Representatieve uitbijters

10

Sabine Krieg, Marc Smeets

Statistische Methoden (10001)

Centraal Bureau voor de Statistiek

Den Haag/Heerlen, 2010

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
2008–2009	=	2008 tot en met 2009
2008/2009	=	het gemiddelde over de jaren 2008 tot en met 2009
2008/'09	=	oogstjaar, boekjaar, schooljaar enz., beginnend in 2008 en eindigend in 2009
2006/'07–2008/'09	=	oogstjaar, boekjaar enz., 2006/'07 tot en met 2008/'09

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

6016510001 X-37

Inhoudsopgave

1. Inleiding op het thema

1.1 Algemene beschrijving

1.1.1 Probleemstelling

In dit document worden schattingsmethoden beschreven die geschikt zijn voor databestanden met representatieve uitbijters. Een uitbijter is een extreme waarneming. In dit document worden alleen uitbijters in de steekproef beschouwd. In de inleiding houden we als voorbeeld de schatting voor een populatietotaal in het achterhoofd die gebaseerd is op een enkelvoudig aselecte steekproef. De schatters die later in dit document beschreven worden, zijn ook geschikt voor complexere steekproefontwerpen en andere populatieparameters. Het totaal op basis van een enkelvoudig aselecte steekproef $y_1,...,y_n$ met *n* elementen wordt geschat als

$$\hat{Y}_t = \frac{N}{n}\sum_{j=1}^{n} y_j \ ,$$

waarbij N het aantal elementen in de populatie is. Elk element in de steekproef wordt dus met N/n vermenigvuldigd. Deze factor wordt ophooggewicht of kortweg gewicht genoemd.

Als er een schatter toegepast wordt die geen rekening houdt met uitbijters (zoals het steekproefgemiddelde of de bovengenoemde totaalschatting) dan heeft een extreme waarneming grote invloed op het schattingsresultaat. Een representatieve uitbijter is een uitbijter waarvan aangenomen wordt dat deze correct waargenomen is en dat in de populatie meer soortgelijke elementen te vinden zijn. Voor een algemene beschrijving van representatieve uitbijters zie ook Smeets (2005).

In de praktijk zullen, ondanks het gaafmaken, niet alle fouten uit de data verwijderd zijn. Dan is dus niet helemaal aan deze aanname voldaan.

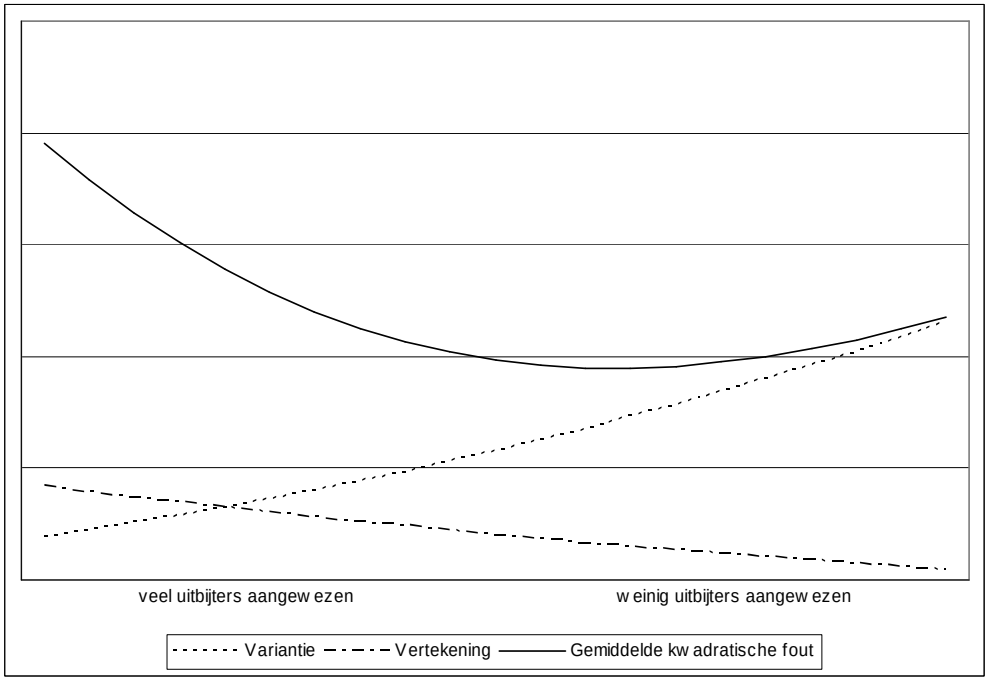
Het wordt vaak niet als wenselijk ervaren wanneer een enkele waarneming grote invloed heeft op het schattingsresultaat. De variantie van de schattingen is dan vaak groot, wat in de praktijk naar voren komt als instabiele schattingen, dat wil zeggen dat de schattingen in verschillende perioden voor dezelfde doelvariabele meer verschillen dan men voor de populatie verwacht.

1.1.2 Oplossingsrichting

Als oplossing wordt dan ervoor gekozen om de invloed van de representatieve uitbijters te beperken. Hierdoor wordt de variantie kleiner, maar wordt er in de meeste gevallen wel een vertekening geïntroduceerd. De behandeling van correcte uitbijters komt daarmee neer op een afweging tussen aan de ene kant een zo klein mogelijke variantie en aan de andere kant een zo klein mogelijke vertekening. De gemiddelde

kwadratische fout van een schatter is gedefinieerd als de verwachting van het kwadraat van het verschil tussen de schatter en de werkelijke waarde van de populatieparameter. Deze kan berekend worden als de som van de variantie en het kwadraat van de vertekening. Een kleine gemiddelde kwadratische fout betekent dat de werkelijke populatieparameter goed benaderd wordt door de schatter. Minimalisatie van de gemiddelde kwadratische fout is daarom een formalisatie van de afweging tussen een kleine variantie en een kleine vertekening. Er wordt dus een niet-zuivere schatter geconstrueerd. De (kleine) vertekening wordt geaccepteerd omdat de variantie kleiner wordt. Figuur 1 illustreert dat de vertekening groot is en de variantie klein als bij veel uitbijters aangewezen worden. Als er weinig of geen uitbijters aangewezen worden is dat andersom. Het optimum ligt bij het minimum van de gemiddelde kwadratische fout.

Figuur 1. Illustratie van het effect van uitbijterbehandeling op variantie, vertekening en gemiddelde kwadratische fout



In de meeste gevallen geldt dat hoe meer elementen er als uitbijter aangewezen worden (waardoor hun invloed beperkt wordt), hoe kleiner de variantie wordt en ook hoe groter de vertekening. Voor kleine steekproeven is het verlies aan nauwkeurigheid vanwege de ontstane vertekening door het aanwijzen van uitbijters relatief klein, vergeleken met de winst vanwege de kleinere variantie. Voor grotere steekproeven weegt de vertekening zwaarder dan de winst door een kleinere variantie. Daarom is het bij grotere steekproeven beter om relatief minder uitbijters aan te wijzen. Voor de precieze invulling hiervan worden in dit document verschillende methoden uitgewerkt.

Als de data symmetrisch verdeeld zijn, ontstaat er door het aanwijzen van uitbijters geen vertekening (tenminste, als er zowel uitzonderlijk grote als uitzonderlijk kleine

waarden als uitbijter aangewezen worden). In dat geval kan het handig zijn om relatief veel waarnemingen als uitbijter aan te wijzen. Deze situatie doet zich voor bij de PS-schatter (zie hoofdstuk 4).

Er zijn verschillende mogelijkheden om de invloed van representatieve uitbijters te beperken:

- De waarneming kan genegeerd worden.
- De waarde van de waarneming kan aangepast worden.
- Het ophooggewicht van de waarneming kan verkleind worden. Het gewicht op 1 zetten betekent dat de eenheid alleen voor zichzelf meetelt en dat het ophogen gebeurt aan de hand van de andere eenheden. De waarde hoeft niet per se uniek te zijn. Ook als de waarde niet uniek is, of als daar niets over bekend is, kan het een goede oplossing zijn om het gewicht op 1 te zetten. Merk op dat op basis van de steekproef het nooit te bepalen is of een waarde uniek is.Als het gewicht op 0 gezet wordt, dan wordt de waarneming genegeerd.

Het is vaak mogelijk om een schatter op twee manieren te schrijven: via het aanpassen van waarden of via het aanpassen van ophooggewichten. Het schattingsresultaat is dan in beide gevallen hetzelfde. In hoofdstuk 2 wordt een schatter uitgewerkt die zowel via het aanpassen van de waarden als ook via het aanpassen van de ophooggewichten toegepast kan worden.

Tot een aantal jaren geleden werd het aanpassen van de gewichten of de waarden in de praktijk altijd handmatig gedaan. De laatste jaren is bij een aantal statistieken een automatische procedure geïntroduceerd. Bij handmatige uitbijterdetectie hangt het uiteindelijke schattingsresultaat af van de (subjectieve) inschattingen van de medewerker.

1.1.3 Robuust schatten

Robuuste schatters kunnen de invloed van uitbijters beperken. Twee eenvoudige robuuste schatters zijn het α-getrimde gemiddelde en het α-gewinsoriseerde gemiddelde. Bij het α-getrimde gemiddelde worden de ½α procent grootste en de ½α procent kleinste waarnemingen genegeerd. Bij het α-gewinsoriseerde gemiddelde worden de ½α procent grootste en de ½α procent kleinste waarnemingen aangepast. Ze krijgen dezelfde waarde als de grootste respectievelijk kleinste niet aan te passen waarde. Een andere bekende, maar complexere schatter is de zogenaamde M-schatter. De drie genoemde robuuste schatters zijn niet specifiek bedoeld voor representatieve uitbijters en worden daarom in dit document niet nader beschreven.

Voor meer informatie over deze schatters zie bijvoorbeeld Huber (1981). In de literatuur over robuust schatten, waarvoor de drie genoemde schatters bedoeld zijn, wordt vaak ervan uitgegaan dat de data vervuild zijn met uitbijters. Zo kan bijvoorbeeld 90% van de populatie normaal verdeeld zijn met gemiddelde μ_1 en variantie σ_1^2, en de overige 10% is normaal verdeeld met gemiddelde μ_2 en variantie σ_2^2. Ook deze 10% is correct, maar wordt beschouwd als afwijkend en niet interessant. Men is dus alleen geïnteresseerd is in het niet-vervuilde gedeelte van de data, dus bijvoorbeeld in μ_1 en σ_1^2. Op het CBS is er geen sprake van vervuiling. Men is juist geïnteresseerd in het gemiddelde of het totaal van de hele populatie, in principe inclusief de bijzonder grote of kleine elementen. Een ander nadeel van deze robuuste schatters is de afhankelijkheid van een parameter, bijvoorbeeld α bij het α-getrimde gemiddelde. Vaak wordt voor α de waarde 5% gekozen. Er is echter geen goede methode om de parameter eenvoudig vast te stellen.
Deze schatters zijn dus niet geschikt voor representatieve uitbijters.

1.1.4 Methoden in dit document

Op het CBS zijn drie schatters ontwikkeld die wel geschikt zijn voor representatieve uitbijters en die in dit document beschreven worden.

- **De eenzijdig gecensureerde schatter**, die als doel heeft de winsorisatie optimaal te bepalen waarbij alleen de grootste (of alleen de kleinste) waarden aangepast worden.
- **De tweezijdig gecensureerde schatter**, die als doel heeft de winsorisatie optimaal te bepalen waarbij zowel de grootste als de kleinste waarden aangepast worden.
- **De PS-methode.** Bij deze methode worden de gewichten van de meest afwijkende waarnemingen op 1 gezet.

In hoofdstuk 11 wordt aangegeven welke afwegingen een rol spelen bij de keuze tussen deze methoden.

Of een afwijkende waarde veel invloed heeft op de schatting, hangt niet alleen van de waarde zelf af maar ook van het ophooggewicht. Een uitzonderlijke waarde met een klein ophooggewicht hoeft geen grote invloed op de schatting te hebben.

Of een waarneming afwijkend is en veel invloed heeft op de schatting, hangt ook af van wat er geschat wordt. In een steekproef van alle bedrijven in Nederland is een omzet van 10 miljoen euro per jaar niet uitzonderlijk. In de deelpopulatie van bedrijven met één medewerker is deze waarde wel uitzonderlijk en waarschijnlijk, afhankelijk van de steekproefgrootte, van sterke invloed op het schattingsresultaat.

Of een waarneming afwijkend is en veel invloed heeft op de schatting, hangt ook af van welke schatter er toegepast wordt. Het bedrijf met een omzet van 10 miljoen euro per jaar heeft een sterke invloed op het steekproefgemiddelde van de deelpopulatie van bedrijven met één medewerker. Als echter de regressieschatter toegepast wordt met BTW als hulpinformatie, en het bedrijf heeft ook een grote BTW-waarde, dan is de invloed van het bedrijf weer minder groot.

De drie schatters die in dit document beschreven worden, zijn bedoeld voor een relatief eenvoudige situatie: er moet een niveauschatting gemaakt worden voor één populatie en voor één doelvariabele, waarbij een representatieve steekproef getrokken is. De schatters zijn ook geschikt (of geschikt te maken) voor complexe steekproefontwerpen.

1.1.5 Open problemen

Op het CBS is echter vaak sprake van complexe situaties. Zo moeten er vaak in plaats van niveauschattingen zo nauwkeurig mogelijke schattingen gemaakt worden van ontwikkelingen, wil men tegelijk consistente schattingen maken voor verschillende doelvariabelen of verschillende doelpopulaties of is de schatting gebaseerd op een register in plaats van een representatieve steekproef. Voor deze situaties (en combinaties ervan) is nog geen valide methodologie ontwikkeld. Als dergelijke methodologie in de toekomst ontwikkeld wordt, kan de beschrijving ervan aan dit document toegevoegd worden. Hiervoor zijn de hoofdstukken 5 t/m 10 gereserveerd. Indien nodig, kunnen er nog meer hoofdstukken toegevoegd kunnen worden.

Een voorlopige en pragmatische oplossing voor deze complexere problemen is het toepassen van één van de in dit document beschreven methodes. Dit gebeurt ook in de praktijk, bijvoorbeeld bij de productiestatistieken, waar de uitbijterafhandeling gericht is op kerncellen en de omzet. De schattingen voor heel Nederland en andere aggregatieniveaus en voor andere doelvariabelen zijn dan niet optimaal. De schattingen voor heel Nederland en bijvoorbeeld de bedrijfstakken worden verkregen door de schattingen van de bijbehorende kerncellen bij elkaar op te tellen. De vertekening voor de schattingen per kerncel is acceptabel in relatie tot de vrij grote variantie. De vertekening van de schattingen voor heel Nederland en voor de bedrijfstakken kan echter onacceptabel hoog zijn in relatie tot de vrij kleine variantie van deze schattingen. Deze pragmatische oplossing houdt dus onvoldoende rekening met verschillende gebruikers.

1.2 Het begrip representatieve uitbijter

Zoals in de voorgaande paragraaf aangegeven is, wordt voor een representatieve uitbijter aangenomen dat in de populatie meer soortgelijke elementen te vinden zijn. De uitbijter is dus representatief voor deze elementen en dat verklaart de term representatief.

Bij het schatten met representatieve uitbijters wordt de invloed van de representatieve uitbijters beperkt, bijvoorbeeld door het verkleinen van het gewicht. Het verkleinen van de gewichten heeft geen betrekking op een aanname over hoeveel soortgelijke elementen er in de populatie te vinden zijn. Er wordt dus niet aangenomen, dat de uitbijter niet volledig representatief is.

Het doel van het beperken van de invloed van representatieve uitbijters is het verkleinen van de variantie. Echter, door het beperken van de invloed wordt een vertekening geïntroduceerd. In dit document wordt uitgewerkt op welke manieren de invloed van representatieve uitbijters beperkt kan worden.

1.3 Afbakening en relatie met andere thema's

Zoals gezegd gaat het in dit document om schattingsmethoden voor databestanden met representatieve uitbijters. De methoden die in dit document beschreven worden, zijn bedoeld voor situaties waarin voor één doelvariabele en voor één doelpopulatie een parameter geschat moet worden. De methoden zijn geschikt als er zo nauwkeurig mogelijke niveauschattingen gemaakt moeten worden, en als deze schattingen gemaakt worden op basis van aselecte steekproeven, zowel voor eenvoudige als voor complexere steekproefontwerpen. Voor complexere situaties is meer onderzoek nodig. Hiervoor is in hoofdstuk 5 t/m 10 ruimte gereserveerd.

Uitbijters spelen niet alleen tijdens het ophogen een rol, maar ook in andere fases van het statistische proces. Tijdens het gaafmaken worden (onder meer) uitbijters aangepast, waarbij vastgesteld of aangenomen wordt dat deze uitzonderlijke waarnemingen fout zijn. Bij het imputeren wordt voor unit- en item-nonrespons een geschikte waarde ingevuld. Deze waarde is dan vaak gebaseerd op het gedeelte van de steekproef dat wel is waargenomen. Deze waarde kan dan beïnvloed zijn door uitbijters, wat niet wenselijk is. Verder kunnen uitbijters ook een verstorende invloed hebben in de analysefase. Ten slotte kunnen ook in de hulpinformatie uitbijters voorkomen, die een storende invloed kunnen hebben.

De in dit document beschreven methoden zijn alleen bedoeld voor de problemen m.b.t. de representatieve uitbijters bij de ophoging. Voor technieken die uitbijters tijdens het gaafmaken behandelen, verwijzen we naar Hoogland e.a. (2009) en voor het omgaan met uitbijters tijdens het imputeren verwijzen we naar Israëls e.a. (2007).

1.4 Plaats in het statistisch proces

De behandeling van representatieve uitbijters maakt deel uit van het ophoogproces. Men kan het proces zo interpreteren dat de waarden of gewichten van uitbijters aangepast worden nadat het databestand is gaafgemaakt en ontbrekende waarden zijn geïmputeerd, en net voordat de steekproef opgehoogd wordt. Het kan ook beschouwd worden als deel van het ophoogproces. Het idee van ophogen is kort be-

schreven in paragraaf 1.1, voor meer informatie zie Banning en Knottnerus (2009) of Särndal e.a. (1992).

1.5 Definities

Begrip	Omschrijving
uitbijter	Een uitbijter is een extreme waarneming. In dit document worden alleen uitbijters in de steekproef beschouwd.
linkeruitbijter	Een linkeruitbijter is een extreem lage waarneming.
rechteruitbijter	Een rechteruitbijter is een extreem hoge waarneming.
representatieve uitbijter	Een representatieve uitbijter is een uitbijter in de steekproef, waarvan aangenomen wordt dat deze correct waargenomen is en dat in de populatie meer soortgelijke elementen te vinden zijn.
niet-representatieve uitbijter	Een niet-representatieve uitbijter is een uitbijter in de steekproef, die niet correct waargenomen is of uniek is in de populatie.
α-getrimd gemiddelde	Het α-getrimde gemiddelde is het gemiddelde van een aantal waarnemingen, waarbij de ½α procent grootste en de ½α procent kleinste waarnemingen genegeerd worden.
α-gewinsoriseerd gemiddelde	Het α-gewinsoriseerde gemiddelde is het gemiddelde van een aantal waarnemingen, waarbij de ½α procent grootste en de ½α procent kleinste waarnemingen aangepast worden. Ze krijgen dezelfde waarde als de grootste respectievelijk kleinste niet aan te passen waarde.

1.6 Algemene notatie

De volgende algemene notaties worden gebruikt:

i : een element (bijvoorbeeld bedrijf of persoon).

n : de steekproefomvang

N : de populatieomvang

h : index om een bepaald stratum aan te geven

L : het aantal strata in de populatie en steekproef

y : de doelvariabele

x : de hulpvariabele

$y_1,...,y_n$: de waarnemingen van de doelvariabele in de steekproef

$w_1,...,w_n$: de insluitgewichten van de waarnemingen in de steekproef

$y_{h1},...,y_{hn_h}$: de waarnemingen van de doelvariabele in stratum h in de steekproef

$\bar{Y}$: schatter voor het gemiddelde van doelvariabele y (met superscript en subscript om de schatter en eventueel de doelpopulatie te specificeren).

2. Eenzijdig gecensureerde schatters

2.1 Korte beschrijving

De gecensureerde schatter is een aantal jaren geleden op het CBS ontwikkeld, zie Renssen e.a. (2004) en Krieg e.a. (2004). Bij de gecensureerde schatter wordt de invloed van uitbijters beperkt door middel van de techniek van winsoriseren (zie paragraaf 1.1). Dat betekent dat de invloed beperkt wordt door de waarde van de uitbijter te vervangen door een grenswaarde. Door de invloed van uitbijters te beperken zal de variantie afnemen, maar zal er meestal een vertekening geïntroduceerd worden. Voor de gecensureerde schatter wordt expliciet een formule afgeleid voor de grenswaarde die de gemiddelde kwadratische fout (in het Engels mean square error of afgekort MSE) minimaliseert en op deze manier een optimum vindt tussen aan de ene kant een lage variantie en aan de andere kant een zo klein mogelijke vertekening. Dit is het optimum voor schatters met een dergelijke grenswaarde, het is mogelijk dat andere type schatters beter zijn. Als er vooraf niets bekend is over de optimale grenswaarde, kan deze geschat worden met de steekproef.

Als er alleen uitbijters met een extreem hoge of een extreem lage waarde aangepast worden, dus slechts aan één kant van de verdeling, spreken we van eenzijdig censureren. Eenzijdig censureren is dus zinvol bij een scheve verdeling met slechts aan één kant extreme waarden. Bij tweezijdig censureren worden er zowel linkeruitbijters (extreem lage waarden) als rechteruitbijters (extreem hoge waarden) aangepast, de zogenaamde linker- en rechteruitbijters, en worden er twee grenswaarden bepaald. De linkeruitbijters worden vervangen door de kleinste grenswaarde en de rechteruitbijters door de grootste grenswaarde. Tweezijdig censureren wordt als aparte methode in hoofdstuk 3 besproken. In dit hoofdstuk werken we de eenzijdig gecensureerde schatter uit, waarbij de grenswaarde met de steekproef geschat wordt.

Als de populatie bekend zou zijn, dan zou een betere schatter gevonden kunnen worden door de grenswaarde uit de populatie te berekenen. Dat is echter geen realistische situatie. Op het CBS worden veel onderzoeken per maand, per kwartaal of per jaar uitgevoerd. Dan kan het nuttig zijn om data uit het verleden te gebruiken om de grenswaarde te berekenen, vooral als de populatie niet te veel verandert. Er is echter niet onderzocht hoeveel een populatie mag veranderen zodat dit zinvol blijft. Zolang dit niet bepaald is, is dit idee geen valide methodologie en wordt daarom hier niet beschreven.

2.2 Toepasbaarheid

Toepassing van de eenzijdig gecensureerde schatter kan interessant zijn wanneer er aan één kant van de verdeling extreme waarden in de waarnemingen voorkomen. In

dat geval levert de eenzijdige gecensureerde schatter naar verwachting betere resultaten dan een schatter die niets met uitbijters doet.

We werken de eenzijdig gecensureerde schatter uit voor enkelvoudige aselecte steekproeven en gestratificeerde steekproeven, waarbij per stratum zonder terugleggen en enkelvoudig aselect getrokken is. In andere situaties kunnen de besproken methoden niet zonder meer toegepast worden en zijn verdere aanpassingen noodzakelijk. Ter illustratie worden aan het eind van de paragraaf 2.3 voor enkele specifieke situaties varianten van de eenzijdig gecensureerde schatter besproken.

Op dit moment wordt de eenzijdig gecensureerde schatter bij een tweetal CBS-enquêtes toegepast. Het gaat om de statistieken Internationale Diensten (Krieg e.a., 2008) en Bouwobjecten in Voorbereiding (Smeets, 2008).

2.3 Uitgebreide beschrijving

2.3.1 Eenzijdig gecensureerde schatter voor enkelvoudige aselecte steekproeven

Eerst bespreken we de eenzijdig gecensureerde schatter voor een enkelvoudig aselecte steekproef. Uit een populatie met N elementen is zonder terugleggen een enkelvoudig aselecte steekproef van n elementen met gelijke kansen getrokken. Stel dat we voor doelvariabele y een populatiegemiddelde willen schatten, waarbij de invloed van uitbijters beperkt is. Voor alle elementen i, met $1 \le i \le n$, in de steekproef is daartoe de doelvariabele y waargenomen. De waarnemingen worden zo gesorteerd dat $y_1 \le y_2 \le \cdots \le y_n$ geldt. Er moet dan een grenswaarde t berekend worden.

We leggen eerst uit hoe het censureren in zijn werk gaat als t gegeven is. Als een waarneming y_i groter is dan de grenswaarde t, wordt de waarneming aangepast en vervangen door deze grenswaarde. Vervolgens wordt het gemiddelde van de populatie geschat door het steekproefgemiddelde te berekenen van de aangepaste waarnemingen. Stel er zijn r waarnemingen kleiner of gelijk aan grenswaarde t. Het gemiddelde wordt dan geschat door

$$\overline{Y}_t^{cens} = \frac{\sum_{j=1}^{r} y_j + (n-r)t}{n}, \qquad (2.3.1)$$

met $y_j \le t$ voor $j = 1, \ldots, r$. Er zijn dus $n-r$ waarnemingen op uitbijter gezet. De r waarnemingen kleiner of gelijk aan t worden ook wel niet-uitbijters genoemd.

Voordat formule (2.3.1) toegepast kan worden, moet t berekend worden. Het doel is om een grenswaarde te vinden, die de gemiddelde kwadratische fout van (2.3.1) minimaliseert. Dit is de optimale grenswaarde. In Renssen e.a. (2004) is een formule

gegeven voor de gemiddelde kwadratische fout als som van de variantie en het kwadraat van de vertekening.

De variantie, de vertekening en de MSE kunnen worden geschreven als

$$\mathrm{Var}(\overline{Y}_t^{cens}) = (1-f)\frac{p\sigma_m^2 + pq(\mu_m - t)^2}{n}, \tag{2.3.2}$$

$$\mathrm{Bias}(\overline{Y}_t^{cens}) = -q(\mu_r - t), \tag{2.3.3}$$

$$\mathrm{MSE}(\overline{Y}_t^{cens}) = (1-f)\frac{p\sigma_m^2 + pq(\mu_m - t)^2}{n} + q^2(\mu_r - t)^2. \tag{2.3.4}$$

Hierin is q de fractie van aangepaste waarden in de populatie, p de fractie van niet-aangepaste waarden in de populatie, σ_m^2 de populatievariantie van de niet-aangepaste waarden, μ_m het populatiegemiddelde van de niet-aangepaste waarden, μ_r het populatiegemiddelde van de aangepaste waarden en $f = \frac{n}{N}$ de steekproeffractie.

Merk op dat voor de variantie een benaderingformule gebruikt is, die geen rekening houdt met de factor $1/(N-1)$ voor enkelvoudig aselecte steekproeven zonder terugleggen. In plaats daarvan is de factor $1/N$ gebruikt. In de praktijk is N heel groot, zodat beide factoren bijna even groot zijn.

De grenswaarde die vergelijking (2.3.4) minimaliseert is een oplossing van de volgende vergelijking

$$\frac{p}{n}(1-f)(t-\mu_m) - q(\mu_r - t) = 0. \tag{2.3.5}$$

Een schatting voor de optimale t wordt vervolgens gevonden door in vergelijking (2.3.5) de parameters op basis van de steekproef in plaats van de populatieparameters te gebruiken.

Dit betekent dat een schatting voor de optimale grenswaarde gevonden kan worden door vergelijking (2.3.5) op te lossen, met $q = \frac{n-r}{n}$ de fractie uitbijters in de steekproef, $p = \frac{r}{n}$ de fractie niet-uitbijters in de steekproef, $\mu_m = \frac{1}{r}\sum_{j=1}^{r} y_j$ het steekproefgemiddelde van de niet-uitbijters en $\mu_r = \frac{1}{n-r}\sum_{j=r+1}^{n} y_j$ het steekproefgemiddelde van de uitbijters en $f = \frac{n}{N}$ de steekproeffractie.

Het oplossen van vergelijking (2.3.5) kan niet rechttoe rechtaan gedaan worden, omdat de parameters p, q, μ_m en μ_r zelf ook van t afhangen. Uit de waarde van t volgt namelijk welke waarnemingen op uitbijter worden gezet. Hieruit volgen r en de waarden van de genoemde parameters. In Renssen e.a. (2004) wordt aangetoond dat er altijd een unieke oplossing van vergelijking (2.3.5) bestaat en dat deze kleiner (of gelijk) is dan de grootste waarde y_n. Er wordt dus altijd ten minste één element als uitbijter aangewezen, behalve in het extreme geval van integraal waarnemen. Dan is het niet zinvol om uitbijters aan te wijzen, immers, alle waarnemingen hebben een gewicht van 1. In dat geval is $t = \mu_r$ de oplossing van (2.3.5). Dit

kan geïnterpreteerd worden dat één element uitbijter wordt, waarvan echter de waarde niet aangepast wordt. De interpretatie dat geen uitbijters aangewezen worden is in dit geval ook mogelijk. Als t nauwelijks kleiner is dan y_n, dan verschilt de gecensureerde schatting nauwelijks van de schatting waarbij geen uitbijters aangewezen worden. In Renssen e.a. (2002) staan algoritmes voor steekproeven die met terugleggen getrokken zijn, voor enkelvoudig aselect trekken en voor gestratificeerde steekproeven. In de bijlage van dit document zijn deze algoritmes aangepast aan de situatie trekken zonder terugleggen, en worden voor meer situaties algoritmes gegeven. Algoritme 1 berekent de optimale grenswaarde voor enkelvoudig aselecte steekproeven. In dit algoritme wordt voor alle mogelijke aantallen uitbijters t uit vergelijking (2.3.5) berekend en dan gecontroleerd of bij deze t het juiste aantal uitbijters hoort. Dit wordt herhaald tot de unieke oplossing van vergelijking (2.3.5) gevonden is.

Dit algoritme kan toegepast worden bij een steekproefgrootte van minimaal 2 elementen. Voor heel kleine steekproeven is het echter niet aan te bevelen om überhaupt schattingen te berekenen, omdat deze een heel grote variantie hebben. Dit geldt ook voor de gecensureerde schatters.

Door de gevonden geschatte optimale grenswaarde t^* in te vullen in vergelijking (2.3.1), vinden we de eenzijdig gecensureerde schatting $\overline{Y}_{t^*}^{cens}$ voor het populatiegemiddelde.

Zoals al gezegd, is bij het afleiden van de formules voor de gecensureerde schatter uitgegaan van bekende populatieparameters, maar worden uiteindelijk schattingen op basis van de steekproef gebruikt. Als de optimale grenswaarde berekend zou zijn op basis van de populatie, dan zou de daarop gebaseerde schatter nauwkeuriger zijn. Maar ook als de optimale grenswaarde geschat is op basis van de steekproef, is de schatter meestal nauwkeuriger dan de schatter zonder uitbijterbehandeling. Dit blijkt uit verschillende simulaties, bijvoorbeeld (Krieg en Smeets, 2005). In Renssen e.a. (2004) zijn benaderingsformules voor variantie en vertekening afgeleid. Ook op basis van deze formules kan geconcludeerd worden dat de gecensureerde schatter in de meeste gevallen nauwkeuriger is dan de schatter zonder uitbijterbehandeling. Een uitzondering is een populatie met meerdere uitbijters met een ongeveer gelijke waarde, in combinatie met een grote trekkingskans. Andere uitzonderingen zijn mogelijk als de benaderingsformules in Renssen e.a. (2004) onvoldoende precies zijn. De benaderingsformules kunnen mogelijk onvoldoende precies zijn voor heel kleine steekproeven, kleine populaties of extreme verdelingen. In de uitzonderingsgevallen is de nauwkeurigheid van de gecensureerde schatter vergelijkbaar met de nauwkeurigheid van de schatter zonder uitbijterbehandeling. Er zijn wel extreme kunstmatige situaties te construeren waar de nauwkeurigheid van de gecensureerde schatter veel kleiner is dan de nauwkeurigheid van de schatter zonder uitbijterbehandeling.

Gecensureerde schatter schrijven met gewichten. De gecensureerde schatter kan geschreven worden als een gewogen gemiddelde van de waarnemingen, waarbij de uitbijters een lager gewicht krijgen en de niet-uitbijters juist een hoger gewicht krijgen:

$$\overline{Y}_{t^*}^{cens} = \frac{1}{n}\left[\sum_{j=1}^{r} g_m y_j + \sum_{j=r+1}^{n} g_r y_j\right], \text{ met}$$

$$\begin{cases} g_m = 1 + \dfrac{q}{p(1+\lambda)} \geq 1, \text{voor de niet - uitbijters} \\ g_r = 1 - \dfrac{1}{1+\lambda} < 1, \text{voor de uitbijters} \end{cases} \quad \text{en}$$

$$\lambda = \frac{nq}{(1-f)p} > 0. \tag{2.3.6}$$

Het is eenvoudig na te gaan dat de gewichten optellen tot n, d.w.z. $rg_m + (n-r)g_r = n$.

Aan de hand van een eenvoudig voorbeeld laten we zien hoe algoritme 1 werkt.

Voorbeeld 2.3.1. De steekproefgrootte is $n = 12$, de populatiegrootte is $N = 120$ (dus $f = 0{,}1$), de waarnemingen zijn 1, 2, 3, 4, 4, 4, 5, 5, 6, 9, 20, 25. Het steekproefgemiddelde is gelijk aan 7,33. In tabel 1 zijn voor alle mogelijke waarden van r de bijbehorende parameters p, η_m, q en η_t berekend. De waarde van t in de zevende kolom is bepaald met vergelijking (B.1). In de laatste twee kolommen zijn de waarden van y_r en y_{r+1} gegeven.

Tabel 1. Voorbeeld berekening grenswaarde t

r	$n-r$	p	q	η_m	η_t	t	y_r	y_{r+1}
11	1	0,92	0,08	5,73	25,00	16,29	20	25
10	**2**	**0,83**	**0,17**	**4,30**	**22,50**	17,54	**9**	**20**
9	3	0,75	0,25	3,78	18,00	15,39	6	9
8	4	0,67	0,33	3,50	15,00	13,50	5	6
7	5	0,58	0,42	3,29	13,00	12,08	5	5
6	6	0,50	0,50	3,00	11,67	11,06	4	5
5	7	0,42	0,58	2,80	10,57	10,18	4	4
4	8	0,33	0,67	2,50	9,75	9,49	4	4
3	9	0,25	0,75	2,00	9,11	8,94	3	4
2	10	0,17	0,83	1,50	8,50	8,40	2	3
1	11	0,08	0,92	1,00	7,91	7,86	1	2

Algoritme 1 doorloopt één voor één de rijen van tabel 1 en controleert of de gevonden t-waarde tussen de waarden van y_r en y_{r+1} ligt. In de tweede rij zijn er twee elementen op uitbijter zijn gezet en geldt $r = 10$. De grenswaarde in deze rij voldoet aan de voorwaarde en de geschatte optimale grenswaarde $t^* = 17{,}54$ is daarmee dus gevonden. De eenzijdig gecensureerde schatter voor het populatiegemiddelde is

dan gelijk aan $\overline{Y}_{t^*}^{\text{cens}} = 6{,}506$. Merk op dat het algoritme eigenlijk stopt bij $r = 10$, maar in de tabel zijn alle mogelijkheden voor r gegeven, om te laten zien dat de oplossing van vergelijking (2.3.5) uniek is.

Opmerking: In het voorbeeld zijn twee elementen als uitbijter aangewezen. In de praktijk blijkt dat in de meeste gevallen weinig uitbijters aangewezen worden en vaak zelfs maar één. Dit geldt ook voor grotere steekproeven. Als er meer elementen als uitbijter aangewezen worden, dan wordt de vertekening steeds groter. Dit speelt al snel een grotere rol dan de winst in de variantie. De geïnteresseerde lezer is uitgenodigd voor meer voorbeelden de geschatte optimale *t* en het aantal uitbijters te berekenen.

2.3.2 Eenzijdig gecensureerde schatter voor gestratificeerde steekproeven

Bij een gestratificeerd steekproefontwerp is de populatie onderverdeeld in L verschillende strata, die bestaan uit $N_1, \ldots, N_L$ elementen. De totale populatieomvang is dan gelijk aan $N = N_1 + \cdots + N_L$. Een steekproef van omvang n is verdeeld over deze strata met steekproefomvangen $n_1, \ldots, n_L$. We gaan ervan uit dat in elk stratum een enkelvoudig aselecte steekproef zonder terugleggen is getrokken. Laat $y_{h1}, \ldots, y_{hn_h}$ de waarnemingen zijn van de doelvariabele y in stratum h zodat $y_{h1} \le y_{h2} \le \cdots \le y_{hn_h}$ geldt. Voor elk stratum wordt een grenswaarde berekend. Laat $\mathbf{t} = (t_1, \ldots, t_h, \ldots, t_L)$ de vector met de te berekenen grenswaarden zijn met waarde t_h in stratum h.

Als de grenswaarden gegeven zijn, dan gaat het censureren per stratum op dezelfde manier als bij een enkelvoudig aselecte steekproef. In ieder stratum h worden de waarden van de doelvariabelen dan vergeleken met de gegeven grenswaarde t_h in het betreffende stratum. Als een waarneming groter is dan de grenswaarde, wordt de waarneming aangepast en vervangen door deze grenswaarde. Vervolgens wordt het stratumgemiddelde van de populatie geschat door het steekproefgemiddelde te berekenen van de aangepaste waarnemingen in stratum h. Stel dat er r_h waarnemingen in stratum h kleiner of gelijk zijn aan grenswaarde t_h, dan wordt het stratumgemiddelde geschat door

$$\overline{Y}_{t_h}^{cens} = \frac{\sum_{j=1}^{r_h} y_{hj} + (n_h - r_h)t_h}{n_h}, \tag{2.3.7}$$

met $y_{hj} \le t_h$ voor $j = 1, \ldots, r_h$ en $h = 1, \ldots, L$. In stratum h zijn er dus $n_h - r_h$ waarnemingen op uitbijter gezet.

Het gemiddelde van de gehele populatie wordt vervolgens geschat door

$$\overline{Y}_{\mathbf{t}}^{cens} = \frac{1}{N}\sum_{h=1}^{L} N_h \overline{Y}_{t_h}^{cens}, \text{ met } \mathbf{t} = (t_1, \ldots, , t_L). \tag{2.3.8}$$

In de gestratificeerde situatie is het doel een vector van grenswaarden $\mathbf{t}^* = (t_1^*, \ldots, t_L^*)$ te vinden, die leidt tot een minimale gemiddelde kwadratische fout van de schatter (2.3.8). De gestratificeerde schatter is dus optimaal voor het gemiddelde over de gehele populatie en niet voor het gemiddelde per stratum.

Voordat formule (2.3.7) toegepast kan worden, moet $\mathbf{t}^* = (t_1^*, \ldots, t_L^*)$ berekend worden. De optimale grenswaarde kan gevonden worden door een stelsel van vergelijkingen op te lossen. Dit stelsel is afgeleid in Renssen e.a. (2004) waarbij uitgegaan wordt van de formule voor de gemiddelde kwadratische fout. In de afleiding is uitgegaan van populatieparameters. Een schatting voor de optimale $\mathbf{t}^* = (t_1^*, \ldots, t_L^*)$ wordt gevonden door in de vergelijking de parameters op basis van de steekproef in plaats van de populatieparameters te gebruiken. Het stelsel is gegeven door:

$$\begin{cases} \dfrac{N_1(1-f_1)p_1(t_1-\mu_{1m})}{n_1} - \displaystyle\sum_{h=1}^{L} N_h q_h(\mu_{hr} - t_h) = 0 \\ \qquad\qquad \vdots \qquad\qquad . \\ \dfrac{N_L(1-f_L)p_L(t_L-\mu_{Lm})}{n_L} - \displaystyle\sum_{h=1}^{L} N_h q_h(\mu_{hr} - t_h) = 0 \end{cases} \tag{2.3.9}$$

Hierin is $q_h = \dfrac{n_h - r_h}{n_h}$ de fractie uitbijters in stratum h in de steekproef, $p_h = \dfrac{r_h}{n_h}$ de fractie niet-uitbijters in stratum h in de steekproef, $f_h = \dfrac{n_h}{N_h}$ de steekproeffractie in stratum h, $\mu_{hm} = (y_{h1} + \cdots + y_{hr_h})/r_h$ het steekproefgemiddelde van de niet-uitbijters in stratum h en $\mu_{hr} = (y_{h(r_h+1)} + \cdots + y_{hn_h})/(n_h - r_h)$ het steekproefgemiddelde van de uitbijters in stratum h.

Net als bij de formule voor enkelvoudig aselecte steekproeven is ook hier voor de variantie een benaderingformule gebruikt is.

Algoritme 2 in de bijlage vindt de unieke oplossing voor stelsel (2.3.9). In principe zou de unieke oplossing van (2.3.9) gevonden kunnen worden door systematisch alle mogelijke combinaties van aantallen uitbijters per stratum door te rekenen. Dit zou echter veel rekentijd kosten. Daarom wordt via een transformatie van de oorspronkelijke data eerst een startpunt bepaald. Deze combinatie van uitbijters wordt dan stapsgewijs aangepast, tot de oplossing gevonden is. Het startpunt wordt bepaald door de oorspronkelijke data te transformeren, waarmee het probleem vereenvoudigd is naar enkelvoudig aselect trekken. In de praktijk blijkt dat uitgaande van dit startpunt de unieke oplossing van stelsel (2.3.9) binnen enkele iteraties gevonden wordt. Voor het bepalen van de tussenstappen is het noodzakelijk dat de strata gesorteerd zijn. De gekozen volgorde heeft geen invloed op de uiteindelijke oplossing, alleen op de tussenberekeningen.

Door de via algoritme 2 gevonden optimale vector van grenswaarden $\mathbf{t}^* = (t_1^*,...,t_L^*)$ in te vullen in vergelijking (2.3.8), vinden we de eenzijdig gestratificeerde gecensureerde schatter voor het populatiegemiddelde $\overline{Y}_{\mathbf{t}^*}^{cens}$.

Gestratificeerde gecensureerde schatter schrijven met gewichten. Ook de gestratificeerde gecensureerde schatter kan geschreven worden als een gewogen gemiddelde van de waarnemingen, waarbij de uitbijters een lager gewicht krijgen en de niet-uitbijters een hoger gewicht:

$$\overline{Y}_{\mathbf{t}^*}^{cens} = \frac{1}{N}\sum_{h=1}^{L}\frac{N_h}{n_h}\left\{\sum_{j=1}^{r_h} g_{hm}y_{hj} + \sum_{j=r_h+1}^{n_h} g_{hr}y_{hj}\right\}, \text{ met}$$

$$\begin{cases} g_{hm} &= 1+\dfrac{q_h}{p_h(1+\lambda)} \geq 1, \text{voor de niet - uitbijters} \\ g_{hr} &= 1-\dfrac{1}{1+\lambda} < 1, \text{voor de uitbijters} \end{cases}$$

en

$$\lambda = \sum_{h=1}^{L}\frac{n_h q_h}{(1-f_h)p_h} > 0\,. \qquad (2.3.10)$$

Het is eenvoudig na te gaan dat de gewichten in ieder stratum optellen tot n_h, d.w.z. $r_h g_{hm} + (n_h - r_h)g_{hr} = n_h$. Als er in een bepaald stratum geen uitbijters gevonden zijn, geldt $q_h = 0$ en $g_{hm} = 1$ voor dat stratum en tellen in dit stratum alle elementen dus even zwaar mee.

Het algoritme 2 kan in principe toegepast worden bij een minimale steekproefgrootte van twee elementen per stratum. Het is echter aan te bevelen om een minimale steekproefgrootte van ongeveer tien elementen per stratum te gebruiken. Bij heel kleine strata kan de uitbijterdetectie instabiel worden. Een minimale steekproefgrootte van vijf tot tien elementen per stratum wordt ook vaak in de steekproeftheorie aanbevolen.

2.3.3 Varianten van de eenzijdig gecensureerde schatter

In deze deelparagraaf bespreken we enkele varianten van de eenzijdig gecensureerde schatter.

Eenzijdig censureren met linkeruitbijters

In paragraaf 2.3.2 is de eenzijdig gecensureerde schatter besproken, waarbij alleen de invloed van extreem grote waarnemingen beperkt wordt. Als de data scheef verdeeld zijn met uitbijters aan de linkerkant (extreem kleine waarnemingen), kan een variant van de besproken schatter toegepast worden waarbij de waarnemingen aan de linkerkant door een grenswaarde begrensd worden. Het doel van deze variant is

om de gemiddelde kwadratische fout te minimaliseren als functie van een grenswaarde s van de schatter

$$\overline{Y}_s^{cens} = \frac{(n - r)s + \sum_{j=n-r+1}^{n} y_j}{n}. \tag{2.3.11}$$

Hierin is r het aantal waarnemingen groter of gelijk aan grenswaarde s en zijn er $n - r$ waarnemingen op uitbijter gezet. Verder werkt de methode op dezelfde wijze als de besproken eenzijdig gecensureerde schatters met een rechter grenswaarde t.

Trekken met ongelijke insluitgewichten

In de praktijk worden steekproeven regelmatig getrokken volgens een complex steekproefontwerp, waarbij de waarnemingen getrokken zijn met ongelijke insluitgewichten. Met een eenvoudige aanpassing kunnen de besproken eenzijdig gecensureerde schatters rekening houden met ongelijke insluitgewichten. Stel dat de waarnemingen $y_1, y_2, ..., y_n$ getrokken zijn met insluitgewichten $w_1, w_2, ..., w_n$, zodat $w_1 + w_2 + \cdots + w_n = N$. De gecensureerde schatter wordt dan toegepast op de waarden $z_i = w_i y_i$ in plaats van op y_i. Het feit dat de insluitgewichten verwerkt zijn in de waarden van $z_i = w_i y_i$ heeft tot gevolg dat de vergelijkingen (2.3.1), (2.3.7) en (2.3.9) aangepast moeten worden.

Vergelijking (2.3.1) wordt dan vervangen door

$$\overline{Y}_t^{cens} = \frac{\sum_{j=1}^{r} z_j + (n - r)t}{N}, \tag{2.3.12}$$

maar vergelijking (2.3.5) en algoritme 1 blijven hetzelfde. Een deel van de formules moet aangepast worden omdat het ophooggewicht verwerkt is in z_i. Merk op dat de parameters p, q, m_n en m_r berekend zijn met z_i in plaats van y_i.

In de gestratificeerde situatie wordt vergelijking (2.3.7) vervangen door

$$\overline{Y}_{t_h}^{cens} = \frac{\sum_{j=1}^{r_h} z_{hj} + (n_h - r_h)t_h}{N_h}. \tag{2.3.13}$$

Het stelsel vergelijkingen (2.3.9) wordt vervangen door

$$\begin{cases} (1 - f_1)p_1(t_1 - m_{1m}) - \sum_{h=1}^{L}(n_h - r_h)(m_{hr} - t_h) = 0 \\ \vdots \\ (1 - f_L)p_L(t_L - m_{Lm}) - \sum_{h=1}^{L}(n_h - r_h)(m_{hr} - t_h) = 0 \end{cases}, \tag{2.3.14}$$

waarbij de parameters als functie van z_{hj} zijn berekend. De oplossing van (2.3.14) kan gevonden worden met algoritme 2, het algoritme waarmee ook de oplossing van (2.3.9) berekend wordt. Alleen de startoplossing wordt op een andere manier berekend, namelijk met algoritme 3.

Opmerking: als een steekproefelement y_i dichtbij het stratumgemiddelde ligt, maar een heel groot insluitgewicht w_i, kan het gebeuren dat dit element als uitbijter gezien wordt. Dat komt onlogisch over. De schatter is in dat geval nog steeds goed. Het gewicht is alleen onnodig verkleind. In de praktijk zal deze situatie zich niet gauw voordoen.

Gebruik van hulpinformatie

Als er voor iedere waarneming y_j een (kolom)vector van hulpinformatie $\mathbf{x}_j$ beschikbaar is en deze informatie ook op populatieniveau bekend is, kan de gegeneraliseerde regressieschatter toegepast worden. Dan wordt de gecensureerde schatter toegepast op de residuen $e_j = y_j - \mathbf{x}_j^{\mathrm{t}}\hat{\boldsymbol{\beta}}$ in plaats van op de waarnemingen zelf. Beschouw bijvoorbeeld de gegeneraliseerde regressieschatter, gegeven door

$$\overline{Y}^{\mathrm{reg}} = \overline{Y} + (\overline{\mathbf{X}}_{pop} - \overline{\mathbf{X}}_{steekpr})^{\mathrm{t}}\hat{\boldsymbol{\beta}} = \overline{\mathbf{X}}_{pop}^{t}\hat{\boldsymbol{\beta}} + (\overline{Y} - \overline{\mathbf{X}}_{steekp}^{t}\hat{\boldsymbol{\beta}}) =: \overline{\mathbf{X}}_{pop}^{t}\hat{\boldsymbol{\beta}} + \overline{E} \ .$$

met $\overline{Y}$ het steekproefgemiddelde van de waarnemingen y_j, $\overline{\mathbf{X}}_{steekp}^{t}$ en $\overline{\mathbf{X}}_{pop}^{t}$ het steekproefgemiddelde en het populatiegemiddelde van de hulpinformatie $\mathbf{x}_j$ en $\overline{E}$ het steekproefgemiddelde van de residuen $e_j = y_j - \mathbf{x}_j^{\mathrm{t}}\hat{\boldsymbol{\beta}}$. De gegeneraliseerde regressieschatter wordt nader beschreven in Banning en Knottnerus (2009) of Särndal e.a. (1992).

Dan is de gecensureerde regressieschatter gegeven door

$$\overline{Y}_{t^*}^{\mathrm{reg,cens}} = \overline{\mathbf{X}}_{pop}^{t}\hat{\boldsymbol{\beta}} + \overline{E}_{t^*}^{\mathrm{reg,cens}} \ . \qquad (2.3.15)$$

2.4 Voorbeeld

Als voorbeeld bespreken we de toepassing van de eenzijdig gecensureerde schatter bij de Internationale Diensten. Over deze toepassing is ook een nota verschenen, zie Krieg e.a. (2008).

Door de Internationale Diensten worden kwartaalcijfers geschat voor de in- en uitvoer van bedrijven voor een aantal diensten, waaronder vervoer, computer- en informatiediensten. De respons bestaat voor een groot deel uit bedrijven die voor geen enkele dienst in- of uitvoer rapporteren. Dit zijn de zogenaamde nulwaarnemingen. Een enkele grote waarde in de respons heeft dan al een grote invloed op de schattingen, waardoor toepassing van een uitbijtermethode zinvol is.

Er wordt een gestratificeerde steekproef getrokken, waarbij de strata ingedeeld zijn naar o.a. de standaardbedrijfsindeling en de grootteklasse. In een telefonisch onder-

zoek door BES is vastgesteld dat het bij de nulwaarnemingen vaak om meetfouten gaat. Om hiervoor te corrigeren, worden de insluitgewichten aangepast en komen er verschillende insluitgewichten voor binnen een stratum. Er is daarom gekozen voor de eenzijdig gestratificeerde gecensureerde schatter met aanpassingen voor de ongelijke insluitgewichten.

Een andere aanpassing betreft het niet meenemen van de nulwaarnemingen bij de uitbijterdetectie. Theoretisch gezien zouden deze waarnemingen wel moeten worden meegenomen, omdat ze tot de steekproef behoren. Om praktische redenen is ervoor gekozen dit toch niet te doen. Zo is het volledige bestand met de nulwaarnemingen, dat erg groot is, in het productieproces niet meer beschikbaar en zou het speciaal bewaard moeten worden voor de uitbijterdetectie.

Als voorbeeld nemen we het eerste kwartaal van 2007. Het steekproefkader bestaat uit 39523 bedrijven en de steekproef bevat 4108 bedrijven (zonder onterechte nulwaarnemingen). De gecensureerde schatter is op de hierboven beschreven manier toegepast op 11 diensten, elke dienst uitgesplitst naar invoer en uitvoer (invoer en uitvoer worden stromen genoemd). De schattingen zijn dus optimaal voor deze 22 doelvariabelen, maar niet voor aggregaten of andere uitsplitsingen. De totale handel wordt dan geschat op 5,88 miljard euro. Als de invloed van de uitbijters niet beperkt wordt dan is de schatting 6,14 miljard euro. De reductie in de schattingen door rekening te houden met uitbijters is dus gemiddeld over alle stroom-dienstcombinaties 4,22%. In totaal zijn er 30 uitbijters gevonden. Voor de meeste stroom-dienstcombinaties is er één uitbijter gevonden, in sommige gevallen zijn het er ook 2 tot maximaal 4.

2.5 Kwaliteitsindicatoren

Aangezien het doel van de gecensureerde schatters is om de gemiddelde kwadratische fout te minimaliseren, ligt het voor de hand de gemiddelde kwadratische fout van de gecensureerde schatter te onderzoeken. In Renssen e.a. (2004) wordt een analytische benadering gegeven voor de schatting van de gemiddelde kwadratische fout, waarbij de variantie geschat wordt met behulp van invloedsfuncties. De invloedsfunctie is een robuustheidsmaat, die de invloed van een specifieke waarneming op de schatter beschrijft. De variantie van de schatter kan benaderd worden door de verwachting te nemen van het kwadraat van de invloedsfunctie (Hampel e.a., 1986). De nauwkeurigheid van deze benaderingsformule hangt af van de steekproefgrootte. Dit is vergelijkbaar met de nauwkeurigheid van variantieschattingen voor andere schatters, zoals voor de gegeneraliseerde regressieschatter. In Renssen e.a. (2004) is voor een aantal varianten van de eenzijdig gecensureerde schatter de

invloedsfunctie afgeleid en is voor enkelvoudig aselecte steekproeven een formule afgeleid voor de schatting van de gemiddelde kwadratische fout.

De gemiddelde kwadratische fout van de gecensureerde schatter kan vergeleken worden met die van schatters die niets met uitbijters doen, in deze context ook wel directe schatters genoemd. Behalve naar de gemiddelde kwadratische fout kan ook alleen gekeken worden naar de vertekening van de gecensureerde schatter. Deze kan in de praktijk alleen benaderd worden als het verschil tussen de schattingen die via de directe schatter en via de gecensureerde schatter verkregen worden. Het is aan de gebruiker om in te schatten of de vertekening acceptabel is. Hierbij moet opgemerkt worden dit een heel onnauwkeurige (maar zonder externe informatie enig mogelijke) benadering van de vertekening is. Een grote benaderde vertekening van de gecensureerde schatter geeft wel aan dat de variantie van zowel de directe schatter als de gecensureerde schatter groot is.

Een andere goede manier om te onderzoeken of de gecensureerde schatters een lagere gemiddelde kwadratische fout hebben dan de directe schatters is het uitvoeren van een simulatiestudie. Het is dan wel belangrijk om te werken met realistische populatiedata om betrouwbare resultaten te krijgen. In paragraaf 4.5 wordt iets meer gezegd over de eisen aan realistische populatiedata.

3. Tweezijdig gecensureerde schatters

3.1 Korte beschrijving

Met de tweezijdig gecensureerde schatter kan een populatiegemiddelde geschat worden, waarbij de invloed van zowel extreem lage als extreem hoge waarnemingen beperkt wordt. De tweezijdig gecensureerde schatter is een aantal jaren geleden op het CBS ontwikkeld, zie Renssen e.a. (2004) en Krieg e.a. (2004). Bij de tweezijdig gecensureerde schatter worden zowel de heel grote als ook de heel kleine extreme waarnemingen aangepast. Dit gebeurt op een soortgelijke manier als bij de eenzijdig gecensureerde schatter. De formules en algoritmes worden wel complexer.

3.2 Toepasbaarheid

Toepassing van de tweezijdig gecensureerde schatter kan interessant zijn wanneer aan beide kanten van de verdeling uitbijters voorkomen. De tweezijdig gecensureerde schatter wijst net als de eenzijdig gecensureerde schatter slechts weinig waarnemingen als uitbijter aan. Ook hier blijkt dat uit de praktijk. Bij een (min of meer) symmetrische verdeling kan het gunstiger zijn om links en rechts op een symmetrische manier meer waarnemingen als uitbijter aan te wijzen dan de tweezijdig gecensureerde schatter doet.

We werken de tweezijdig gecensureerde schatter uit voor enkelvoudige aselecte steekproeven, waarbij zonder terugleggen getrokken is. Voor gestratificeerde steekproefontwerpen wordt de tweezijdig gecensureerde schatter ingewikkelder en is er geen bewijs dat de besproken algoritmes gegarandeerd convergeren. In de praktijk convergeren de algoritmes meestal wel naar een geschatte optimale grenswaarde. Voor toepassing van tweezijdig censureren op gestratificeerde en andere complexe steekproefontwerpen is meer onderzoek nodig.

Tweezijdig censureren wordt op dit moment nog niet toegepast op het CBS.

3.3 Uitgebreide beschrijving

3.3.1 Tweezijdig gecensureerde schatter voor enkelvoudig aselecte steekproeven

In deze deelparagraaf bespreken we de tweezijdig gecensureerde schatter voor enkelvoudig aselecte steekproeven. Uit een populatie met N elementen is zonder terugleggen een enkelvoudig aselecte steekproef van n elementen getrokken. Stel dat we van doelvariabele y het populatiegemiddelde willen schatten, waarbij we de invloed van zowel linkeruitbijters als rechteruitbijters willen beperken. Voor alle elementen i met $1 \le i \le n$ in de steekproef is de doelvariabele y waargenomen. De steekproefelementen worden zo gesorteerd dat $y_1 \le y_2 \le \cdots \le y_n$. We gaan

twee grenswaarden berekenen, een linker grenswaarde s en een rechter grenswaarde t.

Als beide grenswaarden s en t met $s \le t$ gegeven zijn, dan gaat het tweezijdig censureren als volgt. Een waarneming y_i groter dan de grenswaarde t wordt aangepast en vervangen door t en de waarnemingen kleiner dan s worden vervangen door s. Vervolgens wordt het gemiddelde van de populatie geschat door het steekproefgemiddelde te berekenen van de aangepaste waarnemingen. Stel dat er r niet-aangepaste waarnemingen zijn, dat er n_l waarnemingen zijn kleiner dan s en dat er n_r waarnemingen groter zijn dan t. Het gemiddelde wordt dan geschat door

$$\overline{Y}_{s,t}^{cens} = \frac{\sum_{j=1}^{r} y_{n_l+j} + n_l s + n_r t}{n}, \tag{3.3.1}$$

met $s \le y_{n_l+j} \le t$ voor $j = 1, \ldots, r$. Er zijn dus $n - r = n_l + n_r$ waarnemingen op uitbijter gezet.

Net als bij de eenzijdig gecensureerde schatters kan ervoor gekozen worden om de gemiddelde kwadratische fout van (3.3.1) als functie van s en t te minimaliseren. Maar als s en t gelijk worden genomen aan het populatiegemiddelde, dan zijn zowel de variantie als de vertekening van (3.3.1) gelijk aan nul. De beste schatting wordt dus gevonden door zowel s als t gelijk te nemen aan het populatiegemiddelde. Als het populatiegemiddelde bekend zou zijn, dan was dit inderdaad optimaal. In de praktijk is het populatiegemiddelde echter niet bekend en moeten deze optimale grenswaarden geschat worden met een steekproef. Dan wordt voor s en t het steekproefgemiddelde genomen en valt de gecensureerde schatter samen met het steekproefgemiddelde. Deze schatter heeft geen toegevoegde waarde.

In plaats daarvan wordt in Renssen e.a. (2004) voorgesteld om een aangepaste uitdrukking van de gemiddelde kwadratische fout te minimaliseren. In deze uitdrukking is één term uit de formule voor de gemiddelde kwadratische fout weggelaten. Een schatting voor de grenswaarden s^* en t^*, die optimaal zijn voor deze aangepaste uitdrukking, kan gevonden worden door het volgende stelsel van vergelijkingen op te lossen:

$$\begin{cases} \frac{1-f}{n}[p(\mu_m - s) + q_r(t-s)] - q_l(s-\mu_l) = 0 \\ \frac{1-f}{n}[p(t-\mu_m) + q_l(t-s)] - q_r(\mu_r - t) = 0 \end{cases}, \tag{3.3.2}$$

met $f = \frac{n}{N}$, $p = \frac{r}{n}$, $q_l = \frac{n_l}{n}$, $q_r = \frac{n_r}{n}$, $\mu_m = \frac{1}{r}\sum_{j=1}^{r} y_{n_l+j}$, $\mu_l = \frac{1}{n_l}\sum_{j=1}^{n_l} y_j$, $\mu_r = \frac{1}{n_r}\sum_{j=1}^{n_r} y_{n-n_r+j}$.

Er is altijd een unieke oplossing van (3.3.2) waarbij in alle gevallen ten minste één linkeruitbijter en één rechteruitbijter aangewezen wordt (zie Krieg e.a., 2004). Voor de tweezijdig gecensureerde schatter geldt net als voor de eenzijdig gecensureerde schatter dat de toepassing ervan niet nuttig is bij integrale waarneming. Toepassing van formule (3.3.2) zou dan grenswaarden leveren die gelijk zijn aan de kleinste en grootste waarneming in de steekproef.

Boven is opgemerkt dat niet de gemiddelde kwadratische fout maar een andere uitdrukking geminimaliseerd is. Dit lijkt in eerste instantie op een noodgreep. Het resultaat is echter een schatter die ongeveer neerkomt op het tegelijkertijd links en rechts toepassen van de eenzijdig gecensureerde schatter, dat wil zeggen het is vergelijkbaar met links censureren waarbij het resultaat van rechts censureren al bekend is en tegelijk ook rechts censureren waarbij het resultaat van links censureren al bekend is. Deze schatter maakt geen veronderstellingen over symmetrie en is daarom geschikt voor situaties waar de data niet symmetrisch verdeeld zijn. De invloed van uitbijters wordt niet te sterk beperkt, waardoor de hierdoor veroorzaakte vertekening klein blijft, vergelijkbaar met die van de eenzijdig gecensureerde schatter. Als bekend is dat de data symmetrisch verdeeld zijn, dan is de gemaakte keuze niet optimaal.

De tweezijdig gecensureerde schatter kan toegepast worden bij een steekproefgrootte van minimaal 2 elementen. Voor heel kleine steekproeven is het echter niet aan te bevelen om überhaupt schattingen te berekenen, omdat deze een heel grote variantie hebben. Dit geldt ook voor de gecensureerde schatters.

Algoritme 4 in de bijlage vindt de oplossing van (3.3.2). In dit algoritme worden systematisch alle mogelijke combinaties van aantallen linkeruitbijters en rechteruitbijters geprobeerd tot de unieke oplossing voor (3.3.2) gevonden is. De volgorde van proberen is in principe willekeurig. Het is wel gunstig om met weinig uitbijters te beginnen (zoals in stap 4 in het algoritme) omdat dan de oplossing vrij snel gevonden wordt.

Door de gevonden optimale grenswaarden s^* en t^* in te vullen in vergelijking (3.3.1), vinden we de tweezijdig gecensureerde schatting $\overline{Y}^{\text{cens}}_{s^*,t^*}$ voor het populatiegemiddelde.

Tweezijdig gecensureerde schatter schrijven met gewichten. Net als de eenzijdig gecensureerde schatter kan ook de tweezijdig gecensureerde schatter geschreven worden als een gewogen gemiddelde, waarbij de uitbijters een lager gewicht krijgen en de niet-uitbijters een hoger gewicht, d.w.z.

$$\overline{Y}^{\text{cens}}_{s^*,t^*} = \frac{1}{n}\left[\sum_{j=1}^{n_l} g_l y_j + \sum_{j=n_l+1}^{n_l+r} g_m y_j + \sum_{j=n-n_r+1}^{n} g_r y_j\right], \text{ met}$$

$$\begin{cases} g_l & < \quad 1, \text{ voor de linkeruitbijters} \\ g_m & \geq \quad 1, \text{ voor de niet - uitbijters} \quad \text{en} \\ g_r & < \quad 1, \text{ voor de rechteruitbijters} \end{cases}$$

$$n_l g_l + n_m g_m + n_r g_r = n\,. \tag{3.3.3}$$

Voor de exacte uitdrukkingen van g_l, g_m en g_r verwijzen we naar Smeets en Krieg (2004).

3.3.2 Tweezijdig gecensureerde schatter voor andere steekproefontwerpen

Als de tweezijdig gecensureerde schatter toegepast wordt op andere steekproefontwerpen, moet de schatter daarvoor aangepast worden. Voor de tweezijdig gecensureerde schatter moet er nog verder methodologisch onderzoek gedaan worden naar deze aanpassingen. In Krieg e.a. (2004) wordt de tweezijdig gecensureerde schatter uitgewerkt voor gestratificeerde steekproeven zonder terugleggen. Voor deze schatter zijn er echter nog open vragen. Zo is niet duidelijk of en onder welke voorwaarden het stelsel vergelijkingen, dat opgelost moet worden om de geschatte optimale grenswaarden s en t te vinden, een unieke oplossing heeft. Ook is er geen algoritme bekend dat gegarandeerd de optimale oplossing vindt. Het gevonden algoritme werkt wel in de meeste gevallen. Als men de tweezijdig gecensureerde schatter in de praktijk wil toepassen, dan zou dit algoritme geprogrammeerd kunnen worden met een controle of het werkt. Voor het (zeldzame) geval dat dit niet het geval is, kan het algoritme voor eenzijdig censureren toegepast worden, aan de linkerkant en aan de rechterkant (achter elkaar). Dit voorstel moet nog goed onderzocht worden voordat het in de praktijk toegepast kan worden.

Er is nog niet voldoende onderzoek gedaan om de tweezijdig gecensureerde schatter in andere situaties toe te passen.

3.4 Voorbeeld

Op dit moment wordt de tweezijdig gecensureerde schatter nog niet toegepast op het CBS. In de simulatiestudie van Krieg en Smeets (2005) is wel een toepassing van de tweezijdig gecensureerde schatter onderzocht op de productiestatistieken. In deze simulatiestudie is de tweezijdig gestratificeerde gecensureerde schatter toegepast, waarbij de grenswaarden s en t geschat worden met de steekproef. Hierbij is de gegeneraliseerde regressieschatter gebruikt, waarbij de gecensureerde schatter toegepast is op de residuen. Deze variant van de tweezijdig gecensureerde schatter moet nog beter onderzocht worden voordat het in de Methodenreeks beschreven kan worden. Deze schatter wordt vergeleken met de eenzijdig gecensureerde schatter (hoofdstuk 2) en de directe schatter. De directe schatter past geen gewichten of waarden aan van uitbijters. Voor de eenzijdig gecensureerde schatter is de gestratificeerde eenzijdig gecensureerde schatter onderzocht, waarbij de grenswaarde t met de steekproef geschat wordt.

Voor de simulatiestudie is een populatiebestand gecreëerd door steekproefdata van de jaren 2000, 2001 en 2002 en een drietal kerncellen binnen de horeca (restaurants,

cafetaria's en cafés) samen te voegen. Dit populatiebestand wordt in de simulatie beschouwd als populatie voor één kerncel (de restaurants) waaruit steekproeven getrokken worden om schattingen voor deze kerncel te maken. Er zijn 10000 steekproeven getrokken, waarmee de variantie, de vertekening en de gemiddelde kwadratische fout van de verschillende schatters bepaald zijn.

In de PS wordt de gegeneraliseerde regressieschatter toegepast met BTW als hulpinformatie. In een eerste simulatiestudie is de BTW-informatie genegeerd en is een steekproef van $n = 181$ elementen getrokken. In onderstaande tabel staan de resultaten van de simulatie.

Tabel 2. Simulatieresultaten van eenzijdig en tweezijdig censureren zonder hulpinformatie

Schatter	Vertekening	Variantie	MSE
Direct	0	325	325
Eenzijdig gecensureerd	-4,7	249	271
Tweezijdig gecensureerd	-3,2	258	268

We zien dat zowel de eenzijdig gecensureerde schatter als de tweezijdig gecensureerde schatter het beter doet dan de directe schatter en dat de tweezijdig gecensureerde schatter iets beter is dan de eenzijdig gecensureerde schatter.

In een tweede simulatie zijn alleen de elementen geselecteerd, waarvoor BTW-informatie beschikbaar was en is er een steekproef getrokken van $n = 351$ elementen getrokken. In onderstaande tabel staan de resultaten van deze simulatie.

Tabel 3. Simulatieresultaten van eenzijdig en tweezijdig censureren met hulpinformatie

Schatter	Vertekening	Variantie	MSE
Direct	-0,1	27	27
Eenzijdig gecensureerd	-1,3	25	26
Tweezijdig gecensureerd	-0,3	24	24

Ook in deze situatie zijn de gecensureerde schatters iets beter dan de directe schatter en is de tweezijdig gecensureerde schatter beter dan de eenzijdig gecensureerde schatter.

3.5 Kwaliteitsindicatoren

Ook voor tweezijdig gecensureerde schatters ligt het voor de hand om de gemiddelde kwadratische fout te onderzoeken. De variantie kan geschat worden door gebruik te maken van invloedsfuncties (Hampel e.a., 1986). De nauwkeurigheid van deze

benaderingsformule hangt af van de steekproefgrootte. Dit is vergelijkbaar met de nauwkeurigheid van variantieschattingen voor andere schatters, zoals voor de gegeneraliseerde regressieschatter. In Renssen e.a. (2004) wordt voor de tweezijdig gecensureerde schatter toegepast op een enkelvoudig aselecte steekproef de invloedsfunctie afgeleid. Voor andere versies van de tweezijdig gecensureerde schatter zijn er nog geen invloedsfuncties berekend.

De gemiddelde kwadratische fout van de gecensureerde schatter kan vergeleken worden met die van directe schatters. Behalve naar de gemiddelde kwadratische fout kan ook gekeken worden naar de vertekening van de gecensureerde schatter. Deze kan in de praktijk alleen benaderd worden als verschil tussen de schattingen die via de directe schatter en via de gecensureerde schatter verkregen worden. Het is aan de gebruiker om in te schatten of de vertekening acceptabel is. Hierbij moet opgemerkt worden dat een grote vertekening van de gecensureerde schatter betekent dat de variantie van zowel de directe schatter als ook de gecensureerde schatter groot is.

Een andere goede manier om te onderzoeken of de gecensureerde schatters een lagere gemiddelde kwadratische fout hebben dan de directe schatters is het uitvoeren van een simulatiestudie. Het is dan wel belangrijk om te werken met realistische populatiedata om betrouwbare resultaten te krijgen. In paragraaf 4.5 wordt iets meer gezegd over de eisen aan realistische populatiedata.

4. De PS-methode

4.1 Korte beschrijving

Voor de productiestatistieken (PS) is in 2001 een uitbijtermethode ontwikkeld en in productie genomen (zie Vlag e.a., 2001, JVGN, 2001 en Nieuwenbroek en Vlag, 2001). De methode is ontwikkeld voor de ophoging zoals die bij de productiestatistieken toegepast wordt. We noemen deze methode hier PS-methode, de methode wordt ook wel de methode Pieter Vlag genoemd. Bij de ophoging van de PS staat de weegcel centraal. Per weegcel wordt voor elk element in de steekproef een residu als afwijking van de norm berekend. Wat de norm is, hangt af van de situatie. In paragraaf 4.3 wordt beschreven hoe de norm gedefinieerd is.

Als de absolute waarde van een residu te groot is, dus m.a.w. als een residu te sterk afwijkt van de norm, dan wordt het element in de steekproef aangewezen als uitbijter en krijgt het gewicht 1. Het te groot zijn van de absolute waarde wordt getest via een grenswaarde *cut*: als de absolute waarde groter is dan *cut*, dan wordt het element aangewezen als uitbijter. De grenswaarde *cut* is gedefinieerd als

$$cut = upper_f + c(upper_f - lower_f), \tag{4.1.1}$$

met $upper_f$ en $lower_f$ het derde en eerste kwartiel van de absolute waarden van de residuen. Voor de parameter c is de waarde 2,5 gekozen.

De hier beschreven berekeningen worden uitgevoerd voor de omzet als de belangrijkste doelvariabele van de productiestatistieken. De aangepaste gewichten worden dan ook gebruikt voor de ophoging van de overige doelvariabelen.

De PS-methode is in de praktijk ontwikkeld waarbij pragmatische keuzes gemaakt zijn. Het is niet mogelijk om voor deze pragmatische keuzes een methodologische onderbouwing te geven. Er is ook geen garantie dat de gemaakte keuzes optimaal zijn. Integendeel, er zijn soms ook verbeteringen mogelijk. Een drietal mogelijke verbeteringen waar onderzoek naar gedaan is, worden in dit document beschreven.

4.2 Toepasbaarheid

Met de methode wordt de invloed van steekproefelementen die sterk afwijken van de overige elementen beperkt door het ophooggewicht op 1 te zetten.

In een simulatiestudie (zie Krieg en Smeets, 2005) is aangetoond dat deze methode tot nauwkeurigere schattingen leidt dan een schatter zonder uitbijterbehandeling. Ook de gecensureerde schatter (zie hoofdstukken 2 en 3) is minder nauwkeurig dan de PS-methode. Dit resultaat geldt in principe alleen voor het bereik van de simulatiestudie, de horeca. Omdat de structuur van de data bij de andere kerncellen waar-

schijnlijk niet sterk afwijkt, is de PS-methode waarschijnlijk ook voor de overige branches van de productiestatistieken een goede methode.

In Krieg en Smeets (2005) is verder aangetoond dat de PS-methode op verschillende punten verbeterd kan worden. De verbeteringen liggen in de keuze van de parameter c en de manier waarop de residuen berekend worden (geen transformatie in de rest-cellen en rekening houden met verschillende insluitgewichten in de BTW-cellen). Ook zonder deze verbeteringen leidt de PS-methode al tot redelijk nauwkeurige schattingen. Daarom wordt in paragraaf 4.3 zowel de op dit moment toegepaste methode als ook de verbeterde methode beschreven. Merk op dat een slechte keuze van de parameter c tot onnauwkeurige schattingen leidt.

De PS-methode leidt tot nauwkeurige schattingen vanwege het feit dat de residuen voor een groot deel van de weegcellen min of meer symmetrisch verdeeld zijn. Hierdoor leidt het aanwijzen van uitbijters nauwelijks tot vertekening, maar wel tot een veel kleinere variantie. Dit is de reden waarom de PS-methode voor de productiestatistieken een goede methode is.

De PS-methode is toegespitst op de ophoging van de productiestatistieken en kan niet zonder meer toegepast worden in andere situaties. Hiervoor zijn aanpassingen van de methode noodzakelijk. Verder is, zonder nader onderzoek, niet bekend of de PS-methode ook in een andere situatie tot goede schattingen zal leiden. In paragraaf 4.5 worden enkele suggesties gedaan hoe dat onderzocht kan worden.

De PS-methode wordt, met kleine aanpassingen, ook toegepast bij de kortetermijn-statistieken (Handboek productie KS-en, 2004, stand van zaken mei 2009). Er is echter nooit onderzocht of de methode daar ook tot nauwkeurige schattingen leidt. Het doel van de kortetermijnstatistieken is het meten van een ontwikkeling, terwijl het bij de productiestatistieken om niveaucijfers gaat. Verder is bij de kortetermijn-statistieken niet onderzocht of eventuele symmetrie in de data ook leidt tot nauw-keurigere schattingen. Het gebruik van de PS-methode voor de kortetermijnstatistie-ken is daarom geen valide methode.

In de periode 2007-2008 is onderzoek gedaan of de kortetermijnstatistieken geba-seerd kunnen worden op BTW-informatie (zie De Wolf en Van Bemmel, 2007). Ook hierbij is gewerkt met een aangepaste versie van de PS-methode. Door ontwik-kelingen bij de belastingdienst lijkt het erop dat de BTW niet meer in voldoende mate beschikbaar is en kan deze nieuwe opzet waarschijnlijk niet doorgaan (stand van zaken mei 2009).

4.3 Uitgebreide beschrijving

4.3.1 Steekproefontwerp en ophogen bij Impect 1

Zoals al gezegd is de PS-methode specifiek voor de productiestatistieken ontwikkeld. Bij de beschrijving van de methode wordt daarom gebruikgemaakt van de begrippen die bij de productiestatistieken gebruikt worden. Verder sluit de PS-methode aan bij de steekproef en de ophoogmethode. Voor het goede begrip volgt hier een korte uitleg van de productiestatistieken. Voor meer informatie wordt verwezen naar Resing e.a. (2005).

Het doel van de productiestatistieken is de publicatie van jaarcijfers over het Nederlandse bedrijfsleven. De publicaties worden gemaakt voor kerncellen. Een kerncel bestaat uit enkele SBI's. Per kerncel wordt een gestratificeerde steekproef getrokken, waarbij gestratificeerd wordt naar grootteklasse. Vanaf grootteklasse 6 wordt er integraal waargenomen. De ophoging, en daarom ook de uitbijterbehandeling, richt zich op grootteklassen 0 t/m 5. In sommige kerncellen, de zogenaamde PS-plus cellen, wordt een deel van de grootteklassen niet waargenomen, maar via fiscale informatie bijgeschat. De PS-methode wordt daar op dezelfde manier toegepast, maar dan beperkt op de cellen die wel via een steekproef worden geschat.

Er worden naast cijfers op kerncelniveau ook cijfers voor hogere aggregaten en op een gedetailleerder niveau gepubliceerd. Voor deze schattingen worden de gewichten gebruikt die voor de schattingen op kerncelniveau vastgesteld zijn. De simulatie (Krieg en Smeets, 2005) was alleen gericht op kerncellen. Het is dus niet bekend of de PS-methode ook voor andere aggregaten tot goede schattingen leidt. Vermoedelijk zou voor andere aggregaten een andere parameter c optimaal zijn. Ook voor het consistent schatten voor verschillende aggregatieniveaus is op het CBS geen valide methodologie ontwikkeld (zie ook hoofdstuk 5).

Voor de uitbijterdetectie wordt elke kerncel opgesplitst in weegcellen. Er wordt daarbij grotendeels dezelfde indeling gebruikt als bij de schattingsprocedure. Bedrijven waarvoor geen BTW-informatie beschikbaar is, worden ingedeeld in de restcellen, waarbij er voor elke grootteklasse één restcel is. Bedrijven waarvoor wel BTW-informatie beschikbaar is, worden ingedeeld in BTW-cellen. Hierbij vormen grootteklassen 0 t/m 3 één weegcel en grootteklassen 4 en 5 de andere weegcel. Bij het indelen van de kerncel in weegcellen worden minimale celvullingen aangehouden. Als daaraan niet voldaan kan worden, vindt samenvoeging van de weegcellen plaats. Hoe precies wordt samengevoegd staat bijvoorbeeld uitgewerkt in Nieuwenbroek en Vlag (2001). Bij de ophoging worden restcellen voor de grootteklassen lager dan 4 nog verder onderverdeeld naar rechtsvorm: Rechtspersoon (RP) en Natuurlijke persoon (NP).

In de BTW-cellen wordt de gegeneraliseerde regressieschatter toegepast met BTW als hulpinformatie. Via de gegeneraliseerde regressieschatter worden de insluitge-

wichten zodanig aangepast dat het geschatte BTW-totaal precies overeenkomt met de hiervoor bekende populatietotalen. Daarnaast komt ook het geschatte aantal bedrijven overeen met het bekende aantal bedrijven in de populatie. Door het toepassen van de gegeneraliseerde regressieschatter wordt gecorrigeerd voor een eventueel scheve steekproef en wordt de precisie van de schatting verbeterd. Voor een uitvoerige beschrijving van de gegeneraliseerde regressieschatter zie Banning en Knottnerus (2009) of Särndal e.a. (1992).

4.3.2 Uitbijterdetectie in de BTW-cellen, actuele methode

In de weegcel zijn n elementen waargenomen. Gegeven zijn de waarnemingen $y_1,...,y_n$ voor alle steekproefelementen. Verder is voor alle steekproefelementen hulpinformatie (in dit geval BTW-informatie) $x_1,...,x_n$ gegeven. Daarnaast is er ook hulpinformatie over de populatie gegeven. Voor deze elementen wordt de regressielijn berekend en elementen waarvan de absolute waarde van het residu groot is, die dus ver weg liggen van de regressielijn, worden als uitbijter aangewezen. Omdat ook de regressielijn gevoelig is voor uitbijters, wordt er eerst een robuuste regressielijn berekend. Elementen met een grote absolute waarde van het residu ten opzichte van deze robuuste regressielijn worden buiten beschouwing gelaten bij de berekening van de regressielijn in de volgende stap. Het algoritme is beschreven in de bijlage (zie algoritme 5). Door het uitvoeren van dit algoritme wordt de doelvariabele y_i voor sommige elementen aangepast en worden er uitbijters aangewezen. Merk op dat de oorspronkelijke waarde bewaard moet worden, omdat die later nog nodig is.

In de algoritmen in deze paragraaf moeten medianen en kwartielen berekend worden. Deze zijn in eerste instantie alleen gedefinieerd als het aantal elementen oneven is (voor de mediaan), of als het aantal elementen plus 1 deelbaar is door 4 (kwartielen). In andere gevallen worden deze waarden berekend door interpolatie (zie JVGN, 2001). Ook moeten voor de robuuste regressielijn de data verdeeld worden in drie even grote groepen. Als het aantal elementen niet door drie deelbaar is, wordt één van de groepen iets groter of kleiner dan de andere twee groepen. Voor de precieze afleiding van de groepen zie Vlag e.a. (2001).

4.3.3 Uitbijtermethode in de BTW-cellen, verbeterde methode

Voor de elementen in de steekproef zijn ook de insluitgewichten $w_1,...,w_n$ bekend. Deze zijn binnen een weegcel niet allemaal aan elkaar gelijk, omdat de insluitkansen per grootteklasse verschillen en in een weegcel meerdere grootteklassen verenigd zijn. De invloed van een element i in de steekproef wordt bepaald door de grootte van de doelvariabele y_i maar ook door de grootte van het gewicht w_i. Daarom kan de methode die beschreven is in paragraaf 4.3.2 verbeterd worden door als input niet $y_1,...,y_n$ te gebruiken maar $z_1,...,z_n$ met $z_i = w_i y_i$. In de simulatiestudie van

Krieg en Smeets (2005) is aangetoond dat daardoor de nauwkeurigheid van de schattingen verbeterd wordt.

Deze verbetering is niet in productie genomen.

4.3.4 Uitbijterdetectie in de restcellen, actuele methode

In de weegcel zijn n elementen waargenomen. Gegeven zijn de waarnemingen $y_1,...,y_n$ voor alle steekproefelementen. Ook voor deze elementen worden er residuen bepaald. Deze zijn gedefinieerd als het verschil tussen de waarde van het element zelf en de mediaan van alle elementen. De cutoff-waarde wordt niet direct voor deze residuen berekend, maar op getransformeerde waarden. Het algoritme is te vinden in de bijlage (Algoritme 6). Door het uitvoeren ervan wordt de doelvariabele y_i voor sommige elementen aangepast. Merk op dat ook hier de oorspronkelijke waarde bewaard moet worden, omdat die later nog nodig is.

4.3.5 Uitbijtermethode in de restcellen, verbeterde methode

De uitbijtermethode in de restcellen kan op twee manieren verbeterd worden.

Ten eerste zijn ook voor de elementen in de restcellen de insluitgewichten $w_1,...,w_n$ bekend. De methode die beschreven is in paragraaf 4.3.4, kan daarom aangepast worden door als input niet $y_1,...,y_n$ te gebruiken maar $z_1,...,z_n$ met $z_i = w_i y_i$. Dit is alleen dan een verbetering als de gewichten binnen een weegcel niet allemaal gelijk zijn. Dit gebeurt als verschillende grootteklassen in een weegcel bij elkaar genomen worden.

De tweede verbetering houdt in dat de logaritmische transformatie niet uitgevoerd wordt en dat er geen absolute waardes berekend worden. Dat betekent dat $upper_f$ en $lower_f$ berekend worden als het derde en eerste kwartiel van de residuen ε_i. Vervolgens worden de residuen beperkt via

$$\varepsilon_i^{'} = \begin{cases} -cut & als \quad \varepsilon_i < -cut \\ \varepsilon_i & als \quad -cut \le \varepsilon_i \le cut \\ cut & als \quad \varepsilon_i > cut. \end{cases} \tag{4.3.1}$$

Ten slotte worden de *y*-waarden weer aangepast met $y_i^{'} = m_r + \varepsilon_i^{'}$ en wordt het i'de element als uitbijter aangewezen als $y_i \neq y_i^{'}$.

Deze verbeteringen zijn niet in productie genomen.

4.3.6 Minimale celvulling

Bij de productiestatistieken is een minimale celvulling vereist voor het doorvoeren van het uitbijteralgoritme. Als er in een weegcel minder dan 5 elementen zijn, dan

wordt het algoritme voor deze weegcel niet uitgevoerd en worden er geen uitbijters opgespoord.

Aan de ene kant is het een goede beslissing om de schattingen van medianen, kwartielen en regressielijnen niet te baseren op te weinig elementen. Aan de andere kant kunnen juist bij heel kleine steekproeven grote waarnemingen een grote invloed hebben en kan het aanwijzen van een dergelijke waarde als uitbijter de nauwkeurigheid van de schatting verbeteren. Bij de productiestatistieken is het mogelijk om handmatig uitbijters aan te wijzen (zie paragraaf 4.3.8), wat dit probleem op een pragmatische manier opvangt.

4.3.7 *Keuze van de parameter c*

Oorspronkelijk was $c = 1{,}5$ gekozen. Uit de simulatiestudie van Krieg en Smeets (2005) blijkt dat deze waarde te klein is en destijds is voorgesteld deze parameter te verhogen tot $c = 2{,}5$. Dat advies is overgenomen.

Uit de simulatiestudie blijkt ook dat de PS-schatter tot redelijk nauwkeurige schattingen leidt als c ergens binnen een vrij breed bereik gekozen wordt. Het is dus niet noodzakelijk om c nauwkeurig te bepalen. Het goede bereik is echter afhankelijk van de situatie (steekproefgrootte en verdeling van de residuen). Een te kleine c kan tot een sterk vertekende schatting leiden. Dat risico is beperkt door $c = 2{,}5$ te kiezen in plaats van $c = 1{,}5$, zowel voor de BTW-cellen als ook in de restcellen.

De keuze van $c = 2{,}5$ is een pragmatische keuze. Meer onderzoek zou kunnen uitwijzen welke parameter in welke kerncellen optimaal is. Dit is echter heel tijdrovend.

Ten slotte moet nog opgemerkt worden dat $c = 2{,}5$ een slechte keuze zou zijn als de transformatie in de restcellen niet doorgevoerd wordt (deze mogelijke verbetering is beschreven in paragraaf 4.3.5). In dat geval moet voor de restcellen een andere parameter gekozen worden dan voor de BTW-cellen.

4.3.8 Handmatige uitbijterdetectie bij de productiestatistieken

Behalve de automatische uitbijterdetectie zoals beschreven in de paragrafen 4.3.2 en 4.3.4 worden bij de productiestatistieken ook nog handmatig uitbijters aangewezen. Soms worden ook automatisch aangewezen uitbijters handmatig op niet-uitbijter gezet.

Algemeen kan gezegd worden dat men heel terughoudend moet zijn met handmatige aanpassingen in de uitbijterdetectie. Het automatische systeem heeft als doel de nauwkeurigheid van de schattingen te optimaliseren. Handmatige aanpassingen kunnen deze optimalisatie verstoren.

Toch zijn er twee redenen waarom de handmatige uitbijterdetectie noodzakelijk kan zijn. Ten eerste wordt bij kleine cellen geen automatische uitbijterdetectie toegepast.

Ten tweede is de uitbijterdetectie alleen gericht op de totale omzet als belangrijkste doelvariabele. Een extreme waarde voor een andere doelvariabele kan echter ook een grote en storende invloed hebben, die alleen via handmatige uitbijterdetectie beperkt kan worden.

4.3.9 Aanpassen van de gewichten van zowel uitbijters als niet-uitbijters

Nadat de uitbijters gedetecteerd zijn, worden de gewichten van de elementen aangepast. De gewichten van de uitbijters worden op 1 gezet om de invloed van de uitbijters te beperken. De gewichten van de niet-uitbijters worden daarom iets verhoogd zodat de gewichten weer optellen tot de populatieomvang. Het aanpassen gebeurt per stratum en dus per grootteklasse. Merk op dat er niet gestratificeerd is naar het beschikbaar zijn van BTW-informatie. Een stratum kan daarom zowel elementen uit een BTW-cel als elementen uit een restcel bevatten. Stel dat er in een stratum m elementen in de steekproef zijn waargenomen. De insluitgewichten $w_1,...,w_m$ worden nu aangepast. Er worden correctiegewichten $g_1,...,g_m$ berekend. De steekproefelementen die volgens paragrafen 4.3.2 (of 4.3.3), 4.3.4 (of 4.3.5) en 4.3.8 aangewezen zijn als uitbijter, krijgen correctiegewicht $g_i = 1.$ Het aantal uitbijters in het stratum is r. Vervolgens worden de gewichten van de overige elementen aangepast met de volgende formule:

$$g_i = \frac{M - r}{m - r} \tag{4.3.2}$$

met M de populatiegrootte van het stratum.

4.3.10 Andere verbetermogelijkheden

In dit document zijn drie verbetermogelijkheden van de PS-methode beschreven, die onderzocht zijn in de simulatiestudie (Krieg en Smeets, 2005):

- Kiezen van een andere parameterwaarde,
- Geen transformatie in de restcellen,
- Rekening houden met verschillende insluitgewichten.

Het is mogelijk dat de methode ook op andere manieren verbeterd kan worden, bijvoorbeeld door aanpassingen in de formule voor de cutoff-waarde (4.1.1). Dit is echter niet onderzocht.

4.4 Voorbeeld

De PS-methode wordt elk jaar toegepast bij de productiestatistieken. In de simulatiestudie van Krieg en Smeets (2005) zijn de eigenschappen van de methode onderzocht en is de methode in verschillende situaties toegepast. In de simulatiestudie zijn ook de eenzijdig en tweezijdig gecensureerde schatters toegepast, zie de hoofdstuk-

ken 2 en 3, en in het bijzonder 3.4. Daar is ook de opzet van de simulatie beschreven.

In een eerste simulatie is de BTW-informatie genegeerd en is de methode toegepast die in de paragrafen 4.3.4 en 4.3.5 beschreven is. Er is een gestratificeerde steekproef van $n = 181$ elementen getrokken. In de volgende tabel staan de resultaten van de simulatie:

Tabel 4. Simulatieresultaten voor data zonder hulpinformatie

Schatter	c	Transformatie	Vertekening	Variantie	MSE
Direct	Nvt	Nvt	-0,01	325	325
PS	1,5	Ja	-7,12	240	290
PS	1,8	Ja	-4,15	258	275
PS	2,4	Ja	-1,57	285	287
PS	2	Nee	-26,07	169	848
PS	4	Nee	-11,84	197	337
PS	8	Nee	-5,05	213	239
PS	12	Nee	-2,79	239	247

Tabel 4 laat zien dat de directe schatter niet vertekend is, maar de PS-schatter wel. Daar staat tegenover dat de variantie van de directe schatter groter is dan de variantie van de PS-schatter. Bij een kleinere parameter c worden er meer waarnemingen op uitbijter gezet, waardoor de vertekening daar groter is en de variantie kleiner. Het optimum, de kleinste gemiddelde kwadratische fout, ligt in deze situatie voor de PS-schatter met transformatie bij $c = 1{,}8$. Als de PS-schatter zonder transformatie toegepast wordt, zou $c = 1{,}8$ tot een onacceptabel grote vertekening leiden. Ook bij $c = 4$ weegt de reductie van de variantie niet op tegen de vergroting van de vertekening. De optimale parameter ligt hier bij $c = 8$. Een veel grotere parameter (bijvoorbeeld $c = 12$) levert acceptabele schattingen, in tegenstelling tot een veel kleinere parameter (bijvoorbeeld $c = 4$).

In een tweede simulatie zijn alleen elementen in het populatiebestand meegenomen waarvoor wel BTW-informatie beschikbaar is en is de methode toegepast die in de paragrafen 4.3.2 en 4.3.3 beschreven staat. Er is een gestratificeerde steekproef van $n = 351$ elementen getrokken. In tabel 5 staan de resultaten van de simulatie. In de kolom Gewichten is aangegeven of er al dan niet gecorrigeerd is voor de insluitgewichten.

Ook uit deze simulatie blijkt dat de directe schatter niet vertekend is. Onafhankelijk van de parameter is de PS-schatter veel nauwkeuriger dan de directe schatter. Alleen als de parameter heel groot gekozen wordt ($c = 20$, of nog groter, zonder rekening te houden met de gewichten) dan wordt de grenswaarde *cut* zo groot dat de PS-schatter nauwelijks meer verschilt van de directe schatter. Merk op dat voor een heel kleine parameter de vertekening ook weer zo groot kan worden dat de PS-schatter slechter is dan de directe schatter.

Tabel 5. Simulatieresultaten voor data met hulpinformatie

Schatter	c	Gewichten	Vertekening	Variantie	MSE
Direct	nvt	Nvt	0,058	27,0	27,0
PS	2	Nee	-0,17	2,5	2,5
PS	4	Nee	-0,619	3,6	4,0
PS	20	Nee	-0,938	14,2	15,0
PS	0,1	Ja	-0,017	0,7	0,7
PS	1	Ja	0,115	0,9	0,9
PS	4	Ja	-0,455	3,1	3,3

Het is opvallend dat in de verbeterde methode, waar wel rekening wordt gehouden met de gewichten, de vertekening niet groter wordt met een kleinere parameter c. Dit komt doordat de data hier symmetrisch verdeeld zijn. In deze situatie kan de parameter c in principe heel klein gekozen worden. Echter, omdat in de praktijk niet bekend is of de populatie inderdaad perfect symmetrisch verdeeld is, is een heel kleine c niet aan te bevelen.

In de simulatiestudie (Krieg en Smeets, 2005) zijn meer simulatieresultaten te vinden.

4.5 Kwaliteitsindicatoren

De PS-methode is als ad-hocmethode voor de productiestatistieken ontwikkeld, waarbij geen bijbehorende kwaliteitsindicatoren ontwikkeld zijn. Er is bijvoorbeeld geen variantieformule voor de schatter afgeleid. In die zin gaat het hier dus niet om een methodologisch goed onderbouwde methode. Omdat de simulatiestudie heeft laten zien dat de schatter toch nauwkeurig is, is hij desondanks opgenomen in de Methodenreeks.

Er zijn wel enkele aandachtspunten die bij toepassing van de PS-methode onderzocht kunnen worden.

1. Symmetrie. Als de populatiedata symmetrisch verdeeld zijn, dan leidt de PS-schatter tot vrij nauwkeurige schattingen, vergeleken met de directe schatter. Hierbij gaat het niet in alle gevallen erom of de doelvariabele symmetrisch verdeeld is. Bijvoorbeeld als de gereneraliseerde regressieschatter toegepast wordt, dan gaat het om de verdeling van de residuen. Als er sprake is van een gestratificeerde steekproef, dan gaat het om de verdeling per stratum.

Om te onderzoeken of de populatiedata symmetrisch verdeeld zijn, moet men uitgaan van de steekproefdata. Merk op dat een uitbijter een over het algemeen symmetrisch beeld kan verstoren.

De verdeling van de steekproefdata kan via een grafiek onderzocht worden, bijvoorbeeld met een histogram in SPSS. Daarnaast kan de scheefheid S berekend worden, die als volgt gedefinieerd is:

$$S = \frac{m_3}{\sqrt{m_2^3}}, \tag{4.5.1}$$

met

$$m_q = \frac{1}{n}\sum_{i=1}^{n}(y_i - m_1)^q \ , \ q = 2,3, \tag{4.5.2}$$

en

$$m_1 = \frac{1}{n}\sum_{i=1}^{n} y_i . \tag{4.5.3}$$

Voor symmetrisch verdeelde data is $S = 0$. Kleine afwijkingen hiervan zijn geen probleem, ten eerste omdat de scheefheid op basis van een steekproef geschat is, ten tweede omdat de schatter niet slecht hoeft te zijn als de data niet perfect symmetrisch verdeeld zijn.

2. Aantal uitbijters. Als de data symmetrisch verdeeld zijn, dan leidt het aanwijzen van veel uitbijters niet tot vertekening maar vaak wel tot een verkleining van de variantie. In dat geval is het acceptabel als veel uitbijters aangewezen worden.

Als de data echter scheef verdeeld zijn, dan leidt het aanwijzen van uitbijters wel tot een vertekening, die de winst in variantie teniet kan doen. Een algemene vuistregel is dat dan per publicatiecel niet meer dan één uitbijter aangewezen moet worden. Bij de PS bestaat een publicatiecel uit BTW-cellen en restcellen. Vanwege symmetrie in de BTW-cellen kunnen daar vrij veel uitbijters aangewezen worden, en geldt deze vuistregel voor de restcellen. Het gaat er dus om dat in alle restcellen samen niet meer dan één uitbijter aangewezen moet worden. Uit de simulatiestudie (Krieg en Smeets, 2005) blijkt dat bij een goed gekozen parameter c er gemiddeld minder dan één uitbijter in de restcellen aangewezen wordt. Ook de ervaringen met de gecensureerde schatter laten zien dat over het algemeen maar heel weinig uitbijters aangewezen moeten worden.

3. Simulatie. Een goede manier om vast te stellen of de PS-schatter een verbetering is ten opzichte van de directe schatter, is het uitvoeren van een simulatiestudie. Hiermee kan ook de optimale parameter c bepaald worden. Een dergelijke studie kost echter veel tijd. Een ander probleem is dat realistische populatiedata beschikbaar moeten zijn. Een belangrijk aandachtspunt hierbij is dat het populatiebestand veel meer elementen moet omvatten dan de beoogde steekproef. Het gebruik van de steekproef van één periode als populatiebestand, waaruit met terugleggen getrokken wordt, leidt tot onrealistische resultaten van de simulatie. Een mogelijkheid is om de data van meerdere statistiekperioden en meerdere deelpopulaties bij elkaar te nemen. Dit zijn realistische populatiedata als de deelpopulaties op elkaar lijken en er geen grote veranderingen in de tijd zijn. Merk op dat het samenvoegen van meerdere

statistiekjaren minder zinvol kan zijn als het om een panelsteekproef gaat, waar dus in elke periode dezelfde steekproefelementen benaderd zijn.

Opmerking. Als de PS-methode in een andere situatie dan de productiestatistiek toegepast wordt, moet ook aandacht besteed worden aan de keuze van c. Men kan niet zomaar ervan uitgaan dat $c = 2{,}5$ ook in andere situaties een goede keuze is. De bovengenoemde aandachtspunten kunnen helpen bij de keuze van c.

5. Consistent schatten van verschillende niveaus

Vaak moet er niet alleen voor één niveau (bijvoorbeeld heel Nederland) een schatting gemaakt worden, maar ook voor verschillende deelpopulaties. Hierbij is het een vereiste dat de verschillende schattingen consistent zijn, bijvoorbeeld dat de schattingen voor de totalen voor alle deelpopulaties optellen tot de schatting voor het totaal voor heel Nederland. Consistentie wordt niet automatisch bereikt als voor verschillende aggregatieniveaus onafhankelijk van elkaar optimale schattingen gemaakt worden. Immers, op een hoger aggregatieniveau moeten voor een optimale schatting minder uitbijters aangewezen worden.

Voor deze situatie is nog geen valide methodologie ontwikkeld.

Bij de PS doet zich deze situatie voor. Hier is ervoor gekozen de kerncellen centraal te stellen. De uitbijters worden zo aangewezen dat de schattingen voor de kerncellen optimaal zijn (d.w.z. optimaal gegeven de methode). De schattingen voor de overige niveaus volgen dan automatisch door optellen, maar zijn niet optimaal.

6. Consistent schatten voor verschillende doelvariabelen

Vaak moeten er schattingen voor verschillende doelvariabelen gemaakt worden, waarbij de doelvariabelen in een bepaald verband met elkaar staan. Zo is bijvoorbeeld de totale omzet de som van de buitenlandse en binnenlandse omzet.

Voor deze situatie is nog geen valide methodologie ontwikkeld.

Bij de PS doet zich deze situatie voor. Er wordt een pragmatische methode toegepast die gericht is op de omzet als belangrijkste doelvariabele. Bedrijven met een extreme omzet-waarde worden voor alle doelvariabelen op uitbijter gezet. Door deze pragmatische aanpak zijn er geen problemen met consistentie tussen de verschillende doelvariabelen, die wel kunnen ontstaan als per doelvariabele met een apart gewicht wordt gewerkt. In de simulatiestudie van Krieg en Smeets (2005) is voor enkele situaties onderzocht wat de kwaliteit van de schattingen voor enkele andere doelvariabelen is. Hieruit bleek dat de gemiddelde kwadratische fout niet groter is dan van een schatter waarbij geen rekening gehouden wordt met uitbijters.

Deze situatie doet zich ook voor bij Internationale Diensten. Hier is ervoor gekozen invoer en uitvoer per dienst centraal te stellen. De uitbijters worden zo aangewezen dat de schattingen per dienst voor invoer en uitvoer optimaal zijn (d.w.z. optimaal gegeven de methode). De schattingen voor andere doelvariabelen, bijvoorbeeld de totale invoer en de totale uitvoer, volgen dan automatisch door optellen, maar zijn niet optimaal.

Zoals beschreven in hoofdstuk 2, kan de gecensureerde schatter ook in de vorm van aangepaste gewichten geschreven worden. De aanpassingen kunnen berekend worden met omzet als doelvariabele. Deze aangepaste gewichten kunnen dan ook op de andere doelvariabelen toegepast worden. Er wordt dan dezelfde ad-hocoplossing toegepast als bij de PS.

Het is waarschijnlijk dat met een uitbijtermethode die gericht is op meer doelvariabelen verbeteringen mogelijk zijn.

7. Schatten van ontwikkelingen voor één publicatieniveau

Vaak gaat het bij het publiceren van schattingen niet zozeer om het niveau van de schatting, maar meer om de ontwikkeling ten opzichte van een voorgaande periode. Een bekend voorbeeld zijn de kortetermijnstatistieken (KS). Hier worden verschillende verhoudingen berekend, bijvoorbeeld de verhouding van de schattingen van de omzet van de actuele maand en van dezelfde maand van het voorgaande jaar berekend.

In deze situatie is een vertekening van de niveauschattingen minder erg, immers, als de niveauschattingen voor beide periodes even sterk vertekend zijn, dan kan de schatting voor de verhouding (bijna) onvertekend zijn.

Voor deze situatie is op het CBS nog geen valide methodologie ontwikkeld. De methode die momenteel bij de kortertermijnstatistieken toegepast wordt, is een licht aangepaste versie van de PS-methode. Er is echter niet onderzocht of de methode ook voor de kortetermijnstatistieken goed werkt, en hoe de parameter ingesteld zou moeten worden.

In Groot-Brittannië is onlangs onderzoek gedaan naar uitbijterbehandeling bij het meten van ontwikkelingen (zie Lewis, 2008). Dit paper suggereert dat het toepassen van een uitbijtermethode voor niveaucijfers tot betere resultaten leidt dan het toepassen van een specifiek op ontwikkelingen gerichte uitbijtermethode. Onderzocht moet worden of deze conclusie algemeen van toepassing is, of alleen in het daar onderzochte voorbeeld. Daarnaast moet onderzocht worden of de daar gebruikte uitbijtermethode voor ontwikkelingen verbeterd kan worden.

8. Schatten op basis van registraties

Tot nu toe is in dit document uitgegaan van een aselect getrokken steekproef. Het CBS wil echter steeds vaker de publicaties baseren op registraties. De definitie voor representatieve uitbijters in de inleiding gaat uit van steekproeven, maar kan gegeneraliseerd worden naar registraties: een uitbijter is een extreme waarneming in de steekproef of in het register. Een representatieve uitbijter is een uitbijter waarvan aangenomen wordt dat deze correct waargenomen is en dat in de populatie meer soortgelijke elementen te vinden zijn.

Voor registraties geldt vaak dat deze de publicatie dekken. Dan is er geen sprake van representatieve uitbijters in de zin dat alle extreme waarnemingen bekend zijn. In sommige gevallen is het register niet dekkend. Een voorbeeld hiervan is de BTW-registratie voor maandcijfers. Omdat veel bedrijven per kwartaal of per jaar BTW-aangifte doen, kunnen de bedrijven met BTW-informatie per maand gezien worden als een steekproef uit de totale populatie van bedrijven. Er is dan echter geen sprake van een aselect getrokken steekproef. De steekproeftheorie die ontwikkeld is voor aselecte steekproeven is daarom niet van toepassing. Ook de uitbijtermethoden die voor aselecte steekproeven ontwikkeld zijn, kunnen niet zonder meer in deze situatie toegepast worden.

Voor deze situatie is echter nog geen valide methodologie beschikbaar.

9. Schatten van ontwikkelingen op basis van registraties

Er is op het CBS onderzocht hoe de kortetermijnstatistieken geproduceerd kunnen worden op basis van BTW-informatie. Hierbij wordt ook onderzocht hoe in deze situatie met uitbijters omgegaan moet worden. Als een valide methode ontwikkeld is, kan deze hier beschreven worden.

Omdat echter onzeker is in hoeverre de BTW-registratie in de toekomst nog voldoende beschikbaar is, is de invoering hiervan ook erg onzeker.

Net als in hoofdstuk 8 geldt ook bij het schatten van ontwikkelingen op basis van registraties dat bij een volledig dekkend register er geen problemen spelen met representatieve uitbijters.

10. Complexere situaties

In de hoofdstukken 5 t/m 8 is aangegeven voor welke situaties de komende tijd onderzoek gedaan moet worden. Hoofdstuk 9 laat een voorbeeld zien voor nog complexere situaties. Andere complexe situaties of combinaties van de in de hoofdstukken 5 t/m 8 besproken situaties zijn ook denkbaar. Ook voor deze situaties is er geen valide methodologie ontwikkeld.

11. Afsluiting

In dit document zijn drie verschillende methoden voor de behandeling van representatieve uitbijters beschreven. De methoden zijn allemaal gericht op de situatie waarin

- een aselecte steekproef getrokken is,
- een schatting voor de hele populatie gemaakt moet worden, of voor een deelpopulatie waarbij niet naar andere deelpopulaties gekeken wordt,
- een schatting voor één doelvariabele gemaakt moet worden,
- een niveauschatting gemaakt moet worden.

De PS-methode leidt in de productiestatistieken tot veel nauwkeurigere schattingen dan de gecensureerde schatters. Dit komt doordat de data van de productiestatistieken voor een groot deel van de weegcellen symmetrisch verdeeld zijn. Het gaat daarbij om de BTW-cellen, waarbij de residuen t.o.v. de regressielijn symmetrisch verdeeld zijn. Daarom heeft de PS-methode voor de productiestatistieken de voorkeur.

Ook in andere situaties waar sprake is van (min of meer) symmetrisch verdeelde data, gaat de voorkeur uit naar de PS-methode. Dan kan echter ook gekeken worden naar een andere methode die in de literatuur beschreven is, zoals een α-getrimd gemiddelde of een M-schatter (zie Huber, 1981). Deze methoden zijn weliswaar niet voor een situatie met representatieve uitbijters bedoeld, ze kunnen eventueel wel in de praktijk, voor symmetrisch verdeelde data, tot goede schattingen leiden. Deze in de literatuur beschreven methoden hebben als voordeel dat hiervoor wel een variantieformule bekend is. Er is geen vergelijkend onderzoek gedaan voor de PS-methode versus de M-schatter of het α-getrimd gemiddelde. Voordat één van deze methoden in productie genomen kan worden, is er meer onderzoek nodig.

In een situatie met aan één kant van de verdeling extreme waarden in de waarneming gaat de voorkeur uit naar één van de gecensureerde schatters. In de simulatiestudie voor de restcellen is weliswaar de verbeterde PS-schatter iets beter dan de gecensureerde schatters, maar dat geldt alleen voor een goed gekozen parameter voor de PS-schatter. Bij een minder goed gekozen parameter voor de PS-schatter loopt men het risico op sterk vertekende schattingen. Een pragmatisch hulpmiddel voor een goede keuze van de parameter zou het aantal uitbijters zijn dat aangewezen wordt.

De tweezijdig gecensureerde schatter leidt in veel situaties, ook als er maar aan één kant uitbijters zijn, tot iets betere schattingen dan de eenzijdig gecensureerde schatter (tabel 2). Daar staat tegenover dat de tweezijdig gecensureerde schatter een stuk

complexer is, vooral bij complexe steekproefontwerpen. Daarom gaat in een situatie met uitbijters alleen aan één kant de voorkeur uit naar de eenzijdig gecensureerde schatter. Als de data scheef verdeeld zijn, maar wel uitbijters kunnen voorkomen aan beide kanten, dan is aan te bevelen om de tweezijdig gecensureerde schatter te gebruiken. Dan moet men alert zijn en voor het (weinig waarschijnlijke) geval dat het algoritme niet werkt, een pragmatische oplossing vinden.

Bij de gecensureerde schatters wordt de waarde van een uitbijter (of het gewicht) op een voorzichtige manier aangepast, terwijl bij de PS-methode alle uitbijters gewicht 1 krijgen. Hierdoor zijn de PS-schattingen soms niet stabiel. Immers, een kleine verandering van de waarde van één element kan betekenen dat dit element verandert van “geen uitbijter” naar “wel uitbijter”. Dit kan grote invloed hebben op de schattingen.

Door het toepassen van één van de methoden uit dit document wordt de gemiddelde kwadratische fout van de schatters iets kleiner. Merk op dat deze maat, zoals de naam al zegt, een gemiddelde over alle mogelijke steekproeven is uit de populatie. Bij sommige van deze mogelijke steekproeven verschillen de schattingen op basis van een methode met uitbijterbehandeling en de schattingen op basis van een methode zonder uitbijterbehandeling ongeveer evenveel van de echte populatieparameter. Bij andere van deze mogelijke steekproeven zijn de schattingen op basis van een methode met uitbijterbehandeling iets slechter, voor weer andere mogelijke steekproeven zijn de schattingen op basis van een methode met uitbijterbehandeling iets of zelfs veel beter dan de schattingen op basis van een methode zonder uitbijterbehandeling. De mogelijke winst voor sommige steekproeven, met extreme uitbijters, kan daarbij groot zijn. Het toepassen van een methode met uitbijterbehandeling kan daarom ook gezien worden als een soort verzekering tegen slechte steekproeven. Een lagere gemiddelde kwadratische fout geeft aan dat de steekproeven waar uitbijterbehandeling tot een verbetering leidt, domineren.

Voor welke schatter men ook kiest, in ieder geval moet de schatter aangepast worden aan het steekproefontwerp en de ophoogmethode. Daarnaast kunnen eventueel ook aanpassingen zinvol zijn die rekening houden met de datastructuur, bijvoorbeeld als er veel nullen waargenomen worden.

Voor complexere situaties, waar meer dan één doelvariabele geschat wordt, waar optimale schattingen voor verschillende aggregatieniveaus gemaakt moeten worden, waar ontwikkelingen geschat worden, of waar schattingen op basis van registraties gemaakt worden, is op het CBS nog weinig onderzoek gedaan. In deze situaties zou dan als ad-hocaanpak een in dit document beschreven methode gebruikt kunnen worden. Hierbij moeten er dan pragmatische oplossingen bedacht worden om bijvoorbeeld consistentie tussen verschillende doelvariabelen of tussen verschillende deelpopulaties te bereiken. Dit is dan nadrukkelijk een ad-hocoplossing en geen methodologisch onderbouwde oplossing. Als de gekozen oplossing goed beschreven

wordt en onderzocht is dat deze tot betrouwbare schattingen leidt, dan kan ook ervoor gekozen worden de methode in dit document te beschrijven.

Bij de huidige toepassingen van de PS-schatter en de gecensureerde schatter is voor één doelvariabele en één groep van deelpopulaties (bijvoorbeeld de kerncellen bij de PS) gekozen waarop de uitbijtermethodiek toegepast wordt. Voor andere doelvariabelen en andere deelpopulaties (en ook voor een volledige bedrijfstak) volgen de schattingen automatisch uit de aangepaste gewichten oftewel de aangepaste variabelen. Deze andere schattingen zijn echter niet optimaal, omdat de uitbijtermethodiek daarop niet geoptimaliseerd is.

Het is wenselijk om in de nabije toekomst onderzoek te doen om ook voor de complexere situaties goede methoden te ontwikkelen. In dit document is ruimte gelaten om deze methoden te beschrijven (hoofdstukken 5 t/m 10).

In de praktijk kan het nuttig zijn om naast de automatische uitbijterdetectie ook een handmatige uitbijterdetectie mogelijk te maken. Dit is vooral nuttig als de automatische methode niet met alle doelen van het onderzoek rekening houdt, zoals bij het toepassen van een (relatief) eenvoudige methode voor een complexe situatie, zoals de situatie waarin er meer doelvariabelen zijn. Daarnaast kan handmatige uitbijterdetectie ook belangrijk zijn bij lage celvulling, omdat dan sommige automatische methoden instabiel worden. Het zetten van handmatige uitbijters moet echter heel terughoudend gebeuren, omdat anders de automatisch bepaalde optimale oplossing teniet gedaan kan worden.

Het toepassen van een uitbijtermethode maakt het productieproces complexer, het programmeren en beheren ervan kost capaciteit. Het spreekt vanzelf dat een uitbijtermethode alleen dan in een proces ingebouwd moet worden als er inderdaad sprake is van storende uitbijters.

Ten slotte willen we nog een keer benadrukken dat het toepassen van een op uitbijters gerichte schatter vooraf goed onderzocht moet worden. Negatieve effecten door de hierdoor geïntroduceerde vertekening kunnen wel eens pas op termijn naar voren komen. Dit is gebeurd bij de kortetermijnstatistieken, waar verschillende effecten, waaronder de uitbijterbehandeling, leiden tot een oplopende vertekening, de zogenaamde wegloopeffecten (zie van Delden, 2006a en van Delden, 2006b).

12. Literatuur

Banning, R. en P. Knottnerus (2009), *Methodenreeks - Thema: Steekproeftheorie*. Interne CBS-nota, Centraal Bureau voor de Statistiek, Den Haag. (in voorbereiding)

Delden, A. van (2006a), *Wegloopeffecten bij de groeicijfers van de supermarktomzetten: oorzaken en oplossingsrichtingen.* CBS-rapport, Centraal Bureau voor de Statistiek, Voorburg.

Delden, A. van (2006b), *Herberekening KS bij BES*. Interne CBS-notitie, Centraal Bureau voor de Statistiek, Voorburg.

Hampel, F.R., E.M. Ronchetti, P.J. Rousseeuw en W.A. Stahel (1986), *Robust Statistics: the Approach Based on Influence Functions*. Wiley, New York.

Handboek productie KS-en (2004). Interne CBS-nota, Centraal Bureau voor de Statistiek, Heerlen en Voorburg.

Hoogland, J.J., M.P.J. van der Loo, J. Pannekoek en S. Scholtus (2009), *Methodenreeks - Thema: Controle en Correctie*. Interne CBS-nota, Centraal Bureau voor de Statistiek, Den Haag. (in voorbereiding)

Huber, P. J. (1981), *Robust Statistics*. Wiley, New York.

Israëls, A., J. Pannekoek en E. Schulte Nordholt (2007), *Methodenreeks - Thema: Imputatie.* Interne CBS-nota, Centraal Bureau voor de Statistiek, Den Haag.

JVGN (2001), *Specificatie Ophooggewichten bepalen.* Documentatie Impect, project UniEdit, Centraal Bureau voor de Statistiek, Heerlen.

Krieg S., R. Renssen en M.J.E. Smeets (2004), *Enkele uitbreidingen voor de gecensureerde schatter.* Interne CBS-nota, Sector Methoden en Ontwikkeling, Centraal Bureau voor de Statistiek, Heerlen.

Krieg, S. en M.J.E. Smeets (2005), *De Impect-schatter en mogelijke verbeteringen.* Interne CBS-nota, Centraal Bureau voor de Statistiek, Heerlen.

Krieg S., M.J.E. Smeets, R.J. Roos, C.M. Zwaneveld (2008), *Uitbijters bij Internationale Diensten.* Interne nota, Divisie Methodologie en Kwaliteit en divisie Bedrijfseconomische statistieken, Centraal Bureau voor de Statistiek, Heerlen.

Lewis, D. (2008), Winsorisation for Estimation of Change. *Bulletin of the Office for National Statistics* 61, 49-61.

Nieuwenbroek, N.J. en P.A. Vlag (2003), *Evaluatie ophoogmodule binnen Impect.* Interne CBS-nota, Sector Methoden en Ontwikkeling, Centraal Bureau voor de Statistiek.

Renssen, R.H., M.J.E. Smeets, S. Krieg (2002), *Dealing with representative outliers in survey sampling: Algorithms.* Interne CBS-nota, Sector Methoden en Ontwikkeling, Centraal Bureau voor de Statistiek, Heerlen.

Renssen, R.H., M.J.E. Smeets, S. Krieg (2004), *Dealing with representative outliers in survey sampling: Methodology*. Interne CBS-nota, Sector Methoden en Ontwikkeling, Centraal Bureau voor de Statistiek, Heerlen.

Resing, B., R. van der Heijden en D. Debie (2005), *Handboek Impect 1 PS-en*. Interne CBS-nota, Centraal Bureau voor de Statistiek, Heerlen en Voorburg.

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

Smeets, M.J.E. en S. Krieg (2004), *Gecensureerde schatters bij de Productiestatistieken.* Interne CBS-nota, Sector Methoden en Ontwikkeling, Centraal Bureau voor de Statistiek, Heerlen.

Smeets, M.J.E. (2005), *Startcursus Methodologie – Module 11: Representatieve uitbijters.* Interne CBS-nota, Centraal Bureau voor de Statistiek, Heerlen en Voorburg.

Smeets, M.J.E. (2008), *Een uitbijtermethode voor de statistiek Bouwobjecten in Voorbereiding.* Interne CBS-nota, Sector Methodologie Heerlen, Centraal Bureau voor de Statistiek, Heerlen

Vlag, P., G. Heunen, N. Nieuwenbroek, M. Das en H. Pustjens (2001), *Functionaliteit ophoogtraject: technische specificaties detectie uitbijters*. Centraal Bureau voor de Statistiek, Heerlen.

Wolf, P.P. de en K. van Bemmel (2007), *Methode- en procesbeschrijving KS en secundaire bronnen*. Interne CBS-nota, Centraal Bureau voor de Statistiek, Voorburg.

Bijlage: algoritmes

Algoritme 1. Schatten optimale grenswaarde voor enkelvoudig aselecte steekproeven (eenzijdig).

Stap 1. Zet de grootste waarde y_n op uitbijter en laat $r = n-1$ het aantal niet-uitbijters zijn.

Stap 2. Bereken vervolgens de parameters $p = \frac{r}{n}$, $q = 1-p$, $\mu_m = \frac{1}{r}\sum_{j=1}^{r} y_j$ en $\mu_r = \frac{1}{n-r}\sum_{j=r+1}^{n} y_j$.

Stap 3. Los vergelijking (2.3.5) op voor t, waarbij voor de parameters p, q, μ_m en μ_r de in stap 2 berekende waarden gebruikt worden. De oplossing kan geschreven worden als

$$t = \frac{p(1-f)\mu_m + nq\mu_r}{p(1-f)+nq}. \qquad \text{(B.1)}$$

Stap 4. Als $y_r < t \le y_{r+1}$, dan is t de optimale grenswaarde. Anders zetten we de grootste niet-uitbijter ook op uitbijter, verlagen r met 1 en gaan terug naar stap 2.

Algoritme 2. Schatten optimale grenswaarden voor gestratificeerde steekproeven (eenzijdig).

Het algoritme gaat ervan uit dat $f_h < 1$ geldt voor alle strata h. Strata waarvoor geldt dat $f_h = 1$ moeten vooraf uit het bestand verwijderd worden. Daar worden geen uitbijters aangewezen.

Eerst wordt er een startwaarde berekend (stap 1 t/m 4). Daarna worden de optimale grenswaarden berekend volgens een iteratief proces (stap 5 t/m 8).

Stap 1. Bereken eerst met algoritme 1 voor ieder stratum h afzonderlijk een grenswaarde t_h met de waarnemingen $y_{h1},...,y_{hn_h}$. Schat vervolgens voor ieder stratum het stratumgemiddelde $\overline{Y}_{t_h}^{\text{cens}}$ met vergelijking (2.3.7).

Stap 2. Bereken de getransformeerde waarden $\tilde{y}_{hi} = \frac{N_h}{n_h}(y_{hi} - \overline{Y}_{t_h}^{\text{cens}})$.

Stap 3. Laat $u_j = \tilde{y}_{hi}$ zijn, waarbij $j = 1,\ldots,n$, $i = 1,\ldots,n_h$ en $h = 1,\ldots,L$. Sorteer de waarnemingen zodat $u_1 \le \cdots \le u_n$. Bereken met algoritme 1 een grenswaarde $\tilde{t}$ met de waarnemingen $u_1,\ldots,u_n$.

Stap 4. Bepaal voor ieder stratum h het aantal getransformeerde waarden $\tilde{y}_{hi}$ dat kleiner of gelijk is aan $\tilde{t}$ en noem dat aantal r_h. De grootste $n_h - r_h$ waarden y_{hi} in een stratum h worden dus (voorlopig) op uitbijter gezet.

Stap 5. Gegeven $r_1,\ldots,r_L$, bereken voor ieder stratum

$$\mu_{hm} = \frac{1}{r_h}\sum_{i=1}^{r_h} y_{hi}\ ,\ \mu_{hr} = \begin{cases} \dfrac{1}{n_h - r_h}\displaystyle\sum_{i=r_h+1}^{n_h} y_{hi} & \text{als} \quad n_h - r_h > 0 \\ 0 & \text{als} \quad n_h - r_h = 0 \end{cases},$$

$q_h = \dfrac{n_h - r_h}{n_h}$ en $p_h = \dfrac{r_h}{n_h}$.

Stap 6. Los het stelsel van vergelijkingen gegeven door (2.3.9) op voor $\mathbf{t} = (t_1,\ldots,t_L)$, waarbij voor de parameters p_h, q_h, μ_{hm} en μ_{hr} de in stap 5 berekende waarden gebruikt worden. De vector van grenswaarden kan geschreven worden als

$$\mathbf{t}^{\mathrm{t}} = \mathbf{J}^{-1}[\mathbf{D}_n^{-1}\mathbf{P}\overline{\mathbf{Y}}_m + \mathbf{l}\mathbf{q}_r^{\mathrm{t}}\overline{\mathbf{Y}}_r]. \qquad \text{(B.2)}$$

Hierin is $\mathbf{J} = \mathbf{D}_n^{-1}\mathbf{P} + \mathbf{l}\mathbf{q}_r^{\mathrm{t}}$,

waarbij $\mathbf{D}_n = \mathrm{diag}(n_1,...,n_L)$, $\mathbf{P} = \mathrm{diag}(N_1(1-f_1)p_1,...,N_L(1-f_L)p_L)$, $\mathbf{q}_r = (N_1q_1,...,N_Lq_L)^{\mathrm{t}}$, $\overline{\mathbf{Y}}_m = (\mu_{1m},...,\mu_{Lm})^{\mathrm{t}}$, $\overline{\mathbf{Y}}_r = (\mu_{1r},...,\mu_{Lr})^{\mathrm{t}}$ en $\mathbf{l} = (1,1,\ldots,1)^{\mathrm{t}}$. $\mathrm{diag}(...)$ is een diagonaalmatrix waarbij de diagonaalelementen tussen haakjes staan. De superscript t betekent dat de getransponeerde matrix berekend moet worden.

Stap 7. Bereken voor alle strata $1 \le h \le L$ het aantal waarnemingen y_{hi} dat kleiner is aan t_h, en noem dat aantal $\tilde{r}_h$.

Stap 8. Als $r_h = \tilde{r}_h$ voor alle h, dan is de geschatte optimale grenswaarde $\mathbf{t}^* = (t_1^*,\ldots,t_L^*)$ gevonden. Zo niet, neem dan het eerste stratum (d.w.z., neem het stratum met de laagste index h) waarvoor $r_h \neq \tilde{r}_h$ geldt, definieer $h_{\min} = \min\{h : r_h \neq \tilde{r}_h\}$ en bereken voor dat stratum

$r_{h_{\min}} = \begin{cases} r_{h_{\min}} + 1 & \text{als} \quad r_{h_{\min}} < \tilde{r}_{h_{\min}} \\ r_{h_{\min}} - 1 & \text{als} \quad r_{h_{\min}} > \tilde{r}_{h_{\min}} \end{cases}$. Dit wil zeggen dat je één extra uitbijter zet als $r_{h_{\min}} < \tilde{r}_{h_{\min}}$ en één uitbijter minder zet als $r_{h_{\min}} > \tilde{r}_{h\min}$. Ga terug naar stap 5.

Algoritme 3. Bepalen startoplossing voor algoritme 2 bij ongelijke insluitgewichten.

Stap 1. Bereken eerst met algoritme 1 voor ieder stratum h afzonderlijk een grenswaarde t_h met de waarnemingen $z_{h1},...,z_{hn_h}$. Schat voor ieder stratum het stratumgemiddelde $\overline{Z}_h^{\text{cens}} = \frac{1}{n_h}\left[\sum_{j=1}^{r_h} z_{hj} + (n_h - r_h)t_h\right]$.

Stap 2. Bereken de getransformeerde waarden $\tilde{y}_{hi} = (z_{hi} - \overline{Z}_h^{\text{cens}})$.

Stap 3. Laat $u_j = \tilde{y}_{hi}$ zijn, waarbij $j = 1,\ldots,n$, $i = 1,\ldots,n_h$ en $h = 1,\ldots,L$. Sorteer de waarnemingen zodat $u_1 \leq \cdots \leq u_n$. Bereken met algoritme 1 een grenswaarde $\tilde{t}$ met de waarnemingen $u_1,\ldots,u_n$.

Stap 4. Bepaal voor ieder stratum h het aantal getransformeerde waarden $\tilde{y}_{hi}$ dat kleiner of gelijk is aan $\tilde{t}$, en noem dat aantal r_h. De grootste $n_h - r_h$ waarden z_{hi} in een stratum h worden dus (voorlopig) op uitbijter gezet.

Algoritme 4. Schatten optimale grenswaarde voor enkelvoudig aselecte steekproeven (tweezijdig).

Stap 1. Laat $i = 0$ en $j = 0$.

Stap 2. Laat $n_l = i + 1$ het aantal linkeruitbijters zijn en $n_r = j + 1$ het aantal rechteruitbijters. Als $n_l + n_r > n$ dan ga naar stap 4 en anders bereken vervolgens de parameters $r = n - n_l - n_r$, $p = \frac{r}{n}$, $q_l = \frac{n_l}{n}$, $q_r = \frac{n_r}{n}$, $\mu_r = \frac{1}{n_r}\sum_{k=1}^{n_r} y_{n-n_r+k}$, $\mu_m = \frac{1}{r}\sum_{k=1}^{r} y_{n_l+k}$ en $\mu_l = \frac{1}{n_l}\sum_{k=1}^{n_l} y_k$.

Stap 3. Los het stelsel vergelijkingen (3.3.2) op voor s en t, waarbij voor de parameters p, q_l, q_r, μ_l, μ_m en μ_r de in stap 2 berekende waarden gebruikt worden. De grenswaarden zijn gegeven door

$$\begin{pmatrix} s \\ t \end{pmatrix} = \begin{pmatrix} (1-f)(p+q_r)+nq_l & -(1-f)q_r \\ -(1-f)q_l & (1-f)(p+q_l)+nq_r \end{pmatrix}^{-1} \begin{pmatrix} (1-f)p\mu_m + nq_l\mu_l \\ (1-f)p\mu_m + nq_r\mu_r \end{pmatrix} \quad \text{(B.3)}$$

Stap 4. Als $y_{i+1} \leq s \leq y_{i+2}$ en $y_{n-j-1} \leq t \leq y_{n-j}$, dan is de oplossing gevonden. Anders onderscheiden we drie gevallen:

- als $j = 0$, dan $j = i + 1$, $i = 0$ en ga terug naar stap 2,
- als $j > 0$ en $j > i$, dan $i = i + 1$, $j = j$ en ga terug naar stap 2,
- als $j > 0$ en $j \leq i$, dan $i = i$, $j = j - 1$ en ga terug naar stap 2.

Algoritme 5. PS-methode: Aanwijzen uitbijters in BTW-cellen, actuele methode.

Stap 1: bepalen robuuste regressielijn

De datarecords worden verdeeld in drie even grote groepen:

1. de groep met de kleinste *x*-waarden,
2. de groep met de middelste *x*-waarden,
3. de groep met de grootste *x*-waarden.

Van de eerste en derde groep worden de medianen $m_{1,x}$ en $m_{3,x}$ van de *x*-waarden en de medianen $m_{1,y}$ en $m_{3,y}$ van de *y*-waarden berekend. Vervolgens is de robuuste regressielijn gegeven door $y = \alpha x + \beta$, waarbij

$$\beta = \frac{m_{3,y} - m_{1,y}}{m_{3,x} - m_{1,x}} \tag{B.4}$$

en

$$\alpha = m_{1,y} - \beta\, m_{1,x}\,. \tag{B.5}$$

Dit is de rechte lijn die door de punten $(m_{1,x}, m_{1,y})$ en $(m_{3x}, m_{3,y})$ loopt.

Stap 2: berekenen residuen en grenswaarde

Voor elk element *i* wordt het residu berekend:

$$\varepsilon_i = y_i - \alpha - \beta\, x_i \tag{B.6}$$

dan wordt de grenswaarde *cut* berekend als

$$cut = upper_f + c(upper_f - lower_f) \tag{B.7}$$

met $upper_f$ en $lower_f$ het derde en eerste kwartiel van de absolute waarden van de residuen ε_i. Voor de keuze van c zie paragraaf 4.3.7.

Stap 3: begrenzing residuen en herberekening *y*-waarden

Nu worden de residuen begrensd via de volgende formule

$$\bar{\varepsilon}_i = \begin{cases} -cut & als \quad \varepsilon_i < -cut \\ \varepsilon_i & als \quad -cut \le \varepsilon_i \le cut \\ cut & als \quad \varepsilon_i > cut \end{cases}. \tag{B.8}$$

Vervolgens worden de *y*-waarden aangepast:

$$y_i^{'} = \alpha + \beta\, x_i + \bar{\varepsilon}_i\,. \tag{B.9}$$

Stap 4: herberekenen regressielijn

Met $y_1',...,y_n'$ wordt opnieuw een regressielijn berekend. Hierbij wordt de volgende formule gebruikt:

$$b' = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i' - \bar{y}')}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \tag{B.10}$$

$$a' = \bar{y}' - b'\,\bar{x} \tag{B.11}$$

met $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ en $\bar{y}' = \frac{1}{n}\sum_{i=1}^{n} y_i'$ de gemiddelden van de *x*-waarden en de y'-waarden. Het betreft hier dus de rechte lijn door het punt $(\bar{x}, \bar{y}')$. Het resultaat van deze berekening komt overeen met de kleinste kwadratenmethode.

Stap 5: herberekenen residuen, grenswaarde, begrenzing residuen en herberekening *y*-waarden

Nu worden de residuen e_i' berekend zoals beschreven in stap 2, waarbij in plaats van a, b en $y_1,...,y_n$ nu a', b' en $y_1',...,y_n'$ gebruikt wordt. Vervolgens wordt cut' voor e_i' berekend.

Daarna worden de residuen opnieuw begrensd en de *y*-waarden opnieuw berekend, waarbij nu cut', e_i', a', b' en $y_1',...,y_n'$ gebruikt worden in plaats van *cut*, e_i, a, b en $y_1,...,y_n$. De op deze manier begrensde residuen worden $\bar{\bar{e}}_i$ genoemd, de *y*-waarden $y_1'',...,y_n''$.

Stap 6: Herberekening regressielijn, residuen, grenswaarde, begrenzing residuen en herberekening *y*-waarden

Met de formules uit stap 4 wordt opnieuw de regressielijn berekend, waarbij nu $y_1'',...,y_n''$ gebruikt wordt in plaats van $y_1',...,y_n'$. De regressiecoëfficiënten worden nu a'' en b'' genoemd. Vervolgens worden de berekeningen uit stap 5 herhaald, waarbij nu a'' en b'' en $y_1'',...,y_n''$ gebruikt worden. De opnieuw berekende *y*-waarden worden $y_1''',...,y_n'''$ genoemd.

Stap 7: aanwijzen uitbijters

Als $y_i > y_i'''$, dan wordt het i'de element als uitbijter aangewezen.

Opmerking: in de documentatie van de PS-schatter wordt aangegeven, dat alleen bij een duidelijk verschil (meer dan 1%) het element als uitbijter

wordt aangewezen. Deze restrictie is gemaakt om veranderingen door afronding uit te sluiten.

Algoritme 6. PS-methode: Aanwijzen uitbijters in restcellen, actuele methode.

Stap 1: bepalen mediaan

De mediaan m_r van de elementen $y_1,...,y_n$ in de restcel wordt berekend.

Stap 2: berekenen residuen en grenswaarde

Voor elk element *i* wordt het residu berekend:

$$\varepsilon_i = y_i - m_r \,. \tag{B.12}$$

Voor de berekening van de grenswaarde *cut* worden niet deze residuen, maar de getransformeerde residuen τ_i gebruikt. Deze worden berekend als

$$\tau_i = \begin{cases} \log_{10}(abs(\varepsilon_i)) & \text{als} \quad abs(\varepsilon_i) > 0{,}00005 \\ 0 & \text{anders} \end{cases}. \tag{B.13}$$

Dan wordt de grenswaarde *cut* berekend als

$$cut = upper_f + c(upper_f - lower_f) \tag{B.14}$$

met $upper_f$ en $lower_f$ het derde en eerste kwartiel van de getransformeerde residuen τ_i. Voor de keuze van c zie paragraaf 4.3.7.

Stap 3: begrenzing residuen en herberekening *y*-waarden

Nu worden de residuen begrensd via de volgende formule

$$\varepsilon_i^{'} = \begin{cases} -10^{cut} & als \quad \tau_i > cut,\ \varepsilon_i < 0 \\ \varepsilon_i & als \quad -cut \le \tau_i \le cut \\ 10^{cut} & als \quad \tau_i > cut,\ \varepsilon_i > 0 \end{cases}. \tag{B.15}$$

Vervolgens worden de *y*-waarden aangepast:

$$y_i^{'} = m_r + \varepsilon_i^{'} \,. \tag{B.16}$$

Stap 4: aanwijzen uitbijters

Als $y_i \neq y_i^{'}$, dan wordt het i'de element als uitbijter aangewezen.

Opmerking 1: volgens de documentatie van de productiestatistieken wordt de procedure nog een keer herhaald. De mediaan en de kwartielen zijn echter robuust en zullen door de aangepaste waarden van de uitbijters niet wijzigen. Daarom wordt deze herhaling hier niet beschreven.

Opmerking 2: Ook voor deze situatie wordt in de documentatie van de PS-schatter aangegeven, dat alleen bij een duidelijk verschil (meer dan 1%) het

element als uitbijter wordt aangewezen. Deze restrictie is gemaakt om veranderingen door afronding uit te sluiten.

Opmerking 3: deze methode is ontwikkeld voor omzetten. Dit zijn over het algemeen grote positieve waarden. Ook de residuen $\varepsilon_i = y_i - m_r$ zijn dan meestal groter dan 1 of kleiner dan -1. Dan zijn de getransformeerde residuen τ_i allemaal groter dan nul. Als de absolute waarde van ε_i groeit, dan groeit ook het getransformeerde residu τ_i. Onder deze voorwaarde werkt de methode redelijk, zoals de simulatieresultaten in paragraaf 4.4 laten zien. Of de transformatie ook goed werkt als de doelvariabele in een andere orde van grootte is, is twijfelachtig. Zoals de simulatiestudie ook heeft aangetoond heeft het de voorkeur de transformatie niet toe te passen.