geen negatieve gewichten accepteren. Dat kan een reden zijn om af te zien van lineair wegen.

- Aan multiplicatief wegen ligt geen duidelijk model ten grondslag. Daarom is het moeilijk om inzicht te krijgen in de eigenschappen van schatters die zijn gebaseerd op deze weegtechniek. Het is ook bijzonder lastig om de varianties van deze schattingen te bepalen.

- Toepassing van multiplicatief wegen leidt altijd tot positieve gewichten. Als dit een noodzakelijke voorwaarde is, dan dient dus voor deze techniek te worden gekozen.

4.4 Voorbeeld

We kunnen het programma Bascula gebruiken voor het uitvoeren van een multiplicatieve weging. We laten dit zien aan de hand van het zelfde voorbeeld dat we gebruikt hebben bij post-stratificatie en lineair wegen.

Uit de bevolking van Samplonië (N = 1000) hebben we een steekproef (met gelijke kansen en zonder teruglegging) getrokken van n = 100. Er zijn twee hulpvariabelen die we willen gebruiken voor wegen: geslacht (man, vrouw) en leeftijdsklasse (jong, middelbaar, oud). Tabel 3 bevat de verdeling in de steekproef en in de populatie van beide variabelen.

De gang van zaken in Bascula is in grote lijnen hetzelfde als bij lineair wegen. Eerst moeten we het steekproefbestand specificeren. Dat is hier de Blaise-database *Samplon.dbd*.

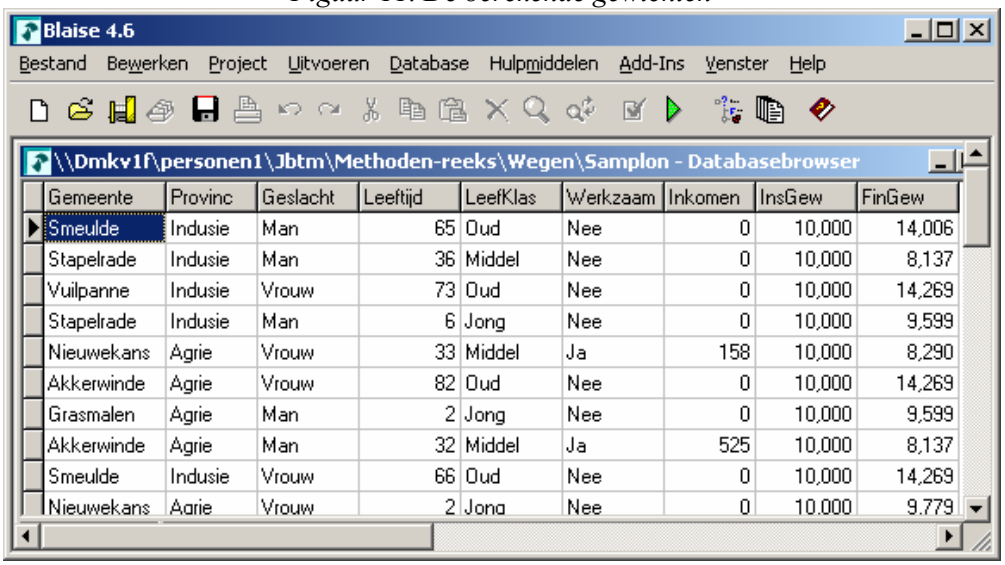
Vervolgens moeten we aangeven welke variabelen kunnen worden gebruikt voor het weegmodel, welke variabele het insluitgewicht bevat, en welke variabele het finale gewicht gaat bevatten.

Daarna moeten we de beschikbare populatie-informatie specificeren. We hebben de gegevens voor de kruising van *Geslacht* en *Leeftijd*. Die moeten we invoeren.

We willen stabiele schattingen en eisen daarom minstens 10 waarnemingen in elke stratum. Door de deze eis valt post-stratificatie als weegtechniek af. Het stratum met de oude mannen bevat immers maar 8 waarnemingen. Bascula reduceert het maximaal mogelijke weegmodel tot *Geslacht* + *LeefKlas*.

Op het tab-blad *Weighting method* biedt Bascula de keuze uit twee weegtechnieken. Dat zijn lineair wegen en multiplicatief wegen. We kiezen voor multiplicatief wegen. De berekeningen kunnen nu worden gestart. Na afloop zijn de berekende gewichten opgeslagen in het steekproefbestand. Zie hiervoor figuur 11.

Figuur 11. De berekende gewichten



Gemeente	Provinc	Geslacht	Leeftijd	LeefKlas	Werkzaam	Inkomen	InsGew	FinGew
Smeulde	Indusie	Man	65	Oud	Nee	0	10,000	14,006
Stapelrade	Indusie	Man	36	Middel	Nee	0	10,000	8,137
Vuilpanne	Indusie	Vrouw	73	Oud	Nee	0	10,000	14,269
Stapelrade	Indusie	Man	6	Jong	Nee	0	10,000	9,599
Nieuwekans	Agrie	Vrouw	33	Middel	Ja	158	10,000	8,290
Akkerwinde	Agrie	Vrouw	82	Oud	Nee	0	10,000	14,269
Grasmalen	Agrie	Man	2	Jong	Nee	0	10,000	9,599
Akkerwinde	Agrie	Man	32	Middel	Ja	525	10,000	8,137
Smeulde	Indusie	Vrouw	66	Oud	Nee	0	10,000	14,269
Nieuwekans	Agrie	Vrouw	2	Jong	Nee	0	10,000	9,779

Een vergelijking van figuur 11 met figuur 10 wijst uit dat de gewichten voor lineair wegen en multiplicatief wegen hier niet veel van elkaar afwijken.

4.5 Kwaliteitsindicatoren

Een weegprocedure is effectief als hij de vertekening van schattingen verwijdert of verkleint. Daarnaast kan hij ook nog de variantie van de schattingen verkleinen. Er valt nooit na te gaan of multiplicatief wegen een vertekening volledig elimineert. Daarvoor zouden de waarden van de te schatten grootheden bekend moeten zijn. Die zijn echter niet bekend. Anders zou het onderzoek niet nodig zijn geweest.

Wel is het zo dat naarmate we meer relevante hulpvariabelen gebruiken in de weging, een schatting steeds meer verschuift in een bepaalde richting. Dit geeft echter nog steeds niet de garantie dat een vertekening volledig is verdwenen.

In uitzonderlijke situaties is het mogelijk om de effectiviteit van een weegprocedure te controleren. Dat is het geval als we onderzoek is gedaan ineen populatie waarvan we een aantal doelvariabelen van te voren al kennen, bijvoorbeeld uit een andere bron.

5. Afsluiting

5.1 Calibratie

Hoewel lineair wegen en multiplicatief wegen zo op het oog twee totaal verschillende technieken lijken te zijn, is het toch mogelijk een soort algemeen raamwerk voor wegen te maken waarvan beide genoemde technieken speciale gevallen zijn. Dit algemene raamwerk is ontwikkeld door Deville en Särndal (1992). Calibratie is echter niet alleen een theoretisch raamwerk om bestaande weegtechnieken opnieuw te beschrijven. Het is ook een raamwerk waarbinnen we nieuwe weegtechnieken kunnen ontwikkelen. Zie hiervoor paragraaf 5.2.

Bij de beschrijving van hun aanpak beperken we ons tot steekproeven zonder teruglegging. Ze introduceren de *calibratieschatter*

$$\overline{y}_{CA} = \frac{1}{N}\sum_{k=1}^{N} w_k a_k Y_k \,. \tag{5.1.1}$$

In deze schatter komen gewichten w_k voor die aan twee voorwaarden moeten voldoen:

- De gewichten w_k moeten zo dicht mogelijk in de buurt liggen van de oorspronkelijke insluitgewichten $d_k = 1/\pi_k$.

- De gewogen steekproefverdeling van de hulpvariabelen moet exact overeen komen met de corresponderende populatieverdeling:

$$\overline{x}_{CA} = \frac{1}{N}\sum_{k=1}^{N} w_k a_k X_k = \overline{X}. \tag{5.1.2}$$

Merk op dat de gewichten w_k anders gedefinieerd zijn dan die in de voorgaande paragrafen. De eerste voorwaarde zorgt ervoor dat de resulterende schatter zuiver, of vrijwel zuiver, is. De tweede voorwaarde zorgt ervoor dat de gewogen steekproef representatief is met betrekking tot de gebruikte hulpvariabelen.

Om aan beide voorwaarden te kunnen voldoen, introduceren Deville en Särndal (1992) een afstandsmaat D. Deze meet het verschil tussen d_k en w_k. Als $D(w_k, d_k)$ de afstand is tussen w_k en d_k, dan komt de oplossing van het probleem neer op het minimaliseren van

$$\sum_{k=1}^{N} a_k D(w_k, d_k) \tag{5.1.3}$$

onder de voorwaarde

$$\frac{1}{N}\sum_{k=1}^{N} w_k a_k X_k = \overline{X}. \tag{5.1.4}$$

Dit probleem kunnen we oplossen met de methode van Lagrange. Hierbij worden de stationaire punten (punten met eerste afgeleide gelijk aan 0) bepaald van de functie

$$L = \sum_{k=1}^{N} a_k D(w_k, d_k) - \lambda' \left(\sum_{k=1}^{N} a_k w_k X_k - \sum_{k=1}^{N} X_k \right), \tag{5.1.5}$$

waarin

$$\lambda = (\lambda_1, \lambda_2, ..., \lambda_p)' \tag{5.1.6}$$

een vector van Lagrange multipliers is.

Door een geschikte afstandsfunctie te kiezen kunnen we hieruit zowel lineair wegen als multiplicatief wegen als speciaal geval terug krijgen. Voor lineair wegen moeten we de afstandsfunctie *D* gelijk nemen aan

$$D(w_k, d_k) = \frac{(w_k - d_k)^2}{d_k}. \tag{5.1.7}$$

Bepalen van de stationaire punten van (5.1.5) voor deze keuze van *D* leidt tot een formule voor de gewichten w_k gelijk is aan die voor lineair wegen.

Voor multiplicatief wegen is de afstandfunctie *D* gedefinieerd als

$$D(w_k, d_k) = w_k \, log(w_k / d_k) - w_k + d_k \tag{5.1.8}$$

Minimaliseren van (5.1.5) voor deze keuze van *D* leidt tot een formule voor de gewichten w_k gelijk is aan die voor multiplicatief wegen.

Deville en Särndal (1992) beschrijven in hun artikel ook nog een aantal andere mogelijke keuzen voor afstandsfunctie *D*. Ze tonen ook aan dat de asymptotische eigenschappen van alle schatters berekend binnen hun raamwerk (en onder volledige respons) hetzelfde zijn. Dit betekent dat het voor grote steekproeven in deze situatie niet zoveel uitmaakt of gekozen wordt voor lineair of multiplicatief wegen. Merk hierbij wel op dat - hoewel de asymptotische eigenschappen hetzelfde zijn - de individuele gewichten die worden berekend met lineair wegen, behoorlijk kunnen afwijken van de gewichten bepaald met multiplicatief wegen.

5.2 Andere aspecten van wegen

5.2.1 Grenzen aan gewichten

Er zijn situaties waarin we als onderzoeker enige controle willen uitoefenen over de waarden die gewichten kunnen aannemen. Eén van die situaties is dat we extreem grote gewichten willen vermijden. Grote gewichten corresponderen meestal met elementen die maar schaars aanwezig zijn in de respons. Grote gewichten leiden tot instabiele schatters voor populatiegrootheden.

Varianties van schatters worden vooral groter in situaties waarin we schattingen maken voor variabelen die geen samenhang vertonen met de weegvariabelen. Kish (1992) toont aan dat de vergroting van de variantie dan met een factor

$$n \frac{\sum_{h=1}^{L} n_h w_h^2}{\left(\sum_{h=1}^{L} n_h w_h \right)^2} \qquad (5.2.1)$$

wordt vergroot. Hierin duidt w_h het gewicht aan dat alle elementen in stratum h hebben. Een andere manier om hier tegen aan te kijken is in termen van de effectieve steekproefomvang. De *effectieve steekproefomvang* is gedefinieerd als de steekproefomvang die nodig is om met een gewone eenvoudige aselecte steekproef zonder teruglegging dezelfde nauwkeurigheid te brengen. Als de effectieve steekproefomvang veel kleiner is dan werkelijke steekproefomvang, dan is dit een duidelijke indicatie dat de gebruikte schattingsmethode niet zo effectief is. We moeten relatief veel waarnemingen doen om een gewenste nauwkeurigheid te bereiken. In bovenstaande situatie is de effectieve steekproefomvang gelijk aan

$$n_{eff} = \frac{\left(\sum_{h=1}^{L} n_h w_h \right)^2}{\sum_{h=1}^{L} n_h w_h^2}$$

Grote gewichten leiden ertoe dat de effectieve steekproefomvang laag wordt, en dus dat de nauwkeurigheid van de schattingen veel minder groot is dan we op grond van de werkelijke steekproefomvang zouden mogen verwachten.

Al eerder is gesignaleerd dat toepassing van lineair wegen kan leiden tot negatieve gewichten. Om die negatieve gewichten te vermijden, zou ook hier controle over de gewichten wenselijk kunnen zijn.

Het is mogelijk de theorie van de calibratieschatter uit te breiden door het opleggen van de randvoorwaarde dat de gewichten binnen zekere grenzen moeten liggen. Dit leidt tot een iteratief proces waarbij een lineair weegprocedure net zolang wordt herhaald tot de gewichten binnen de gewenste grenzen liggen. Als te strikte grenzen worden opgelegd, dan kan het gebeuren dat er geen geschikte gewichten kunnen worden berekend.

Het is mogelijk in Bascula grenzen op te leggen aan de gewichten. Daarvoor is een techniek geïmplementeerd die ontwikkeld is door Huang & Fuller (1978).

5.2.2 Gewichten voor personen en huishouden

Er zijn surveys waarbij personen in twee stappen worden geselecteerd. Er worden dan eerst adressen geloot, en vervolgens op elk geselecteerd adres alle relevante personen. We spreken dan van een *clustersteekproef*.

Als het doel van het onderzoek is om uitspraken te doen over de populatie van individuele personen, dan is er verder geen probleem. We kunnen een weegprocedure uitvoeren en aan elke waargenomen persoon een gewicht toekennen.

Het kan ook zijn dat het doel van het onderzoek is om uitspraken te doen over de populatie van huishoudens, dan kunnen we een weging uitvoeren op de waargenomen huishoudens. Elke huishouden krijgt dan een gewicht. Lastig is hierbij wel dat er weinig hulpvariabelen beschikbaar zijn waarvoor de verdeling in de populatie van alle huishoudens bekend is.

Aangezien het heel goed mogelijk is om gewichten voor personen uit te rekenen, kunnen we onze afvragen of het mogelijk is om met die persoonsgewichten gewichten voor huishoudens te bepalen. Mogelijke aanpakken zouden kunnen zijn: (1) neem het gewicht van het hoofd van het huishouden, (2) neem het gewicht van een willekeurig gekozen lid van het huishouden, of (3) neem het gemiddelde van de gewichten van de leden.

Welke van deze drie aanpakken ook wordt genomen, er ontstaan altijd problemen. Als de huishoudgewichten worden toegepast bij het schatten van kenmerken van individuele personen, dan zullen die gewogen schattingen niet overeenkomen met de bekende waarden voor de populatie. Ook kunnen er inconsistenties ontstaan. We zouden bijvoorbeeld het totale inkomen via de huishoudens of via de individuele personen kunnen schatten. Beide schattingen zullen niet hetzelfde zijn.

De algemene regressieschatter biedt een oplossing. Stel dat de populatie uit N personen bestaat. Die vormen samen $M < N$ huishoudens. Het onderdeel uitmaken van een huishouden leggen we vast in een $N \times M$-matrix H. Iedere rij stelt een persoon voor en iedere kolom een huishouden. Element H_{ij} neemt de waarde 1 aan als persoon i deel uitmaakt van huishouden j. In alle andere gevallen geldt $H_{ij} = 0$.

Als we de matrix H vermenigvuldigen met de $N \times p$-matrix X van waarden voor de personen van de p dummy-hulpvariabelen, dan is het resultaat een $M \times p$-matrix $Z = H'X$. Elke rij van Z representeert een huishouden. Rij i van Z bevat de som van de hulpvariabelen over alle leden van het desbetreffende huishouden.

We kunnen nu de theorie van de algemene regressieschatter toepassen om huishoudgewichten te maken. De doelpopulatie is de populatie van de M huishoudens. In plaats van de matrix X gebruiken we de matrix Z. We krijgen dan gewichten die zodanig zijn dat de gemiddelde waarde per huishouden voor een hulpvariabele in de steekproef overeenkomt met de gemiddelde waarde in de populatie.

Alle leden in het huishouden krijgen hetzelfde gewicht als het huishouden. Hierdoor zijn schattingen op basis van individuele personen consistent met schattingen op basis van huishoudens. Meer over deze vorm van *consistent wegen* kan worden gevonden in Lemaître & Dufour (1987) en Nieuwenbroek (1993).

5.2.3 Variantieschattingen

De theorie van lineair wegen maakt het mogelijk om formules op te stellen voor de variantie van gewogen schattingen. Bij complexe steekproefontwerpen is het niet echt eenvoudig om deze formules te implementeren, omdat daarvoor de tweede orde

insluitkansen nodig zijn. Vooral bij grote steekproeven vraagt dat veel computertijd en computergeheugen.

Aan de theorie van multiplicatief wegen ligt geen duidelijk expliciet model ten grondslag. Daarom bestaan voor deze vorm van wegen geen formules voor de varianties van gewogen schattingen.

Er zijn echter andere manieren om toch schattingen voor de varianties te berekenen, zonder dat daar de tweede orde insluitkansen voor nodig zijn. Een voorbeeld daarvan is de methode van de '*Balanced Half Samples*'. We leggen de methode uit aan de hand van een gestratificeerde tweetrapssteekproef.

Stel dat een gestratificeerde tweetrapssteekproef is getrokken. Het aantal strata wordt aangegeven met L. In de eerste trap zijn in elk stratum twee clusters geloot. Vervolgens wordt in de tweede trap uit elke geselecteerde cluster een aantal elementen geloot.

Met de steekproefgegevens kunnen we een schatting t maken voor een populatie-grootheid T. We kunnen echter ook in elk stratum één van de twee clusters uitkiezen, en op basis van die gegevens een schatting bereken. We gebruiken dan dus maar de helft van de clusters. Dat wordt een 'half sample' genoemd. Er zijn in totaal 2^L manieren om uit de L strata één van de twee clusters te kiezen. Zij nu t_α de schatting op basis van 'half sample' α. Dan kunnen we de variantie van de oorspronkelijke schatter t zuiver schatten met

$$\frac{1}{K}\sum_{\alpha=1}^{K}(t_\alpha - t)^2 , \qquad (5.2.1)$$

waarin K het aantal 'half samples' is.

Een complicatie is nu dat niet elk steekproefontwerp uitgaat van een stratificatie waarbij twee clusters per strata worden getrokken. Daarom zijn aanpassingen vereist. In het programma Bascula is dit geïmplementeerd. Daardoor kan het programma een flink aantal verschillende steekproefontwerpen aan.

Een andere complicatie is het grote aantal mogelijke 'half samples' als er veel strata zijn. Dit kan heel veel rekentijd vergen. Een oplossing hiervoor om slechts een beperkt deel van de 'half samples' te gebruiken. Maar die moeten dan wel zorgvuldig worden geselecteerd. Dit wordt 'balanced half samples' genoemd. Meer informatie over het schatten van varianties kan worden gevonden in bijvoorbeeld Wolter (1985).

5.3 Wegen van steekproeven van bedrijven

De theorie van het wegen als correctie voor non-respons is hier vooral beschreven in de context van onderzoek onder personen en huishoudens. Deze methoden zijn echter in grote lijnen ook heel goed toepasbaar voor onderzoek bij bedrijven. De theorie blijft hetzelfde. Er is alleen sprake van andere onderzoekseenheden. Ook bij bedrijfsenquêtes treedt non-respons op. En hiervoor kan worden gecorrigeerd mits voldoende relevante hulpvariabelen beschikbaar zijn.

Bedrijfsenquêtes verschillen wel in één aspect van persoonsenquêtes, en dat is dat niet elke onderzoekseenheid even groot is. Er zijn grote bedrijven en kleine bedrijven. Het is heel goed voorstelbaar dat het ontbreken van een groot bedrijf (als gevolg van non-respons) een veel groter effect heeft op de uitkomsten van het onderzoek dan een klein bedrijf. In dergelijke situaties ligt het niet voor de hand om zonder meer een weging uit te voeren. Er zou bijvoorbeeld kunnen worden overwogen om de waarden van de variabelen van door non-respons ontbrekende bedrijven alsnog te verkrijgen door ermee contact op te nemen. Ook zouden die waarden kunnen worden geschat met behulp van een imputatietechniek.

6. Literatuur

Bethlehem, J.G. (1988), Reduction of Nonresponse Bias through Regression Estimation. *Journal of Official Statistics* 4, blz. 251-260.

Bethlehem, J.G. (2002), Weighting nonresponse adjustments based on auxiliary information. In: R.M. Groves, D.A. Dillman., J.L. Eltinge en R.J.A. Little. (red), *Survey Nonresponse*. Wiley, New York, blz. 275-288.

Bethlehem, J.G. en Keller, W.J. (1987), Linear weighting of sample survey data. *Journal of Official Statistics* 3, blz. 141-154.

Deville, J.C. en Särndal, C.E. (1992), Calibration estimation in survey sampling. *Journal of the American Statistical Association* 87, blz. 376-382.

Huang, E.T. en Fuller, W.A. (1978), Nonnegative regression estimation for sample survey data. *Proceedings of the Social Statistics Section, American Statistical Association*, blz. 300-305.

Kish, L. (1992), Weighting for unequal $P_{i.}$ *Journal of Official Statistics* 8, blz. 183-200.

Lemaître, G. & Dufour J. (1987), An integrated method for weighting persons and families. *Survey Methodology* 13, blz. 199-207.

Nieuwenbroek, N.J. (1993), *An integrated method for weighting characteristics of persons and households using the linear regression estimator*. Report 8445-93-M1-1, Statistics Netherlands, Voorburg, The Netherlands.

Särndal, C.E. en Lundström, S. (2005), *Estimation in Surveys with Nonresponse*. John Wiley & Sons, Chichester, England.

Wolter, K.M. (1985), *Introduction to variance estimation*. Springer Verlag, New York.