Imputatie

10

Abby Israëls, Léander Kuijvenhoven, Jan van der Laan, Jeroen Pannekoek en Eric Schulte Nordholt

Statistische Methoden (201101)

Centraal Bureau voor de Statistiek

Den Haag/Heerlen, 2011

Verklaring van tekens

.	=	gegevens ontbreken
*	=	voorlopig cijfer
**	=	nader 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
2010–2011	=	2010 tot en met 2011
2010/2011	=	het gemiddelde over de jaren 2010 tot en met 2011
2010/'11	=	oogstjaar, boekjaar, schooljaar enz., beginnend in 2010 en eindigend in 2011
2008/'09–2010/'11	=	oogstjaar, boekjaar enz., 2008/'09 tot en met 2010/'11

In geval van afronding kan het voorkomen dat het weergegeven totaal niet overeenstemt met de som van de getallen.

Colofon

Uitgever
Centraal Bureau voor de Statistiek
Henri Faasdreef 312
2492 JP Den Haag

Prepress
Centraal Bureau voor de Statistiek - Grafimedia

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

60165201101 X-37

Inhoudsopgave

1. Inleiding op het thema

1.1 Algemene beschrijving

1.1.1 Beschrijving van het thema

Bij enquêtes komt het voor dat respondenten op één of meer vragen geen antwoord geven, terwijl dit wel van ze wordt verlangd. Men spreekt dan van *item-nonrespons* (of partiële nonrespons) en van (ten onrechte) *ontbrekende waarden* (*missing values*). Redenen om een vraag niet te beantwoorden zijn het niet kunnen of willen geven van het antwoord. Op ingewikkelde of moeilijk te begrijpen vragen kan men vaak geen antwoord geven, op gevoelige vragen wil men het dikwijls niet. Ook bij registers kunnen gegevens ontbreken die het CBS wel had willen hebben.

Er zijn een aantal manieren om met ontbrekende waarden om te gaan. Eén daarvan is *imputeren* van een geldige waarde voor de ontbrekende waarde in het data-bestand. We spreken dan van *imputeren* of *imputatie* (zie paragraaf 1.4 voor de definitie) als processtap en van een *geïmputeerde waarde* of *imputatie* als resultaat.

Een alternatief voor imputeren is om het achterwege te laten. De ontbrekende waarden blijven dan onbekend. Dit zal men in de eerste plaats doen bij terecht ontbrekende waarden. Mensen zonder baan hoeven geen antwoord te geven op vragen over de werkkring; meestal zorgt de routing in de vragenlijst er al voor dat deze vragen alleen aan de werkzame personen worden gesteld. Antwoorden als ‘weet niet’, ‘geen mening’ of ‘onbekend’ zal men ook zo laten wanneer ze iets zeggen over de kennis of oordeel van de respondent. Maar zelfs in het geval van ten onrechte ontbrekende waarden kan men besluiten niet te imputeren en het probleem niet in het data-bestand op te lossen, maar bij de schatting of analyse. Speciaal bij kwalitatieve variabelen heeft men als alternatief het introduceren van een categorie ‘onbekend’. Imputatie wordt dan ook vaker toegepast bij kwantitatieve dan bij kwalitatieve variabelen, en daarom ook vaker bij bedrijfsstatistieken dan bij sociale statistieken.

Redenen om te imputeren, in plaats van het veld leeg te laten, zijn:

1. het verkrijgen van een ‘volledig’ (geheel gevuld) data-bestand;
2. verhoging van de kwaliteit van het micro-bestand en/of van de parameterschattingen.

Ad 1. Het verkrijgen van een volledig bestand, met volledige records, vergemakkelijkt het aggregeren en tabelleren, en voorkomt inconsistenties tussen tabellen. Zo leiden ontbrekende waarden op een variabele Opleiding (in klassen) ertoe dat de leeftijdsverdeling in de tabel ‘Leeftijd × Opleiding’ afwijkt van de leeftijdsverdeling in de tabel ‘Leeftijd × Geslacht’, tenzij je ‘onbekend’ als categorie meeneemt; men zou de inconsistenties ook kunnen opheffen door ‘consistent en herhaald’ wegen (zie Methodenreeks, thema ‘Steekproeftheorie’, deelthema

‘Herhaald wegen’). Wanneer er bij een steekproefonderzoek scores ontbreken op de (kwantitatieve) variabele Inkomen, dan kan men het gemiddelde inkomen alleen schatten voor de (deel)populatie van personen die bij ondervraging gerespondeerd zouden hebben, en dat is een weinig relevante parameter. Imputatie verhelpt dit probleem, maar is natuurlijk alleen bruikbaar wanneer de imputaties van voldoende kwaliteit zijn.

Ad 2. Wanneer men imputatie wil toepassen om de kwaliteit te verhogen moet duidelijk zijn ‘de kwaliteit waarvan’. Vaak is het voornaamste doel gemiddelden en totalen nauwkeurig te bepalen, zoals bij de productiestatistieken, waar totale omzetten de belangrijkste output zijn. Maar men kan ook de verdeling van een variabele zo goed mogelijk willen bepalen; denk bijvoorbeeld aan een inkomensverdeling en bijbehorende ongelijkheidsmaatstaven. Bij leefsituatie-onderzoeken is het ook belangrijk een goed micro-bestand te hebben, waarop onderzoekers allerlei analyses kunnen uitvoeren. Verschillende doelstellingen kunnen tot verschillende ‘optimale’ imputaties leiden. Voor het maken van statistieken wil men echter maximaal één imputatie per ontbrekende waarde hebben, omdat anders de onderzoeksresultaten niet meer intern consistent zijn. In het algemeen kan het CBS betere imputaties voor algemeen gebruik leveren dan externe gebruikers, omdat zij vaak niet over alle achtergrondkenmerken beschikken die bij het imputeren van nut zijn.

1.1.2 Probleem en oplossingen

1.1.2.1 Leeswijzer

Soms valt bij het ontbreken van een score de ‘werkelijke’ waarde met 100% zekerheid af te leiden uit andere kenmerken van het object. Men kan dan *deductieve imputatie* (hoofdstuk 2) toepassen door die waarde te imputeren. Men maakt hiervoor gebruik van afleidingsregels die dikwijls ook bij het editen worden gebruikt. Indien toepasbaar, heeft deze methode voorrang boven alle andere imputatiemethoden. Men gebruikt deze imputatiemethode ook wel wanneer men iets minder dan 100% zekerheid over de juistheid heeft.

Ook als een dergelijke afleiding niet mogelijk is, beschikt men vaak toch over extra informatie (hulpvariabelen, *x*-variabelen) die een nauwkeurige schatting van de ontbrekende waarde (op de *y*-variabele) mogelijk maakt. Door het zoeken naar een geschikt, goed verklarend model, kan men proberen via *modelmatige imputatie* de kwaliteit van het bestand of van de te schatten populatie-parameters te verbeteren. Het gekozen model genereert dan de in te vullen waarde(n). Exacte toetsing van de kwaliteit van de imputaties is echter niet mogelijk: de echte waarden zijn immers onbekend, tenzij men in staat is uit andere bronnen of enquêtes informatie te krijgen. Schatting van het model is alleen mogelijk voor de item-respondenten. Er zal dan ook meestal sprake zijn van een imputation bias (onzuiverheid in de uitkomsten doordat men feilbare imputaties creëert), want het gefitte model met de parameters zal meestal niet exact gelden voor de nonrespondenten.

Wanneer er geen zekerheid bestaat omtrent de te imputeren waarde $\tilde{y}_i$, kan men dus proberen deze modelmatig te schatten. Men zoekt dan naar een model voor *y* waarmee de ontbrekende waarde y_i zo goed mogelijk kan worden voorspeld. Vaak gebruikt men hiervoor een regressiemodel en spreekt dan van *regressie-imputatie* (hoofdstuk 5). Dit past men vooral toe voor kwantitatieve *y*-variabelen. De in hoofdstuk 3 en 4 te behandelen *mean imputation* en *ratio-imputatie* zijn speciale gevallen van regressie-imputatie. Bij mean imputation (imputeren van het gemiddelde) wordt geen hulpinformatie gebruikt, meestal omdat deze niet beschikbaar is; bij ratio-imputatie wordt slechts één kwantitatieve hulpvariabele gebruikt. Deze methoden worden apart behandeld vanwege hun eenvoud en veelvuldige toepassing. Daarnaast bestaan er *donor-imputatiemethoden (hot deck)*: *random hot deck*, *sequentiële hot deck* en *nearest neighbour* (incl. *predictive mean matching*); zie hoofdstuk 6. Qua doelstelling zijn deze methoden vergelijkbaar met regressie-imputatie Maar ze kunnen wat eenvoudiger worden toegepast indien er meerdere ontbrekende waarden op één record geïmputeerd moeten worden, terwijl de verbanden tussen de variabelen daarbij beter worden geschat. Bij donor-imputatie wordt voor iedere nonrespondent *i* een donor-record *d* gezocht met zoveel mogelijk dezelfde kenmerken als de *recipiënt i*, voor zover de kenmerken van invloed worden geacht op de doelvariabele *y*. Vervolgens wordt de donor-score, y_d, gebruikt als imputatie: $\tilde{y}_i = y_d$. In hoofdstuk 7 volgt het probleem van *multivariate imputatie*, waarbij meerdere ontbrekende waarden bij één object voorkomen, en enkele oplossingen hiervoor. In hoofdstuk 8 wordt aandacht besteed aan imputatie bij longitudinale data (w.o. panels). Hierbij kan gebruik worden gemaakt van gegevens van hetzelfde object op andere tijdstippen, eventueel zonder gebruik te maken van gegevens van de andere objecten.

In de rest van paragraaf 1.1.2 bespreken we een aantal zaken die mede bepalend zijn voor de keuze van de imputatiemethode of voor de wijze van toepassen van de methoden. Overigens kunnen verschillende experts tot verschillende keuzen komen, of tot verschillende uitwerkingen van dezelfde methode.

1.1.2.2 Imputatievariabele kwantitatief of kwalitatief

Donor-imputatie (hoofdstuk 6) kan voor ieder type *y*-variabele worden toegepast. Regressie-imputatie (hoofdstuk 5) wordt vooral toegepast als *y* een kwantitatieve variabele is. Meestal gebruikt men hiervoor het lineaire regressiemodel, maar er is geen bezwaar om andere dan lineaire functies van *y* te hanteren. Ook als *y* een kwalitatieve variabele is kan men regressie-analyse toepassen, maar dan worden het aangepaste modellen zoals binaire of multinomiale logistische regressie.

1.1.2.3 Wel/geen hulpinformatie beschikbaar

Indien men bij een kwantitatieve *y*-variabele geen hulpinformatie (x-variabelen) gebruikt, omdat er geen beschikbaar is of omdat er nauwelijks winst mee wordt behaald, gaat regressie-imputatie over in mean imputation (hoofdstuk 3). We behandelen deze methode apart vanwege de populariteit ervan.

Indien bij een kwalitatieve *y*-variabele geen hulpvariabelen beschikbaar zijn, kan men de meest voorkomende waarde (de modus) imputeren, wat normaliter niet is aan te bevelen, of loten uit de categorieën met kansen evenredig aan de geobserveerde categoriefrequenties. Dit laatste komt overeen met het imputeren met een random donor (hoofdstuk 6) uit de gehele populatie. Imputatie zonder gebruik van hulpinformatie is alleen te rechtvaardigen wanneer het weinig item-nonrespondenten betreft en de imputaties weinig invloed hebben op de te schatten parameters.

1.1.2.4 Imputatie per deelpopulatie

Men kan een imputatiemodel maken voor de hele populatie, of per deelpopulatie zoals per SBI × Grootteklasse bij bedrijfsstatistieken. Het onderscheiden van dergelijke imputatieklassen (imputatiestrata) heeft zin wanneer er binnen de klassen weinig variatie is in de scores op imputatievariabele *y* (intern homogeen) en de scores tussen de klassen wel sterk verschillen. Omdat bij regressie-analyse ook kwalitatieve *x*-variabelen in het imputatiemodel kunnen worden meegenomen, kan men het onderscheiden van deelpopulaties daar ook zien als onderdeel van het modelleren, namelijk het selecteren van hulpvariabelen die sterk samenhangen met doelvariabele *y* en het in het model meenemen van deze variabelen met alle interactietermen. Hot deck donor-imputatie (hoofdstuk 6) is per definitie alleen bedoeld voor kwalitatieve *x*-variabelen, dus voor deelpopulaties. De y-variabelen mogen kwalitatief of kwantitatief zijn.

1.1.2.5 Selectie van hulpvariabelen of deelpopulaties

Selectie van variabelen en interacties wordt hier niet uitgebreid behandeld. Het is net als regressie-analyse een onderdeel van de multivariate analyse waarover veel literatuur bestaat. Men zoekt naar hulpvariabelen die sterk zijn gecorreleerd met de doelvariabele *y* én bij voorkeur het selectie-effect zo goed mogelijk verklaren. Het is veelal een kwestie van trial and error en gezond verstand, maar men kan ook met forward of backward zoekprocedures automatisch *x*-variabelen aan het model toevoegen of eruit weglaten. Voor het selecteren van homogene imputatieklassen (*x*-variabelen kwalitatief) bestaan ook automatische zoekprocedures, zoals het mede op het CBS ontwikkelde WAID en de SPSS-module Answer trees. In het slothoofdstuk worden zullen enkele richtlijnen gegeven.

Men kan een norm stellen aan de fractie verklaarde variantie van het model voor de respondenten (R^2). Meestal zal een dergelijke maat een kwaliteitsnorm zijn voor de sterkte van de lineaire relatie tussen *y* en de *x*-variabelen.

1.1.2.6 Imputatie met of zonder storingsterm (y kwantitatief)

Men kan voor een ontbrekende waarde op *y* de best mogelijke voorspelling volgens het regressiemodel imputeren. Doet men dit voor alle ontbrekende waarden, dan wordt er ‘te mooi’ geïmputeerd. Alle geïmputeerde records voldoen dan perfect aan het imputatiemodel. De imputaties worden dan vaak onbruikbaar bij nadere analyses op het microbestand, of ook al soms bij eenvoudige tabellen, hetgeen een reden is

om de geïmputeerde waarden te ‘vlaggen’ (paragraaf 9.1). Een bekend voorbeeld is een nationale bevolkingsstatistiek, waar bij onbekende leeftijd van man of vrouw de imputatieregel werd gebruikt dat de man twee jaar ouder is dan de vrouw. Een dergelijk imputatiemodel kan best goed zijn voor de leeftijdsverdeling van zowel mannen als vrouwen. Maar onderzoekers die het datamateriaal gebruikten kwamen tot de verrassende ontdekking dat er een piek zat in het leeftijdsverschil tussen man en vrouw.

In het algemeen zorgt het imputeren van de best mogelijke voorspelling volgens het regressiemodel voor een onderschatting van de variatie in de scores (‘regressie naar het gemiddelde’). Het leidt tot te gepiekte verdelingen en te dunne staarten, met name wanneer *y* veel ontbrekende waarden heeft en de regressie weinig van de variantie van *y* verklaart (lage R^2). Bij mean imputation is dit effect het sterkst. Voor het schatten van gemiddelden of totalen is dit allemaal geen bezwaar, maar wel voor het schatten van verdelingen (bijv. een inkomensverdeling) en spreidingsmaten.

Wanneer men de verdeling goed wil bepalen, is het raadzaam om aan de best mogelijke voorspelling een random storing toe te voegen. Men kan bij regressie-analyse kiezen tussen (1) trekking uit een (normale) kansverdeling, en (2) toevoeging van het residu van een aselect getrokken donor. In hoofdstuk 5 onderscheiden we regressie-impuatie zowel met als zonder toevoeging van zo’n residu. In hoofdstukken 3 en 4, bij mean imputation en ratio-imputatie, bespreken we alleen de variant zonder storingsterm. Toevoegen van een storingsterm valt dan onder regressie-imputatie.

Bij donor-imputatie wordt impliciet ook een residu gebruikt, namelijk die van de al dan niet random gekozen donor. De spreiding in de verdeling van *y* blijft dus behouden.

Rubin (1987) merkt op dat na het toevoegen van een random storing de variantie van *y* toch nog wordt onderschat, als gevolg van de onzekerheid van het imputatiemodel. Men kan ook die onderschatting tenietdoen via *multipele imputatie*. Men brengt dan meerdere imputaties per ontbrekende waarde aan door meerdere parameterschattingen, random storingen of modellen te creëren. Toevoeging van de variantie tussen de imputaties per record zorgt dan voor een zuivere schatting van de variantie van y.

1.1.2.7 Deterministische of stochastische imputatie

Wanneer er random getrokken wordt uit donoren of uit een verdeling van residuen, dan spreekt men van *stochastische imputatie*. Vanwege deze stochastiek zijn de imputaties dan niet reproduceerbaar. Bij *deterministische imputatie* zijn de imputaties wel reproduceerbaar, gegeven het gekozen imputatiemodel. Het onderscheid tussen stochastische en deterministische imputatie gaat in veel gevallen gepaard met het in de vorige deelparagraaf behandelde onderscheid tussen wel/niet gebruik maken van een residu. Nearest neighbour imputatie, inclusief predictive mean matching, is echter deterministisch, omdat de donor via een bepaalde afstandsfunctie vastligt.

1.1.2.8 Keuze tussen regressie- en donor-imputatie / x-variabelen kwalitatief of kwantitatief

De keuze tussen regressie- en donor-imputatie is vaak niet vanzelfsprekend. Dit komt vooral door het onbekend zijn van de werkelijke, ontbrekende waarden. Er valt niet te toetsen welk model beter is. We geven hier toch een aantal zaken die een invloed op de keuze kunnen hebben.

- Bij regressie-analyse en nearest neighbour zijn zowel kwalitatieve als kwantitatieve *x*-variabelen mee te nemen. Bij hot deck donor-imputatie kan men alleen kwalitatieve variabelen meenemen, tenzij met de kwantitatieve variabelen vooraf discretiseert. Maar dan gaat daarbij het kwantitatieve aspect van de variabele deels verloren.
- Bij hot deck donor-imputatie is men soms beperkt in het opnemen van belangrijke *x*-variabelen in het model t.o.v. regressie-imputatie. Men is namelijk gedwongen alle interacties tussen de kwalitatieve variabelen mee te nemen, waardoor het aantal parameters groot kan worden t.o.v. de steekproefomvang. Bij het regressiemodel kan men zuiniger omgaan met het aantal parameters.
- Door kwantitatieve *x*-variabelen te categoriseren, door ze te vervangen door series dummy-variabelen (één per categorie) verliest men informatie. Maar als er sprake is van een sterk niet-lineaire relatie met y, levert dit categoriseren wel een grotere verklaarde variantie op.
- Bij donor-imputatie is de geïmputeerde donor-score altijd een geldige waarde. Indien bijvoorbeeld *y* alleen gehele getallen kan aannemen, zal de regressie-voorspelling vrijwel nooit geheeltallig zijn, terwijl bij donor-imputatie wel alleen gehele getallen kunnen worden geïmputeerd. Bij donor-imputatie voldoet het ontvangende record ook automatisch aan de controleregels indien het donor-record daaraan voldoet en de koppeling van donor en ontvanger exact is op de *x*-variabelen.
- Wanneer er op een record meerdere waarden ontbreken, is donor-imputatie makkelijker toepasbaar; zie hoofdstuk 7 over multivariate imputatie.

1.1.2.9 Wel / niet wegen

Bij de meeste te behandelen methoden bestaat de mogelijkheid om bij het imputeren de item-respondenten ongelijk te wegen, bijvoorbeeld door hun gewichten omgekeerd evenredig aan de insluitkansen (kans om in de steekproef te zitten) te geven, of gewichten die volgen uit de herweging ter compensatie van selectieve unit-nonrespons.[1] Bij (lineaire) regressie-imputatie houdt dit in dat men een gewogen kleinste-kwadraten-schatting uitvoert, en bij hot-deck donor-imputatie dat potentiële donoren met een kleine insluitkans, en dus een groot insluitgewicht, een grotere kans hebben om donor te worden dan potentiële donoren met een grote insluitkans. Op deductieve imputatie en bij nearest neighbour is wegen niet van invloed.

[1] Overigens hebben ook de item-nonrespondenten een ophooggewicht.

Er is geen eenduidig advies te geven over het gebruik van de gewichten. Modelmatig gezien is iedere uitkomst even betrouwbaar gemeten, wanneer men uitgaat van identiek verdeelde storingen, ongeacht de insluitkans of responskans. Geloof in het imputatiemodel betekent derhalve dat men niet hoeft te wegen, en het zelfs beter achterwege kan laten, omdat wegen de standaardfouten groter maakt. Indien men de variabele met gewichten, of de aan de weging ten grondslag liggende variabelen, als verklarende variabelen in het model kan opnemen, is wegen ook overbodig. Een optie is dan ook om hier bij de selectie van *x*-variabelen voor te zorgen. Meer hierover is te vinden in Pannekoek en Israëls (2000). Echter, vanuit de steekproeftheorie gezien zijn de antwoorden van een steekproefelement 'representatief' voor populatie-elementen die niet zijn getrokken, als zouden zij hetzelfde antwoord hebben gegeven. Vanuit dit principe (of uitgaand van aselecte unit nonrespons) is wegen nodig om steekproeftechnisch zuivere schatters te krijgen. Voor donor-imputatie geeft Kalton (1983) een aantal methoden waarbij de kans op donorschap evenredig is met het gewicht. Het kan handig zijn om er tevens voor te zorgen dat donor en recipiënt een ongeveer even groot gewicht krijgen, om te voorkomen dat een object met een zeer klein gewicht donor wordt voor een recipiënt met een heel groot gewicht, waardoor het gewicht van de donor onevenredig toeneemt (teveel gewicht krijgt). Men kan dit ook trachten te voorkomen door de wegingsvariabele of de aan de weging ten grondslag liggende hulpvariabelen als categoriale *x*-variabele(n) mee te nemen.

Soms wil men niet alleen voor de item-nonrespondenten een score imputeren, maar voor alle niet in de steekproef voorkomende objecten. We noemen dit 'massa-imputatie', ook al is er maar één doelvariabele *y* in het geding. Men moet dan natuurlijk wel over een register of steekproefkader beschikken. Voor de imputatie van de niet-steekproefelementen geldt ook dat wegen minder nodig is naarmate de wegingsvariabelen als *x*-variabelen in het model zijn opgenomen. Maar het kan voorkomen dat dit niet mogelijk is omdat de wegingsvariabelen alleen voor de steekproefelementen bekend zijn. Dan is wegen een optie. Na massa-imputatie kan men eenvoudig totalen en gemiddelden voor *y* berekenen. Bij een weighted hot-deck procedure komt dit overeen met het gebruik de poststraficatieschatter, en bij de gewogen kleinste kwadratenschatting met de regressieschatter; zie het thema 'Steekproeftheorie', deelthema's 'Steekproefontwerpen en Ophoogmethoden'. Dergelijke schatters worden ook 'synthetische schatters' genoemd en worden besproken in het deelthema 'Synthetische schatters en kleine-domeinschatters' van het thema 'Modelmatig schatten'. Alleen worden de schatters daar rechtstreeks berekend, zonder imputaties in het data-bestand aan te brengen.

1.1.2.10 Overige zaken

De volgende zaken, die niet direct de keuze van de methode beïnvloeden maar wel aandacht verdienen, worden in hoofdstuk 9 kort behandeld:

1. vlaggen / documenteren;
2. omgaan met uitbijters (uitschieters);
3. selectie van hulpvariabelen;

4. niet-negatieve variabelen met veel nullen;
5. combinatie van methoden (hiërarchie).

1.2 Afbakening en relatie met andere thema's

Item-nonrespons onderscheidt zich van *(unit) nonrespons*, waarbij iemand helemaal niet aan de enquête meedoet, of een deel van de objecten in een register ontbreekt. De onderzoeker moet bepalen of er in het geval van gedeeltelijke respons voldoende antwoorden zijn gegeven om het record mee te nemen of het als unit nonrespons te bestempelen. In dat geval van 'echte' nonrespons is wegen een optie; zie het thema 'Wegen als correctie voor non-respons'. Zoals in paragraaf 1.1.2.9 is beschreven, komen sommige totaalschatters na imputatie overeen met bepaalde wegingsmethoden en zijn tevens te beschouwen als synthetische schatters; zie Methodenreeks, thema 'Modelmatig schatten', deelthema 'Synthetische schatters en Kleine-domeinschatters'.

We maken verder onderscheid tussen imputeren en afleiden/typeren. Bij het afleiden van variabelen worden nieuwe variabelen gecreëerd als functie van in het bestand reeds bestaande variabelen. Bij imputeren worden ontbrekende waarden op een bestaande variabele gecreëerd.

Bij het gaafmaakproces (zie Methodenreeks, thema 'Controle en correctie') worden eventuele fouten opgespoord en gecorrigeerd. Wanneer de oorspronkelijke, fout geachte waarde geen rol speelt bij het corrigeren, zien we de correctie ook als een imputatie. Er is dan feitelijk eerst een ontbrekende waarde gecreëerd, door de foute waarde op missing te zetten. Het komt echter ook voor dat de oorspronkelijke waarde invloed heeft op de toe te kennen waarde, bijvoorbeeld bij de 'duizendfouten' bij economische statistieken. De in paragraaf 1.4 te geven definitie van imputeren (vrij naar Begrippenlijst van Proces-metadata), maakt duidelijk dat er dan geen sprake is van een imputatie.

De definitie van imputeren impliceert niet dat het bestand na imputatie intern consistent is, in de zin van dat aan alle gaafmaakregels is voldaan. Men kan wel als extra eis bij het imputatieproces meenemen dat de geïmputeerde waarden aan alle (of aan bepaalde) controleregels voldoen, zodat er geen verboden inconsistenties of niet-toegelaten waarden ontstaan als gevolg van het imputeren. Men kan aan deze eis voldoen door controleregels als restricties bij het imputeren mee te nemen, of door de ongerestricteerde imputaties achteraf gaaf te maken. Dit laatste leidt soms tot een iteratief proces. Bij grote surveys met veel variabelen en met records met meerdere onbrekende waarden, zijn inconsistenties niet altijd te vermijden, ook al gebruikt men methoden van multivariate imputatie.

1.3 Plaats in het statistisch proces

Imputatie is een onderdeel van het verwerkingsproces (throughput). Het is geen noodzakelijke processtap: men kan besluiten om de velden leeg te laten en het probleem bij de ophoging (wegen of herhaald wegen) of analyse op te lossen.

Van belang is dat de ontbrekende waarden bij eerdere processtappen duidelijk in het bestand zijn aangegeven. Dit kan doordat het veld is opengelaten, of via speciale codes als -1, 9 en 99 wanneer dat niet tot verwarring leidt. Problematischer is het wanneer nullen zijn ingevuld voor ontbrekende waarden, wat wel gebeurt bij bedrijfsstatistieken. Dan is geen onderscheid meer te maken tussen ontbrekende waarden en echte nullen. Ook voor het gaafmaken geeft dit problemen.

Vaak is het imputeren een vervolg op het detecteren van fouten, zoals in de inleiding van dit themarapport is beschreven. Zoals gezegd beschouwen wij het corrigeren van dergelijke fouten slechts als imputeren, wanneer bij de correctiestap de oorspronkelijke waarde geen rol meer speelt. Na gaafmaken en imputeren is het microbestand geschikt voor aggregeren en tabelleren. Bij steekproefonderzoeken zullen schattingprocedures nodig zijn.

Verder in het proces zal men meestal blij zijn te kunnen werken met geïmputeerde bestanden, en dankbaar van de imputaties gebruik maken. Toch bestaan er situaties waarin men de imputaties wil negeren, bijvoorbeeld bij het doen van secundaire analyses op micro-bestanden, maar ook bij het bepalen van betrouwbaarheidsmarges. Door de geïmputeerde waarden tijdens het imputatie-proces te ‘vlaggen’ (zie hoofdstuk 9), kan men ook aan deze wensen tegemoet komen. Dit vlaggen van geïmputeerde waarden zou dan ook een verplichting moeten zijn.

1.4 Definities

Begrip	Omschrijving
Item-nonrespons	het ten onrechte ontbreken van een antwoord van een respondent (op één of meerdere vragen)
Item-nonrespondent	een object dat op een bepaalde variabele ten onrechte niet heeft gerespondeerd
Imputatie, imputeren	bepalen en introduceren van een (nieuwe) waarde op een plaats waar een waarde ontbreekt of op ‘onbekend’ (ontbrekend) is gezet
Geïmputeerde waarde, imputatie	waarde die wordt ingevuld voor een ontbrekende waarde
Imputatievariabele	de variabele waarop ontbrekende waarden worden geïmputeerd
Imputatieklasse, imputatiestratum	een deelpopulatie waarvoor imputaties worden uitgevoerd, zonder gebruik te maken van informatie uit de overige deelpopulaties. Men kan verschillende imputatiemethoden gebruiken voor verschillende imputatieklassen.
Deductieve imputatie (logische imputatie)	imputatie waarbij op logische gronden een waarde wordt geïmputeerd zonder kansmechanisme, ook wanneer de waarde niet 100% zeker juist is
Donor-imputatie	Imputatie waarbij de ontbrekende waarde is overgenomen van een donor-record met zoveel mogelijk dezelfde kenmerken als de ontvanger.
Multivariate imputatie	imputatie van meerdere ontbrekende waarden per record
Massa-imputatie	imputatie voor alle in de populatie ontbrekende waarden op een bepaalde variabele
Longitudinale imputatie	Imputatie, waarbij gebruik wordt gemaakt van waarden voor dezelfde variabele op andere tijdstippen van hetzelfde object of andere objecten. Deze imputatie kan ook multivariaat zijn.

1.5 Algemene notatie

We hanteren in dit thema de volgende algemene notatie:

i = index voor object (record);

y = doelvariabele;

y_i = score van object i op doelvariabele y; we gaan er hier vanuit dat de waargenomen score geen meetfout bevat;

obs = verzameling objecten waarvan y_i is waargenomen (*obs*erved);

mis = verzameling objecten waarvan y_i *niet* is waargenomen (*mis*sing);

$\tilde{y}_i$ = geïmputeerde waarde voor ontbrekende y_i.

Bij de meeste methoden zal ook enige specifieke notatie worden geïntroduceerd.

2. Deductieve imputatie

2.1 Korte beschrijving

In het algemeen zijn imputaties voorspellingen voor de ontbrekende waarden, op grond van een model. In sommige gevallen kunnen imputaties echter ook direct afgeleid worden uit de wél waargenomen waarden in hetzelfde record met behulp van afleidingsregels die geen te schatten parameters bevatten zoals bij modellen het geval is.

Voorbeeld 1. Burgerlijke staat is onbekend, maar de persoon in kwestie is 10 jaar. Dan kan men met zekerheid zeggen dat deze persoon ongehuwd is.

Voorbeeld 2. Bij een bedrijfsenquête wordt gevraagd naar de totale omzet (O), omzet uit hoofdactiviteit (O1) en omzet uit nevenactiviteiten (O2). Indien één van deze drie vormen van omzet ontbreekt is die uit te rekenen via de regel: O1+O2=O.

Bovenstaande imputatieregels zijn voorbeelden van *deductieve* of *logische imputatie*. Bij deze imputatiemethode wordt gekeken of het mogelijk is om op grond van logische of wiskundige relaties tussen de variabelen, de waarde van één of meer van de ontbrekende variabelen eenduidig af te leiden uit de wel waargenomen waarden. Voor de ontbrekende variabelen waarvoor dit mogelijk is, is deze unieke waarde de deductieve imputatie.

Imputatieregels kunnen ook toegepast worden als de regel niet noodzakelijk altijd moet opgaan, maar slechts met grote waarschijnlijkheid opgaat. We spreken ook dan van deductieve of logische imputatie.

2.2 Toepasbaarheid

Voor deductieve imputatie is het niet nodig om modellen te specificeren of te schatten. Met als input alleen de controleregels, kan het proces geheel automatisch uitgevoerd worden. Deductieve imputaties zijn bovendien in zekere zin de best mogelijke imputaties. Ze zijn exact gelijk aan de werkelijke waarden als de andere waarden in het record correct zijn. Gezien deze laatste voorwaarde is het van belang om de methode uit te voeren nadat zoveel mogelijk fouten opgespoord zijn en vervolgens gecorrigeerd zijn (systematische fouten) of op ‘missing’ zijn gezet. Vervolgens is deductieve imputatie echter de meest logische volgende stap. Modelmatige en donor-methoden kunnen daarna toegepast worden. Deze methoden kunnen voor het schatten van de parameters profiteren van de reeds deductief ingevulde waarden.

Gezien de voordelen van de methode zal altijd nagegaan moeten worden welke mogelijkheden er zijn om deductief te imputeren.

2.3 Uitgebreide beschrijving

2.3.1. Eenvoudige imputatieregels

Veel deductieve imputaties kunnen uitgevoerd worden via eenvoudige regels van de 'als-dan' vorm, bijvoorbeeld:

als *burgerlijke staat = onbekend* **en** *leeftijd < 15* **dan** *burgerlijke staat = ongehuwd.*
Of
als *totaal arbeidskosten = onbekend* **en** *werknemers op de loonlijst* = 0 **dan** *totaal arbeidskosten* = 0.

Deze regels worden door inhoudelijk deskundigen opgesteld en kunnen stuk voor stuk toegepast worden met veel verschillende software.

2.3.2 Het gebruik van gelijkheidsrestricties

Een bijzonder rijke bron voor deductieve imputaties vormen de uitgebreide stelsels van gelijkheden die voor de productiestatistieken moeten gelden. Dit kan oplopen tot rond de 100 variabelen met 30 gelijkheidsrestricties. De meeste van deze gelijkheidsrestricties zijn van de vorm 'Totaalvariabele' = 'som van de Subtotalen (of deelposten of specificaties)'. Als in zo'n geval één van de subtotalen of het totaal ontbreekt, is het direct duidelijk met welke waarde de ontbrekende variabele geïmputeerd moet worden. Er is één vergelijking met één onbekende. In de praktijk komen veel variabelen voor in meerdere gelijkheden. We hebben dan te maken met een stelsel vergelijkingen met meestal meerdere ontbrekende variabelen, waarbij het niet direct inzichtelijk is of de waarden van sommige ontbrekende variabelen door dit stelsel uniek bepaald worden en welke die unieke waarden dan zijn. Hieronder wordt een methode beschreven waarmee automatisch de deductieve imputaties voor zulke stelsels vergelijkingen kunnen worden gegenereerd.

Stel dat een record bestaat uit p variabelen en dat er op deze p variabelen q lineaire gelijkheidsrestricties van toepassing zijn. Deze restricties kunnen weergegeven worden in de vorm

$$\mathbf{R}y = 0 \qquad (2.3.1)$$

met y de p-vector met variabelen, $\mathbf{R}$ een q×p matrix waarvan iedere rij één restrictie weergeeft. Bijvoorbeeld, het blok bedrijfsopbrengsten bevat de volgende vijf variabelen:

Tabel 1. Vijf variabelen uit het blok bedrijfsopbrengsten

Netto omzet uit hoofdactiviteit	y_1
Netto omzet uit overige activiteiten	y_2
Totaal netto omzet	y_3
Totaal overige bedrijfsopbrengsten	y_4
Totaal bedrijfsopbrengsten	y_5

Op deze variabelen zijn twee restricties van toepassing: $y_3 = y_1 + y_2$ en $y_5 = y_4 + y_3$. Deze restricties kunnen geformuleerd worden in de vorm (2.3.1) met

$$\mathbf{R} = \begin{bmatrix} 1 & 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & 1 & -1 \end{bmatrix}.$$

Als de vector met variabelen *y* bestaat uit *o* geobserveerde waarden en *m* ontbrekende waarden dan kan, na een permutatie van elementen, deze vector gepartitioneerd worden als $\mathbf{y} = (\mathbf{y}_o', \mathbf{y}_m')'$, waarbij $\mathbf{y}_o$ de *o*-vector is met de geobserveerde waarden van *y* en $\mathbf{y}_m$ de *m*-vector met de ontbrekende waarden. Als we **R** partitioneren in overeenstemming met de partitionering van **y**, kunnen we schrijven

$$\begin{bmatrix} \mathbf{R}_o & \mathbf{R}_m \end{bmatrix} \begin{bmatrix} \mathbf{y}_o \\ \mathbf{y}_m \end{bmatrix} = \mathbf{0}, \tag{2.3.2}$$

zodat

$$\mathbf{R}_m \mathbf{y}_m = -\mathbf{R}_o \mathbf{y}_o = \mathbf{a}, \text{ zeg.} \tag{2.3.3}$$

Deze laatste uitdrukking is een stelsel lineaire vergelijkingen in de ontbrekende waarden $\mathbf{y}_m$. De bedoeling van deductief imputeren is om uit dit stelsel zoveel mogelijk ontbrekende waarden op te lossen.

Voor een stelsel lineaire vergelijkingen is het gebruikelijk om drie gevallen te onderscheiden: I) er zijn geen oplossingen (het stelsel is strijdig), II) er is precies één oplossing. en III) er zijn oneindig veel oplossingen.

Geval I doet zich voor als de rang van $\mathbf{R}_m$ ongelijk is aan de rang van [$\mathbf{R}_m$ **a**]. Als de formulering van de restricties zodanig is dat daardoor geen tegenstrijdigheden ontstaan, kan geval I zich alleen voordoen bij fouten in de data. Dit soort fouten, die schendingen van de restricties veroorzaken, worden bij economische statistieken echter eerst opgespoord. Vervolgens worden een aantal waarden als fout aangemerkt en vervolgens op ‘missing’ gezet. De nieuwe ontbrekende waarden worden op een zodanige wijze aangewezen dat er imputaties bestaan voor de ontbrekende waarden die aan de restricties voldoen. Als er op bovenstaande wijze omgegaan is met schendingen van restricties kan geval I zich dus niet meer voordoen.

Geval II doet zich voor als de rang van $\mathbf{R}_m$ gelijk is aan het aantal ontbrekende waarden *m*. Alle ontbrekende waarden kunnen dan deductief geïmputeerd worden, er is maar één waarde voor $\mathbf{y}_m$ die aan de restricties voldoet.

In het algemeen zullen we echter te maken hebben met geval III; er zijn oneindig veel oplossingen voor $\mathbf{y}_m$. In dit laatste geval is het echter mogelijk dat sommige elementen van $\mathbf{y}_m$ in alle mogelijke oplossingen dezelfde waarden hebben. Deze elementen kunnen deductief geïmputeerd worden.

De verzameling oplossingen voor $\mathbf{y}_m$, zeg $\tilde{\mathbf{y}}_m$ is gegeven door (zie bijvoorbeeld Rao (1973), pag. 24)

$$\tilde{\mathbf{y}}_m = \mathbf{R}_m^- \mathbf{a} + (\mathbf{R}_m^- \mathbf{R}_m - \mathbf{I})\mathbf{z} = \mathbf{b} + \mathbf{C}\mathbf{z} \qquad (2.3.4)$$

met $\mathbf{R}_m^-$ een gegeneraliseerde inverse van $\mathbf{R}_m$ (d.w.z. een m × q matrix waarvoor geldt: $\mathbf{R}_m \mathbf{R}_m^- \mathbf{R}_m = \mathbf{R}_m$), en **z** een willekeurige m-vector. Doordat **z** willekeurig gekozen kan worden genereert (2.3.4) in het algemeen oneindig veel oplossingen voor $\mathbf{y}_m$, er is alleen een unieke oplossing als $\mathbf{R}_m$ van volledige rang is en $\mathbf{R}_m^-$ dus de reguliere inverse is. Als sommige elementen van $\mathbf{y}_m$ gelijk zijn voor alle mogelijke oplossingen, dus voor iedere willekeurige waarde van **z**, dan moet gelden dat de corresponderende rijen van **C** alleen nullen bevatten**.** Deze elementen zijn dus eenvoudig op te sporen en kunnen deductief geïmputeerd worden met de corresponderende waarden van **b**.

2.3.3 Het gebruik van niet-negativiteit

Een andere mogelijkheid om deductief te imputeren is door gebruik te maken van de niet-negativiteit van veel variabelen. Stel bijvoorbeeld dat slechts twee deelposten van een optelling van acht posten geobserveerd zijn, maar dat deze wel optellen tot het gerapporteerde totaal. Als de ontbrekende deelposten niet negatief mogen zijn, kunnen ze allen met nul worden geïmputeerd omdat hun som nul moet zijn.

Om dit soort oplossingen te vinden, beschouwen we weer de gelijkheid $\mathbf{R}_m \mathbf{y}_m = \mathbf{a}$. Veronderstel nu dat er een element a_j van **a** is dat gelijk is aan nul. Voor de corresponderende rij, $\mathbf{r}_{m.j}$, van $\mathbf{R}_m$ geldt dan dus $\mathbf{r}'_{m.j} \mathbf{y}_m = 0$. Als nu voor alle elementen van $\mathbf{y}_m$ die corresponderen met de niet-nul elementen van $\mathbf{r}_{m.j}$ geldt dat

i) deze elementen kunnen niet negatief zijn,

ii) de corresponderende niet-nul elementen van $\mathbf{r}_{m.j}$ zijn allen negatief of allen positief,

dan zijn deze elementen van $\mathbf{y}_m$ gelijk aan nul.

De op deze wijze afgeleide deductieve 0-imputaties voor de ontbrekende waarden $\mathbf{y}_m$ zijn dus gegeven door

$$\tilde{y}_{mj} = 0 \quad \text{als} \quad a_j = 0 \quad \text{en aan voorwaarden i en ii voldaan is.} \qquad (2.3.5)$$

2.4 Voorbeeld

Een voorbeeld van het toepassen van deductieve imputatie bij bedrijfsstatistieken is beschreven in Pannekoek en Tempelman (2005). Dit voorbeeld betreft gegevens van Productiestatistieken die betrekking hebben op de Groothandel en de Detailhandel. De data met betrekking tot de Groothandel bestaan uit 875 bedrijven in grootteklasse (4 t/m 9) en 102 variabelen. Op deze variabelen zijn 30 gelijkheidsrestricties van toepassing en daarnaast zijn er nog 26 eenvoudige imputatieregels geformuleerd door gebruik te maken van een relatie van de vorm ‘als $y_1 = 0$ dan $y_2 = 0$’ en is gebruikgemaakt van de niet-negativiteit van bijna al deze variabelen. De data van de

Detailhandel bestaan uit 1242 records (grootteklasse 0 t/m 3) en 54 variabelen waarop 15 gelijkheidsrestricties van toepassing zijn en daarnaast zijn er nog 21 eenvoudige imputatieregels geformuleerd van dezelfde vorm als bij de Groothandel en is ook van de niet-negativiteit gebruik gemaakt.

De data van deze Productiestatistieken hebben reeds enkele bewerkingsstappen ondergaan, waarbij wordt gecorrigeerd voor overduidelijke fouten. Hieronder vallen bijvoorbeeld uniforme duizendfouten of waarnemingen die onterecht negatief zijn. Bovendien worden tijdens deze stap ook lege totalen en subtotalen ingevuld als de deelposten daarvan tenminste wel zijn ingevuld. Dit laatste is een eerste deductieve imputatiestap. Daarnaast zijn door het foutenlokalisatie algoritme van het programma CherryPi alle editregels gecontroleerd en bij schendingen van editregels de nodige waarden als foutief aangemerkt en vervolgens op 'missing' gezet. De ontbrekende waarden in deze bestanden zijn dus zowel het gevolg van partiële non-response als van opgespoorde fouten.

Op deze data zijn met behulp van de gelijkheidsrestricties en de eenvoudige imputatieregels alle mogelijke deductieve imputaties uitgevoerd. De resultaten zijn weergegeven in tabel 2.

Tabel 2. Aantallen deductieve imputaties bij de Groothandel en de Detailhandel

	Groothandel		Detailhandel	
Aantal ontbrekende waarden	35068		27693	
Aantal deductieve imputaties	24048 (69%)		12927 (47%)	
Waarvan gelijk aan nul		22647 (94%)		11708 (91%)
Waarvan niet gelijk nul		1401 (6%)		1219 (9%)
Resterende ontbrekende waarden	11020		14766	

Uit deze tabel blijkt dat deductieve imputatie zeer effectief is, voor een groot deel van de ontbrekende waarden (69% en 47%) kan op deze wijze, zonder imputatiemodel en zonder aanpassingen van imputaties, geïmputeerd worden met de enig mogelijke waarde die aan alle editregels voldoet.

De deductieve imputaties in tabel 2 zijn meestal (voor meer dan 90%) gelijk aan nul. Hierbij moet wel opgemerkt worden dat dit niet de enige deductieve imputaties zijn. In de T040 stap hebben al een aantal deductieve imputaties plaatsgevonden die niet-nul zijn: het invullen van lege subtotalen. Veel van de nul-imputaties hebben er mee te maken dat berichtgevers de vragen naar specifieke kostenposten waar zij geen uitgaven aan hebben gehad leeg laten in plaats van met 0 te beantwoorden. Het zelfde geldt voor inkomsten uit specifieke onderdelen van de bedrijfsopbrengsten. Met deductieve imputatie kunnen een groot aantal van deze niet ingevulde nul-waarden teruggevonden worden. Het is overigens niet aan te bevelen om een niet

ingevuld veld altijd met de waarde nul te imputeren als dat niet in strijd is met de editregels. Pannekoek en Tempelman (2005) laten zien dat dat soms tot substantiële vertekening in de publicatietotalen kan leiden.

3. (Group) mean imputation / imputatie van het (groeps)gemiddelde

3.1 Korte beschrijving

Bij *mean imputation* (‘imputatie van het gemiddelde’) wordt een ontbrekende waarde vervangen door de gemiddelde score op de desbetreffende variabele bij de objecten die wel een geldige score hebben.

Bij *group mean imputation* (‘imputatie van het groepsgemiddelde’) wordt een ontbrekende waarde vervangen door de gemiddelde score op de desbetreffende variabele bij de objecten die een geldige score hebben én in dezelfde deelpopulatie zitten als de item-nonrespondent.

Mean imputation leidt tot een piek in de verdeling, omdat voor iedere ontbrekende waarde hetzelfde gemiddelde wordt geïmputeerd. Bij group mean imputation is sprake van een aantal kleinere pieken.

3.2 Toepasbaarheid

Bij zuivere mean imputation wordt geen hulpinformatie gebruikt. Deze methode is daarom alleen aan te raden indien er geen hulpinformatie beschikbaar is of wanneer de beschikbare hulpvariabelen nauwelijks samenhangen met de imputatievariabele y. Als de fractie ontbrekende waarden op een variabele zeer klein is en de imputaties nauwelijks effect zullen hebben op de te schatten parameter (bijv. het populatietotaal), kan mean imputation uit efficiency-overwegingen toelaatbaar zijn. Men moet echter zeer zuinig zijn met het toepassen van deze wat al te eenvoudige methode.

Bij group mean imputation wordt wel hulpinformatie gebruikt, namelijk een indeling in groepen (deelpopulaties, imputatieklassen, imputatiestrata) op grond van één of meer kwalitatieve variabelen. Hoe homogener naar de te imputeren variabele de deelpopulaties, des te beter de imputaties, uitgaande van de assumptie dat de indeling in deelpopulaties niet alleen goed discrimineert bij de respondenten, maar ook bij de item-nonrespondenten (zie in paragraaf 1.1.2.8).

Zoals vermeld leidt zuivere mean imputation tot een gepiekte verdeling. De methode is daarom eventueel geschikt wanneer de output zich beperkt tot geschatte populatiegemiddelden en –totalen. Het feit dat door de imputatie een volledig data-bestand wordt verkregen, garandeert consistentie van geaggregeerde uitkomsten. Zuivere mean imputation is echter ongeschikt voor het schatten van een (inkomens)verdeling of voor het schatten van een spreidingsmaat als de standaarddeviatie. Het leidt ook doorgaans niet tot kwalitatief goede individuele imputaties, maar bij geen enkele imputatiemethode bestaat een dergelijke garantie.

Bij group mean imputation is de gepiektheid van de verdeling meestal veel kleiner, omdat de variatie tussen de groepen bij de imputatie wordt meegenomen; alleen de variatie binnen de groepen wordt verwaarloosd. Indien de verhouding van deze

tussen- en binnenvariantie groot is, kan men met deze methode ook spreidingsmaten redelijk schatten, gegeven het gelden van het imputatiemodel.

3.3 Uitgebreide beschrijving

Conform de notatie uit paragraaf 1.5 is de geïmputeerde waarde $\tilde{y}_i$ voor een ontbrekende score y_i bij mean imputation gelijk aan het geobserveerde gemiddelde

$$\tilde{y}_i = \bar{y}_{obs} = \frac{\sum_{k \in obs} y_k}{n_{obs}}, \qquad (3.3.1)$$

met y_k de waargenomen score van de k^e respondent en n_{obs} het aantal item-respondenten voor variabele *y*.

Indien gewenst kunnen de objecten ongelijk worden gewogen, bijvoorbeeld vanwege verschillen in insluitkans; zie deelparagraaf 1.1.2.9 en eigenschap 3 in paragraaf 3.5. Er wordt dan niet opgehoogd met een vaste ophoogfactor N/n (met N de populatieomvang, en n de steekproefomvang of het aantal respondenten), maar met individuele gewichten w_i die onderling verschillen. De resulterende imputatie

$$\tilde{y}_i = \bar{y}_{obs}^{(w)} = \frac{\sum_{obs} w_k y_k}{\sum_{obs} w_k} \qquad (3.3.2)$$

is dan doorgaans een betere (minder vertekende) schatter van het populatiegemiddelde.

Mean imputation kan worden toegepast voor de nonrespons in de steekproef of voor de ontbrekende waarden in de populatie. Voor iedere ontbrekende waarde wordt hetzelfde gemiddelde geïmputeerd. Men kan de methode veelal zinvoller toepassen na eerst imputatieklassen te hebben bepaald. Bij deze group mean imputation wordt (3.3.1) vervangen door

$$\tilde{y}_{hi} = \bar{y}_{h;obs} = \frac{\sum_{hk \subset obs} y_{hk}}{n_{h;obs}}, \qquad (3.3.3)$$

waarbij y_{hk} de waargenomen score van de k^e respondent in klasse *h* en $n_{h;obs}$ het aantal item-respondenten voor variabele *y* in *h*.

Er is geen ingewikkelde software nodig om een (groeps)gemiddelde te imputeren. Met SPSS14.0 kan (group) mean imputation eenvoudig worden toegepast via Transform \ Replace Missing values \ Method Series mean. De procedure Replace Missing values is bedoeld voor tijdreeksen, en is daarmee bruikbaar voor ontbrekende waarden bij longitudinale imputatie (hoofdstuk 8).

Men kan mean imputation dus toepassen:
- met het gemiddelde van de hele steekproef of populatie, of per imputatieklasse;
- ongewogen, of gewogen met gewichten w_i.

De mogelijkheid om de methode toe te passen met een storingsterm behandelen we bij regressie-imputatie in hoofdstuk 5.

3.4 Voorbeeld

Voorbeeld 1. *Energiestatistiek-1*[2]

Voor het schatten van het energieverbuik van bedrijven in Nederland werd tot voor kort gebruik gemaakt van de enquête 'Energieverbruik bij bedrijven'. Door het wegvallen van deze enquête wordt er gewerkt aan het opzetten van een secundair waarnemingstraject, waarbij van energiebedrijven afkomstige 'verbruikgegevens per bedrijf' worden gebruikt voor het schatten van het totale energieverbuik. Hiervoor worden verbruikgegevens op basis van NAW-gegevens (naam, adres, woonplaats) gekoppeld aan bedrijfseenheden in het ABR.

Een voorbeeld van group mean imputation is het gemiddelde elektriciteitverbruik per bedrijf te gebruiken per bedrijfssector/SBI, zoals de glastuinbouw.

Voorbeeld 2. *Productiestatistiek-1*

Bij de Productiestatistieken (PS) worden deelpopulaties gevormd op basis van SBI en GK. De steekproefomvang is te klein om alle cellen SBI × GK te kunnen onderscheiden. De imputatieprocedure verschilt enigszins tussen grote en kleinere bedrijven.

Indien hulpinformatie over een bedrijf met onvolledige nonrespons beschikbaar is, bijvoorbeeld in de vorm van omzet uit het vorige jaar of uit de Korte-termijn Statistiek (KS), dan moet deze natuurlijk worden gebruikt. Voorbeeld 2 in paragraaf 4.4 toont hoe dat gebeurt. Wanneer echter dergelijke informatie niet beschikbaar is, wordt Group mean imputation gebruikt. Men kan dan bij een ontbrekende omzet de gemiddelde omzet in de imputatieklasse (stratum) imputeren. Dit zal onder andere veel gebruikt worden voor nieuwe bedrijven, waarvoor geen gegevens uit een vorige periode beschikbaar zijn.

3.5 Eigenschappen

1. Na toepassen van mean imputation volgens (3.3.1) voor alle item-nonrespondenten is het ongewogen steekproefgemiddelde gelijk aan het ongewogen responsgemiddelde. Zou men massa-imputatie toepassen door het responsgemiddelde niet alleen als geïmputeerde waarde te gebruiken voor de eventuele item-nonrespondenten, maar ook voor degenen die niet in de steekproef zitten, dan wordt ook het hiermee geschatte populatiegemiddelde gelijk aan het responsgemiddelde, en ook gelijk aan de directe schatter voor het populatiegemiddelde (met ophooggewichten N/n).

[2] Met dank aan Edgar Soufan.

2. Group mean imputation leidt evenzo tot dezelfde overall-totalen en -gemiddelden als de (post-)stratificatieschatter, wanneer de strata als imputatieklassen worden gebruikt.

3. Na toepassen van gewogen mean imputation volgens (3.3.2) voor alle item-nonrespondenten is het met insluitkansen gewogen steekproefgemiddelde gelijk is aan het met insluitkansen gewogen responsgemiddelde, ongeacht het gewicht van de item-nonrespondenten. Ophoging zorgt er ook nu voor dat de populatieschatting niet door de imputaties wordt beïnvloed. Evenzo is na massa-imputatie het populatiegemiddelde gelijk aan het gewogen responsgemiddelde.

3.6 Kwaliteitsindicatoren

Mean imputation leidt tot een onderschatting van de variantie S_y^2 van imputatievariabele *y*,

$$\hat{S}_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 \quad , \tag{3.6.1}$$

omdat voor de item-nonrespondenten de bijdrage in de teller een nul is. Zou men voor $V(\bar{y})$, de variantie van het steekproefgemiddelde $\bar{y}$, de naïeve schatter

$$\hat{V}(\bar{y}) = (1-\frac{n}{N})\frac{1}{n}\hat{S}_y^2 = (1-\frac{n}{N})\frac{1}{n(n-1)}\sum_{i=1}^{n}(y_i - \bar{y})^2 \tag{3.6.2}$$

gebruiken met y_i hetzij bekend, hetzij geïmputeerd met $\bar{y}$, dan wordt ook deze variantie (kwadraat van de standaardfout) onderschat, en daarmee ook de betrouwbaarheidsmarge. Men doet dan namelijk ten onrechte of men gegevens heeft van alle n objecten, in plaats van alleen van degenen die op *y* gerespondeerd hebben. De juiste schatter voor $V(\bar{y})$ verkrijgt men door de steekproefomvang *n* in formule (3.6.2) te vervangen door het aantal item-respondenten n_{obs} en S_y^2 alleen te bepalen over de item-respondenten. Vanzelfsprekend is het steekproefgemiddelde $\bar{y}$ gelijk aan het responsgemiddelde. Voor group mean imputation geldt het bovenstaande per groep.

Zie paragraaf 5.6 voor verdere kwaliteitsindicatoren.

4. Ratio-imputatie

4.1 Korte beschrijving

Bij *ratio-imputatie* voor variabele *y* wordt één hulpvariabele *x* gebruikt die sterk met *y* samenhangt, in die zin dat *x* bij (redelijke) benadering evenredig is met *y*. Indien *R* de verhouding tussen *y* en *x* weergeeft, wordt de ontbrekende waarde y_i vervangen door

$$\tilde{y}_i = Rx_i \ . \qquad (4.1.1)$$

Een voorbeeld is het bepalen van een onbekende bedrijfsomzet (y) uit het aantal werkzame personen (x). Voor *R* zal men dan de gemiddelde bedrijfsomzet per werkzame persoon nemen. De meest voorkomende situatie is dat *x* hetzelfde meet als *y*, maar in een eerdere waarnemingsperiode. We noteren de variabelen *y* en *x* dan als respectievelijk y^t en y^{t-1}. Formule (4.1.1) gaat dan over in

$$\tilde{y}_i^t = Ry_i^{t-1} \ , \qquad (4.1.2)$$

met *R* de relatieve toename van de variabele van tijdstip *t-1* op *t*. Doorgaans wordt *R* uit de data geschat.

4.2 Toepasbaarheid

Ratio-imputatie kan worden toegepast voor ontbrekende waarden op een kwantitatieve variabele *y*, wanneer er een kwantitatieve (hulp)variabele *x* is te vinden die een min of meer vaste verhouding heeft met de doelvariabele *y*. Men kan formule (4.1.1) zien als een enkelvoudige regressievergelijking waarbij de regressielijn door de oorsprong gaat. Er wordt dus geen constante term meegenomen. Ratio-imputatie is dus een speciaal geval van regressie-analyse (geschat met gewogen kleinste kwadraten). Indien een model met een constante term beter fit, of wanneer men extra variabelen aan model (4.1.1) zou willen toevoegen, *kan* de algemene regressie-imputatie geschikter zijn.

Doorgaans wordt er op het CBS geen residu aan (4.1.1) toegevoegd. Bij veel statistieken waar ratio-imputatie wordt toegepast, zijn gemiddelden en totalen de belangrijkste output. In het verleden werd er bij sommige omzetstatistieken als uitzondering een tabel geproduceerd met het aantal bedrijven dat een hogere vs. lagere omzet heeft dan het jaar daarvoor. Indien imputatie volgens (4.1.2) wordt toegepast en R wordt geschat op 1,01, dan wordt voor alle item-nonrespondenten aangenomen dat ze van tijdstip *t-1* op *t* een omzetgroei hebben gehad, hetgeen in deze situatie onwaarschijnlijk is. Voor díe tabel is het derhalve nodig een residu aan (4.1.2) toe te voegen. We bespreken dit toevoegen van een residu verder in hoofdstuk 5; zie ook deelparagraaf 1.1.2.6.

Net als bij mean imputation kan men ratio-imputatie apart per deelpopulatie (imputatieklasse) toepassen. Men doet dit vooral wanneer de ratio's tussen de deelpopulaties sterk verschillen. Deze mogelijkheid wordt in de volgende paragraaf besproken.

4.3 Uitgebreide beschrijving

Vaak beschikt men over een hulpvariabele *x* die min of meer evenredig is met y. Wanneer y_i ontbreekt maar x_i wel bekend is, kan men (4.1.1) als imputatie gebruiken, met *R* de evenredigheidsconstante. Doorgaans is *R* niet bekend en schat men *R* uit de records waarvoor *x* én *y* bekend zijn:

$$\hat{R} = \sum_{ibs} y_i \Big/ \sum_{obs} x_i \quad . \tag{4.3.1}$$

Invullen in (4.1.1) geeft

$$\tilde{y}_i = \hat{R} x_i = \frac{\sum_{obs} y_i}{\sum_{obs} x_i} x_i \quad . \tag{4.3.2}$$

De evenredigheidsconstante is dus gelijk aan het quotiënt (ratio) van de gemiddelden op *y* en *x* voor de item-respondenten van variabele *y*.

In het geval dat *x* en *y* alleen in het tijdstip verschillen, gaat formule (4.1.3) over in

$$\tilde{y}_i^t = \hat{R} y_i^{t-1} = \frac{\sum_{obs} y_i^t}{\sum_{obs} y_i^{t-1}} y_i^{t-1} \quad . \tag{4.3.3}$$

De te schatten parameter R is nu de relatieve toename van de variabele van *t-1* op *t*.

Men kan model (4.3.2) ook apart voor verschillende deelpopulaties toepassen. Iedere deelpopulatie *h* heeft dan een eigen ratio R_h. Men kan dit *group ratio imputatie* noemen. Het toepassen hiervan heeft alleen zin als de lineaire relatie tussen *x* en *y* sterk, en op zijn minst significant, verschilt tussen de deelpopulaties. De deelpopulaties mogen ook niet te klein zijn, want dat kan leiden tot onzuiverheid en eventueel grote standaardfouten voor totaalschatters. Het werken met groepen levert bij ratio imputatie doorgaans minder winst op dan bij *group mean imputation*; ratio's van groepen meestal homogener zijn dan groepsgemiddelden.

Voor het bepalen van de ratio *R* bestaat weer de mogelijkheid om item-respondenten te wegen met insluitgewichten.

Voor ratio-imputatie is geen ingewikkelde software nodig. Formules (4.3.2) en (4.3.3) zijn eenvoudig te berekenen na de ratio *R* te hebben geschat.

4.4 Voorbeeld

Voorbeeld 1. *Energiestatistiek-2* [3]

Voor ratio-imputatie lijkt het totaal aantal werkzame personen of de omzet per bedrijf een goede indicator voor de hoogte van het energieverbruik. Men zou kunnen onderzoeken of er verschillende verhoudingsfactoren bestaan voor verschillende bedrijfssectoren. Ook kan men onderzoeken of uitbreiding tot een algemener regressiemodel winst oplevert.

Voorbeeld 2. *Productiestatistiek-2*

Bij ontbrekende waarden in de Productiestatstiek bestaat een automatische imputatieprocedure voor de kleinere (niet-cruciale) bedrijven, waarbij voornamelijk gebruik wordt gemaakt van ratio-imputatie. Er wordt een vaste volgorde gebruikt voor de beschikbaarheid van hulpinformatie. Deze hiërarchie, afnemend in kwaliteit van de hulpinformatie, is:

1. waarneming bij hetzelfde bedrijf in jaar *t*-1 (voor alle variabelen);
2. waarneming bij hetzelfde bedrijf uit de Korte-termijn Statistiek (KS) van jaar *t* (alleen voor *y* = omzet);
3. waarneming van stratumgenoten (GK × SBI) in jaar *t*.

Wanneer een bedrijf item-nonrespons heeft, wordt dus allereerst gekeken of het bedrijf het voorafgaande jaar wél een geldige score op die variabele had. Zo ja, dan wordt formule (4.3.3) toegepast, met y^t de desbetreffende variabele in jaar *t*, y^{t-1} in het jaar ervoor en $\hat{R}$ een trendcorrectie. Voor de omzetvariabelen geeft de trendcorrectie de omzetontwikkeling weer. Dit alles vindt plaats binnen een combinatie van GK en SBI (3-digit) met een minimale celvulling van 15 bedrijven.

Wanneer echter y_i^{t-1} onbekend is, bijvoorbeeld doordat het bedrijf het jaar ervoor niet in de steekproef zat, wordt voor de tweede of derde optie gekozen, afhankelijk van de doelvariabele. Deze opties zijn echter geen ratio-imputaties. Bij de tweede optie wordt voor bedrijven die ook aan de KS van jaar *t* hebben deelgenomen, de getotaliseerde jaaromzet exact overgenomen; geïmputeerde omzetten worden ook nu niet toegelaten. Dit kopiëren van de waarde uit een ander bestand wordt 'cold deck' genoemd; zie hoofdstuk 6. Optie 3 is een group mean imputation, met een combinatie van GK en SBI als imputatieklasse. Voor nieuwe bedrijven zal vaak optie 3 gehanteerd worden.

4.5 Eigenschappen

- Een speciaal geval van ratio-imputatie verkrijgt men door *R=1* te nemen. Dit betekent dat de imputatie $\tilde{y}_i$ gelijk is aan x_i. Variabele *x* is dan een 'proxy-variabele' voor y. Indien *x* uit een externe bron afkomstig is, wordt dit 'cold deck imputatie' genoemd (zie hoofdstuk 6). Een voorbeeld is dat voor een

[3] Met dank aan Edgar Soufan.

ontbrekende waarde y_i^t de waarde uit een vorige periode, y_i^{t-1}, wordt overgenomen. Met variabelen die stabiel in de tijd zijn valt dit te overwegen, maar vaak zal men de voorkeur geven aan het schatten van *R*, in plaats van deze gelijk aan 1 te stellen.

- De ratio $\sum y_i / \sum x_i$ verandert niet door ratio-imputatie. Wanneer men de ratio-schatter (zie Banning e.a. 2010) gebruikt voor ophoging van steekproef naar populatie met *x* als hulpvariabele voor y, dan verandert de populatieschatting niet door meenemen van de geïmputeerde waarden.

4.6 Kwaliteitsindicatoren

Ratio-imputatie leidt tot een onderschatting van de spreiding van de waarden van $y_i - Rx_i$ wanneer geen storingsterm in het model wordt meegenomen. Zou men voor de variantie van het geschatte populatiegemiddelde met behulp van de ratio-schatter

$$\hat{V}(\hat{\bar{Y}}_R) \equiv \hat{V}(\hat{R}\bar{X}) = (1-\frac{n}{N})\frac{1}{n(n-1)}\sum_{i=1}^{n}\left(y_i - \hat{R}x_i\right)^2 \qquad (4.6.1)$$

de naïeve schatter gebruiken, met y_i hetzij bekend, hetzij geïmputeerd, dan wordt ook deze variantie onderschat, en daarmee de betrouwbaarheidsmarge. Men doet dan namelijk ten onrechte of men *y*-scores heeft van alle n objecten, in plaats van alleen van degenen die op *y* gerespondeerd hebben. De juiste schatter voor $V(\hat{\bar{Y}}_R)$ verkrijgt men door de steekproefomvang *n* in de formules voor de ratio-schatter te vervangen door het aantal item-respondenten n_{obs} en alleen te sommeren over de item-respondenten

Zie paragraaf 5.6 voor verdere kwaliteitsindicatoren.

5. Regressie-imputatie

5.1 Korte beschrijving

Bij regressie-*imputatie* wordt voor een ontbrekende waarde y_i de optimale voorspelling geïmputeerd die volgt uit een geschikt gekozen regressiemodel dat *y* voorspelt uit één of meer *x*-variabelen. De parameters van het model worden geschat met behulp van de objecten met een geldige score op *y* en op de (meeste) *x*-variabelen.

Soms voegt men aan deze optimale voorspelling een random storingsterm toe, om te voorkomen dat de geïmputeerde data-set te goed aan het regressiemodel voldoet.

5.2 Toepasbaarheid

Bij regressie-imputatie is de doelvariabele *y* kwantitatief. De verklarende hulpvariabelen van het regressiemodel zijn kwantitatief, maar door het gebruik van dummy-variabelen kunnen ook kwalitatieve variabelen in het model worden opgenomen. Lineaire regressie-analyse wordt dan ook 'variantie-analyse' genoemd. Dergelijke regressies kunnen wel leiden tot waarden die theoretisch niet kunnen voorkomen, zoals niet-gehele getallen wanneer het waardenbereik van *y* alleen de gehele getallen bevat. Donor-imputatie, dat met enige goede wil ook als een vorm van regressie-analyse kan worden opgevat, voorkomt dit probleem.

Regressie-imputatie is ook toepasbaar voor een binaire (dichotome) doelvariabele. Men kan dan bijvoorbeeld een logistisch regressiemodel gebruiken; zie voorbeeld 3 in paragraaf 5.4.

In deelparagraaf 1.1.2.6 is reeds uitgelegd dat voor iedere item-nonrespondent op *y* hetzij de beste voorspelling kan worden geïmputeerd, hetzij dat hieraan een aselecte storingsterm wordt toegevoegd. De keuze hangt af van het doel van de imputatie. Voor het schatten van (gemiddelden en totalen) is zo'n residu niet nodig, maar wil men dat de spreiding in *y* ook na imputatie behouden blijft, dan verdient het de voorkeur om een residu eraan toe te voegen.

In deelparagraaf 1.1.2.9 is gewezen op de mogelijkheid van het uitvoeren van een gewogen regressie-analyse, wanneer men respondenten met een hoger steekproefgewicht zwaarder wil laten meetellen. Heterogeniteit van de storingen kan een andere reden zijn voor zo'n schatting met gewogen kleinste-kwadraten.

5.3 Uitgebreide beschrijving

We behandelen in de Methodenreeks geen theorie over regressie-analyse, maar zien dat als algemene kennis. Er bestaat genoeg literatuur over lineaire en andersoortige regressie. Voor modelselectie beperken we ons tot opmerkingen in deelparagraaf 1.1.2.5 en paragraaf 9.3.

Bij r*egressie-imputatie* wordt een regressiemodel verondersteld voor de voorspelling van *y* door een set hulpvariabelen $x_1,\ldots,x_p$. Het regressiemodel luidt

$$y = \alpha + \beta_1 x_1 + \ldots + \beta_p x_p + \varepsilon = \alpha + \beta' x + \varepsilon \qquad (5.3.1)$$

met *x* een p-vector met variabelen $x_1,\ldots,x_p$, α een scalaire parameter, β een *p*-vector met parameters en $\varepsilon \sim N(0,\sigma^2 I)$ een vector met n_{obs} onafhankelijke, normaal-verdeelde storingen met variantie σ^2; *I* is de identiteitsmatrix. Men kan ook het model zonder constante term beschouwen, door weglating van α.

De parameters α en $\beta_1,\ldots,\beta_p$ worden geschat met behulp van de records waarvoor zowel *y* als de hulpvariabelen zijn waargenomen. Dit resulteert in parameter-schatters *a*, $b_1,\ldots$, b_p. Meestal wordt de kleinste-kwadratenmethode als schattingsmethode gebruikt. Dit resulteert in een predictor-variabele

$$\hat{y} = a + b' x \ , \qquad (5.3.2)$$

met kleinste-kwadratenschatters *a* en *b* voor resp. α en β. Deze predictor-variabele is zowel voor item-respondenten als item-nonrespondenten gedefinieerd.

Er zijn nu twee manieren om een imputatie $\tilde{y}_i$ voor de item-nonrespondenten te bepalen:

1. zonder storingsterm:

$$\tilde{y}_i = \hat{y}_i = a + b' x_i \ , \qquad (5.3.3)$$

2. met storingsterm

$$\tilde{y}_i = \hat{y}_i + e_i = a + b' x_i + e_i \ . \qquad (5.3.4)$$

Conform deelparagraaf 1.1.2.6 zijn er twee manieren om de storingsterm e_i te bepalen:

a. $e_i = e_d$ met e_d het residu van een willekeurige of speciaal geselecteerde donor.

b. e_i is een trekking uit de normale verdeling met verwachting 0 en variantie σ^2.

In beide gevallen wordt het residu bepaald via het regressiemodel.

Niet-lineaire modellen hebben een algemenere gedaante:

$$y = f(\beta' x) \ . \qquad (5.3.5)$$

De storingsterm ε kan aan dit model worden toegevoegd, of kan er impliciet inzitten.

In het geval van een binaire *y*-variabele met scores 0 en 1 kan een logistisch regressiemodel worden gebruikt:

$$\ln \frac{p}{1-p} = \alpha + \beta_1 x_1 + \ldots + \beta_p x_p \equiv \alpha + \beta' x \ , \qquad (5.3.6)$$

met p de kans dat *y* de score 1 aanneemt, gegeven de *x*-variabelen en het model. In het geval van een ontbrekende *y*-waarde kan men de β-parameters schatten,

bijvoorbeeld door middel van maximum likelihood, en vervolgens de imputatiekans *p* op de score 1 via

$$\hat{p} = \frac{e^{\hat{\alpha}+\hat{\beta}'x}}{1+e^{\hat{\alpha}+\hat{\beta}'x}} = \frac{1}{e^{-(\hat{\alpha}+\hat{\beta}'x)}+1} \ . \qquad (5.3.7)$$

Bij SPSS14.0 \ Analyze \ Regression kunnen de voorspelde waarden volgens (5.3.2) worden opgeslagen via SAVE \ Unstandardized predicted values, zowel voor linear als nonlinear regression. Er wordt dan een variabele gecreëerd met de default-naam PRE_1, die zowel voor de item-nonrespondenten als de item-respondenten de waarde $\hat{y}_i$ bevat. De variabele *y* na imputatie wordt dan verkregen door voor iedere item-respondent de $\hat{y}$-score te vervangen door de echte score y_i. (Men kan natuurlijk ook in de oorspronkelijke y-variabele de ontbrekende waarden vervangen door de model-scores $\hat{y}_i$, al zal men dan de geïmputeerde waarden moeten vlaggen.) Bij 'binary logistic' en 'multinomial logistic' kunnen de voorspelde categorie-kansen worden opgeslagen. Dit kan niet alleen voor de item-respondenten, maar ook voor de item-nonrespondenten, zodat de imputaties volgens (5.3.7) meteen worden verkregen.

5.4 Voorbeelden

Voorbeeld 1. *Energiestatistiek*[4]

Voor het toepassen van regressie-imputatie om ontbrekende energieverbruiken te bepalen, kan men denken aan een regressiemodel met als *x*-variabelen aantal werkzame personen, omzet en SBI. Het kan blijken dat het geen zin heeft om zowel aantal werkzame personen als omzet mee te nemen. Zo zou men het energieverbruik kunnen laten afhangen van alleen de omzet binnen iedere bedrijfssector. Dit is algemener dan in voorbeeld *x* in paragraaf 4.4 bij de ratio-imputatie per bedrijfssector, omdat ook de constante term in de regressievergelijking kan worden opgenomen en omdat ook niet-lineaire verbanden mogelijk zijn.

Voorbeeld 2. *Productiestatistiek-2*

Bij de ophoging van de Productiestatistiek wordt per 'kerncel' (combinatie van GK × SBI) de regressieschatter gebruikt voor bedrijven waarvan de naast de opgegeven omzet (*y*) ook de btw-omzet (*x*) bekend is. Dit ophogen van respons naar populatie maakt imputatie voor ontbrekende *y*-omzetten overbodig. Maar men verkrijgt dezelfde uitkomsten wanneer men eerst regressie-imputatie volgens formule (5.3.3) toepast met dezelfde regressie van *y* op *x*, en vervolgens de steekproef (inclusief de item-nonrespons) naar de populatie ophoogt, althans wanneer men op dezelfde manier omgaat met de steekproefgewichten.

[4] Met dank aan Edgar Soufan.

Voorbeeld 3. *Huishoudenstatistiek*

Het CBS ontvangt ieder jaar een afslag van de GBA (Gemeentelijke basisadministratie) op 1 januari. De GBA bevat per adres gegevens over de bewoners, waaronder hun gezinsrelaties. De huishoudensamenstelling ontbreekt echter. Voor de Jaarlijkse Huishoudenstatistiek is het essentieel te weten welke personen op een adres volgens de gangbare definitie één huishouden vormen. Tot het statistiekjaar 1999 was de statistiek gebaseerd op de huishoudbox van de EBB. Vanaf 1999 is de GBA de basis en worden de variabelen 'aantal huishoudens' en 'huishoudensamenstelling' afgeleid uit de gezinsstructuur (Harmsen en Israëls, 2000). Voor meer dan 90% van de GBA-adressen zijn de gegevens op deze afgeleide variabelen daarmee bekend. Voor de overige adressen is echter noch het aantal huishoudens noch de exacte samenstelling bekend. Voor deze adressen vinden imputaties plaats, met aparte imputatiemodellen voor verschillende situaties.

We bespreken hier het eenvoudigste type adressen met onbekende huishoudensamenstelling: adressen met twee niet-in-gezinsverband-levende personen (kortweg: adressen met 2 losse personen). Het is voor deze adressen onbekend of de twee bewoners samen één huishouden vormen of beiden alleenstaand zijn. Eerst wordt een deductieve imputatie (zie paragraaf 2.1) toegepast, met behulp van een afleidingsregel: wanneer beide personen volgens de GBA op dezelfde datum op het adres zijn komen wonen, dan wordt '1 huishouden' geïmputeerd. Dit zal een lichte onderschatting van het aantal huishoudens geven. De resterende adressen worden gekoppeld aan de EBB-steekproef. Dit leverde voor 1999 een koppelingssteekproef EBB × 'GBA met 2 losse personen' op van 1662 adressen. Op grond van deze steekproefadressen werd een imputatiemodel gemaakt.

Met behulp van de bezoekenverantwoordingen en de huishoudbox van de EBB werd voor ieder steekproefadres bepaald of het 1 of 2 huishoudens bevatte. Dit was soms gecompliceerd vanwege nonrespons of door afwijking tussen feitelijke en geregistreerde bewoning. De kans op 2 huishoudens bleek sterk samen te hangen met leeftijd van beide personen (met name het verschil in leeftijd), het al dan niet behoren tot hetzelfde geslacht, stedelijkheidsgraad en het aantal ongehuwden op het adres. Het logistische-regressiemodel (5.3.6) met p de kans op 2 huishoudens was voor 1999:

$$\ln \hat{p}/(1-\hat{p}) = \textit{(.1470 * DIFLFT) + (.0527 * GEMLFT) – (.3916 * STED) + (.7513 * NONGEHUW) + (.0888 * MM) – (6.4201* MV) – (5.7154 * VM) – (DIFLFT * ZELFGESL) - (.0631 GEMLFT * ZELFGESL) – (.9184 * NONGEHUW * ZELFGESL) + constante.}$$

Hierbij is

- DIFLFT = abs. leeftijdsverschil;
- GEMLFT = gemiddelde leeftijd
- STED = stedelijkheidsgraad (scores 1-5, met 1 = hoog en 5 = laag);
- NONGEHUW = aantal ongehuwden (0, 1 of 2);
- ZELFGESL = 2 als twee personen van hetzelfde geslacht, anders 1.

De combinatie van geslacht oudste en jongste persoon bevat vier categorieën die zijn ingevoerd als dummy-variabelen MM, MV, VM en VV met scores 1 (behorend tot

de desbetreffende categorie) en 0. VV is als referentiecategorie uit de vergelijking weggelaten.

De opzet was om voor deze en andere groepen de logistische regressie gewogen uit te voeren met de insluitgewichten (o.a. oversampling van als werkloos geregistreerden) of de EBB-ophooggewichten. Bij de adressen met 2 losse personen is hiervan afgezien, omdat deze gewicht-variabelen geen significante bijdrage gaven aan het model. Voor sommige andere koppelingsgroepen zijn de regressies wel gewogen uitgevoerd, wat steekproeftechnisch beter is, omdat bij deze imputatie de steekproef wordt aangevuld tot de populatie; zie deelparagraaf 1.1.2.9.

Formule (5.3.7) is tenslotte gebruikt om voor ieder niet-steekproefadres uit de koppelingssteekproef de kans op twee huishoudens te schatten, waarna met een lotingmechanisme voor ieder record '1' of '2' is geïmputeerd. Om te voorkomen dat er te vaak naar boven of beneden wordt afgerond is er cumulatief afgerond.

Het bovenstaande voorbeeld is een geval van register-imputatie: voor alle adressen met ontbrekende score op 'aantal huishoudens' vindt een imputatie plaats. De ontbrekende scores zijn bovendien zeer selectief. 'Aantal huishoudens' is namelijk een afgeleide variabele die alleen voor specifieke groepen niet uit de GBA is af te leiden. Slechts door koppeling met een extern steekproefbestand is informatie over het aantal huishoudens voor die groepen beschikbaar gekomen.

5.5 Eigenschappen

1. Indien in formule (5.3.1) geen hulpvariabelen *x* worden gebruikt, gaat deze formule over in $y = \mu + \varepsilon$ met μ de verwachte waarde van *y*, en gaat formule (5.3.3) over in $\tilde{y}_i = \hat{\mu} = \bar{y}$. Dit is mean imputation (hoofdstuk 3).

2. Indien geen constante term wordt gebruikt en alleen de kwantitatieve hulpvariabele x_1, dan gaat formule (5.3.1) over in $y = Rx + \varepsilon$, en gaat (5.3.3) over in formule (4.1.1). Onder bepaalde heterogeniteitveronderstellingen leidt de gewogen kleinste kwadratenschatter tot ratio-imputatie volgens formule (4.3.2).

3. Wanneer $\tilde{y}_i$ wordt geïmputeerd volgens formule (5.3.3), dan heeft het meenemen van de imputaties geen invloed op de schatting van het populatietotaal, wanneer hiervoor de regressieschatter wordt gebruikt met hetzelfde model als het imputatiemodel; zie het thema Steekproeftheorie (Banning e.a., 2010). Zoals in deelparagraaf 1.1.2.9 is besproken worden dergelijke schatters ook 'synthetische schatters' (Boonstra en Buelens, 2007) genoemd.

4. Wanneer de imputatie periodiek wordt herhaald, worden de individuele mutaties sterk overschat (zie hoofdstuk 8).

5.6 Kwaliteitsindicatoren

Het is van belang de kwaliteit van een imputatie in de gaten te houden. Een probleem hierbij is dat meestal de werkelijke waarde onbekend is. Vaak verschillen

gemiddelden voor en na imputatie. Dit is niet noodzakelijk verontrustend want de item-nonrespons kan dan selectief zijn geweest. Indien een overlap met andere onderzoeken bestaat, kunnen externe validaties worden uitgevoerd om een indruk te krijgen van de kwaliteit van de geleverde imputatie. Veelal bestaan er echter definitie- en populatieverschillen tussen de verschillende onderzoeken zodat de mogelijkheden tot dergelijke validaties beperkt zijn.

Omdat doorgaans geen echte toetsing van de kwaliteit van imputaties mogelijk is, zijn de nu volgende kwaliteitsindicatoren voor regressie-imputatie alleen gebaseerd op het model zoals dit is gefit voor de item-respondenten.

- Fit-maten. Voor lineaire regressie-analyse met de kleinste-kwadratenschatter kan R^2 worden gebruikt om de sterkte van het model bij de respondenten te kwantificeren, en daarmee verschillende imputatiemodellen met elkaar te kunnen vergelijken. Men moet de winst in R^2 dan wel afzetten tegen het extra aantal vrijheidsgraden. Ook bij donor-imputatie (hoofdstuk 6) is deze fit-maat van toepassing, daar dit te beschouwen is als regressie op dummy-variabelen. Bij sommige niet-lineaire modellen kan de likelihood als indicator worden gebruikt, of een van de likelihood afgeleide grootheid, zoals AIC of Nagelkerke's R^2. Overigens is het theoretisch mogelijk dat model A ondanks een betere fit dan model B bij de item-respondenten, bij de item-nonrespondenten slechter fit, d.w.z. gemiddeld grotere residuen heeft.

- Validatie/simulatie. Een andere mogelijkheid om een indruk te krijgen van de kwaliteit van een imputatiemethode is het uitvoeren van een simulatie-experiment. Geldige waarden worden dan tijdelijk weggelaten en vervolgens worden nieuwe geldige waarden voor deze weggelaten waarden geïmputeerd. Men kan alle item-respondenten één voor één of in groepjes weglaten, maar kan het weglaten ook beperken tot een deel van de item-respondenten. Als de dan geïmputeerde waarden $\tilde{y}_i$ lijken op of, bij kwalitatieve *y*-variabelen, zelfs gelijk zijn aan de oorspronkelijke waarden y_i, dan geeft dit vertrouwen in de imputatiemethode. Door een geschikte afstandsfunctie te definiëren kan men de geschiktste methode of het geschiktste model kiezen. Een voorbeeld van een afstandsfunctie is de gemiddelde absolute afwijking van geïmputeerde en werkelijke waarden, $\frac{1}{I}\sum_{i=1}^{I}|\tilde{y}_i - y_i|$ met I het aantal imputaties dat wordt beschouwd. Op geaggregeerd niveau kan men als afstandsfunctie nemen de over de simulaties gemiddelde absolute afwijking tussen de geaggregeerde waarden met en zonder imputatie, $\frac{1}{T}\sum_{t=1}^{T}|\tilde{Y}_t - Y_t|$. Een dergelijk experiment is uitgevoerd in Schulte Nordholt (1998).

- IJking aan externe gegevens is doorgaans niet of moeilijk toe te passen, zowel voor de individuele geïmputeerde waarden als op geaggregeerd niveau. Het verkrijgen van de ontbrekende gegevens via (her)benadering van item-nonrespondenten is evenmin eenvoudig te verwezenlijken.

- Berekening van variantie en onzuiverheid is doorgaans ingewikkeld. Men kan te maken hebben met steekproeffouten, selectieve nonrespons, systematische fouten in het imputatiemodel en onzekerheid in het imputatiemodel (door toevoeging van residuen of het random aanwijzen van donoren). Meer over variantieberekeningen is te vinden in bijvoorbeeld Rao (1996). Soms kunnen exacte varianties alleen worden berekend via *multipele imputatie* (Rubin, 1987). Voor elke ontbrekende waarde worden dan verschillende waarden geïmputeerd. Toevoeging van de variantie tussen de imputaties van eenzelfde record zorgt dan voor een zuivere schatting van de variantie van het populatiegemiddelde. Er zijn wel praktische problemen met meervoudige imputatie, zoals dataopslag, ingewikkeldere berekeningen van eenvoudige populatieparameters en complexere analyses op de data). Bovendien is de onderschatting met 'enkelvoudige' imputatie vaak niet zo groot. Mogelijk dat deze techniek in de toekomst toch meer gebruik zal worden.

6. Donor-imputatie (hot deck imputatie)

6.1 Korte beschrijving

Bij *donor imputatie* (hot deck imputatie) wordt voor iedere item-nonrespondent *i* een donor-record *d* in het bestand gezocht met zoveel mogelijk dezelfde kenmerken, voor zover deze van invloed worden geacht op de imputatievariabele(n) *y*. Van deze donor wordt de score, y_d, gebruikt als imputatie:

$$\tilde{y}_i = y_d \,. \tag{6.1.1}$$

De item-nonrespondent wordt de 'recipiënt' (ontvanger) genoemd.
Er bestaan verschillende manieren om een donor te vinden. Deze zijn in te delen in

1. methoden die gebruik maken van imputatieklassen;
2. methoden die een donor zoeken door het minimaliseren van een afstandsfunctie (nearest neighbour hot deck).

Voorbeelden van de 1e klasse van methoden zijn de *random hot deck* en de *sequentiële hot deck* imputatie. Bij random hot deck imputatie worden imputatieklassen gevormd op basis van in klassen ingedeelde hulpvariabelen (achtergrondkenmerken). Uit de overgebleven groep potentiële donoren, met dezelfde kenmerken (x-variabelen) als de item-nonrespondent, wordt er dan aselect één als donor voor de desbetreffende imputatie geloot. Bij de sequentiële hot deck imputatie worden niet daadwerkelijk groepen gevormd, maar wordt voor iedere item-nonrespondent de score op de doelvariabele geïmputeerd van het eerstvolgende record in het databestand met dezelfde scores op bepaalde achtergrondkenmerken.

Een speciaal geval van de 2e klasse is *predictive mean matching*, waarbij nearest-neighbour-donor wordt bepaald door via de voorspelde waarde van *y* voor een gekozen regressiemodel.

Naast hot deck imputatie bestaat ook *cold deck imputatie*. Hierbij wordt de te imputeren waarde uit een *ander* bestand overgenomen, bijvoorbeeld een waarde van hetzelfde object op dezelfde variabele op een vorig tijdstip. Cold deck is in die zin geen echte donor-imputatie. We zullen cold deck imputatie niet als een gevalideerde methode beschouwen. De methode wordt ook weinig meer gebruikt. Als de imputatie uit een ander bestand een correcte waarde is, kunnen we dit zien als een deductieve of logische imputatie (hoofdstuk 2). Als het een waarde uit een eerdere periode betreft, dan is het sec overnemen van de waarde zelden goed te verdedigen. Men zal er dan meestal een trendfactor aan toevoegen, waardoor sprake is van ratio-imputatie (hoofdstuk 4).

6.2 Toepasbaarheid

Random en sequentiële hot deck imputatie worden toegepast als de hulpvariabelen categoriaal zijn. Zijn de meeste variabelen van nature kwalitatief, dan zal men de overige, kwantitatieve variabelen tevoren in klassen indelen. Bij zeer grote bestanden waarop hot deck imputatie wordt toegepast, wordt uit praktische overwegingen wel eens gekozen voor de sequentiële hot deck methode. De doorlooptijd zou anders namelijk erg toenemen, terwijl de kwaliteit van de imputatie (zie paragraaf 5.6) niet noemenswaardig verandert. Om een random donor te krijgen, moet men dan wel eerst de records in een random volgorde in het bestand zetten, maar men hoeft geen lotingmechanisme meer toe te passen.

Nearest neighbour imputatie wordt veeleer toegepast bij de imputatie met behulp van kwantitatieve *x*-variabelen, wanneer informatie verloren zou gaan wanneer deze variabelen tijdelijk in klassen zouden worden ingedeeld. Toch is het ook mogelijk om kwalitatieve hulpvariabelen mee te nemen, zolang de afstandsfunctie daar maar op een verstandige manier mee omgaat. Omdat er bij nearest neighbour een afstandsfunctie tussen potentiële donor en recipiënt wordt geminimaliseerd, is het essentieel dat het belang van iedere *x*-variabele wordt gekwantificeerd in de vorm van een weegfactor; zie hierover paragraaf 6.3.

Donor-imputatie wordt ook gebruikt wanneer per record meerdere waarden ontbreken op aan elkaar gerelateerde variabelen. Door hiervoor één donor aan te wijzen, voorkomt men onderlinge inconsistentie van de imputaties. Dit kan worden gezien als een specifieke oplossing voor het probleem van multivariate imputatie (hoofdstuk 7).

6.3 Uitgebreide beschrijving

6.3.1 Random en sequentiële hot deck imputatie

De bedoeling bij hot deck imputatie is een object in hetzelfde bestand te vinden met soortgelijke achtergrondkenmerken, bijvoorbeeld een individu van hetzelfde geslacht, in dezelfde leeftijdsklasse, woonachtig in dezelfde provincie en werkzaam in dezelfde bedrijfstak. De gedachte is weer dat als een aantal achtergrond-kenmerken van twee individuen overeenstemt de waarden van de te imputeren variabele beter met elkaar overeen zullen komen. Bij random en sequentiële hot deck moeten de donoren exact dezelfde waarden op de achtergrondkenmerken hebben, d.w.z. in dezelfde imputatieklasse zitten. Bij nearest neighbour (paragraaf 6.3.2) worden geen imputatieklassen gevormd en is enige discrepantie in de scores op de *x*-variabelen tussen donor en recipiënt toegelaten.

Bij random en sequentiële hot deck moeten de scores op de achtergrondkenmerken dus identiek zijn. Indien er in het bovenstaande voorbeeld geen respondent is te vinden met dezelfde vier kenmerken als de item-nonrespondent, dan is de imputatieklasse kennelijk te beperkt. Men zal dan voor de imputatie voor deze item-

respondent minstens één van de vier kenmerken moeten laten vallen, of klassen moeten samenvoegen. Indien er echter meer dan één potentiële donor in de relevante imputatieklasse zit, dan moet er worden geloot. In plaats van loten kan men ook een kenmerk toevoegen, in de hoop uiteindelijk één donor over te houden. Men moet wel zien te voorkomen dat één object donor wordt van vele recipiënten. Dergelijk multipele donorschap vergroot namelijk de standaardfouten van gemiddelden en totalen van y, vanwege het risico dat uitschieters worden 'uitvergroot'. Men kan dit bijvoorbeeld voorkomen door binnen een imputatieklasse multipele donors pas toe te staan als alle objecten aan de beurt zijn geweest.

In paragraaf 6.1 is al beschreven dat bij de sequentiële hot deck voor iedere item-nonrespondent de score op *y* wordt geïmputeerd van het eerstvolgende respondent-record in het databestand met dezelfde achtergrondkenmerken. Men kan natuurlijk ook het voorgaande record met die achtergrondkenmerken nemen. Indien een aantal item-nonrespondenten uit dezelfde imputatieklasse vlak achter elkaar in het bestand voorkomen, bestaat het risico dat ze allemaal dezelfde donor krijgen. Ter voorkoming hiervan kan men de sequentiële hot deck methode aanpassen door niet telkens één record, maar de eerste *m* records te selecteren, en daar dan één uit te loten. Sequentiële hot deck kan worden toegepast na een random sortering van de records, in welk geval de methode de 'random sequentiële hot deck methode' wordt genoemd. Sequentiële hot deck kan ook worden uitgevoerd zonder sortering vooraf of alleen na sortering op de uitgekozen achtergrondkenmerken. Men is dan afhankelijk van hoe het bestand in elkaar zit, waardoor onzuiverheid kan ontstaan. In alle gevallen hangen de imputaties af van de volgorde van de records.

Selectie van de hulpvariabelen is een lastig proces. Zowel inhoudelijke als statistische argumenten spelen bij dit proces een rol. Meer hierover is te vinden in de paragrafen 1.1.2.6 en 9.3.

Tot dusverre hielden we geen rekening met eventuele steekproefgewichten. De ramdom (sequential) hot deck wordt echter ook vaak gewogen toegepast; zie Kalton (1983) en paragraaf 1.1.2.10.

6.3.2 Nearest neighbour imputatie

Bij nearest neighbour (hot deck) imputatie wordt een afstand *d(i,j)* gedefinieerd tussen twee objecten *i* en *j*, met *i* de item-nonrespondent en *j* een willekeurige item-respondent. De afstandsfunctie *d* kan op vele manieren worden gedefinieerd. Een veel-gebruikte functie is de Minkowski-afstand $d(i,j) = (\sum_k | x_{ki} - x_{kj} |^z)^{1/z}$,

waarbij de x-variabelen kwantitatief zijn. De respondent *j* met de kleinste waarde van *d(i,j)* is de nearest neighbour van item-nonrespondent *i* en wordt diens donor. Voor *z = 2* gaat de Minkowski-afstand over in de Euclidische afstand, en voor *z = 1* in de zogenaamde city-block-afstand. Hoe groter *z*, des te meer 'straf' komt er op grote verschillen tussen x_{ki} en x_{kj} .

Een betere, algemenere afstandsfunctie is de gewogen afstandsfunctie

$$d_v(i,j) = (\sum_k v_k \mid x_{ki} - x_{kj} \mid^z)^{1/z} \quad . \qquad (6.3.2.1)$$

De extra factor v_k representeert het gewicht (belang) van variabele x_k. Omdat alleen het relatieve gewicht relevant is, mogen we zonder verlies aan algemeenheid aannemen dat $\sum_k v_k = 1$. Het is van essentieel belang dat het gewicht van iedere *x*-variabele vooraf wordt bepaald. Feitelijk kan dit gewicht niet los worden gezien van het waardenbereik of de spreiding van de *x*-variabelen. Praktisch gezien zijn de gewichten vaak makkelijker te bepalen, wanneer men de *x*-variabelen eerst zo heeft genormaliseerd dat ze variantie 1 hebben.

Het is ook mogelijk om bij de definitie van *d(i,j)* rekening te houden met de covarianties tussen de variabelen, maar dit bemoeilijkt doorgaans de bepaling van de gewichten. Een andere mogelijke afstandsfunctie is $\max_k v_k \mid x_{ki} - x_{kj} \mid$ of, wat algemener, $\max_k v_k d(x_{ki}, x_{kj})$. Hiermee zoekt men naar een donor die op geen enkele *x*-variabele sterk van de recipiënt verschilt. Deze afstandsfunctie volgt overigens uit formule (6.3.2.1) met z gelijk aan oneindig.

Een specifiek geval van nearest neighbour is de in Little (1988) beschreven methode van *predictive mean matching*. Bij deze imputatiemethode wordt eerst een lineaire regressie uitgevoerd van de imputatievariabele *y* op verschillende kwantitatieve verklarende *x*-variabelen, op basis van de records zonder item non-respons op de in de regressie gebruikte variabelen. Vervolgens wordt de resulterende regressievergelijking gebruikt om voor alle records waarden te voorspellen voor de imputatievariabele *y*, conform formule (5.3.2). Item-nonrespondent *i* krijgt dan die item-respondent *j* als donor waarvan de voorspelde waarde $\hat{y}_j$ zo dicht mogelijk ligt bij de voorspelde waarde $\hat{y}_i$ van de item-nonrespondent. Ten slotte wordt de *waargenomen* waarde y_j van donor *j* geïmputeerd, d.w.z. $\tilde{y}_i = y_d \circ y_j$ conform formule (6.1.1). Dat predictive mean matching een specifiek geval van nearest neighbour imputatie is volgt uit de afstandsfunctie:

$$d(i,j) = \mid \hat{y}(x_i) - \hat{y}(x_j) \mid \quad . \qquad (6.3.2.2)$$

Bij nearest neighbour, inclusief predictive mean matching, kan men ook de dichtstbijzijnde *m* records selecteren en daaruit random eentje trekken, precies zoals bij sequentiële hot deck is beschreven; eventueel kan men donoren met een kleinere score op de afstandsfunctie een grotere kans geven om in te loten. Het meenemen van steekproefgewichten, zoals bij de gewogen random hot deck methode, heeft geen invloed op de nearest neighbour, wanneer men zich beperkt tot één neighbour. Bij predictive mean matching zal het wegen bij de regressie-analyse ook niet veel invloed hebben.

Men kan de random en nearest neighbour hot deck methoden combineren door eerst klassen te vormen op grond van één of meer achtergrondkenmerken, en vervolgens binnen die ‘blokken’ de nearest neighbour methode toe te passen. Dit is één van de

manieren om nearest neighbour toe te passen met zowel kwalitatieve als kwantitatieve variabelen. In dit geval wegen de kwalitatieve variabelen (oneindig veel) zwaarder dan de kwantitatieve variabelen. Algemener kan men aan afstandsfunctie (6.3.2.1) een afstandsfunctie voor kwalitatieve variabelen toevoegen en een gewogen som van beide als gecombineerde afstandsfunctie nemen. De kwalitatieve variabelen kunnen daarbij ook onderling worden gewogen.

In paragraaf 1.2 hebben we een onderscheid gemaakt tussen 'imputeren' en het ruimere begrip 'corrigeren'. Bij imputeren wordt een ontbrekende waarde vervangen door een geldige waarde; corrigeren van een foutieve waarde door een geldige waarde wordt slechts als imputatie beschouwd wanneer de oorspronkelijke, fout geachte waarde geen rol speelt bij het corrigeren. Nearest neighbour kan eenvoudig worden uitgebreid tot correctie waarbij de oorspronkelijke waarde wel een invloed heeft. De te kiezen afstandsfunctie wordt dan uitgebreid via een restrictie dat de nieuwe waarde weinig mag afwijken van de oorspronkelijke, foutieve waarde Zie het themarapport 'Controle en correctie' in de Methodenreeks (Hoogland e.a., 2010) en Scholtus (2008).

6.4 Voorbeeld

Voorbeeld. *Woningbehoeftenonderzoek (WBO)*
Donor-imputatie is in het verleden op het CBS veelvuldig toegepast bij het WBO. Het ging daar onder meer om het imputeren van inkomensvariabelen en variabelen die de woning betreffen, zoals de verkoopwaarde van de woning. Bij deze variabelen komen veel ontbrekende waarden voor. Als achtergrondkenmerken konden allerlei persoonskenmerken worden gebruikt, maar ook aantal kamers en bezit van een tuin. Vanwege het kwalitatieve karakter van de meeste *x*-variabelen is vooral gebruik gemaakt van donor-imputatie (random hot deck en Predictive mean matching), hoewel ook regressie-imputatie had kunnen worden gebruikt. Gebruik is gemaakt van het programma SURFOX van ABF Research te Delft.

6.5 Eigenschappen

De hot deck en cold deck methoden zijn deterministische imputatiemethoden (paragraaf 1.1.2.7). Na random sortering van het bestand zijn hot deck methoden stochastische methoden geworden. Zoals de naam al aangeeft, is ook de random hot deck methode een stochastische methode. Ook door het toevoegen van een storingsterm (meestal wordt $\varepsilon_i \sim N(0,\sigma^2)$ gekozen) worden deterministische imputatiemethoden tot stochastische methoden.

6.6 Kwaliteitsindicatoren

Zie paragraaf 5.6.

7. Multivariate imputatie

7.1 Korte beschrijving

Tot nu toe was er sprake van telkens één doelvariabele waarvoor waarden ontbraken. Vaak zijn er bij één record ontbrekende waarden op meerdere variabelen, waarbij samenhang bestaat tussen die variabelen. Het imputeren van alle ontbrekende variabelen is dan een multivariaat probleem. In dit hoofdstuk worden verschillende aanpakken voor multivariate imputatie besproken.

Donor-imputatie (hoofdstuk 6) is eenvoudig toe te passen voor meerdere ontbrekende variabelen. Eén donor-record levert dan alle ontbrekende waarden van de recipiënt. Men moet in een dergelijk geval wel imputatieklassen creëren die homogeen zijn voor meerdere doelvariabelen of, in het geval van nearest neighbour imputatie, zorgen voor hulpvariabelen in de afstandsfunctie die samenhangen met meerdere doelvariabelen. Het overnemen van alle ontbrekende doelvariabelen uit hetzelfde donor-record zorgt er ook voor dat de geïmputeerde waarden onderling consistent zijn. Consistentie tussen de geïmputeerde waarden en de oorspronkelijke waarden van de recipiënt is in het algemeen niet gegarandeerd. Het is wel mogelijk om ook consistentie te krijgen tussen geïmputeerde en oorspronkelijke waarden door daar bij de selectie van de donor rekening mee te houden. Deze vorm van donor-imputatie wordt beschreven in hoofdstuk 6 van het Methodenreeks-thema ‘Controle en correctie’ (Hoogland e.a., 2010). Toepassingen van deze methode op gegevens uit het Gemeentelijk Basis Register (GBA) zijn beschreven in Pannekoek e.a. (2008) en Scholtus (2008).

Als er meerdere variabelen zijn met ontbrekende waarden, zal het bij regressie-imputatie (en als bijzonder geval ratio-imputatie) vaak voorkomen dat de predictor(en) ontbrekende waarden bevatten. Er zijn twee op het CBS vaak toegepaste oplossingen voor dit probleem. De ene oplossing is gebaseerd op een vooraf bepaalde volgorde van de doelvariabelen. De eerste doelvariabele wordt geïmputeerd met behulp van een model dat alleen predictoren bevat zonder ontbrekende waarden. Voor de volgende doelvariabele kunnen predictoren gekozen worden uit de variabelen zonder ontbrekende waarden én de in de vorige stap geïmputeerde variabele enzovoort. Bij de tweede oplossing wordt geen gebruik gemaakt van geïmputeerde waarden in de predictoren, maar wordt voor iedere doelvariabele een aantal optionele modellen met verschillende predictoren gespecificeerd. De keuze van het toe te passen model voor een bepaalde doelvariabele in een bepaald record wordt bepaald door de modellen in een van te voren bepaalde volgorde af te lopen. Als de predictoren van het eerste model geen ontbrekende waarden bevatten wordt dat model toegepast, anders wordt het tweede model toegepast als de predictoren daarvan tenminste geen ontbrekende waarden bevatten enzovoort. Deze methoden worden verder toegelicht in deelparagraaf 7.3.1.

Bij de bedrijfsstatistieken doet zich vaak de situatie voor dat er restricties gelden tussen verschillende doelvariabelen. Zo kunnen de totale omzet en de omzetten van een aantal deelposten bekend zijn, maar andere deelposten niet zijn ingevuld. Een simultane vorm van ratio-imputatie kan de ontbrekende deelposten zodanig imputeren dat er een consistent record ontstaat waarbij de (geïmputeerde) deelposten optellen tot het totaal. Aparte ratio-imputaties leiden in het algemeen tot een inconsistent record. Deze methode wordt (nog) niet toegepast op het CBS maar hier wel behandeld omdat het een eenvoudige en nuttige uitbreiding is van ratio-imputatie.

In dit hoofdstuk wordt ervan uitgegaan dat de hulpvariabelen voor het imputeren van een doelvariabele zelf ook ontbrekende waarden kunnen bevatten en dus ook doelvariabelen kunnen zijn. Omdat het onderscheid tussen hulpvariabelen (x-variabelen) en doelvariabelen (y-variabelen) daarom niet meer van toepassing is, worden in dit hoofdstuk alle variabelen met *y* aangeduid.

7.2 Toepasbaarheid

Voor de toepasbaarheid van donor- en regressie-imputatietechnieken voor multivariate problemen geldt wat betreft het meetniveau van de variabelen hetzelfde als wat bij de univariate toepassing van deze technieken in hoofdstuk 5 en 6 is vermeld.

7.3 Uitgebreide beschrijving

7.3.1 Sequentiële imputatie; volgorde van variabelen en volgorde van modellen.

In paragraaf 5.3 is regressie-imputatie besproken voor één doelvariabele. Nu gaan we er vanuit dat er meerdere doelvariabelen met regressie-imputatie geïmputeerd moeten worden. De eenvoudigste methode is om het probleem op te lossen met het herhaald toepassen van de methode voor één doelvariabele. Als de hulpvariabelen voor iedere doelvariabele geen ontbrekende waarden bevatten is dit een eenduidige methode, maar als de hulpvariabelen zelf ook ontbrekende waarden bevatten zijn er verschillende keuzes nodig om tot een toepasbare oplossing te komen.

Eén mogelijkheid is om de variabelen in een bepaalde volgorde te imputeren, zodanig dat de predictoren voor iedere doelvariabele eerst zelf geïmputeerd worden. Er zijn dan altijd waarden voor de predictoren beschikbaar. Deze methode wordt ondermeer toegepast bij de productiestatistieken.

Een andere mogelijkheid is om voor iedere doelvariabele modellen met verschillende predictoren te specificeren. Bij de imputatie kan dan een model gekozen worden waarvoor de predictoren in het betreffende record waargenomen zijn, er wordt dan geen gebruik gemaakt van geïmputeerde waarden in de predictoren. Deze methode is ondermeer toegepast bij het imputeren voor de statistiek Bouwobjecten In Voorbereiding (BIV), (zie Van der Loo en Pannekoek, 2007).

De methode waarbij de predictoren eerst geïmputeerd worden, wordt hieronder toegelicht aan de hand van een vereenvoudigde weergave van de gehanteerde imputatieprocedure voor de productiestatistieken. Bij de productiestatistieken wordt net als bij veel andere economische statistieken gebruik gemaakt van ratio-imputatie. In tabel 3 is voor een aantal doelvariabelen weergegeven welke hulpvariabele gebruikt wordt voor het imputeren van ontbrekende waarden met behulp van ratio-imputatie.

Tabel 3. Imputatieschema voor variabelen uit een productiestatistiek

Variabele	*Hulpvariabele*
y_1: Omzet	-
y_2: Totale bedrijfslasten	Omzet
y_3: Totale personeelslasten	Totale bedrijfslasten
y_4: Kosten huisvesting	Totale bedrijfslasten
y_5: Kosten energie	Totale bedrijfslasten
y_6: Kosten overig	Totale bedrijfslasten
y_7: Kosten vast personeel	Totale personeelslasten
y_8: Kosten overig personeel	Totale personeelslasten

De variabele *Omzet* wordt niet geïmputeerd. Records waarvoor deze centrale variabele ontbreekt worden als non-respons opgevat. Voor de te imputeren records is *Omzet* dus altijd waargenomen. De overige variabelen worden geïmputeerd met de ratio-methode zoals beschreven in hoofdstuk 4. De geïmputeerde waarde $\tilde{y}_{ij}$ voor een doelvariabele y_j in een record *i* kan dan weergegeven worden als:

$$\tilde{y}_{ij} = y^*_{ik}\hat{R}_{jk},$$

met y^*_{ik} de waarde van de hulpvariabele y_k voor de doelvariabele y_j als deze geobserveerd is en anders de geïmputeerde waarde $\tilde{y}_{ik}$, en $\hat{R}_{jk}$ de schatting voor de evenredigheidsconstante R_{jk} behorende bij de variabelen y_j en y_k. Deze imputatiemethode wordt toegepast binnen strata gevormd door combinaties van grootteklasse en bedrijfstak (*group ratio-imputatie,* zie paragraaf 4.3).

De volgorde waarin de doelvariabelen geïmputeerd worden is als volgt: eerst wordt y_2 geïmputeerd met y_1, vervolgens y_3-y_6 met y_2 en tenslotte y_7 en y_8 met y_3. Iedere variabele die gebruikt wordt als hulpvariabele wordt eerst geïmputeerd voordat hij als hulpvariabele wordt gebruikt. Op deze wijze is er altijd een waarde voor de hulpvariabele beschikbaar: ofwel een waargenomen waarde ofwel een geïmputeerde waarde.

Ratio imputatie met gebruik van geïmputeerde waarden voor de hulpvariabele is vergelijkbaar met een methode waarbij met verschillende modellen wordt geïmputeerd en geen gebruik gemaakt wordt van geïmputeerde waarden voor de hulpvariabele. Deze relatie wordt hieronder beschreven. Als de hulpvariabele geïmputeerd is geldt voor de imputatie van de doelvariabele

$$\tilde{y}_{ij} = y^{*}_{ik}\hat{R}_{jk} = \tilde{y}_{ik}\hat{R}_{jk} = y_{il}\hat{R}_{kl}\hat{R}_{jk},$$

met y_l de hulpvariabele voor y_k. Hierbij is aangenomen dat y_{il} is waargenomen. Dit laat zien dat voor de records waarvoor y_k geïmputeerd is, de imputaties niet variëren met y_k maar wel met y_l. Het product $\hat{R}_{kl}\hat{R}_{jk}$ kan opgevat worden als een schatter voor de verhouding R_{jl}. Als de schattingen voor de ratio's R_{jl}, R_{kl} en R_{jk} gebaseerd zijn op dezelfde records geldt exact $\hat{R}_{jl} = \hat{R}_{kl}\hat{R}_{jk}$ en wordt de geïmputeerde waarde gelijk aan een ratio-imputatie met y_l als hulpvariabele. De bovenbeschreven methode is vergelijkbaar met: imputeer y_j met de hulpvariabele y_k als deze waargenomen is en anders met de hulpvariabele y_l. Dit is een voorbeeld van het specificeren van verschillende modellen voor één doelvariabele.

Het specificeren van verschillende modellen voor iedere doelvariabele en vervolgens een model kiezen waarvoor de predictoren waargenomen zijn is toepasbaar op regressie-imputatie in het algemeen. Het nadeel van deze methode is dat er meer modellen gespecificeerd moeten worden dan bij het imputeren van de predictoren. Een voordeel is echter dat er meer mogelijkheden zijn om zo goed mogelijk voorspellende modellen te specificeren. Als bijvoorbeeld bij de productiestatistiek voor een zekere branche de variabele *totale personeelslasten* sterk samenhangt met de variabele *totale bedrijfslasten*, kan er voor gekozen worden om ontbrekende waarden in *totale personeelslasten* te imputeren met *totale bedrijfslasten* als hulpvariabele en ontbrekende waarden in *totale bedrijfslasten* te imputeren met *totale personeelslasten* als hulpvariabele. Pas als beide variabelen ontbreken kan dan voor ieder van deze variabele teruggevallen worden op *omzet* als hulpvariabele.

7.3.2 Ratio-imputatie van deelposten

In het voorbeeld van de vorige deelparagraaf, werd ratio-imputatie toegepast voor deelvariabelen met als hulpvariabele het desbetreffende totaal. Deze situatie doet zich vaak voor bij economische statistieken.

In het algemeen gaat het om variabelen y_j; $j = 0, \cdots, J$, waarvoor de restrictie (of edit-regel) geldt: $y_0 = \sum_{j=1}^{J} y_j$.

Als één van de deelvariabelen y_j ontbreekt, kan deze ene ontbrekende waarde eenvoudig worden geïmputeerd met een deductieve methode (zie hoofdstuk 2). Ook als de som van de geobserveerde variabelen gelijk is aan de 'totaalvariabele' is deductieve imputatie mogelijk, namelijk met de waarde nul voor ieder van de ontbrekende variabelen. Is de som van de geobserveerde deelvariabelen echter kleiner dan de waarde van de totaalvariabele en zijn er meerdere deelvariabelen met ontbrekende waarden dan blijft er nog een deel van het totaal over dat verdeeld moet worden over de ontbrekende waarden.

Een methode om deze verdeling te bepalen is door gebruik te maken van de ratio's van de deelvariabelen tot het totaal maar deze zodanig te herschalen dat de som van de geïmputeerde waarden gelijk is aan het verschil tussen het totaal en de som van de waargenomen deelvariabelen. Als we de waargenomen deelvariabelen in record *i* indiceren met $j = 1, \cdots, J_{i,obs}$ en de ontbrekende deelvariabelen met $j = J_{i,obs} + 1, \cdots, J$, dan is de som van de geobserveerde deelvariabelen in record *i*

$$S_{i,obs} = \sum_{j=1}^{J_{i,obs}} y_{ij}$$

en de som van de ontbrekende deelvariabelen in dat record

$$S_{i,mis} = y_{i0} - S_{i,obs}\,.$$

De imputaties voor de deelvariabelen met behulp van de herschaalde ratio's tot het totaal zijn dan gegeven door

$$\tilde{y}_{ij} = S_{i,mis} \frac{\hat{R}_j}{\sum_{j=J_{i,obs}+1}^{J} \hat{R}_j}\,.$$

Omdat de herschaalde ratio's optellen tot 1 is de som van de geïmputeerde waarden gelijk aan $S_{i,mis}$ en voldoet het geïmputeerde record aan de edit-regel.

Deze vorm van ratio-imputatie, waarbij gebruik gemaakt wordt van de extra informatie dat de som van de ontbrekende waarden bekend is, zal tot betere resultaten leiden dan de gebruikelijke ratio-imputatie die geen gebruik maakt van het bekende totaal van de ontbrekende waarden. Een hot-deck variant van deze methode wordt besproken in Pannekoek en De Waal (2005). In deze variant worden de ratio's niet geschat met behulp van schattingen van de totalen van hulp- en doelvariabele (zoals beschreven in paragraaf 4.3), maar worden zij geschat met de overeenkomstige ratio's zoals die zijn waargenomen in een donor-record (de ratio hot-deck methode).

7.3.3 Simultane regressie-imputatie.

Een algemene multivariate regressie-methode die in veel literatuur over imputatiemethoden wordt beschreven is een methode die gebaseerd is op de veronderstelling dat de simultane verdeling van de betrokken doel- en hulpvariabelen multivariaat normaal is. Met deze methode is het mogelijk om stochastische imputaties te genereren waardoor niet alleen de varianties van geïmputeerde variabelen maar ook de correlaties tussen alle variabelen zo goed mogelijk behouden blijven.

Het uitgangspunt bij deze methode is dat iedere ontbrekende variabele geïmputeerd wordt met een regressiemodel met alle waargenomen variabelen als predictoren. Als bijvoorbeeld de eerste drie variabelen in een record ontbrekende waarden hebben, imputeren we met de drie regressiemodellen (analoog aan formule 5.3.1).

$$y_{i1} = \alpha_1 + \beta_1' y_{i,obs} + \varepsilon_{i1}$$
$$y_{i2} = \alpha_2 + \beta_2' y_{i,obs} + \varepsilon_{i2} ,$$
$$y_{i3} = \alpha_3 + \beta_3' y_{i,obs} + \varepsilon_{i3}$$

met $y_{i,obs}$ de vector met de waarden van de in record *i* geobserveerde variabelen.

Meer algemeen kunnen de regressievergelijkingen voor de ontbrekende waarden in een record *i* worden samengevat in de vorm

$$y_{i,mis} = \alpha_{i,mis} + \beta_{m.o.(i)} y_{i,obs} + \varepsilon_{i,mis} \qquad (7.3.1)$$

met $y_{i,mis}$ de vector met ontbrekende waarden in record *i* en $\alpha_{i,mis}$ de vector met de constanten voor de regressies, $\beta_{m.o.(i)}$ de $q_i \times p_i$-matrix met de regressiecoëfficiënten voor de regressie van de q_i variabelen die ontbreken voor record *i* op de p_i (predictor) variabelen die waargenomen zijn voor record *i* en $\varepsilon_{i,mis}$ de vector met storingen voor de q_i regressies. De matrix met regressie-coëfficiënten hangt af van *i*, maar alleen omdat de variabelen die ontbreken per record kunnen verschillen. Voor records waarin dezelfde variabelen ontbreken is de matrix $\beta_{m.o.(i)}$ gelijk. De storingen zullen in het algemeen gecorreleerd zijn zodat we voor de storingen aannemen dat ze normaal verdeeld zijn met verwachting 0 en een niet-diagonale covariantiematrix: $\varepsilon_{i,mis} \sim N(0, \Sigma_{\varepsilon_{i,mis}})$.

Als er geen ontbrekende waarden zijn, kunnen de parameters van een multivariaat regressiemodel zoals (7.3.1) verkregen worden met de kleinste-kwadraten-methode, analoog aan de schattingsprocedure voor univariate regressiemodellen. Zijn er wel ontbrekende waarden dan zouden de parameters geschat kunnen worden op basis van de records waarin alle variabelen geobserveerd zijn. Het aantal volledige records kan echter beperkt zijn, vooral bij veel variabelen. Een alternatief in zulke gevallen is schattingen te berekenen via het zogenaamde EM-algoritme. Dit is een iteratieve procedure waarmee de parameters geschat kunnen worden op basis van incomplete data; alle gegevens (ook uit de records met non-respons) worden hierbij gebruikt (zie Little and Rubin, 1987).

Met de schattingen $a_{i,mis}$ en $b_{m.o(i)}$ voor de parameters $\alpha_{i,mis}$ en $\beta_{m.o.(i)}$ kunnen de ontbrekende waarden in record *i* geïmputeerd worden volgens

$$\tilde{y}_{i,mis} = a_{i,mis} + b_{m.o(i)} y_{i,obs} . \qquad (7.3.2)$$

Dit is een imputatie zonder storingen, alleen gericht op het reproduceren van de gemiddelden maar niet van de varianties of covarianties. Willen we de (co)varianties van de variabelen na imputatie ook zo goed mogelijk behouden, dan kunnen we een vector met storingen $e_{i,mis}$ die getrokken is uit de multivariate normale verdeling met verwachting 0 en covariantiematrix $\Sigma_{\varepsilon_{i,mis}}$. Het EM-algoritme levert ook een schatting voor deze covariantiematrix.

8. Methoden voor longitudinale imputatie

8.1 Korte beschrijving

We spreken van longitudinale data wanneer dezelfde variabelen meerdere keren worden gemeten bij dezelfde objecten. Panels, waarbij door een steekproef geselecteerde objecten gedurende langere tijd gevolgd worden, zijn hier een speciaal geval van. Maar de in dit hoofdstuk beschreven methoden voor longitudinale imputatie zijn evenzeer van toepassing op andere typen longitudinale data, zoals registers die met enige regelmaat beschikbaar komen. CBS-voorbeelden van (roterende) panels zijn de KS (Korte-termijnstatistiek voor bedrijfsomzetten) en de EBB (Enquête Beroepsbevolking). De GBA is een longitudinaal register dat ieder jaar wordt vernieuwd, terwijl tevens gegevens worden verkregen over tussentijdse verhuizingen en veranderingen in bijvoorbeeld de burgerlijke staat. De meeste registers leveren longitudinale informatie op wanneer exemplaren van verschillende datums onderling worden gekoppeld. Bijvoorbeeld bestanden met banen, uitkeringen en inkomens. In het bijzonder kunnen longitudinale bestanden worden samengesteld uit het SSB (Sociaal Statistisch Bestand). Deze bestanden zullen echter vaak longitudinaal geïmputeerd moeten worden. Met deze bestanden kan men bijvoorbeeld de levensloop van individuen volgen. In het kader van EU-SILC is Nederland verder verplicht paneldata aan Eurostat op te leveren. Dit gebeurt op basis van tal van bestanden waaronder het paneldeel van de EBB.

Longitudinale imputatie onderscheidt zich van de andere in dit rapport beschreven methoden doordat er bij de imputatie gebruik wordt gemaakt van gegevens van hetzelfde object op andere tijdstippen, soms zonder gebruik te maken van gegevens van andere objecten. Per object heeft men dus een tijdreeks met één of meer ontbrekende waarden, waarvoor moet worden geïmputeerd.

Ontbrekende waarden bij longitudinale data kennen twee gedaanten:

1. Verspreid voorkomende ontbrekende waarden doordat objecten gedurende één of meerdere perioden niet worden waargenomen of doordat niet alle variabelen bij de objecten worden waargenomen.
2. Panel-uitval; objecten willen op een zeker moment niet meer meedoen en dus zijn er vanaf een bepaald tijdstip geen waarnemingen meer van het object.

Opgemerkt moet worden dat sterfte en migratie geen ontbrekende waardes teweeg brengen. Deze personen of bedrijven horen niet meer tot de doelpopulatie en moeten dus niet geïmputeerd worden. Lepkowski (1989) geeft een meer gedetailleerde uiteenzetting van verschillende verschijningsvormen van ontbrekende waarden bij longitudinale gegevens.

8.2 Toepasbaarheid

Longitudinale imputatie is bruikbaar wanneer er ontbrekende waarden voorkomen bij longitudinale data. Zij y_{it} een ontbrekende score van object *i* op tijdstip *t* op variabele *y*. Dan kunnen waarnemingen van *y* van object *i* op vorige en volgende tijdstippen worden gebruikt om een geïmputeerde waarde $\tilde{y}_{it}$ te creëren. Vaak is de informatie over *y* beperkt tot eerdere tijdstippen, d.w.z. $y_{it-1}, y_{it-2}, \ldots$ Deze informatie is zowel bruikbaar voor de behandeling van uitval als van verspreid voorkomende ontbrekende waarden. Informatie over latere tijdstippen is alleen bruikbaar wanneer men tijd heeft op de resultaten ervan te wachten of wanneer men imputaties maakt voor een aantal tijdstippen tegelijk, teneinde een zo goed en volledig mogelijk longitudinaal databestand te verkrijgen.

Er zijn twee belangrijke redenen om longitudinale imputatietechnieken te gebruiken in plaats van de cross-sectionele methoden die in voorgaande hoofdstukken besproken zijn.

1. Ten eerste, zullen eerdere of latere waarnemingen aan hetzelfde object erg goede voorspellers zijn voor de missende waarde. De kwaliteit van de imputatie kan dus sterk verbeterd worden. Om dit te bereiken kunnen in feite ieder van de hiervoor besproken methodes gebruikt worden, waarbij voorgaande en toekomstige waarnemingen gebruikt worden als hulpvariabele.
2. Ten tweede, bekijkt men longitudinale data in het algemeen niet alleen cross-sectioneel (bijvoorbeeld het aantal samenwonenden op een bepaald tijdstip), maar is men ook geïnteresseerd in veranderingen in de tijd (bijvoorbeeld het aantal personen dat is gaan samenwonen). Om deze veranderingen correct te kunnen schatten is het belangrijk dat bij de imputatie rekening gehouden wordt met voorgaande en toekomstige waarden.

Panel-uitval bij steekproeven is doorgaans ook op te lossen door middel van wegen. Wanneer men een standcijfer op een bepaald tijdstip wil schatten, kan men de recente uitval als unit-nonrespons beschouwen en bij de nonrespons van eerdere tijdstippen voegen. Panel-uitval bij registers is meestal terecht: die komt door sterfte en emigratie. Voor literatuur over panel-uitval zie bijvoorbeeld Fitzmaurice e.a. (2004, Hfdst. 14).

Veel methoden voor longitudinale data kunnen omgaan met missende data. Zie bijvoorbeeld Van der Laan en Kuijvenhoven (2008) voor een aantal van deze methodes en een literatuurlijst over longitudinale analysemethoden. Het is verder niet altijd noodzakelijk om ontbrekende gegevens te imputeren. Afhankelijk van het doel van de analyses heeft het soms de voorkeur om niet te imputeren.

8.3 Uitgebreide beschrijving

Aangezien de longitudinale imputatiemethoden niet uit één methode bestaan, wordt ieder van de methoden apart besproken in de volgende paragrafen. In deze paragraaf

wordt daarom alleen een aantal kenmerken van longitudinale imputatiemethoden besproken.

De verschillende methoden hebben een aantal kenmerken waarmee ze worden gekarakteriseerd.

- Gebruik van informatie van andere objecten. Een aantal methoden gebruikt alleen voorgaande en toekomstige waarnemingen aan een gegeven object bij de imputatie. Het voordeel hiervan is dat de imputatiemethode vaak redelijk eenvoudig is en ook gemakkelijk is toe te passen op grote datasets. Een nadeel hiervan is echter dat de additionele informatie van andere objecten niet wordt meegenomen, zodat er informatieverlies op kan treden. Zo kan bijvoorbeeld het inkomen overgenomen worden uit de voorgaande periode, waarbij gecorrigeerd wordt voor de gemiddelde inkomensstijging. Gebruik van deze informatie, indien beschikbaar, zal in het algemeen tot een betere imputatie leiden.
- Geschiktheid voor continue en/of categoriale data. Alle besproken methoden zijn geschikt voor continue data. Niet alle methoden zijn echter geschikt voor categoriale data.
- Multivariaat/univariaat. Bij longitudinale data zal het regelmatig voorkomen dat er bij één object meerdere waarnemingen aan *y* ontbreken. Sommige methodes imputeren in één keer meerdere ontbrekende waarden en zullen hierdoor vaak beter in staat zijn om de samenhang tussen waarnemingen op de verschillende tijdstippen te bewaren. Andere methodes kunnen telkens maar één ontbrekende waarde imputeren. Bij meerdere ontbrekende waardes moeten deze methodes meerdere keren toegepast worden. Bij deze handelwijze is het op voorhand niet gegarandeerd dat de samenhang tussen waarnemingen op verschillende tijdstippen behouden is.

Tabel 4 toont voor ieder van de methodes de hierboven genoemde kenmerken. De methodes worden in de volgende paragrafen verder toegelicht.

Tabel 4. Kenmerken van de imputatiemethoden

Methode	**Continu**	**Categoriaal**	**Gebruik informatie andere objecten**	**Multivariaat**
Interpolatie	+	-	-	-
Last observation carried forward	+	+	-	-
Ratio-imputatie	+	-	+	-
Regressie-imputatie	+	+	+	+
Cold deck	+	+	-	+/-
Hot deck	+	+	+	+
Little en Su	+	-	+	+

8.4 Interpolatie

8.4.1 Korte beschrijving

Bij interpolatie worden ontbrekende waarnemingen geschat uit voorgaande en toekomstige waarnemingen. Hierbij wordt geen gebruik gemaakt van informatie van andere individuen of van hulpvariabelen. Voor individu *i* wordt $\tilde{y}_{it}$ dus geschat door

$$\tilde{y}_{it} = f(y_{it-1}, y_{it-2}, \ldots, y_{it-K}, y_{it+1}, y_{it+2}, \ldots, y_{it+L}) . \qquad (8.4.1)$$

Hierbij worden K waarnemingen uit het verleden en L waarnemingen uit de toekomst gebruikt.

8.4.2 Toepasbaarheid

Interpolatie is toepasbaar voor kwantitatieve variabelen in de situatie waar het lastig is om modelaannames te doen en waarbij de overige objecten geen informatie geven over de te imputeren waarde. Als de overige objecten wel informatie bevatten over het te imputeren object, dan zijn methodes die deze informatie meenemen (zoals regressie-imputatie, ratio-imputatie en de methode van Little en Su) aan te bevelen.

8.4.3 Uitgebreide beschrijving

Voor kwantitatieve *y*-variabelen bestaat de volgende, vrij algemene interpolatieformule voor $\tilde{y}_t$ op basis van de waarnemingen $y_{t-K}, \ldots, y_{t-1}, y_{t+1}, \ldots, y_{t+L}$:

$$\tilde{y}_t = \frac{\sum_{k=1}^{K} w_{-k} y_{t-k} + \sum_{\ell=1}^{L} w_{\ell} y_{t+\ell}}{\sum_{k=1}^{K} w_{-k} + \sum_{\ell=1}^{L} w_{\ell}}, \qquad (8.4.2)$$

met gewichten $w_{-1} \geq w_{-2} \geq \ldots \geq w_{-K}$ en $w_1 \geq w_2 \geq \ldots \geq w_L$; y_T weegt dus in beide richtingen vanaf tijdstip *t* minder zwaar mee, naarmate tijdstip *T* verder van tijdstip *t* af ligt. De gewichten zijn vrij te kiezen. Men kan bijvoorbeeld voor $w_k = w_{-k} = 1/k$ kiezen.

Formule (8.4.2) is ook bruikbaar wanneer er meerdere scores van object *i* ontbreken. Indien bijvoorbeeld y_{t+k} niet bekend is en we $\tilde{y}_t$ willen bepalen, definiëren we $w_k = 0$. De formule kan ook worden gebruikt wanneer er alleen informatie uit het verleden bekend is ($w_1 = w_2 = \ldots = w_L = 0$), wat bij panel-uitval het geval is.

Speciale gevallen van formule (8.4.2) zijn:

1. *lineaire interpolatie tussen de vorige en eerstvolgende waarneming.*

   Wanneer y_{t-1} en y_{t+1} beide beschikbaar zijn, dan gaat formule (8.4.2) over in het rekenkundig gemiddelde van beide:

$$\tilde{y}_t = \frac{w_1(y_{t-1} + y_{t+1})}{2w_1} = \frac{y_{t-1} + y_{t+1}}{2}, \qquad (8.4.3)$$

indien $w_{-1} = w_1$. Als y_{t-1} of y_{t+1} ontbreekt, neemt men de dichtst bij tijdstip *t* liggende waarnemingen y_{t-k} en $y_{t+\ell}$ met respectievelijk de gewichten $w_{-k} = 1/k$ en $w_\ell = 1/\ell$. Formule (8.4.2) gaat dan over in

$$\tilde{y}_t = \frac{\ell y_{t-k} + k y_{t+l}}{k + \ell}. \qquad (8.4.4)$$

Stel bijvoorbeeld dat y_t en y_{t+1} onbekend zijn en dat $y_{t-1} = 3$ en $y_{t+2} = 4$, dan volgt uit formule (8.4.4) dat $\tilde{y}_t = \big((2\times 3) + (1\times 4)\big)/\big(1+2\big) = 10/3 = 3{,}333$. Ook op y_{t+1} kan men met behulp van formule (8.4.4) de interpolatie toepassen: $\tilde{y}_{t+1} = \big((1\times 3) + (2\times 4)\big)/\big(2+1\big) = 11/3 = 3{,}667$. We krijgen voor $\tilde{y}_{t+1}$ dezelfde waarde wanneer we de eerder geïmputeerde waarde $\tilde{y}_t$ als bekend veronderstellen:

$$\tilde{y}_{t+1} = \big((1\times \tilde{y}_t) + (1\times y_{t+2})\big)/\big(1+1\big) = \big(10/3 + 4\big)/2 = 3{,}667.$$

2. *gemiddelde van voorafgaande en eerstvolgende p waarnemingen.*

   Men kan een ongewogen gemiddelde van de voorafgaande en eerstvolgende p waarnemingen bepalen, of de waarnemingen ongelijke gewichten geven zoals bij formule (8.4.2). Lineaire interpolatie is dan een bijzonder geval met p=1, $w_{-k} = 1/k$ en $w_\ell = 1/\ell$.

3. *lineaire trend (regressie van y op T)*

   De regressievergelijking $y = \alpha + \beta T + \varepsilon$ kan worden geschat met behulp van de y-waarnemingen die men wil meenemen. Voor tijdstip *T=t*, waarop niet is waargenomen, verkrijgt men dan de regressie-imputatie

   $$\tilde{y}_t = \hat{y}_t = a + bt \qquad (8.4.5)$$

   met *a* en *b* de kleinste-kwadratenschatters, conform formule (5.3.3). Uit de theorie van regressie-analyse is bekend dat $\hat{y}_t$ een lineaire combinatie is van de waarnemingen $y_{t'}$. De lineaire trend gaat over in lineaire interpolatie wanneer de hulpinformatie alleen is gebaseerd op het vorige en eerstvolgende tijdstip. We hebben hier de waarnemingen niet gewogen.

   Natuurlijk kunnen de parameters uit formule (8.4.5) met andere verliesfuncties worden geschat, of kan een vorm van niet-lineaire-regressie worden gebruikt. Dan is de imputatie niet noodzakelijk meer van de vorm van (8.4.2).

Van de bovenstaande drie methodes verdient in het algemeen de eenvoudige lineaire interpolatie (tussen de vorige en eerstvolgende waarneming) de voorkeur. Dit is

zeker zo wanneer de data een geheugenloos proces volgen. De waarnemingen op het vorige en volgende tijdstip bevatten dan alle informatie; die op de andere tijdstippen zijn irrelevant. Als er echter grote meetfouten in de data voorkomen, zullen ook de scores op andere tijdstippen van belang zijn.

Binnen SPSS bestaat de module RMV voor het schatten van ontbrekende waarden in een tijdreeks. Deze module bevat de volgende methoden, met tussen haakjes hoe de methode uit onze formules volgt:

- Linear interpolation (formule (8.4.4)),
- Mean of p nearest preceding and p subsequent values (methode 2),
- Median of p nearest preceding and p subsequent values (variant op methode 2),
- Series mean, d.w.z. gemiddelde van alle waarden uit tijdreeks (specificatie van formule (8.4.2)),
- Linear trend (methode 3).

8.4.4 Eigenschappen

- Interpolatie is eenvoudig toe te passen op grote datasets, omdat er bij de interpolatie slechts informatie van één object wordt gebruikt. Objecten kunnen daarom één voor één verwerkt worden.
- Aangezien er geen informatie van andere objecten gebruikt wordt, kan deze methode minder nauwkeurige schattingen opleveren dan methoden die dat wel doen.
- Doordat er geen storingsterm in de imputatie gebruikt wordt, kan de reeks ‘te mooi’ worden. De significantie van verbanden tussen de verschillende tijdstippen kan overschat worden. Dit kan voorkomen worden door een storingsterm toe te voegen, zie paragraaf 1.1.2.6.

8.5 Last observation carried forward/backward

8.5.1 Korte beschrijving

Last observation carried forward (LOCF) is een methode die vaak wordt gebruikt in de praktijk buiten het CBS. De methode is echter niet zonder problemen, maar wordt vaak toegepast omdat die erg eenvoudig toe te passen is. Bij deze methode wordt de laatst waargenomen waarde van een individu gebruikt voor de waarden van alle latere tijdsstippen die geïmputeerd moeten worden. Bij de uitgebreidere beschrijving worden varianten op deze methode besproken.

8.5.2 Toepasbaarheid

Deze methode is vooral toepasbaar op categorische variabelen waarvan het bekend is dat ze niet of nauwelijks veranderen in de tijd. Een voorbeeld van een dergelijke variabele is geslacht. Bij andere categorische en kwantitatieve variabelen levert deze

methode vaak ten onrechte een te stabiel beeld van de werkelijkheid op. Bijvoorbeeld bij indexcijfers kan deze methode leiden tot het observeren van een onjuiste prijsstabiliteit.

8.5.3 Uitgebreide beschrijving

Bij LOCF wordt de laatst waargenomen waarde y_{it-1} gebruikt om de ontbrekende waarde y_{it} te imputeren. Een andere variant is last observation carried backward, waarbij de eerstvolgende waargenomen waarde y_{it+1} wordt teruggelegd voor de te imputeren waarde y_{it}. Net als bij LOCF kan deze waarde voor meerdere opeenvolgende ontbrekende waarden gebruikt worden.

Bij random carry-over (Williams and Bailey, 1996) wordt een ontbrekende tussenliggende waarde y_{it} geïmputeerd door of y_{it-1} of y_{it+1} te gebruiken. Dit betekent overigens dat de methode niet gebruikt kan worden als waardes bij twee of meer aansluitende tijdspunten ontbreken. Ook kan deze niet toegepast worden als de eerste waarneming en/of laatste waarneming ontbreekt. Voor deze gevallen moet men uitwijken naar andere imputatiemethoden.

8.5.4 Eigenschappen

LOCF heeft als probleem dat het vaak niet realistisch is om aan te nemen dat de laatste waarde door de tijd heen niet meer verandert. Deze assumptie moet dan ook onderzocht worden. Normaliter hebben de gegevens van een individu enige variatie door random fluctuaties (of meetfouten) door de tijd heen. LOCF ontkent deze variantie. Op deze manier wordt echter ook de imputatie-onzekerheid niet adequaat meegenomen met als gevolg verkeerde statistische gevolgtrekkingen. Een eenvoudige oplossing is om een storingsterm toe te voegen (zie paragraaf 1.1.2.6). Hetzelfde geldt ook voor de LOCB methode, waarbij ook onderzocht moet worden of er niet sprake is van een te stabiele tijdreeks die afwijkt van de werkelijkheid.

8.6 Ratio-imputatie

De methode is al besproken in hoofdstuk 4. Zoals daar ook is aangegeven wordt deze methode veel gebruikt bij longitudinale data waar het vaak redelijk is om aan te nemen dat de waarneming op tijdstip *t* proportioneel is met de waarneming op tijdstip *t-1*. Deze methode kan gezien worden als een verfijning van last-observation-carried-forward, waarbij tevens gecorrigeerd wordt voor algehele veranderingen in de tijd. Opgemerkt moet worden dat de storingsterm voor elke tijdsperiode elke keer anders gekozen kan worden. Voor een verdere bespreking van deze methode wordt verwezen naar hoofdstuk 4. Deze vorm van imputatie wordt veelvuldig toegepast bij de economische statistieken.

8.7 Regressie-imputatie

8.7.1 Korte beschrijving

Regressie-imputatie is al in hoofdstuk 5 besproken en wat daar besproken is, is in het algemeen ook geldig voor de longitudinale situatie. We zullen daarom in deze paragraaf alleen ingaan op zaken die te maken hebben met het longitudinale karakter van de data. Aangezien longitudinale data in feite multivariaat is, is de analyse hiervan vaak complexer. Longitudinale data heeft echter wel als voordeel dat in het algemeen in het verleden en/of toekomst waargenomen waarden van een variabele erg goede voorspellers zijn voor missende waarden.

In hoofdstuk 5 wordt de situatie besproken waar we de waarde van één variabele y willen voorspellen aan de hand van een aantal variabelen x_j. We zijn dan voornamelijk geïnteresseerd in de variabele y. In het geval van longitudinale data hebben we voor ieder individu i meerdere waarnemingen y_{it} waarbij t loopt van 1 tot en met M. Bij één individu kunnen meerdere y_{it} missen. Bij analyses op longitudinale data is men in het algemeen geïnteresseerd in de samenhang tussen de waarnemingen op de verschillende tijdstippen – men wil bijvoorbeeld verandering bestuderen. Het is daarom belangrijk om bij de imputatie de samenhang tussen de waarnemingen te behouden. De imputatie is dus multivariaat; multivariate imputatie wordt in hoofdstuk 7 besproken.

Een mogelijkheid, die niet multivariaat is, is om voor iedere missende y_{it} apart een univariaat model op te stellen, waarbij y_{it} niet alleen afhangt van een set covariaten x_{ij}, maar ook van voorgaande en toekomstige waarnemingen van y_{it}:

$$\mathrm{E}[y_{it}] = f(x_{i1},\ldots,x_{ip},y_{it-1},y_{it-2},\ldots,y_{it+1},y_{it+2},\ldots)\,. \qquad (8.7.1)$$

Voor iedere ontbrekende waarneming moet dus een model worden opgesteld en in het geval van meerdere ontbrekende waarnemingen en ontbrekende covariaten moeten aparte modellen opgesteld worden. Dit kan erg complex worden en het is erg lastig om de samenhang tussen de waarnemingen te bewaren.

Een andere mogelijkheid is het gebruik van een multivariaat model (zie bijvoorbeeld Verbeke en Molenberghs (2000)). Hierbij wordt één model opgesteld dat alle waarnemingen beschrijft. De verschillende waarnemingen aan individu i worden geschreven als vector $\mathbf{y_i}$ en er wordt een model opgesteld dat deze vector beschrijft. Bijvoorbeeld een lineair model

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \boldsymbol{\varepsilon}_i\,, \qquad (8.7.1)$$

waarbij de vector $\boldsymbol{\varepsilon}_i$ een multivariate normale verdeling volgt. In het geval van longitudinale data is het belangrijk om de correlatie die bestaat tussen de verschillende waarnemingen (bijvoorbeeld iemand met een hoog inkomen bij een bepaalde waarneming zal waarschijnlijk ook een hoog inkomen hebben bij de volgende waarneming) te modelleren.

Het multivariaat modelleren van longitudinale data valt buiten het thema Imputatie en zal hier dus verder niet worden besproken. Voor meer informatie zie bijvoorbeeld Van der Laan en Kuijvenhoven (2008), Verbeke en Molenberghs (2000) en Molenberghs en Verbeke (2005) voor discrete longitudinale gegevens. Deze geven ook een overzicht van literatuur over deze materie. Gelman en Hill (2006) en Longford (2005) geven meer gedetailleerde beschrijvingen van hiërarchische of multi-level-modellen.

8.7.2 Toepasbaarheid

- Regressie-imputatie is toepasbaar voor zowel kwantitatieve variabelen als categoriale variabelen. In het laatste geval kan echter geen gebruik gemaakt worden van (multivariate) lineaire regressie, maar moet bijvoorbeeld logistische regressie gebruikt worden.
- De hiervoor besproken multivariate regressiemodellen kunnen vaak omgaan met verschillende waarnemingstijdstippen voor de verschillende individuen. De meeste andere methoden besproken in dit hoofdstuk gaan ervan uit dat alle individuen op vaste waarneemmomenten worden waargenomen (bijvoorbeeld ieder jaar of ieder kwartaal).

8.7.3 Eigenschappen

Bij de analyse van longitudinale data is men in het algemeen geïnteresseerd in veranderingen door de tijd. Zoals besproken in 1.1.2.6 kan men bij het imputeren ervoor kiezen om wel of geen storingsterm toe te voegen. Als men in het geval van longitudinale data de storingsterm weglaat, zal de significantie van de veranderingen sterk overschat worden.

8.8 Cold deck

Cold deck imputatie is al besproken in hoofdstuk 6. Daar is de methode als niet gevalideerde methode beschouwd. We zullen deze methode verder niet bespreken. Opgemerkt moet worden dat we de methode Last observation carried forward/backward niet beschouwen als cold deck methode. Bij cold deck imputatie wordt namelijk gebruik gemaakt van informatie uit een externe bron. Bij Last observation carried forward/backward wordt echter gebruik gemaakt van een bestand van een eerdere of respectievelijk een latere tijdsperiode. Dit bestand wordt niet als externe bron gezien.

8.9 Hot deck

Hot deck imputatie of donor-imputatie is al besproken in hoofdstuk 6. In dat hoofdstuk werd al vermeld dat donor-imputatie gebruikt wordt wanneer per record meerdere waarden ontbreken. Dit maakt donor-imputatie bijzonder geschikt om toe te passen op longitudinale gegevens. Bij de hot deck methode kunnen namelijk meerdere waarden voor één individu geïmputeerd worden. Hiervoor wordt in de

regel één donor aangewezen om zodoende onderlinge consistentie van imputaties te verkrijgen. Bij longitudinale gegevens wordt op deze manier de correlatie tussen opvolgende waarden in de tijd beter behouden. Hoofdstuk 7, dat handelt over multivariate imputatie, gaat verder in op dit onderwerp.

8.10 Methode van Little en Su

8.10.1 Korte beschrijving

De methode van Little en Su (Little en Su, 1989) neemt zowel het individuele niveau als de gemiddelde trend door de tijd mee in de imputatie. Hierbij wordt het volgende model gehanteerd

$$\text{(imputatie)} = \text{(rijeffect)} \times \text{(kolomeffect)} \times \text{(residu)}. \qquad (8.10.1)$$

Het kolomeffect beschrijft de gemiddelde verandering door de tijd en wordt dus ook wel periode-effect genoemd, terwijl het rijeffect het individuele niveau gecorrigeerd voor het periode-effect beschrijft. Bij de methode van Little en Su wordt het residu overgenomen van een ander individu die wat betreft rijeffect het meest lijkt op het individu dat geïmputeerd wordt. De aanname is dat individuen die wat betreft rijeffect op elkaar lijken ook wat betreft residuen op elkaar lijken.

Als een individu meerdere gerelateerde waardes mist (bijv. bruto en netto loon, binnen SURFOX wordt dit record matching genoemd), worden deze waardes in één keer geïmputeerd, waarbij één donor gebruikt wordt voor de residuen.

8.10.2 Toepasbaarheid

De methode van Little en Su kan worden toegepast voor ontbrekende waarden in een kwantitatieve positieve variabele *y*, die te modelleren is als een periode-effect maal een individueel effect en waarbij stochastische imputatie gewenst is. De methode is redelijk eenvoudig toe te passen en kan omgaan met verschillende patronen van missende data, waaronder meerdere missende waarden per individu.

De methode heeft problemen met individuen waarbij de waargenomen waarden alle gelijk zijn aan nul. Deze individuen kunnen niet geïmputeerd worden met de methode van Little en Su.

8.10.3 Uitgebreide beschrijving

Het kolomeffect c_t geeft de gemiddelde verandering van de objecten in de tijd en wordt geschat door

$$c_t = \frac{\bar{y}^t}{\frac{1}{M}\sum_{t=1}^{M} \bar{y}^t}, \qquad (8.10.2)$$

waarin $\bar{y}^t$ het gemiddelde is over de waargenomen y_i^t op tijdtip *t*, *M* is het aantal perioden. Het rijeffect r_i voor individu *i* wordt gegeven door

$$r_i = \frac{1}{m_i} \sum_t \frac{y_i^t}{c_t}, \tag{8.10.3}$$

waarbij gesommeerd wordt over de m_i beschikbare y_i^t voor individu *i*.

Het residu wordt overgenomen van een ander individu *j* waarvan de tijdstippen die bij individu *i* ontbreken wel zijn waargenomen. Individu *j* wordt geselecteerd door eerst alle individuen te sorteren op rijeffect en vervolgens het individu te kiezen waarvan het rijeffect het dichtst bij die van *i* ligt. Het residu van individu *j* wordt gegeven door

$$e_j^t = \frac{y_j^t}{r_j c_t}. \tag{8.10.4}$$

Invullen in vergelijking (8.10.1) geeft

$$\tilde{y}_i^t = r_i c_t e_j^t = r_i c_t \frac{y_j^t}{r_j c_t} = \frac{r_i}{r_j} y_j^t. \tag{8.10.5}$$

De donor (van de residuen) heeft in het ideale geval zoveel mogelijk dezelfde eigenschappen als de ontvanger. De standaardmethode zoals hiervoor besproken probeert dit te bereiken door donor en ontvanger met behulp van het rijeffect aan elkaar te koppelen. Het is echter ook mogelijk om de methode uit te breiden, door de standaardmethode toe te passen binnen strata. Hiermee wordt ook toegestaan dat de kolomeffecten verschillen tussen de strata, dus dat het gemiddelde verloop door de tijd verschilt tussen de strata. Deze methode wordt ook wel extended Little en Su genoemd en is onder andere toegepast in HILDA (Starick en Watson, 2006).

8.10.4 Eigenschappen

- Deze methode kan door de manier waarop de residuen bepaald worden ook de waarde nul imputeren, zelfs als de waargenomen waarden ongelijk aan nul zijn. De frequentie waarmee nul geïmputeerd zal worden zal van dezelfde orde zijn als de fractie nullen in de complete data. Veel andere methoden zoals regressie-imputatie en ratio-imputatie missen deze eigenschap.
- Deze methode neemt impliciet aan dat het rijeffect groter dan nul is. Voor een individu waarvan de waargenomen waarden gelijk aan nul zijn, kan nooit een waarde ongelijk aan nul geïmputeerd worden. In het algemeen is dit niet realistisch.
- Donoren waarvan het rijeffect gelijk is aan nul geven een probleem, omdat in formule (8.10.5) hierdoor gedeeld wordt. In het algemeen zullen deze vooral gekoppeld worden aan ontvangers waarvan het rijeffect ook gelijk is aan nul,

wat zoals in het vorige punt is besproken ook problematisch is. Voor individuen met rijeffect gelijk aan nul kan deze methode dus niet gebruikt worden.

8.10.5 Kwaliteitsindicatoren

- De residuen kunnen redelijk gemakkelijk uitgerekend worden met formule (8.10.4). Als het model goed past, dan zijn de residuen ongeveer gelijk aan één. Hierbij moet wel rekening worden gehouden met het feit dat de residuen niet symmetrisch verdeeld zijn. De residuen zijn namelijk altijd groter dan nul.
- Validatie/simulatie. Zie hiervoor paragraaf 5.6.
- Er zijn geen formules bekend waarmee de variantie en onnauwkeurigheid bepaald kunnen worden. Multipele imputatie (zie 5.6 en Rao, 1996) is met de methodes zoals hier besproken niet uit te voeren. Hiervoor is het namelijk noodzakelijk om meerdere verschillende waardes te imputeren. De hier besproken methodes imputeren echter altijd dezelfde waarde. Misschien dat het aanpassen van de donorselectie multipele imputatie mogelijk maakt.

8.11 Afsluiting

Een punt waarmee bij longitudinale data rekening gehouden moet worden is op welke manier met nieuwe informatie moet worden omgegaan. Bij een longitudinaal databestand wordt de best mogelijke imputatie op microniveau verkregen als zoveel mogelijk informatie uit het verleden en toekomst wordt meegenomen. Als er dus nieuwe informatie binnenkomt, zoals een nieuwe golf data bij een panel, kan deze nieuwe informatie gebruikt worden om de reeds geïmputeerde waarden te herzien of verbeteren. Er moet dus besloten worden hoe ver men informatie mee terug neemt:

- Er kan besloten worden om de nieuwe informatie niet te gebruiken om eerder gemaakte imputaties te verbeteren. De eerder gemaakte imputaties zijn dan niet zo goed als ze misschien zouden kunnen zijn, maar men voorkomt dat men verschillende versies krijgt van hetzelfde bestand. Een nadeel is dat de vergelijkbaarheid van de data in de tijd in het geding komt. De nieuwe informatie zou bijvoorbeeld in strijd kunnen zijn met de reeds geïmputeerde waarden. Het is dan moeilijk om longitudinale analyse uit te voeren.
- Als men de nieuwe informatie wel gebruikt om eerdere imputaties te herzien, krijgt men te maken met verschillende versies van gegevens. Men krijgt dus bijvoorbeeld een bestand over 2008 met de informatie die beschikbaar was in 2008 en een bestand over 2008 met de informatie die beschikbaar was tot en met 2009.

9. Afsluiting

9.1 Vlaggen / documenteren

Het is verplicht te documenteren welke waarden geïmputeerd zijn en welke methoden hiervoor zijn gebruikt, inclusief de in het model gebruikte hulpvariabelen en parameters. Dit is nodig om het proces reproduceerbaar te maken. Er zijn diverse mogelijkheden om geïmputeerde waarden in het bestand aan te geven:

- 'vlaggen' van de geïmputeerde waarden;
- werken met ongeïmputeerde en geïmputeerde bestanden;
- variabelen voor en na imputatie onderscheiden.

Een dergelijke documentatie is ook noodzakelijk voor onderzoekers die nadere analyses willen doen op het microbestand. Voor hen kan het ongewenst zijn om de imputaties te gebruiken, omdat dat zou kunnen leiden tot verkeerde conclusies. Ook voor het bepalen van standaardfouten is het nodig te weten welke scores echt zijn en welke scores geïmputeerd, inclusief de imputatiemethode.

Een bij sommige statistieken gebruikte werkwijze is om meteen van een geïmputeerd bestand uit te gaan, ook wanneer er nog geen data binnen zijn. De geïmputeerde waarden kunnen dan aanvankelijk gebaseerd worden op de waarden van periode *t-1*. Telkens zodra nieuwe gegevens binnenkomen, vervangen deze de imputaties, waarna de resterende imputaties worden geüpdatet. Een dergelijke werkwijze is alleen procesmatig afwijkend (men kan op ieder moment snel schattingen leveren), maar niet qua methodiek.

9.2 Omgaan met uitbijters

Indien er bij de respondenten uitbijters (uitschieters) op de variabele *y* voorkomen, kan men overwegen de invloed hiervan bij het imputeren te beperken. Zo kan men een robuuste vorm van regressie-analyse uitvoeren, of een potentiële donor met een, gegeven de hulpvariabelen, extreme waarde op *y* een kleinere kans geven om als donor te fungeren. Het aldus bij het imputeren rekening houden met uitbijters verkleint de betrouwbaarheidsmarges, maar introduceert een (extra) onzuiverheid. Men moet hiermee dus zeer voorzichtig zijn en goed weten welke parameterschatters het onderzoek moet opleveren. Zo zal men dergelijke robuuste methoden eerder willen toepassen voor kleine (deel)populaties, omdat de standaardfouten anders te groot worden, dan voor zeer grote populaties. Inhoudelijke kennis moet mede bepalend zijn voor de beslissing hoe met uitbijters om te gaan. Wanneer bijvoorbeeld iemand 400 dagtochten in een jaar naar zijn volkstuin heeft gemaakt, hoeft dat nog geen reden te zijn om die persoon niet als donor mee te nemen. Maar stel dat het een 50-jarige man uit Assen betreft, dan is er weinig reden om die uitbijter te versterken door een hem als donor aan te wijzen voor een ongeveer even oude man uit Assen.

9.3 Selectie van hulpvariabelen

Als toevoeging aan deelparagraaf 1.1.2.5 geven we hier enkele richtlijnen voor het selecteren van hulpvariabelen.

- Selecteer *x*-variabelen waarvan je verwacht dat ze ook voor de item-nonrespondenten relevant zijn. In de regel zal men toch kijken of de variabelen veel verklarende waarde hebben voor de item-respondenten, omdat toetsing van het model voor de item-nonrespondenten niet mogelijk blijkt.
- Neem niet te veel variabelen op in een regressiemodel. De parameters worden dan slecht geschat. Voor goede voorspellingen (imputaties) is een redelijk zuinig model te verkiezen.
- Bij donor-imputatie hindert het echter niet als er veel deelpopulaties (veel variabelen met veel categorieën) worden onderscheiden. Zelfs het toevoegen van nonsens-variabelen met het doel een unieke donor over te houden is geen probleem, maar hoogstens een alternatieve manier om uiteindelijk één random donor uit een deelpopulatie te trekken. Men moet dan wel oppassen voor multipele donoren; zie paragraaf 6.3.
- De volgorde van in het invoeren van variabelen in het model is een kwestie van modelselectie. Gebruik kwaliteitsmaatstaven om de winst van het toevoegen van een variabele te kunnen kwantificeren; bijvoorbeeld toename van R^2, F-toets, AIC, BIC.

9.4 Niet-negatieve variabelen met veel nullen

Bij activiteiten waaraan niet iedereen deelneemt, ontstaat een verdeling waarbij een deel van de populatie, de non-participanten, nul scoort en de overigen, de participanten, allerlei positieve waarden. Voorbeelden zijn vakantiebesteding in euro's, aantal gereden kilometers met de auto en omzet aan een bepaalde nevenactiviteit. Hot deck methoden werken goed bij dit type variabelen, in de zin dat zij de verdeling behouden. Past men echter mean imputation toe, dan zal geen enkele nul worden geïmputeerd. Ook regressie-imputatie geeft problemen. Negatieve imputaties kunnen voorkomen bij dergelijke niet-negatieve variabelen. Indien het doel alleen is populatiegemiddelden te schatten, dan is het niet zo'n probleem. Maar wanneer men ook de spreiding van de variabelen 'redelijk' wil houden, of de fractie participanten goed wil kunnen schatten, kan men deze technieken niet toepassen. Een mogelijkheid is om de imputatie dan in twee stappen te doen. Men kan dan bijvoorbeeld eerst via een logistische regressie beslissen (imputeren) of item-nonrespondenten wel of niet participeren, en vervolgens voor de veronderstelde participanten de score bepalen via een lineair regressiemodel.

9.5 Combinatie van methoden (hiërarchie)

Wanneer men voor ontbrekende waarden op een variabele *y* wil imputeren, hanteert men soms een strategie met verschillende methoden of modellen, afhankelijk van de beschikbare hulpinformatie voor het record; zie voorbeeld 2 in paragraaf 4.4. In dat voorbeeld wordt eerst gekeken naar informatie over dezelfde variabele in een

voorgaande periode, vervolgens naar informatie uit een andere bron en ten slotte naar informatie over dezelfde variabele van de item-respondenten.

10. Literatuur

Banning, R., Camstra, A. en Knottnerus, P. (2010), *Methodenreeks: Thema: Steekproeftheorie, Deelthema's Steekproefontwerp en Ophoogmethoden*. Centraal Bureau voor de Statistiek, Den Haag.

Boonstra H.J. en Buelens, B. (2007), *Methodenreeks: Thema: Modelmatig schatten. Deelthema's: Synthetische schatters en Kleine-domeinschatters*. CBS, Heerlen.

Fitzmaurice, G.M., Laird, N.M. en Ware, J.H. (2004), *Applied Longitudinal Analysis*. Wiley, New York.

Gelman, A. en Hill, J. (2006), *Data Analysis Using Regression and Multilevel/Hierarchical Models.* Cambridge University Press, Cambridge.

Harmsen, C.N. en Israëls A.Z. (2000), *Nieuwe Huishoudensstatistiek: Nieuw versus Oud*. Interne nota. Centraal Bureau voor de Statistiek, Voorburg.

Hoogland, J., Loo, M.P.J. van der, Pannekoek, J. en Scholtus, S. (2010), *Controle en correctie*. Rapport Methodenreeks, CBS, Den Haag.

Kalton, G. (1983), *Compensating for Missing Survey Data*. Survey Research Center Institute for Social Research, The University of Michigan.

Laan, D.J. van der en Kuijvenhoven, L. (2008), *Longitudinale analyse: Multilevel-modellen voor paneldata.* Interne nota. Centraal Bureau voor de Statistiek, Voorburg.

Lepkowski, J.M. (1989), *Treatment of wave nonresponse in panel surveys* In: Kasprzyk, D., Duncan, G.J., Kalton, G., Singh, M.P., eds., *Panel Surveys*. Wiley, New York.

Little, R.J.A. (1988), Missing-data adjustments in large surveys. *Journal of Business & Economic Statistics*, 6, 287-296.

Little, R.J.A. en Rubin, D.B. (1987), *Statistical Analysis with Missing data*. John Wiley & Sons, New York.

Little, R. J. A. en Su, H.L. (1989), Item Non-response in Panel Surveys. In: Kasprzyk, D., Duncan, G.J., Kalton, G., Singh, M.P., eds., *Panel Surveys*. Wiley, New York.

Longford, N. (2005), *Missing Data and Small-Area Estimation: Modern Analytical Equipment for the Survey Statistician.* Springer-Verlag, New York.

Loo, M. van der, en Pannekoek, J. (2007), *Advies gaafmaken en imputeren van de statistiek Bouwobjecten in Voorbereiding*. Interne nota, Centraal Bureau voor de Statistiek, Voorburg.

Molenberghs, G. en Verbeke, G. (2005), *Models for Discrete Longitudinal Data*. Springer-Verlag, New York.

Pannekoek, J., Harmsen, C., Huis, M. van, en Prins, K. (2008) *Automatisch gaafmaken van GBA-gegevens met de "Nearest-neighbor Imputation Methodology"*. Interne nota, Centraal Bureau voor de Statistiek, Den Haag.

Pannekoek, J. en Israëls, A.Z. (2000), Effecten van Steekproefontwerp op (regressie) analyses. *Kwantitatieve Methoden* 65, 113-131.

Pannekoek, J. en D.C.G. Tempelman (2005). *Imputatiemethoden voor Impect statistieken: deductieve imputatie en correctie voor overduidelijke fouten*. Interne nota. Centraal Bureau voor de Statistiek, Voorburg.

Pannekoek, J. en Waal, T. de (2005). Automatic Edit and Imputation for Business Surveys: The Dutch Contribution to the EUREDIT Project. *Journal of Official Statistics*, 21, 257-286.

Rao, C.R. (1973*), Linear statistical inference and its applications 92nd ed.*). Wiley, New York.

Rao, J.N.K. (1996). On variance estimation with imputed survey data. *Journal of the American Statistical Association* 91, 499-506.

Rubin, D.B. (1987), *Multiple imputation for nonresponse in surveys*. John Wiley & Sons, New York.

Schulte Nordholt, E. (1998), Imputation: methods, simulation experiments and practical examples. *International Statistical Review*, 66, 157-180.

Scholtus, S. (2008), *Automatisch gaafmaken van GBA-gegevens met CANCEIS*. Interne nota, Centraal Bureau voor de Statistiek, Den Haag.

Starick, R. and Watson, N. (2006), *An Evaluation of Alternative Income Imputation Methods in the HILDA Survey*, HILDA Project Technical Paper Series, Melbourne Institute of Applied Economic and Social Research, The University of Melbourne.

Verbeke, G. en Molenberghs, G. (2000), *Linear mixed models for longitudinal data*. Springer-Verlag, New York.

Williams, T.R., en Bailey, L. (1996), *Compensating for Missing Wave Data in the Survey of Income and Program Participation (SIPP).* Proceedings of the American Statistical Association, Survey Research Methods Section, pp. 305–310.