Modelmatig schatten

Deelthema's: Synthetische schatters en Kleine-domeinschatters

08

Harm Jan Boonstra en Bart Buelens

Statistische Methoden (08001)

Centraal Bureau voor de Statistiek

Voorburg/Heerlen, 2008

Verklaring van tekens

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

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

Colofon

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

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

Prepress
Centraal Bureau voor de Statistiek - Facilitair bedrijf

Omslag
TelDesign, Rotterdam

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

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

Internet
www.cbs.nl

ISSN: 1876-0333

6016508001 X-37

Inhoudsopgave

1. Inleiding op het thema

1.1 Algemene beschrijving en leeswijzer

Het thema modelmatig schatten kan zeer breed geïnterpreteerd worden. De meeste takken van de statistiek zijn modelmatig in de zin dat er uitspraken worden gedaan op basis van *expliciete* (kans)modellen die ontbrekende informatie relateren aan beschikbare informatie. Het doel van deze beschrijving is te laten zien hoe relatief eenvoudige modellen gebruikt kunnen worden om schattingen voor populatiegrootheden te maken. We concentreren ons op schattingen van gemiddelden of totalen voor (eindige) populaties of deelpopulaties.

De steekproeftheorie wordt doorgaans beschreven vanuit een design-gebaseerd standpunt. Daarbij staat het kansmechanisme dat gebruikt wordt om de steekproef te trekken centraal. Er bestaan echter situaties waarin een design-gebaseerde aanpak niet of niet goed werkt. Twee van dergelijke situaties zijn die waarin:

1. er geen (bekend) kanssteekproefontwerp is, zoals bij administratieve data uit (onvolledige) registers of bij sommige vormen van internetwaarneming;
2. er te weinig steekproefdata beschikbaar is om betrouwbare schattingen te maken. Dit speelt vooral als het detailniveau waarop cijfers gemaakt moeten worden hoog is, zodat de steekproefomvang in de verschillende deelpopulaties (te) klein is.

In deze situaties kunnen modelmatige schattingsmethoden gebruikt worden. Eerst wordt besproken hoe populatietotalen geschat kunnen worden aan de hand van lineaire regressiemodellen. Deze modellen hangen niet expliciet van een steekproefontwerp af en kunnen daarom gebruikt worden in situatie 1. Er worden een aantal zaken genoemd waar op gelet moet worden bij gebruik van deze modellen. De schatters die uit het gebruik van deze modellen volgen worden ook wel synthetische schatters genoemd. Als er voldoende geschikte hulpvariabelen als input voor het regressiemodel zijn, dan kunnen synthetische schatters ook gebruikt worden voor het maken van schattingen voor kleine domeinen.

In het tweede deelthema wordt een eenvoudig type model besproken dat specifiek geschikt is voor kleine-domeinschattingen, ook in situaties waar synthetische schattingen tekortschieten. Kleine domeinen zijn deelpopulaties waarbij de steekproefomvang te klein is om betrouwbare directe (design-gebaseerde) schattingen te maken. Het model, dat de verschillende domeinen met elkaar verbindt, eventueel door gebruik te maken van relevante hulpinformatie op domeinniveau, zorgt voor betere schattingen. Het deelthema kleine-domeinschatters behandelt dus situatie 2.

1.2 Afbakening en relatie met andere thema's

We onderscheiden twee deelthema's bij het thema modelmatig schatten: synthetische schatters en kleine-domeinschatters. Ook andere thema's in de methodenreeks, over onder andere macro-integratie en seizoenscorrectie/tijdreeksmodellen, maken gebruik van modelmatige schattingsmethoden.

Het begrip synthetische schatter wordt gebruikt voor schatters gebaseerd op regressiemodellen met alleen vaste effecten, zonder random effecten. Vaste effecten zijn de gebruikelijke regressiecoëfficiënten, terwijl random effecten geïnterpreteerd kunnen worden als een groep regressiecoëfficiënten met de vooraf opgelegde beperking dat ze volgens een gemeenschappelijke kansverdeling rond 0 gespreid liggen. We zullen hier alleen synthetische schatters gebaseerd op lineaire regressiemodellen beschrijven. Dus synthetische schatters gebaseerd op bijvoorbeeld een logistisch regressiemodel worden niet besproken. Onder het deelthema kleine-domeinschatters wordt verder ingegaan op modellen met zowel vaste als random effecten, ook wel gemengde modellen genoemd. De random effecten corresponderen hier met de domein-indicatoren, en kunnen verschillen tussen domeinen verklaren die niet door de andere gebruikte hulpvariabelen worden verklaard.

Het model voor kleine-domeinschattingen dat in het tweede deelthema aan bod komt is geformuleerd op domeinniveau. Dit betekent dat de te schatten domeingemiddelden direct gemodelleerd worden in termen van hulpvariabelen op domeinniveau, en als inputdata worden directe schattingen en bijbehorende variantieschattingen voor domeinen gebruikt. Kleine-domeinmodellen geformuleerd op eenheidsniveau, zeg persoonsniveau, worden niet besproken.

In het geval van kanssteekproeven zal de synthetische schatter voor het populatietotaal gebaseerd op lineaire regressiemodellen vaak overeenkomen met de algemene regressieschatter, zeker bij steekproefontwerpen met gelijke insluitkansen. Paragraaf 2.5 gaat daar iets verder op in. De methodologie van de algemene regressieschatter en het daaraan gerelateerde wegen komen aan bod bij andere thema's in de methodenreeks: 'Steekproeftheorie' en 'Wegen als correctie voor nonrespons'. Er is ook een sterke relatie met de bij 'Controle en correctie/imputatie' behandelde methode van regressie-imputatie. Daar worden op micro-niveau ontbrekende waarden vervangen door geïmputeerde waarden op basis van een regressiemodel. In tegenstelling tot bij synthetische schatters hoeft het schatten van populatietotalen of –gemiddelden daarbij geen (primair) doel te zijn.

1.3 Plaats in het statistisch proces

In het statistisch proces neemt schatten een plaats in na het controleren en gaafmaken van de data. Voor synthetisch schatten is dat niet anders. Het zal dan bijvoorbeeld gaan om registerdata die eerst gekoppeld worden aan een populatieregister ("ruggengraat") en vervolgens worden gaafgemaakt.

De kleine-domeinschatters die in het tweede deelthema behandeld worden gebruiken als inputdata directe schattingen en bijbehorende variantieschattingen op

domeinniveau. Deze directe schattingen kunnen uit een weging van de steekproefdata voortkomen. Het weegprogramma Bascula kan zowel schattingen voor deelpopulaties als bijbehorende variantieschattingen berekenen, zie Nieuwenbroek en Boonstra (2002).

1.4 Algemene notatie

We gaan uit van een doelvariabele *y* waarvoor het populatietotaal of gemiddelde geschat moet worden. Deze variabele neemt de waarden $y_1,\ldots,y_N$ aan voor de populatie-eenheden $U=\{1,\ldots,N\}$. Voor een deelverzameling *s* van n<N unieke eenheden zijn de waarden van *y* bekend. We veronderstellen dat deze waarden foutloos zijn. De deelverzameling *s* kan de respons van een (kans)steekproef zijn of het kan de verzameling eenheden van een onvolledig register zijn, maar we zullen het meestal de steekproef noemen. Het complement van de steekproef in de populatie, dus alle eenheden waarvoor de doelvariabele niet bekend is, geven we aan met $r=U\backslash s$. Deze bestaat uit *N-n* eenheden.

Daarnaast veronderstellen we de aanwezigheid van een vector van hulpvariabelen *x* die bekend zijn voor de gehele populatie *U*. De vector *x* heeft dimensie *p*.

Steekproefgemiddelden worden genoteerd als $\bar{y},\bar{x}$ en populatietotalen als t_y,t_x. Dus $t_y=\sum_{i\in U} y_i$ en $\bar{y}=(1/n)\sum_{i\in s} y_i$, enzovoort. Populatiegemiddelden worden verkregen door t_y,t_x door *N* te delen, $\theta_y=t_y/N$ en $\theta_x=t_x/N$, waarbij de populatieomvang *N* bekend verondersteld wordt.

Voor *m* deelpopulaties of domeinen gebruiken we een index $d = 1, \ldots, m$. Zo is bijvoorbeeld $t_{y;d}=\sum_{i\in U_d} y_i$ het populatietotaal van *y* in domein *d*, en $\theta_{y;d}=t_{y;d}/N_d$ het populatiegemiddelde van y in domein *d*, met N_d de populatieomvang van domein *d*. Steekproefgemiddelden worden genoteerd als $\bar{y}_d,\bar{x}_d$; dit zijn gemiddelden over de steekproef s_d van omvang n_d die binnen deelpopulatie *d* valt. De overige, niet waargenomen, N_d-n_d eenheden in deelpopulatie *d* worden aangeduid met r_d.

2. Synthetische schatters

2.1 Korte beschrijving

Het lineaire regressiemodel voor de doelvariabele *y* gegeven hulpvariabelen *x* is

$$y_i = \beta^T x_i + \varepsilon_i \ , \tag{2.1.1}$$

met β een *p*-vector van regressiecoëfficiënten en ε_i een normaal verdeelde foutterm, onafhankelijk voor *i=1,...,N*.

Het model wordt gefit aan de hand van de steekproefdata *(y, x)* voor de eenheden in *s*. Vervolgens worden de eenheden in *r=U\s* bijgeschat (voorspeld) op basis van het gefitte model. De schatting voor het populatietotaal van *y* wordt dan

$$\hat{t}_y = \sum_{i \in s} y_i + \sum_{i \in r} \hat{y}_i = n\bar{y} + \hat{\beta}^T (t_x - n\bar{x}) \ , \tag{2.1.2}$$

met $\hat{\beta}$ de vector van geschatte regressiecoëfficiënten. Dus in (2.1.2) spreekt de waargenomen steekproef voor zich en de rest wordt bijgeschat volgens het gefitte model. Merk op dat de populatietotalen t_x bekend moeten zijn, naast de steekproefgemiddelden $\bar{x}$. Door de lineaire structuur van het model hoeven de individuele waarden van *x* voor eenheden in *r=U\s* niet voorhanden te zijn.

Andere populatiegrootheden kunnen op een vergelijkbare manier worden geschat. Totalen voor deelpopulaties, bijvoorbeeld, worden geschat volgens

$$\hat{t}_{y;d} = \sum_{i \in s_d} y_i + \sum_{i \in r_d} \hat{y}_i = n_d \bar{y}_d + \hat{\beta}^T (t_{x;d} - n_d \bar{x}_d) \, , \tag{2.1.3}$$

met $t_{x;d}$ de populatietotalen van *x* in deelpopulatie *d*.

In paragraaf 2.3 volgt een meer uitgebreide beschrijving. Aanvullende informatie over modelmatig schatten bij steekproeven uit eindige populaties is te vinden in Rubin (1987), Ghosh en Meeden (1997), Vaillant e.a. (2000) en Rao (2003).

2.2 Toepasbaarheid

Het lineaire regressiemodel en de daaruit voortvloeiende synthetische schatters zijn breed toepasbaar. In het bijzonder kunnen deze modelmatige schattingstechnieken gebruikt worden ter vervanging van design-gebaseerde methoden om populatietotalen te schatten wanneer er geen sprake is van een bekend kanssteekproefontwerp. Dit is het geval bij schatten uit onvolledige registers, waarvan het BTW omzetregister een voorbeeld is. De onvolledigheid van de data uit dit register wordt voor een deel veroorzaakt door de noodzaak om tijdig te publiceren wanneer nog niet alle aangiften binnen zijn, en ook door koppeling met het Algemeen Bedrijvenregister (ABR).

Daarnaast kunnen de synthetische schatters gebruikt worden voor het schatten van totalen of gemiddelden van deelpopulaties waarbinnen de steekproefomvangen te klein zijn om per deelpopulatie directe schatters te gebruiken. In die situatie spreken we van synthetische schatters zolang er *geen* domeinspecifieke hulpvariabelen in de vector x worden meegenomen. De synthetische schatters zijn daarmee een eenvoudig soort kleine-domeinschatters.

Het lineaire regressiemodel veronderstelt een kwantitatieve doelvariabele y. Soms kan een transformatie van de doelvariabele worden uitgevoerd om het model beter te doen passen. Voor een doelvariabele die altijd positief is, kan bijvoorbeeld een logaritmische transformatie nuttig zijn. De getransformeerde data worden dan gebruikt om het model te fitten. Het voorspellen van populatietotalen wordt in dat geval iets ingewikkelder omdat de gefitte waarden teruggetransformeerd moeten worden alvorens ze op te tellen.

Voor categoriale data, met per categorie een 0/1 variabele y, wordt soms een logistisch regressiemodel gebruikt. Maar ook in deze gevallen kan het lineaire regressiemodel interessant zijn. Een voordeel van het lineaire regressiemodel is dat het eenvoudiger te fitten is. Zolang de gefitte waarden $\hat{y}_i = \hat{\beta}^T x_i$ binnen het interval [0,1] blijven, op mogelijk enkele uitzonderingen na, lijkt het lineaire regressiemodel een redelijke keus.

Bij het gebruik van synthetische schatters voor populatietotalen of gemiddelden is het belangrijk om zoveel mogelijk rekening te houden met eventuele selectie-effecten. Selectie-effecten zijn effecten die systematische verschillen in de doelvariabele tussen de steekproef en de rest van de populatie veroorzaken. Hierdoor wordt het model (2.1.1), dat voor de gehele populatie geponeerd wordt, minder bruikbaar om het niet waargenomen deel van de populatie te voorspellen; er kunnen vertekeningen ontstaan.

Om selectie-effecten te verminderen is het belangrijk om het model uit te breiden met hulpvariabelen die deze selectie-effecten zo goed mogelijk verklaren. Dit kan onderzocht worden door de steekproefgemiddelden van de hulpvariabelen af te zetten tegen de populatiegemiddelden. Hulpvariabelen die grote verschillen laten zien zouden in het model opgenomen moeten worden, tenzij duidelijk is dat ze niet samenhangen met de doelvariabele y.

Bij onvolledige registers kunnen selectie-effecten optreden door de registratieprocedures, of door het koppelen aan andere bronnen. Deze effecten zijn niet altijd gemakkelijk te achterhalen, maar het is belangrijk om de mogelijk met y samenhangende hulpvariabelen te onderzoeken op eventuele systematische verschillen tussen geregistreerde en ontbrekende eenheden.

Ook bij kanssteekproeven ontstaan bijna altijd selectie-effecten ten gevolge van non-respons. Bij modelmatig schatten geldt hetzelfde als bij het wegen van steekproefdata, namelijk dat om vertekening te voorkomen getracht moet worden hulpvariabelen in het model op te nemen die zowel verklarend zijn voor de non-respons als voor de doelvariabele.

Geplande selectiviteit ontstaat door een steekproefontwerp met ongelijke insluitkansen. De design-gebaseerde methodologie corrigeert hiervoor door te wegen met de inverse insluitkansen. Bij modelmatige methoden wordt in die situatie de insluitkansvariabele zelf als hulpvariabele op een passende manier meegenomen in het regressiemodel.

Bij het gebruik van synthetische schatters voor kleine domeinen is het erg belangrijk dat er goede hulpinformatie beschikbaar is. Als de gebruikte hulpinformatie nauwelijks voorspellend is voor de doelvariabele dan worden de schattingen van domeingemiddelden te veel naar het algehele steekproefgemiddelde getrokken. Als dat het geval is dan zijn de specifieke kleine-domeinmethoden gebaseerd op modellen met random domeineffecten, zoals beschreven in het volgende deelthema, geschikter.

2.3 Uitgebreide beschrijving

2.3.1 Het lineaire regressiemodel

Het lineaire regressiemodel is gegeven in vergelijking (2.1.1). De fouttermen ε_i worden onafhankelijk verondersteld, en normaal verdeeld volgens

$$\varepsilon_i \sim N(0, v_i\sigma^2)\,. \tag{2.3.1}$$

De hulpvariabele *v* wordt ook wel variantiestructuur genoemd. De modelvarianties $v_i>0$ geven het model meer flexibiliteit. Zo is het bekend dat bij veel variabelen in bedrijfsstatistieken er sprake is van heteroscedasticiteit, waarbij de spreiding rond de lineaire voorspeller $\beta^T x_i$ toeneemt met de omvang van de bedrijven. Voor v_i kan dan een bepaalde positieve macht van de omvang (aantal werknemers of een andere maat) worden genomen, zie ook Hedlin et al. (2001). De waarden v_i moeten bekend zijn voor de steekproefeenheden, en om varianties te kunnen schatten moet bovendien het populatietotaal t_v bekend zijn. Als er geen informatie is die er op duidt dat de modelvarianties verschillend zijn dan nemen we $v_i = 1$ voor alle eenheden, en daarmee $t_v = N$.

2.3.2 Fitten van het model

De standaardschatting voor de vector van regressiecoëfficiënten is

$$\hat{\beta} = \left(\sum_{i\in s} x_i x_i^T / v_i \right)^{-1} \sum_{i\in s} x_i y_i / v_i\,. \tag{2.3.2}$$

Onder de veronderstelling (2.3.1) is deze schatting optimaal, in de zin dat de verwachte kwadratische fout (onder het model) wordt geminimaliseerd. Merk op dat $\hat{\beta}$ niet afhangt van de variantieparameter σ^2. Deze is wel van belang voor de

variantieschattingen die in paragraaf 2.3.5 worden besproken. Een schatting voor σ^2 is

$$\hat{\sigma}^2 = \frac{1}{n-p}\sum_{i\in s}(y_i - \hat{\beta}^T x_i)^2 / v_i, \tag{2.3.3}$$

met *p* de dimensie van de vector *x*.

2.3.3 Modelgebaseerde schattingen voor populatietotalen

De schatting voor het populatietotaal t_y op basis van het gefitte model en de populatietotalen van de hulpvariabelen is

$$\hat{t}_y = n\bar{y} + \hat{\beta}^T(t_x - n\bar{x}), \tag{2.3.4}$$

met $\hat{\beta}$ zoals in (2.3.2). Soms wordt met synthetische schatter alleen de term $\hat{\beta}^T t_x$ bedoeld. Voor kleine steekproeffracties *n/N* is dit bijna hetzelfde als (2.3.4). Onder de conditie dat de variantiestructuur *v* ook in de vector van hulpvariabelen *x* is opgenomen kan aangetoond worden dat $\bar{y} - \hat{\beta}^T \bar{x} = 0$ en in dat geval is (2.3.4) exact gelijk aan $\hat{\beta}^T t_x$.

2.3.4 Synthetische schatters voor deelpopulaties

De schattingen voor deelpopulatietotalen $t_{y;d}$ gebaseerd op het gefitte model zijn gegeven in (2.1.3) met $\hat{\beta}$ zoals in (2.3.2). Bij kleine steekproeffracties n_d / N_d worden de schattingen goed benaderd door $\hat{\beta}^T t_{x;d}$.

2.3.5 Modelgebaseerde variantieschattingen

De onder het model verwachte variantie van de fout $t_y - \hat{t}_y$ is

$$Var(t_y - \hat{t}_y) = \left((t_x - n\bar{x})^T \left(\sum_{i\in s} x_i x_i^T / v_i \right)^{-1} (t_x - n\bar{x}) + t_v - n\bar{v} \right) \sigma^2, \tag{2.3.5}$$

met t_v en $\bar{v}$ het populatietotaal respectievelijk steekproefgemiddelde van de variabele *v*. Een variantieschatting van $\hat{t}_y$ krijgen we door de schatting (2.3.3) voor σ^2 in te vullen. De eerste term van (2.3.5) is de variantie ten gevolge van onzekerheid in de geschatte coëfficiënten $\hat{\beta}$, en de tweede term, $(t_v - n\bar{v})\sigma^2 = \sum_{i\in r} v_i \sigma^2$, is de voorspelvariantie die er altijd is, ook al zou β bekend zijn.

De foutvarianties van de domeintotaalschattingen zijn

$$Var(t_{y;d} - \hat{t}_{y;d}) = \left((t_{x;d} - n_d \bar{x}_d)^T \left(\sum_{i \in s} x_i x_i^T / v_i \right)^{-1} (t_{x;d} - n_d \bar{x}_d) + t_{v;d} - n_d \bar{v}_d \right) \sigma^2,$$

met $t_{v;d}$ en $n_d \bar{v}_d$ het populatie- respectievelijk steekproeftotaal van de variabele *v* in domein *d*. Schattingen worden verkregen door $\hat{\sigma}^2$ voor σ^2 in te vullen. Als de verschillen tussen domeinen niet voldoende verklaard worden door de verschillen in hulpvariabelen dan zullen deze modelgebaseerde variantieschattingen al snel te laag uitvallen. Een model met random effecten zou in dat geval beter zijn, maar in de literatuur worden ook alternatieve, design-gebaseerde, MSE-schatters beschreven, die ook een term bevatten voor de vertekening van de synthetische domeinschatters, zie Rao (2003), sectie 4.2.4.

2.3.6 Software

De besproken modelmatige schattingen en variantieschattingen kunnen relatief eenvoudig worden berekend in standaard statistische softwarepaketten. In een paket met uitgebreide regressie- en matrixfaciliteiten zoals R/S-Plus is het benodigde programmeerwerk beperkt.

Voor een grote klasse van modellen komen de synthetische schatters voor populatietotalen (maar niet voor domeintotalen) overeen met algemene regressieschatters, zie paragraaf 2.5. Deze kunnen met het weegprogramma Bascula berekend worden, zie Nieuwenbroek en Boonstra (2002) voor een handleiding. Bascula biedt echter niet de mogelijkheid om een variantiestructuur *v* te kiezen.

2.3.7 Modelkeuze

We beperken ons hier tot lineaire regressiemodellen, dus de keuze voor een model komt vooral neer op de keuze van een vector met geschikte hulpvariabelen *x*, en eventueel een variantiestructuur *v*. Daarnaast is er de mogelijkheid om de doelvariabele te transformeren naar een variabele die beter beschreven wordt door het lineaire regressiemodel. We gaan hier verder in op de keuze van hulpvariabelen *x*.

De vector van hulpvariabelen *x* moet bestaan uit variabelen die iets zeggen over de doelvariabele *y*. Hoe beter een hulpvariabele correleert met *y*, des te belangrijker is het om deze variabele mee te nemen in de vector *x*. Daarnaast dient onderzocht te worden welke hulpvariabelen in de steekproef een andere verdeling hebben dan in de rest van de populatie; de steekproef is niet representatief voor deze variabelen. Ook als deze variabelen niet heel sterk met *y* correleren, is het toch beter om ze mee te nemen in de vector *x*, om mogelijke vertekening terug te dringen. Niet alle selectie-effecten kunnen op deze manier worden gecorrigeerd. Het wel/niet ontbreken van eenheden in de steekproef *s* kan, ondanks het toevoegen van vele verklarende hulpvariabelen, nog steeds gerelateerd zijn aan de doelvariabele zelf. Voor dat deel van de selectie-effecten kan niet gecorrigeerd worden met het hier beschreven regressiemodel.

Bij het uitbreiden van de vector van hulpvariabelen is het van belang hoeveel een nieuwe hulpvariabele nog *toevoegt* qua voorspelkracht voor *y* en/of het steekproefselectiemechanisme. Het aantal hulpvariabelen mag niet te groot worden ten opzichte van de steekproefomvang *n*, anders bestaat het gevaar op "overfitten" waardoor het model zijn voorspelkracht verliest.

Het opnemen van de variantiestructuur *v* als een van de componenten van de vector *x* van hulpvariabelen zorgt niet alleen voor de vereenvoudiging van sommige uitdrukkingen, maar ook voor een bepaalde robuustheid tegen sommige soorten van misspecificatie. Zo is het bij homoscedastische modellen ($v_i = 1$ voor alle eenheden) gebruikelijk om een intercept in het model te hebben, corresponderend met een constante component in de vector *x*.

Om een redelijke variantiestructuur *v* te selecteren kan de samenhang van residuen gebaseerd op het model met $v_i = 1$ met relevante hulpvariabelen bestudeerd worden. In de literatuur worden verschillende tests voor heteroscedasticiteit besproken, zie bijvoorbeeld Greene (1997). Vooral variantieschattingen zoals (2.3.4) zijn gevoelig voor misspecificatie van *v*. Verschillende meer robuuste variantieschatters worden besproken in Valliant et al. (2000), Hoofdstuk 5.

2.4 Voorbeeld

Aan de hand van een mogelijke toepassing op de Kortetermijnstatistieken (KS) hopen we de bovenstaande beschrijving te verduidelijken.

Bij de KS wordt de omzetontwikkeling van maand op maand geschat. Nu nog wordt hier een panelsteekproef voor gebruikt, maar het is de bedoeling dat in de nabije toekomst wordt overgegaan op het gebruik van de BTW-omzetregistratie van de belastingdienst. De BTW-omzet vervangt dan de primair waargenomen omzet. We gaan hier niet in op mogelijke verschillen tussen BTW-omzet en werkelijke omzet, maar nemen de BTW-omzet als uitgangspunt om omzetontwikkelingen te schatten. Voor het gemak negeren we hier ook de problematiek rond jaar- kwartaal- en maandaangevers.

Niet voor alle bedrijfseenheden van het ABR zijn BTW-data over de relevante periode beschikbaar. Die onvolledigheid heeft een aantal oorzaken, o.a. het niet tijdig beschikbaar zijn van sommige BTW-aangiften en miskoppelingen met het ABR. Er is echter geen bekend kanssteekproefmechanisme dat zegt wat de kansen zijn dat voor bedrijfseenheden de BTW-data beschikbaar zijn. Design-gebaseerde schattingen zijn dus niet mogelijk. De modelmatige aanpak is niet afhankelijk van een kanssteekproefmechanisme maar probeert de redenen voor het ontbreken van data wel mee te modelleren als deze samenhangen met de doelvariabelen.

We gaan uit van een publicatiecel, d.w.z. een deelpopulatie waarover gepubliceerd moet worden, op twee tijdstippen *t* en *t-1*. De eenheden in de populatie geven we aan met U_t en U_{t-1} (deelpopulaties van het ABR op tijdstippen *t* en *t-1*) en de BTW-

omzetvariabele met *y* op tijdstip *t* en *z* op tijdstip *t-1*. De omzetontwikkeling is gedefinieerd als

$$O_t = \frac{t_y}{t_z} = \frac{\sum_{i \in U_t} y_i}{\sum_{i \in U_{t-1}} z_i} \ . \qquad (2.4.1)$$

BTW-omzetdata uit perioden *t-1* en *t* zijn beschikbaar voor bedrijfseenheden $s_{t-1} \subset U_{t-1}$ en $s_t \subset U_t$, en ontbreken voor de overige eenheden.

Beide populatietotalen in (2.4.1) kunnen modelmatig worden geschat. Daarbij kan hulpinformatie uit het ABR worden gebruikt. Een belangrijke hulpvariabele is het aantal werkzame personen (WP). Een eenvoudig model zou dan zijn

$$y_i = \beta_1 + \beta_2 \ \mathrm{WP_i} + \varepsilon_i \qquad \text{met} \qquad \varepsilon_i \sim N(0, \mathrm{WP_i}\sigma^2) \,,$$

en hulpvariabelen $x_i = (1, \mathrm{WP_i})^T$. Het verband tussen omzet en aantal werkzame personen hoeft niet precies lineair te zijn, en er kan geëxperimenteerd worden met verschillende machten van WP, zowel in de vector *x* als in de variantiestructuur.

Andere mogelijke hulpvariabelen die toegevoegd kunnen worden aan de vector *x* zijn categoriale variabelen zoals een indeling naar bedrijfstakken of rechtsvorm. De ABR-variabele grootteklasse (GK) is afgeleid van WP en voegt niets aan het model toe, *mits* de afhankelijkheid van WP goed gespecificeerd is. Dit laatste zal vaak niet zo zijn en dan kan het toch nuttig zijn om GK toe te voegen.

Na het fitten van het model (afzonderlijk voor *t* en *t*-1) wordt (2.4.1) geschat volgens

$$\hat{O}_t = \frac{\hat{t}_y}{\hat{t}_z} = \frac{n_t \bar{y} + \hat{\beta}_t^T (t_{x;t} - n_t \bar{x}_t)}{n_{t-1} \bar{z} + \hat{\beta}_{t-1}^T (t_{x;t-1} - n_{t-1} \bar{x}_{t-1})} . \qquad (2.4.2)$$

Op deze manier worden populatietotalen in teller en noemer elk geschat op basis van een eigen model, met normaal gesproken overeenkomstige hulpvariabelen. Deze cross-sectionele schattingen kunnen eventueel verder verbeterd worden met behulp van multivariate of tijdreeksmodellen. Varianties voor teller en noemer van (2.4.2) kunnen worden geschat met behulp van (2.3.5). Een variantieschatting voor de ratio kan worden verkregen met behulp van linearisatie.

Als er in een publicatiecel te weinig data aanwezig is om het model goed te kunnen fitten, dan kunnen meerdere (liefst vergelijkbare) cellen worden samengevoegd. De schattingen voor de afzonderlijke publicatiecellen worden dan berekend met formule (2.1.3).

2.5 Eigenschappen

Hoewel dit onderdeel van de methodenreeks niet in eerste instantie geschreven is voor de situatie van een bekend steekproefontwerp, is het toch instructief om in die situatie het verband van de synthetische schatter met de design-gebaseerde algemene regressieschatter voor een populatietotaal te schetsen. Zoals eerder opgemerkt, komt

een modelmatige schatter gebaseerd op een lineair regressiemodel soms overeen met een design-gebaseerde schatter. De algemene regressieschatter (GREG) voor het populatietotaal t_y onder een steekproef getrokken met insluitkansen π_i is (Särndal e.a., 1992)

$$\hat{t}_y^{GREG} = \hat{t}_y^{HT} + \hat{\gamma}^T (t_x - \hat{t}_x^{HT}) , \qquad (2.5.1)$$

$$\hat{\gamma} = \left(\sum_{i \in s} x_i x_i^T / \pi_i \right)^{-1} \sum_{i \in s} x_i y_i / \pi_i ,$$

met $\hat{t}_y^{HT} = \sum_{i \in s} y_i / \pi_i$ en $\hat{t}_x^{HT} = \sum_{i \in s} x_i / \pi_i$ de Horvitz-Thompson schatters voor de populatietotalen van *y* en *x*. Dit wordt ook wel een model-assisted schatter genoemd omdat impliciet gebruik wordt gemaakt van een model, maar wel zodanig dat de design-zuiverheid bij benadering gehandhaafd blijft. Een voldoende voorwaarde voor exacte gelijkheid van de synthetische schatter (2.3.4) en de GREG (2.5.1) is dat (1) $v_i = \pi_i$ (voor alle *i*, tot op een constante factor) zodat de coëfficiënten $\hat{\beta}$ en $\hat{\gamma}$ hetzelfde zijn, en (2) $1 - \pi_i = c^T x_i$ voor een constante *p*-vector *c*, voor alle *i*. De tweede conditie zegt dat de variabele $1 - \pi_i$ in de vector van hulpvariabelen *x* moet zitten. Dit is zeker het geval als zowel de constante als de insluitkans in *x* zitten. Om de gelijkheid aan te tonen vermenigvuldigen we eerst beide kanten van de conditie (2) met $x_i^T \hat{\gamma} / \pi_i$ en sommeren over $i \in s$. Dit geeft

$$(\hat{t}_x^{HT} - n\bar{x})^T \hat{\gamma} = \sum_{i \in s} c^T x_i y_i / \pi_i = \hat{t}_y^{HT} - n\bar{y} ,$$

waarbij voor de laatste gelijkheid conditie (2) nog een keer is toegepast. Samen met $\hat{\beta} = \hat{\gamma}$ geeft dit $\hat{t}_y = \hat{t}_y^{GREG}$. Zie Boonstra (2005) voor deze en andere relaties tussen design- en modelgebaseerde schatters, en voor verdere verwijzingen naar de literatuur.

Een van de implicaties van deze overeenkomst tussen design- en modelgebaseerde schatters is dat we er altijd voor kunnen zorgen dat een synthetische schatter voor een populatietotaal bij benadering design-zuiver is onder een bekend kanssteekproefontwerp: kies de variantiestructuur $v_i = \pi_i$ en zorg dat naast de constante ook de insluitkansvariabele (=variantiestructuur) in de vector *x* van hulpvariabelen zit. Andersom is het nuttig te weten welke expliciete modelveronderstellingen ten grondslag liggen aan design-gebaseerde schatters. Een model dat de werkelijke doelvariabele slecht beschrijft zal niet leiden tot een grote design-vertekening maar mogelijk wel tot een grote design-variantie. Gelukkig kan deze laatste in toom gehouden worden door de steekproefomvang voldoende groot te kiezen.

2.6 Kwaliteitsindicatoren

Naast het beoordelen van de modelgebaseerde variantieschattingen, die een maat zijn voor de nauwkeurigheid van de schattingen gegeven het model (en dus zelf ook gevoelig zijn voor de modelkeuze), zou ook het model zelf aan verschillende toetsen onderworpen moeten worden.

Een plot van de residuen $e_i = y_i - \hat{\beta}^T x_i$ tegen de gefitte waarden $\hat{y}_i = \hat{\beta}^T x_i$ verschaft vaak inzicht in eventuele misspecificatie van het model, en mogelijke verbeteringen. Voor een goed model zullen de residuen ongeveer normaal verdeeld zijn rond 0 en zal er verder geen afhankelijkheid zijn met de gefitte waarden. Goodall (1983) geeft een uitgebreide beschrijving van modeldiagnose aan de hand van residuen.

Een andere strategie is het met elkaar vergelijken van verschillende modellen aan de hand van bepaalde modelselectiematen, om uiteindelijk het beste model te kiezen. Deze modelselectiematen wegen modelfit af tegen complexiteit van het model. Modellen met relatief veel coëfficiënten hebben weliswaar kleine residuen, maar de voorspelkracht kan door een teveel aan coëfficiënten sterk afnemen. De voorspelkracht van een model kan ook op een meer directe manier worden onderzocht door het model te fitten op een deel van de data en vervolgens de fouten in de voorspellingen van de overige data te bepalen. Een gevoeligheidsanalyse waarbij een model op verschillende redelijke manieren wordt gewijzigd is ook een manier om inzicht te verkrijgen in de kwaliteit van de modelschattingen.

Gangbare maten voor variabeleselectie in regressiemodellen zijn AIC en BIC, terwijl cross-validatie een directe maat is voor voorspelkracht van een model, zie bijvoorbeeld Hastie e.a. (2003). Het onderwerp modelselectie en evaluatie is echter zeer breed en redelijk complex. Bovendien heeft dit onderwerp binnen de steekproeftheorie een extra dimensie, namelijk dat van selectie-effecten, zoals kort beschreven in paragraaf 2.3.7. We zullen er hier niet verder op ingaan.

3. Kleine-domeinschatters

3.1 Korte beschrijving

Bij een steekproef spreken we van kleine domeinen wanneer we het hebben over deelgroepen van de populatie met een te kleine steekproefomvang om betrouwbare directe schattingen te kunnen maken. Met modelmatige schattingsmethoden wordt informatie uit andere domeinen aangewend om de schatting voor elk klein domein te verbeteren. Er moet een modelveronderstelling gedaan worden, en een schattingsmethode gekozen om schattingen onder het model te maken.

In dit deelthema beschrijven we de EBLUP schatter (Empirical Best Linear Unbiased Predictor), uitgaande van een linear gemengd model waarin hulpinformatie op domeinniveau kan worden meegenomen. Dit model is gekend als een Fay-Herriot (FH) model (Fay and Herriot, 1979), en is gedefinieerd als

$$\begin{aligned} \hat{\theta}_{y;d} &= \theta_{y;d} + \varepsilon_d \\ \theta_{y;d} &= \theta_{x;d}^T \beta + v_d \end{aligned} \qquad (3.1.1)$$

waarbij $\varepsilon_d \sim N(0, \psi_d)$ en $v_d \sim N(0, \sigma_v^2)$ voor d = 1, …, m en m het aantal domeinen. Het populatiegemiddelde van y voor domein d is $\theta_{y;d}$. $\hat{\theta}_{y;d}$ is een directe, design-gebaseerde schatter voor $\theta_{y;d}$ met fout ε_d. Directe schattingen zijn alleen gebaseerd op informatie van het domein zelf. We veronderstellen dat de schattingen $\hat{\theta}_{y;d}$ niet vertekend zijn, met variantieschattingen ψ_d. De vector $\theta_{x;d}$ bestaat uit domeinspecifieke hulpvariabelen. De random effecten v_d hebben variantie σ_v^2 en zijn onafhankelijk van ε_d.

Een lineair gemengd model onderscheidt zich van de lineaire regressiemodellen zoals ze in het deelthema synthetische schatters zijn gebruikt door de aanwezigheid van zogenaamde random effecten, in dit geval random domeineffecten. De variaties in de domeinschattingen die niet verklaard worden door de hulpvariabelen of de steekproeffouten worden gevat in de random effecten van model (3.1.1). De vector β zal normaal gesproken een intercept μ bevatten. De effecten $\mu + v_i$ vormen dan een set van domein-intercepts, met een gemeenschappelijke onderliggende verdeling $N(\mu, \sigma^2)$. Dit geeft een alternatieve, hiërarchische of multi-level formulering van het model. Gelman en Hill (2006) en Longford (2005) bevatten uitgebreide beschrijvingen van hiërarchische of multi-level modellen.

In Boonstra e.a. (2007) worden een aantal alternatieve modellen voor kleine-domeinschattingen onderzocht, onder andere modellen geformuleerd op eenheidsniveau versus domeinniveau, en lineaire versus logistische modellen. Op basis van onderzoek en simulatiestudies (Boonstra e.a., 2007) is in eerste instantie

voor de benadering met een lineair gemend model op domeinniveau gekozen omdat die het evenwicht bewaart tussen eenvoud en nauwkeurigheid. Deze methode is ook geïmplementeerd in een prototype software tool (Buelens, 2007).

De in dit hoofdstuk beschreven methoden en formules worden in detail gegeven en afgeleid in Rao (2003).

3.2 Toepasbaarheid

Het doel is om schattingen voor kleine domeinen te maken. Per definitie zijn directe schattingen voor kleine domeinen onbetrouwbaar. De modelmatige methode is dan ook interessant om nauwkeurigere schattingen voor kleine domeinen te berekenen. Dit is meestal het geval bij detaillering, wanneer schattingen voor kleine deelgroepen van de populatie moeten berekend worden, en wanneer er bij het steekproefontwerp hiermee geen rekening was gehouden.

De aanwezigheid van random effecten in de modelvergelijkingen (3.1.1) zorgt ervoor dat domeinen van elkaar kunnen verschillen onder het model, nog afgezien van verschillen veroorzaakt door verschillen in hulpvariabelen. Om goede schattingen te kunnen maken op basis van het model is het wel belangrijk dat goede verklarende variabelen $\theta_{x;d}$ beschikbaar zijn als hulpinformatie. Als de hulpinformatie niet goed correleert met de doelvariabele zullen de random effecten aan invloed winnen, en heeft het model minder voorspelkracht. Impliciet wordt er dus verwacht dat de domeinen gelijkaardig zijn. Wanneer domeinen onderling zeer sterk verschillen, en deze verschillen niet gevat worden door de hulpinformatie, leidt dit tot grotere random effecten. Het selecteren van goede hulpvariabelen komt neer op het kiezen van een geschikt model; dit proces staat bekend onder de naam modelselectie. Over dit aspect wordt kort iets gezegd in paragrafen 2.6 en 3.6.

De hulpinformatie hoeft voor model (3.1.1) alleen op domeinniveau beschikbaar te zijn. Als de hulpinformatie op eenheidsniveau beschikbaar is dan kunnen steekproefgemiddelden $\bar{x}_d$ i.p.v. populatiegemiddelden als hulpvariabelen in (3.1.1) gebruikt worden. Op die manier kan beter voor non-respons worden gecorrigeerd, zie Boonstra e.a. (2007).

3.3 Uitgebreide beschrijving

3.3.1 Lineair gemengd model op domeinniveau

We gebruiken een FH-model, een lineair gemengd model op domeinniveau, zoals gedefinieerd in (3.1.1). De uitdrukking (3.1.1) kan ook geschreven worden als

$$\hat{\theta}_{y;d} = \theta_{x;d}^T \beta + v_d + \varepsilon_d . \qquad (3.3.1)$$

Bij het fitten van dit model maken we gebruik van directe, design-gebaseerde schattingen $\hat{\theta}_{y;d}$ en van variantieschattingen ψ_d. De schatter $\hat{\theta}_{y;d}$ kan bijvoorbeeld een Horvitz-Thompson of een regressieschatter zijn.

Omdat we te maken hebben met kleine domeinen kunnen de variantieschattingen ψ_d instabiel zijn. Een oplossing is het poolen van deze schattingen. Als $\hat{\theta}_{y;d}$ het steekproefgemiddelde in domein *d* is, dan is de bijbehorende schatting van de design variantie, uitgaande van een enkelvoudig aselecte steekproef,

$$\psi_d = \frac{1}{n_d}(1-\frac{n_d}{N_d})s_d^2,$$

met steekproefvariantie

$$s_d^2 = \frac{1}{n_d - 1}\sum_{i\in S_d}(y_i - \hat{\theta}_{y;d})^2 ,$$

Bij poolen berekenen we de steekproefvariantie voor meerdere domeinen samen, of zelfs alle domeinen samen. In dit laatste geval berekenen we een gepoolde steekproefvariantie,

$$s_{pooled}^2 = \frac{1}{n-m}\sum_{d=1}^{m}(n_d - 1)s_d^2 ,$$

en gebruiken deze in plaats van de domeinspecifieke varianties s_d^2 in bovenstaande uitdrukking voor ψ_d.

De modelvariantie σ_v^2 wordt geschat met de Fay-Herriot-momentenschatter (Rao, 2003). Als uitgangspunt van deze methode wordt opgemerkt dat

$$E\left(\sum_d \frac{(\hat{\theta}_{y;d} - \theta_{x;d}^T\tilde{\boldsymbol{\beta}})^2}{\psi_d + \sigma_v^2}\right) = E\left(h(\sigma_v^2)\right) = m - p \qquad (3.3.2)$$

met *m* het aantal kleine domeinen, *p* de dimensie van de vector van hulpvariabelen $\theta_{x;d}$, en

$$\tilde{\beta} = \tilde{\beta}(\sigma_v^2) = \left(\sum_d \theta_{x;d}\theta_{x;d}^T/(\psi_d + \sigma_v^2)\right)^{-1}\left(\sum_d \theta_{x;d}\hat{\theta}_{y;d}/(\psi_d + \sigma_v^2)\right). \qquad (3.3.3)$$

De schatting $\hat{\sigma}_v^2$ wordt verkregen door het iteratief oplossen van

$$h(\sigma_v^2) = m - p . \qquad (3.3.4)$$

Hiervoor nemen we als startwaarde $\sigma_v^{2(a=0)} = 0$ en berekenen

$$\sigma_v^{2(a+1)} = \sigma_v^{2(a)} + \frac{1}{h'_*(\sigma_v^{2(a)})}\left(m - p - h(\sigma_v^{2(a)})\right) \qquad (3.3.5)$$

met $h'_*(\sigma_v^2) = -\sum_d \frac{(\hat{\theta}_{y;d} - \theta_{x;d}^T\tilde{\beta})^2}{(\psi_d + \sigma_v^2)^2}$ een benadering van de afgeleide van $h(\sigma_v^2)$.

Dit iteratieve proces convergeert snel, meestal in minder dan 10 iteraties. Indien geen positieve oplossing wordt gevonden, wordt $\hat{\sigma}_v^2 = 0$ genomen. In dit laatste geval zijn er dus geen random effecten en zullen we een synthetische schatter op domeinniveau krijgen.

De vertekening en variantie van deze schatting worden gegeven door respectievelijk B($\hat{\sigma}_v^2$) en V($\hat{\sigma}_v^2$), met

$$B(\hat{\sigma}_v^2) = \frac{2\left(m\sum_d (\psi_d + \hat{\sigma}_v^2)^{-2} - \left(\sum_d (\psi_d + \hat{\sigma}_v^2)^{-1} \right)^2 \right)}{\left(\sum_d (\psi_d + \hat{\sigma}_v^2)^{-1} \right)^3}, \tag{3.3.6}$$

$$V(\hat{\sigma}_v^2) = 2m\left(\sum_d (\psi_d + \hat{\sigma}_v^2)^{-1} \right)^{-2}. \tag{3.3.7}$$

3.3.2 Empirical Best Linear Unbiased Predictor (EBLUP)

De op model (3.3.1) gebaseerde Empirical Best Linear Unbiased Predictor (EBLUP) schatter wordt gegeven door

$$\hat{\theta}_{y;d}^{eblup} = \gamma_d \hat{\theta}_{y;d} + (1-\gamma_d)\theta_{x;d}^T \hat{\beta} \tag{3.3.8}$$

met

$$\hat{\beta} = \tilde{\beta}(\hat{\sigma}_v^2) = \left(\sum_d \gamma_d \theta_{x;d} \theta_{x;d}^T \right)^{-1} \left(\sum_d \gamma_d \theta_{x;d} \hat{\theta}_{y;d} \right) \tag{3.3.9}$$

en

$$\gamma_d = \frac{\hat{\sigma}_v^2}{\psi_d + \hat{\sigma}_v^2}. \tag{3.3.10}$$

De EBLUP-schatter (3.3.8) is een gewogen combinatie van een directe schatter $\hat{\theta}_{y;d}$ en een synthetische schatter $\theta_{x;d}^T \hat{\beta}$. De directe schatter krijgt een groot gewicht γ_d (3.3.10) indien de variantie ψ_d klein is. Met andere woorden, de EBLUP-schattingen zijn vooral gebaseerd op de directe schattingen wanneer die nauwkeurig zijn, en op de modelmatige schattingen in het andere geval.

De gemiddelde kwadratische fout of mean square error (MSE) van de EBLUP-schattingen (3.3.8) wordt geschat als

$$mse(\hat{\theta}_{y;d}^{eblup}) = g_{1d}(\hat{\sigma}_v^2) - B(\hat{\sigma}_v^2)(1-\gamma_d)^2 + g_{2d}(\hat{\sigma}_v^2) + 2g_{3d}(\hat{\sigma}_v^2), \tag{3.3.11}$$

met

$$g_{1d}(\hat{\sigma}_v^2) = \gamma_d \psi_d ,$$

$$g_{2d}(\hat{\sigma}_v^2) = (1-\gamma_d)^2 \theta_{x;d}^t \left(\sum_d \theta_{x;d} \theta_{x;d}^T / (\psi_d + \hat{\sigma}_v^2) \right)^{-1} \theta_{x;d} ,$$

$$g_{3d}(\hat{\sigma}_v^2) = {\psi_d}^2 (\psi_d + \hat{\sigma}_v^2)^{-3} V(\hat{\sigma}_v^2) .$$

De eerste term, g_{1d}, is de inherente voorspelvariantie die aanwezig is ook als β en σ_v^2 gekend zouden zijn; g_{2d} en g_{3d} zijn de bijdragen door de onzekerheid in respectievelijk β en σ_v^2.

3.3.3 Prototype software SmallAreaEstimator

Het in paragraaf 3.3.1 beschreven model en de in paragraaf 3.3.2 beschreven EBLUP-schatter zijn bij de afdeling Methodologie Heerlen geimplementeerd in een prototype software tool, *SmallAreaEstimator* (Buelens, 2007). Deze tool is een plugin voor SPSS en biedt de gebruiker een grafische user interface die toelaat om kleinedomeinschattingen te maken binnen de SPSS software omgeving.

3.4 Voorbeeld

Een voorbeeld uit de Enquête-Beroepsbevolking (EBB) is het schatten van jaarcijfers van de werkloosheid op gemeentelijk niveau. Er is vraag naar deze cijfers, maar de opzet van de bestaande EBB-steekproef laat niet toe om betrouwbare schattingen te maken op dit niveau. Voor vele gemeenten zijn er immers niet voldoende waarnemingen, en voor sommige gemeenten zelfs helemaal geen.

Als voorbeeld nemen we de CAPI steekproef van de EBB van 2005. Het gaat om 86.589 personen, en 454 gemeenten. We gebruiken de in paragraaf 3.3.3 genoemde software tool *SmallAreaEstimator (SAE)*. De directe schatter die in de huidige versie van de software is geïmplementeerd is het ongewogen steekproefgemiddelde. Met SAE berekenen we onder andere deze directe schattingen $\hat{\theta}_{y;d}$ en bijhorende varianties ψ_d. De variatiecoëfficiënt *vc* = $\sqrt{\psi_d} / \hat{\theta}_{y;d}$ kan gebruikt worden als een maat voor aanvaardbare nauwkeurigheid. Wanneer we een maximumwaarde voor de *vc* van 0,2 als criterium gebruiken, zijn de directe schattingen voor slechts 38 gemeenten voldoende nauwkeurig.

Als hulpinformatie voor de modelmatige schattingen gebruiken we per gemeente het aantal mensen ingeschreven bij het Centrum voor Werk en Inkomen (CWI), en de populatieomvangen uitgesplitst naar drie leeftijdsgroepen. Met behulp van SAE berekenen we de EBLUP schattingen en bijhorende MSEs. De EBLUP schattingen zijn voor 437 gemeenten voldoende nauwkeurig (ze hebben een *vc* kleiner dan 0,2).

Dit voorbeeld toont aan dat er veel winst aan nauwkeurigheid is te bereiken door gebruik te maken van de EBLUP schatter in plaats van de eenvoudige directe design-based schatter.

3.5 Eigenschappen

Modelmatige kleine-domeinschattingen hebben een *smoothing* effect. De verdeling van de schattingen zal een kleinere spreiding hebben dan de verdeling van de werkelijke waarden. Bijgevolg zullen hoge extreme waarden vaak onderschat en lage extreme waarden overschat worden. Het is dus best mogelijk dat kleine-domeinschattingen die in het algemeen als goed worden aangenomen, voor specifieke individuele domeinen toch helemaal niet goed zijn. Deze situatie ontstaat bijvoorbeeld bij atypische domeinen: domeinen die om welke reden dan ook essentieel verschillen van alle andere domeinen en waarbij die verschillen niet verklaard worden door de gebruikte hulpvariabelen.

De EBLUP schatter (3.3.8) is een combinatie van een directe en een synthetische schatter. Asymptotisch is de EBLUP schatter design zuiver omdat (3.3.8) voor grote n_d nadert tot de directe schatter, die design zuiver is.

Voor domeinen die goed vertegenwoordigd zijn in de steekproef zullen de directe schattingen nauwkeurig zijn, en de gewichten γ_d (3.3.10) groot. Met andere woorden, de EBLUP schatter is voor deze domeinen vooral, tot soms bijna helemaal, gebaseerd op de directe schatter. Het is dan ook belangrijk om een goede directe schatter te kiezen, ook omdat de directe schattingen gebruikt worden om het model te fitten. Het in paragraaf 3.3.3 beschreven prototype gebruikt voorlopig steeds het ongewogen steekproefgemiddelde als directe schatter, en biedt dus geen keuzemogelijkheden. Verwacht wordt dat volgende versies van dit prototype zulke keuzes wel zullen bieden, via het opnemen van gewichten. Op die manier zullen ook regressieschatters als directe schatters kunnen gebruikt worden in de software, door het aanbieden van bijvoorbeeld in Bascula berekende gewichten.

3.6 Kwaliteitsindicatoren

In eerste instantie kan gekeken worden naar de standaardfouten en variatiecoëfficiënten van de modelmatige schattingen in vergelijking met die van de directe schattingen. Het is echter zo dat de standaardfouten van de modelmatige schattingen zelf ook schattingen zijn gebaseerd op het model, en er dus voorzichtig mee omgegaan moet worden. Dankzij de random effecten in het model en de asymptotische design zuiverheid van de EBLUP schatter zijn ze echter wel redelijk robuust.

Er zijn een aantal andere maten en tests die kunnen aangeven of het gekozen model en de bijhorende schattingen plausibel zijn, zie paragraaf 2.6. Het onderzoek naar deze maten en tests voor modelselectie is nog lopende bij DMH.

4. Literatuur

Boonstra, H.J. (2005), *Model-based estimation of a finite population total: a Bayesian approach*. BPA-nr TMO-R&D-2005-08-16-HBTA, Centraal Bureau voor de Statistiek, Heerlen.

Boonstra, H.J., Buelens, B. en Smeets, M. (2007), *Estimation of municipal unemployment fractions – a simulation study comparing different small area estimators.* Onderzoeksrapport, BPA-nr DMK-DMH-2007-04-20-HBTA, Centraal Bureau voor de Statistiek, Heerlen.

Buelens, B. (2007), *Methodologie van de kleinedomeinschattingen in het prototype 'SmallAreaEstimator'*. Interne nota, BPA-nr DMH-2007-01-19-BBUS, Centraal Bureau voor de Statistiek, Heerlen.

Fay, R.E. en Herriot, R.A. (1979), Estimation of income for small places: an application of James-Stein procedures to census data. *Journal of the American Statistical Association* 74, 268-277.

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

Ghosh, M. en Meeden, G. (1997), *Bayesian Methods for Finite Population Sampling*. Chapman & Hall.

Goodall, C. (1983), Examining Residuals. In: *Understanding Robust and Exploratory Data Analysis*, D.C. Hoaglin, F. Mosteller en J.W. Tukey (editors), John Wiley & Sons.

Greene, W.H. (1997), *Econometric Analysis*. Prentice Hall.

Hastie, T., Tibshirani, R., en Friedman, J.H. (2003), *The Elements of Statistical Learning*. Springer-Verlag.

Hedlin, D., Falvey, H., Chambers, R., en Kokic, P. (2001), Does the Model Matter for GREG Estimation? A Business Survey Example. *Journal of Official Statistics* 17 (4), 527-544.

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

Nieuwenbroek, N. en H.J. Boonstra (2002), *Bascula 4.0 Reference Manual*. BPA-nr 279-02-TMO, Centraal Bureau voor de Statistiek, Heerlen.

Rao, J.N.K. (2003), *Small Area Estimation*. John Wiley & Sons.

Rubin, D.B. (1987), *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons.

Särndal, C.-E., Swensson, B., en Wretman, J. (1992), *Model Assisted Survey Sampling*. Springer-Verlag.

Valliant, R., Dorfman, A.H. en Royall, R.M. (2000), *Finite Population Sampling and Inference – A Prediction Approach*. John Wiley & Sons.