Effecten van het steekproefontwerp op (regressie-)analyses

Jeroen Pannekoek en Abby Israëls[1]

Samenvatting:

In dit artikel wordt ingegaan op de gevolgen van een complex steekproefontwerp bij het analyseren van de gegevens. In de traditionele literatuur over analysemethoden worden sterke (model)veronderstellingen gemaakt, zoals het onafhankelijk en identiek verdeeld zijn van de waarnemingen. Schattingsprocedures voor de modelparameters zijn uitsluitend gebaseerd op het veronderstelde stochastische proces dat de data heeft voortgebracht, het steekproefontwerp speelt geen rol. Dit staat lijnrecht tegenover de klassieke steekproeftheorie waarbij schattingsprocedures juist alleen op het steekproefontwerp gebaseerd zijn. In dit artikel worden de verschillen tussen deze twee benaderingen uitgewerkt voor regressieanalyse en wordt getoond dat model-veronderstellingen bij data afkomstig uit complexe steekproefontwerpen vaak niet plausibel zijn en dat schattingsprocedures gebaseerd op de steekproeftheorie een robuust alternatief bieden.

Trefwoorden: complexe steekproef, multivariate analyse, varianties, wegen

1. Inleiding

Steekproeftheorie en analysemethoden zijn binnen de statistiek twee gescheiden disciplines. Beide hebben hun eigen stochastiek en veronderstellingen. Het steekproefontwerp is er meestal op gericht om eenvoudige populatie-parameters als totalen en gemiddelden zo nauwkeurig mogelijk te kunnen schatten. Hiervoor zijn soms ingewikkelde steekproefontwerpen nodig, zoals steekproeven met ongelijke kansen, stratificatie, clustering etc. (Bethlehem, 2000). Bij analyses heeft men te maken met ingewikkelder parameters die relaties tussen variabelen beschrijven voor een oneindige populatie. Anderzijds gaan de analysemethoden uit van eenvoudige veronderstellingen ten aanzien van de herkomst van de data, zoals van het

---

[1] Centraal Bureau voor de Statistiek, Sector Methoden en Ontwikkeling, Postbus 4000, 2270 JM Voorburg, E-mail: jpnk@cbs.nl

De in dit rapport weergegeven opvattingen zijn die van de auteurs en komen niet noodzakelijk overeen met het beleid van het Centraal Bureau voor de Statistiek.

onafhankelijk en identiek verdeeld zijn van de waarnemingen, hetgeen sterke overeenkomsten vertoont met een enkelvoudige, aselecte steekproef met teruglegging. Wanneer men rekening wil houden met de complexiteit van de getrokken steekproeven biedt de standaard software voor analyses geen uitkomst. Zo zijn de standaardfouten en betrouwbaarheidsintervallen niet meer correct.

Pas de laatste twee decennia vindt er onderzoek plaats naar de combinatie van beide disciplines: de analyse van complexe-steekproefdata (zie bijvoorbeeld Skinner e.a., 1989). Het gaat daarbij vooral om de invloed van

- het schatten van parameters bij ongelijke steekproefgewichten, en
- het algehele steekproefontwerp op standaardfouten en toetsen.

In paragraaf 2 wordt ingegaan op het verschil in stochastiek bij steekproeftheorie (design based benadering) en analysemethoden (model-benadering) en de combinatie van beide (model based benadering). Ook wordt de invloed bekeken van het wel of niet corrigeren (wegen) voor ongelijke steekproefgewichten bij het schatten van parameters, m.n. regressiecoëfficiënten. Paragraaf 3 gaat in op de verschillen in variantieschatters van regressiecoëfficiënten bij de design based en de modelmatige benadering. In paragraaf 4 volgt een illustratie en paragraaf 5 sluit af met enkele conclusies.

2. Design based versus model (based)

2.1 Design based benadering

Zij U een eindige populatie van N elementen. Een eindige-populatieparameter is gedefinieerd als een functie van scores $y_1,\ldots,y_k,\ldots,y_N$. Voorbeelden zijn

- populatietotaal $t = \sum_{k\in U} y_k$,
- populatiegemiddelde $\bar{y}_U = \frac{1}{N}\sum_{k\in U} y_k$,
- regressiecoëfficiënt (hellingshoek)

$$B_U = \frac{\sum_{k\in U}(y_k - \bar{y}_U)(x_k - \bar{x}_U)}{\sum_{k\in U}(x_k - \bar{x}_U)^2} \qquad (2.1)$$

Het steekproefontwerp legt voor ieder element k uit U de kans π_k vast om in de steekproef terecht te komen, en voor ieder paar elementen (k,k') de kans $\pi_{kk'}$ dat beide worden getrokken. Het resultaat van het trekken is een steekproef S van omvang n (meestal tevoren gekozen). S wordt bepaald door de indicatorvariabele a met

- a_k=1 als element k in de steekproef terechtkomt, en
- a_k=0 als element k niet wordt getrokken (k=1,…,N).

Er geldt: $P(a_k=1)= P(k\in S) = \pi_k$ en $P(a_k=1$ én $a_{k'}=1)=\pi_{kk'}$ met $\sum_{k=1}^{N} a_k = n$.

De elementen uit de steekproef kunnen worden genummerd van i=1 t/m n.[2]

Het eenvoudigste steekproefontwerp is de enkelvoudig aselecte steekproef (SRS) zonder teruglegging. Hierbij zijn zowel π_k als $\pi_{kk'}$ constant (π_k=n/N). Het populatietotaal $t = \sum_{k\in U} y_k$ wordt (design-)zuiver geschat door $\hat{t} = \frac{N}{n}\sum_{i\in S} y_i$, terwijl het steekproefgemiddelde $\bar{y}_S = \frac{1}{n}\sum_{i\in S} y_i$ een zuivere schatter is voor $\bar{y}_U$. We kunnen dit laatste noteren als

$$E_{a=SRS}\,\bar{y}_S = \bar{y}_U . \qquad (2.2)$$

waarbij E_a de verwachting is over het trekkingsmechanisme a. Ook bij andere 'zelfwegende' designs, d.w.z. designs met π_k=n/N, zoals een proportioneel gestratificeerde en een clustersteekproef, zijn deze ongewogen schatters zuiver.

Vaak echter worden de populatie-elementen met ongelijke kansen π_k getrokken, Redenen hiertoe kunnen zijn dat men over bepaalde deelpopulaties nauwkeuriger resultaten wil verkrijgen, of dat de steekproef uit een adressenkader is getrokken en men per adres of huishouden slechts één persoon wil enquêteren. Om toch design-zuivere schatters voor de populatieparameters te krijgen wordt ter correctie voor deze ongelijke insluitkansen ieder record i uit de steekproef gewogen met $1/\pi_i$. Zo is de 'Horvitz-Thompson-schatter' (π-schatter)

$$\hat{t}_\pi = \sum_{i\in S} y_i / \pi_i \qquad (2.3)$$

een zuivere schatter voor het populatietotaal t van y, en $\hat{t}_\pi / N$ voor het populatie-gemiddelde: $E_a \hat{t}_\pi / N = \bar{y}_U$.[3]

Meer algemeen verkrijgt men design-consistente schatters door iedere som over de steekproefelementen gewogen te nemen (Särndal, e.a., 1991). Zo wordt B_U uit (2.1) geschat door

$$\hat{B}_\pi = \frac{\sum_{i\in S}(y_i - \bar{y}_S)(x_i - \bar{x}_S)/\pi_i}{\sum_{i\in S}(x_i - \bar{x}_S)^2/\pi_i} . \qquad (2.4)$$

Bij analyses schaalt men de gewichten $1/\pi_i$ meestal zodanig dat ze tot n optellen, waardoor het gemiddelde gewicht gelijk wordt aan 1.

---

[2] Ter onderscheid gebruiken we index *k* voor de populatie, en *i* voor de steekproef.

[3] De schatter $(\sum_S y_i/\pi_i)/(\sum_S 1/\pi_i)$ is overigens een iets nauwkeuriger schatter voor $\bar{y}_U$. Deze schatter is 'slechts' design-consistent (asymptotisch design-zuiver).

Gewichten worden overigens niet alleen gebruikt om te corrigeren voor ongelijke insluitkansen, maar ook ter correctie van verschillen in nonresponsfracties tussen categorieën van achtergrondvariabelen. Hiermee hoopt men de onzuiverheid ten gevolge van nonrespons te verkleinen.

Bij een SRS-steekproef zonder teruglegging is niet alleen π_i constant, maar ook π_{ij}. Hierdoor is de variantie van $\bar{y}_S$ eenvoudig te schatten:

$$\mathrm{var}_a(\bar{y}_S) = \frac{1}{n(n-1)}(1-f)\sum_{i\in S}(y_i - \bar{y}_S)^2 , \qquad (2.5)$$

met f=1-n/N de 'eindigheidscorrectie'. Ook voor wat ingewikkelder designs bestaan standaardformules voor varianties van gemiddelden en totalen; zie par. 3.1.

2.2 Model-benadering

Bij modelmatige analyses gaat men uit van een geheel andere stochastiek. De waarnemingen y_i (i=1,...,n) worden nu beschouwd als realisaties van stochastische variabelen Y_i met een gemeenschappelijke verdeling ξ. Soms neemt men aan dat de Y_i onafhankelijk en identiek verdeeld (i.i.d.) zijn. De Y-variabele is nu stochastisch, in tegenstelling tot bij de design based benadering waar de y-scores vastliggen en de stochastiek (randomisatie) n elementen uit populatie U levert met vaste waarden x_i, y_i, etc. Bij de model-benadering is geen sprake van eindige-populatieparameters, maar van parameters die het data-genererend proces beschrijven, zoals de verwachting μ van Y, of regressiecoëfficiënt β van het regressiemodel ξ: $y_i = \alpha + \beta x_i + \varepsilon_i$ met ε_i i.i.d. Bij de model-benadering is het steekproef-gemiddelde $\bar{y}_S$ een zuivere schatter voor μ, d.w.z.

$$E_\xi \bar{y}_S = \mu , \qquad (2.6)$$

in plaats van (2.2). De ξ-variantieschatter van $\bar{y}_S$ is dus gelijk aan (2.5), afgezien van de eindigheidscorrectie (1-f). Deze ξ-variantie is dus gelijk aan de design-variantie van $\bar{y}_S$ voor een SRS-design met teruglegging. Bij ingewikkelde parameterschatters zal de ξ-variantie wel verschillen van de design-variantie. Merk op dat de standaard-software voor analyses, zoals SPSS, S-PLUS en SAS, uitgaat van de model-benadering, en dus niet de gewenste standaardfouten en betrouwbaarheidsintervallen oplevert wanneer men rekening wil houden met het steekproefontwerp.

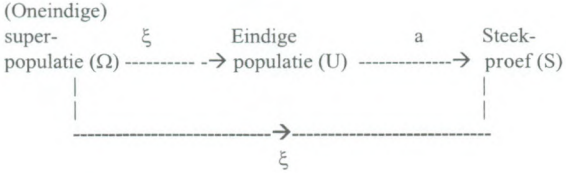
Waar wegen, of meer algemeen het rekening houden met het steekproefdesign, bij de design based benadering haast vanzelfsprekend is wanneer elementen met ongelijke kansen zijn getrokken, wordt dit bij modelmatige analyses veel minder toegepast. Dit komt deels doordat de modelveronderstellingen zo stringent kunnen zijn dat er geen plaats meer voor is. Een model-eis van i.i.d. impliceert dat alle waarnemingen even waardevol zijn, hetgeen wegen (of rekening houden met het steekproefontwerp) overbodig maakt, en dat alle waarnemingen onafhankelijk zijn,

hetgeen een clustereffect uitsluit. Versoepelt men de modelveronderstellingen, bijvoorbeeld door voor de storingen van een regressie niet te eisen dat ze ongecorreleerd met de predictoren zijn, dan is er wel ruimte voor het rekening houden met het steekproefontwerp. De model based benadering biedt dan een theoretische fundering voor het combineren van steekproef- en modeleigenschappen.

2.3 Model based benadering

Uit het voorgaande volgt dat men kan kiezen voor de steekproefeigenschappen met randomisatie als stochastiek (variabele a) of voor de model-veronderstellingen. De model based benadering brengt deze twee benaderingen tezamen (zie figuur 1).

Figuur 1.



De y_k (k=1,…,N) uit U worden nu beschouwd als realisaties van stochastische variabelen Y_k (k=1,…,N) met gemeenschappelijke verdeling ξ. Alle populatie-elementen voldoen dus aan het individueel model ξ. De eindige populatie U is dus als het ware getrokken uit een oneindige 'super-populatie' Ω; men had dus een net iets andere set van N elementen kunnen trekken. Uit populatie U wordt weer een steekproef S getrokken volgens een bepaald design, zoals in paragraaf 2.1 beschreven. Bij een SRS-design zijn de n waarnemingen uit S ook rechtstreeks te beschouwen als realisaties van Y met verdeling ξ. De onderste pijl in figuur 1 is hierop van toepassing. We zijn dan in de model-benadering beland.

Ter vergelijking van de drie benaderingen: stel we hebben een serie waarnemingen $y_1,\ldots,y_i,\ldots,y_n$. Uitgaande van een model ξ zit de stochastiek in de variabele Y, en geldt $E_\xi \bar{y}_S = \mu$. Bij design based analyse zit de stochastiek in de randomisatie, en geldt $E_a \bar{y}_S = \bar{y}_U$. Bij model based analyse heeft men beide vormen van stochastiek en geldt $E_\xi E_a \bar{y}_S = E_\xi \bar{y}_U = \mu$. De nadruk ligt dan vaak meer op modelparameters van de super-populatie (μ of β) dan op de parameters van de eindige populatie ($\bar{y}_U$ of B_U). Het verschil tussen deze parameters is echter gering, namelijk van de orde O(1/N). Zelfs als men zuiver geïnteresseerd is in modelparameters, kan men dus de corresponderende π-schatters voor de eindige populatie gebruiken en dus gewichten hanteren.

Hieronder laten we zien dat het gebruik van π-schatters bij het schatten van modelparameters kan corrigeren voor vertekening ten gevolge van ongewenste

selectie-effecten; zie ook Ten Cate (1986) en Nathan & Smith (1989). Het is tevens een robuust alternatief voor het gebruik van modellen.

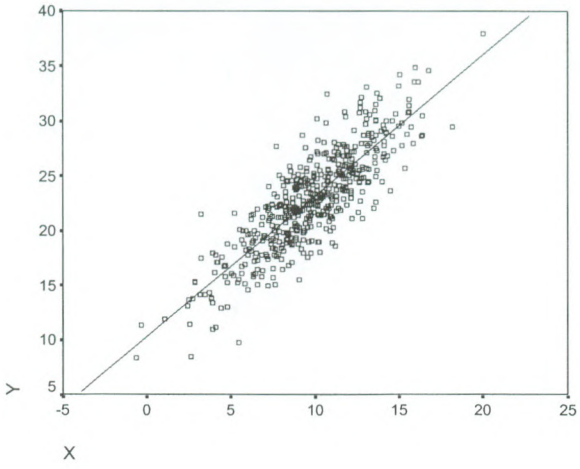
2.4 Wegen bij enkelvoudige regressie?

Design based is het haast vanzelfsprekend om rekening te houden met het steekproef-design. Door te wegen verkrijgt men (asymptotisch) design-zuivere schatters. Alleen wanneer de variantie te zeer toeneemt vanwege de verschillen in gewichten kan men besluiten tot het niet-wegen of tot een minder robuuste weging. Ondanks dat bij design based analyse geen modelveronderstellingen worden gemaakt, kan het zinvol zijn om een regressiecoëfficiënt B_U (bijvoorbeeld voor een trend) voor een eindige populatie te beschouwen. Men doet dan overigens niet veel meer dan 'een lijn fitten'. In paragraaf 3 zal hier aandacht aan worden besteed.

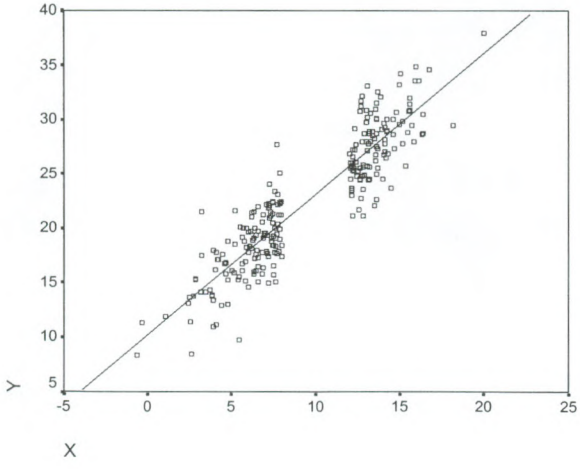
Bij modellen ligt de keuze tussen wel of niet wegen wat gecompliceerder. Dat verwachtingen systematisch te hoog of te laag kunnen worden geschat wanneer de selectie niet zuiver is zal duidelijk zijn. Ook bij regressie-analyse doen zich dit soort problemen zich voor, zoals de figuren 2a-2d laten zien.

We hebben met behulp van SPSS 500 cases gegenereerd met x normaal verdeeld (μ=10, σ=3). Vervolgens is y gedefinieerd als een lineaire functie van x, waaraan een normaal verdeelde storing is toegevoegd. Figuur 2a toont de puntenwolk met n=500. De regressiecoëfficiënt β (of eigenlijk B_U vanwege de eindigheid van de populatie) is gelijk aan 1,288. Bij figuur 2b is geselecteerd op x door alleen de hoogste en laagste 125 waarden mee te nemen. Zoals bekend heeft selectie op x geen systematische invloed op de regressiecoëfficiënten α en β: de ongewogen kleinste-kwadratenschatter $\hat{\beta}_{OLS}$ is een zuivere schatter voor β. We vinden $\hat{\beta}_{OLS}$=1,295. De selectie heeft wel invloed op de standaardfout van $\hat{\beta}_{OLS}$. Selectie van extreme waarden zorgt voor een stabiele regressielijn, dus kleine standaardfouten, en een hoge R^2. Een dergelijke stratificatie naar x is als opzet voor een regressie-analyse dus te verkiezen boven een willekeurige steekproef uit x, gegeven de lineariteit. Anders wordt het wanneer we op de doelvariabele selecteren. In figuur 2c zijn de 250 'extreme' waarden op y weggehaald. Omdat de selectievariabele gecorreleerd is met y|x, dus met de residuen, is de geschatte regressiecoëfficiënt $\hat{\beta}_{OLS}$ geen zuivere schatter meer voor β. De 250 overgebleven punten liggen systematisch boven of onder de regressielijn uit figuur 2a, omdat de verwachting van de storingen niet meer gelijk aan 0 is. Het weggooien van de hoge en lage waarden van y leidt ertoe dat de variatie in y sterk wordt gereduceerd t.o.v. de variatie in x; $\hat{\beta}_{OLS}$ wordt dus veel kleiner (0,426).

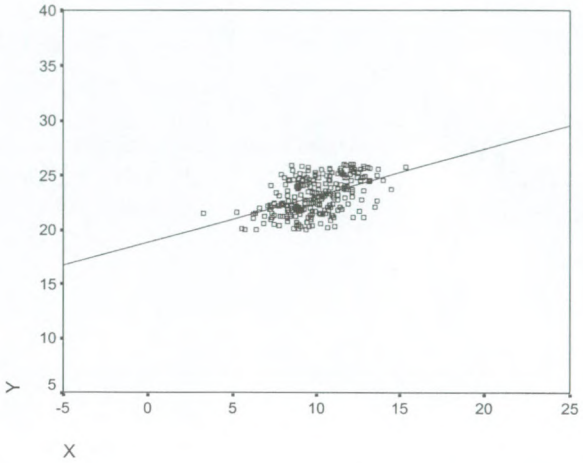
Figuur 2a. Regressie van y op x voor alle 500 cases ($\beta = 1{,}288$)



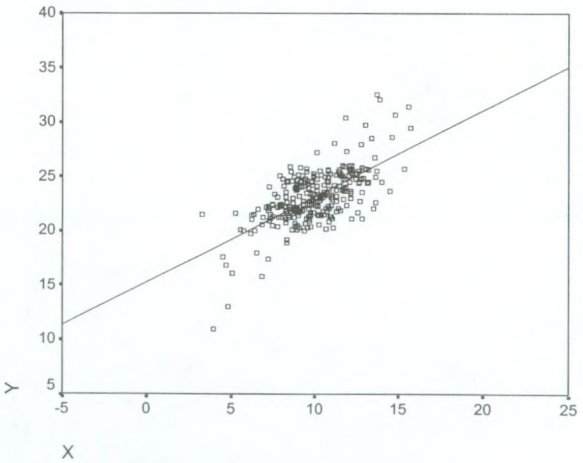
Figuur 2b. Regressie van y op x voor alle 250 cases uit uiterste kwartielen van x ($\hat{\beta}_{OLS} = 1{,}295$)



Figuur 2c. Regressie van y op x voor alle 250 cases uit middelste kwartielen van y ($\hat{\beta}_{OLS} = 0{,}426$)



Figuur 2d. Regressie van y op x voor 275 cases: alle 250 cases uit middelste twee kwartielen van y + 25 cases erbuiten ($\hat{\beta}_{OLS} = 0{,}793$; $\hat{\beta}_{\pi} = 1{,}389$)



In figuur 2c is het niet mogelijk via weging een model-zuivere of een design-consistente schatter voor β te verkrijgen, omdat er in het stratum 'hoogste + laagste kwartiel' geen enkele waarneming zit; er valt dus niets 'op te hogen'. In figuur 2d is dit wel mogelijk. We hebben nu een disproportioneel gestratificeerde steekproef getrokken waarbij alle waarnemingen uit figuur 2c zijn meegenomen én 25 van de 250 daar weggelaten cases. De ongewogen schatter $\hat{\beta}_{OLS}$ is nu gelijk aan 0,793, hetgeen nog steeds veel kleiner is dan 1,288. We kunnen echter de 25 'extremen' een gewicht 10 geven. De resulterende gewogen schatter $\hat{\beta}_{\pi}$ is gelijk aan 1,389. Uitgaande van een lineair model is deze π-schatter zuiver, design based is het een consistente schatter voor B_U. Dat de schatting redelijk afwijkt van β komt doordat de variantie sterk toeneemt wanneer bepaalde records, die niet betrouwbaarder zijn gemeten dan de andere records, tien keer zo zwaar meetellen. Hoe de variantie van deze gewogen schatter kan worden berekend volgt in paragraaf 3. Merk op dat het gebruikte trekkingsmechanisme, met ongelijke kansen π_i die een functie zijn van y|x, een i.i.d.-veronderstelling voor de storingen uitsluit. Zou men desondanks veronderstellen dat de 275 storingen i.i.d. verdeeld zijn, dan negeert men daarmee ten onrechte het steekproefontwerp.

Ook bij figuur 2b zou men 10% van de weggelaten punten kunnen toevoegen. Er ontstaat dan een duidelijk verschil tussen de design based en de modelbenadering. De 'model-aanhanger' zal de toegevoegde waarnemingen gewoon meenemen, maar niet ook nog een extra gewicht geven. Hij heeft immers al een zuivere schatter en de toegevoegde waarnemingen hoeven geen kleinere storingen te hebben. Een 'steekproefman' wil wel wegen; de lineariteitsveronderstelling is hem vreemd; het wegen zou zelfs kunnen aantonen dat er van lineariteit geen sprake is.

We hebben gezien dat weging overbodig is wanneer het selectiemechanisme verdisconteerd is in x. Wanneer het selectiemechanisme wél met de storingen samenhangt treedt een bias op in de regressiecoëfficiënten. Dit laat zich veralgemeniseren tot multipele regressie. Wanneer de selectie wordt veroorzaakt door een variabele die (incl. de interacties met alle andere x-variabelen) als predictor in het model is opgenomen, dan is wegen overbodig. Omgekeerd kan men er dus voor zorgen een ongewogen analyse uit te voeren door de wegingsvariabele met al zijn interacties als x-variabelen toe te voegen. Praktisch gezien wil men dit vaak niet. Immers, het model bevat veel meer parameters en men is niet altijd in die extra parameters geïnteresseerd. Dus blijft een gewogen analyse dan gewenst.

3. Design- en model-variantie van schatters voor regressiecoëfficiënten

In deze paragraaf bespreken we de design-variantie en de model-variantie van de parameters in het lineaire regressiemodel. Omdat de benadering voor de design variantie die we zullen bespreken gebaseerd is op de variantie van de Horvitz-Thompson schatter voor totalen, besteden we eerst aandacht aan variantieschatters voor totalen. Vervolgens gaan we in op enkele standaardresultaten met betrekking

tot de modelmatige benadering van het lineaire regressiemodel. Daarna wordt de design-variantie van de parameters van het lineaire regressiemodel besproken en vergeleken met de modelmatige aanpak.

3.1 Variantie-schatters voor totalen en andere lineaire grootheden

De variantie van de Horvitz-Thompson schatter of π-schatter (2.3) voor het populatietotaal van y ($\sum_{i\in S} y_i/\pi_i$) is bekend en in de standaardwerken over steekproeftheorie (Cochran, 1977, Särndall e.a., 1992) worden schatters voor deze variantie uitgewerkt voor verschillende steekproefontwerpen. Voor een aantal steekproefontwerpen kan deze variantie geschat worden met (bijzondere gevallen van) een benaderingsformule voor een meertrapssteekproef met stratificatie bij de eerste trap (Särndal e.a., 1992, pag. 154). Een voorbeeld van een meertrapssteekproef is een tweetrapssteekproef waarbij eerst huishoudens worden getrokken (eerste trap) en vervolgens binnen de geselecteerde huishoudens personen worden getrokken (tweede trap). De groepen elementen (huishoudens in dit voorbeeld) waaruit bij de eerste trap getrokken wordt, worden primaire eenheden of clusters genoemd. Als bovendien Nederland eerst opgedeeld is in bepaalde regio's en binnen iedere regio is een onafhankelijke steekproef van huishoudens getrokken, is er sprake van een gestratificeerde tweetrapssteekproef.

Als we de index i voor de steekproefelementen vervangen door de samengestelde index hjk, met $h = 1\ldots H$ de index voor de strata, $j = 1\ldots n_h$ de index voor de eerste trap (primaire steekproefeenheden, PSU's) in stratum h en $k = 1\ldots m_{hj}$ de index voor de tweede trap (steekproefelementen in PSU j en stratum h), dan geldt

$$\hat{\text{var}}(\hat{t}_\pi) = \sum_{h=1}^{H}(1-f_h)\frac{n_h}{n_h-1}\sum_{j=1}^{n_h}(\hat{t}_{hj}-\hat{\bar{t}}_h)^2\,, \tag{3.1}$$

met $1-f_h = 1-n_h/N_h$ de eindigheidscorrectie, N_h het aantal populatie-elementen en n_h de steekproefomvang in stratum h en

$$\hat{t}_{hj} = \sum_{k=1}^{m_{hj}} y_{hjk}/\pi_{hjk} \tag{3.2}$$

en

$$\hat{\bar{t}}_h = \frac{1}{n_h}\sum_{j=1}^{n_h}\hat{t}_{hj}\,. \tag{3.3}$$

Bijzondere gevallen van (3.1) zijn bijvoorbeeld:

Gestratificeerde meertrapssteekproef met teruglegging bij de eerste trap.

In dit geval is de eindigheidscorrectie gelijk aan 1 en (3.1) een zuivere schatter voor $\text{var}(\hat{t}_\pi)$.

Gestratificeerde eentrapssteekproef.

De PSU's zijn de steekproefelementen, $m_{hj} = 1$. Ook in dit geval is de schatter (3.1) zuiver.

Als er in plaats van één populatietotaal meerdere populatietotalen geschat worden dan kan de covariantiematrix van deze schattingen geschat worden met een eenvoudige generalisatie van (3.1). Verzamelen we de schatters voor de populatietotalen in een vector $\hat{\mathbf{t}}_\pi$, dan kan deze generalisatie geschreven worden als

$$\hat{\text{var}}(\hat{\mathbf{t}}_\pi) = \sum_{h=1}^{H} (1 - f_h) \frac{n_h}{n_h - 1} \sum_{j=1}^{n_h} (\hat{\mathbf{t}}_{hj} - \hat{\bar{\mathbf{t}}}_h)(\hat{\mathbf{t}}_{hj} - \hat{\bar{\mathbf{t}}}_h)', \qquad (3.4)$$

met $\hat{\mathbf{t}}_{hj}$ de vector met π-schatters voor de totalen in PSU j in stratum h en $\hat{\bar{\mathbf{t}}}_h$ het gemiddelde van de n_h vectoren $\hat{\mathbf{t}}_{hj}$ in stratum h.

Deze variantieformules zijn niet alleen van toepassing op schattingen voor het populatietotaal maar meer algemeen op schatters die geschreven kunnen worden als een som over de steekproefelementen, dus in de vorm $\hat{t}_\pi = \sum_{i \in S} z_i / \pi_i$, waarbij z_i een functie is van de waarden van de variabelen voor element i. Voorbeelden van deze *lineaire schatters* zijn naast de schatter voor het populatiegemiddelde van y ($z_i = y_i / N$) ook schatters voor verschillen tussen populatiegemiddelden ($z_i = (x_i - y_i)/N$) en schatters voor populatie-kwadraatsommen en kruisproducten ($z_i = x_i^2$ en $z_i = x_i y_i$).

De schatters voor een aantal voor analyses belangrijke grootheden zijn echter geen lineaire schatters. Bijvoorbeeld, een design-consistente schatter voor de populatievariantie van x is gegeven door $\frac{1}{N}\sum_{i \in S}(x_i^2 / \pi_i) - (\frac{1}{N}\sum_{i \in S} x_i / \pi_i)^2$, dus de π-schatter voor het gemiddelde van x-kwadraat minus de π-schatter voor x-gemiddeld in het kwadraat. Deze schatter is niet te schrijven als een lineaire schatter, zoals hierboven gedefinieerd. Ook design based schatters voor eindige-populatie-covarianties, correlatie- en regressie-coëfficiënten zijn geen lineaire schatters. Voor deze schatters zijn de standaardformules voor totalen niet van toepassing en moeten andere methoden gebruikt worden om de variantie te schatten. Een mogelijkheid hiervoor is om eerst een lineaire benadering voor de niet-lineaire schatter te vinden, en vervolgens de variantie van deze lineaire benadering (die volgens de standaard methode geschat kan worden) op te vatten als een benadering voor de variantie van de niet-lineaire schatter. In paragraaf 3.3 zal deze methode besproken worden aan de hand van de π-schatter voor de eindige-populatie regressiecoefficient. Om de notatie voor het multipele regressiemodel te introduceren en om de vergelijking te kunnen maken met de meer traditionele modelmatige benadering volgen eerst enkele resultaten m.b.t. de model-variantie van de parameters van het lineaire model.

3.2 Model-variantie van de schatter voor de regressiecoëfficiënt

Het standaard lineaire regressiemodel kan geschreven worden als

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon, \tag{3.5}$$

met $\mathbf{y}$ de n-vector met de waarden van de te verklaren variabele, $\mathbf{X}$ de n×k matrix met de waarden van de regressoren, β een k-vector met regressiecoefficiënten, ε een stochastische n-vector met storingstermen waarvan de (model)verwachting nul is en n het aantal waarnemingen. Door de stochastiek in ε, is $\mathbf{y}$ ook een stochastische n-vector met verwachting $E_\xi \mathbf{y} \mid \mathbf{X} = \mathbf{X}\beta$ en covariantiematrix $\mathrm{var}_\xi(\mathbf{y} \mid \mathbf{X}) = \mathrm{var}_\xi(\varepsilon \mid \mathbf{X})$.

De ongewogen kleinste kwadraten (OLS) schatter voor β is $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$. Gebruikmakend van $\beta = E_\xi\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E_\xi\mathbf{y}$ en $\varepsilon = \mathbf{y} - E_\xi\mathbf{y}$ kan de covariantiematrix van de OLS-schatter voor β geschreven worden als

$$\mathrm{var}_\xi(\hat{\beta}) = E_\xi\{(\hat{\beta}-\beta)(\hat{\beta}-\beta)'\} = E_\xi(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\varepsilon\varepsilon'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}. \tag{3.6}$$

In de standaard regressie-theorie worden de waarden van de regressoren (de $\mathbf{X}$-matrix) als gegeven (niet stochastisch) beschouwd, de enige bron van variantie zijn de stochastische storingstermen ε. We kunnen daarom de verwachting naar binnen halen en (3.6) schrijven als

$$\mathrm{var}_\xi(\hat{\beta}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E_\xi(\varepsilon\varepsilon')\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \tag{3.7}$$

Als bovendien de (standaard) veronderstelling gemaakt wordt dat de storingstermen gelijke varianties hebben (homoscedastisch zijn), zodat $E_\xi\varepsilon\varepsilon' = \sigma^2\mathbf{I}$, met $\mathbf{I}$ de identiteitsmatrix, dan reduceert (3.7) tot de bekende uitdrukking

$$\mathrm{var}_\xi(\hat{\beta}) = (\mathbf{X}'\mathbf{X})^{-1}\sigma^2, \tag{3.8}$$

en deze variantie kan geschat worden met

$$\hat{\mathrm{var}}_\xi(\hat{\beta}) = (\mathbf{X}'\mathbf{X})^{-1}s^2, \tag{3.9}$$

waarin $s^2 = \mathbf{e}'\mathbf{e}/(n-p)$, met $\mathbf{e}$ de vector met geobserveerde residuen en p het aantal parameters.

3.3 Design-variantie van de π-schatter voor de regressiecoëfficiënt

De design based benadering heeft betrekking op een wel omschreven eindige populatie. Ieder populatie-element k heeft een waarde y_k op de te verklaren variabele en een waarde $\mathbf{x}_k$ op de vector met regressoren. De y-waarden zowel als de x-waarden zijn vaste (niet-stochastische) grootheden. In deze benadering is de definitie van β als de parameter van de verdeling van y niet van toepassing. Parameters van eindige populaties geven een beschrijving van een specifieke

populatie en zijn functies van de vaste populatie-waarden van de onderzochte variabelen. De eindige-populatie regressiecoefficient is gedefinieerd als

$$\mathbf{B}_U = (\mathbf{X}'_U \mathbf{X}_U)^{-1} \mathbf{X}'_U \mathbf{y}_U , \tag{3.10}$$

met $\mathbf{X}_U$ en $\mathbf{y}_U$ de matrix met x-waarden en de vector met y-waarden voor alle N populatie-elementen, dus als de waarde die de OLS-schatter zou aannemen als de hele populatie geobserveerd was.

Voor de matrix $\mathbf{X}'_U \mathbf{X}_U$ geldt $\mathbf{X}'_U \mathbf{X}_U = \sum_{k=1}^{N} \mathbf{x}_k \mathbf{x}'_k$ en voor de vector $\mathbf{X}'_U \mathbf{y}_U$ geldt $\mathbf{X}'_U \mathbf{y}_U = \sum_{i=1}^{N} \mathbf{x}_k y_k$. Ieder element van de matrix $\mathbf{X}'_U \mathbf{X}_U$ en ieder element van de vector $\mathbf{X}'_U \mathbf{y}_U$ is dus een populatie-totaal en de eindige populatie parameter $\mathbf{B}_U$ is een (niet-lineaire) functie van populatietotalen. Een design-consistente schatter $\hat{\beta}_\pi$ voor $\mathbf{B}_U$ wordt verkregen door π-schatters te substitueren voor deze onbekende populatietotalen (Särndal et al., 1992, hfdst. 5). Dit leidt tot een gewogen kleinste kwadraten (WLS) schatter waarbij gewogen is met de inversen van de insluitkansen:

$$\hat{\beta}_\pi = (\mathbf{X}' \prod^{-1} \mathbf{X})^{-1} \mathbf{X}' \prod^{-1} \mathbf{y} = \hat{\mathbf{T}}^{-1} \hat{\mathbf{t}} , \tag{3.11}$$

met $\prod$=diag(π) en π de vector met insluitkansen voor de steekproefelementen.

De grootheid $\hat{\mathbf{T}} = \mathbf{X}' \prod^{-1} \mathbf{X} = \sum_{i \in S} \mathbf{x}_i \mathbf{x}'_i / \pi_i$ is een π-schatter voor de populatiematrix $\mathbf{T} = \mathbf{X}'_U \mathbf{X}_U$ en $\hat{\mathbf{t}} = \mathbf{X}' \Pi^{-1} \mathbf{y} = \sum_{i \in S} \mathbf{x}_i y_i / \pi_i$ is een π-schatter voor de populatievector $\mathbf{X}'_U \mathbf{y}_U$. De schatters $\hat{\mathbf{T}}$ en $\hat{\mathbf{t}}$ zijn lineaire grootheden, maar $\hat{\beta}_\pi$ is een niet-lineaire functie van deze lineaire grootheden. Een benadering voor de variantie van $\hat{\beta}_\pi$ kan gevonden worden door deze niet-lineaire functie te benaderen door een lineaire functie (via een eerste orde Taylor benadering).

De lineaire benadering voor $\hat{\beta}_\pi$ kan geschreven worden als (zie Särndal et al., 1992, Zeelenberg, 1996):

$$\hat{\beta}_\pi \approx \mathbf{B}_U + \mathbf{T}^{-1} (\hat{\mathbf{t}} - \hat{\mathbf{T}} \mathbf{B}_U) . \tag{3.12}$$

De design-variantie van $\hat{\beta}_\pi$ is dan bij benadering de variantie van de benadering voor $\hat{\beta}_\pi$, dus

$$\mathrm{var}_a(\hat{\beta}_\pi) \approx \mathbf{T}^{-1} \mathrm{var}_a(\hat{\mathbf{t}} - \hat{\mathbf{T}} \mathbf{B}) \mathbf{T}^{-1} \tag{3.13}$$

De stochastische component in (3.12) en (3.13) is te schrijven als

$$\begin{aligned} \hat{\mathbf{z}} &= \hat{\mathbf{t}} - \hat{\mathbf{T}} \mathbf{B} = \mathbf{X}' \prod^{-1} (\mathbf{y} - \mathbf{X} \mathbf{B}) = \sum_{i \in S} \mathbf{x}_i (y_i - \mathbf{x}'_i \mathbf{B}) \pi_i^{-1} \\ &= \sum_i \mathbf{z}_i / \pi_i \end{aligned} \tag{3.14}$$

De vector $\hat{\mathbf{z}}$ is een gewogen steekproeftotaal en kan opgevat worden als een π-schatter voor een vector met de overeenkomstige populatietotalen. De design-variantie van $\hat{\mathbf{z}}$ kan geschat worden met behulp van formules voor de varianties van

steekproeftotalen zoals (3.4), waarbij de onbekende **B** vervangen wordt door de schatter $\hat{\beta}_\pi$. Een schatter voor de variantie van $\hat{\beta}_\pi$ wordt vervolgens gevonden door voor de onbekende **T** in (3.14) de schatter $\hat{\mathbf{T}}$ te substitueren, dus

$$\hat{\text{var}}_a(\hat{\beta}_\pi) = \hat{\mathbf{T}}^{-1}\,\hat{\text{var}}_a(\hat{\mathbf{z}})\hat{\mathbf{T}}^{-1}. \tag{3.15}$$

Als voorbeeld kunnen we (3.15) uitwerken voor een eentrapssteekproef zonder stratificatie, maar met ongelijke kansen. Volgens (3.4) geldt dan

$$\hat{\text{var}}_a(\hat{\mathbf{z}}) = (1-f)\frac{n}{n-1}\sum_i(\hat{\mathbf{z}}_i-\hat{\bar{\mathbf{z}}})(\hat{\mathbf{z}}_i-\hat{\bar{\mathbf{z}}})' = (1-f)\frac{n}{n-1}\sum_i\hat{\mathbf{z}}_i\hat{\mathbf{z}}_i', \tag{3.16}$$

met $\hat{\mathbf{z}}_i = \mathbf{x}_i(y_i - \mathbf{x}_i'\hat{\beta}_\pi)\pi_i^{-1} = \mathbf{x}_i e_i \pi_i^{-1}$ en gebruik makend van $\hat{\bar{\mathbf{z}}} = \frac{1}{n}\sum_i\hat{\mathbf{z}}_i = \mathbf{0}$.

De geschatte design-variantie van $\hat{\beta}_\pi$ wordt dan volgens (3.15)

$$\begin{aligned}\hat{\text{var}}_a(\hat{\beta}_\pi) &= (1-f)\frac{n}{n-1}\mathbf{T}^{-1}\left(\sum_i\hat{\mathbf{z}}_i\hat{\mathbf{z}}_i'\right)\mathbf{T}^{-1}\\ &= (1-f)\frac{n}{n-1}\mathbf{T}^{-1}\left(\sum_i\mathbf{x}_i\pi e_i^2\pi_i\mathbf{x}_i'\right)\mathbf{T}^{-1}\\ &= (1-f)\frac{n}{n-1}(\mathbf{X}'\Pi^{-1}\mathbf{X})^{-1}\mathbf{X}'\Pi^{-1}diag(e^2)\Pi^{-1}\mathbf{X}(\mathbf{X}'\Pi^{-1}\mathbf{X})^{-1},\end{aligned} \tag{3.17}$$

met $diag(e^2)$ de diagonaalmatrix met op de diagonaal de gekwadrateerde residuen van de gewogen regressie. Als, in het bovenstaande voorbeeld de steekproef bovendien is getrokken met gelijke kansen, dus een SRS-design met teruglegging, dan wordt de schatter $\hat{\beta}_\pi$ gelijk aan de OLS-schatter $\hat{\beta}$ en reduceert (3.17) tot:

$$\hat{\text{var}}_a(\hat{\beta}) = (1-f)\frac{n}{n-1}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'diag(e^2)\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}. \tag{3.18}$$

Een vergelijking van (3.18) en (3.9) laat zien dat hoewel, in het geval van een SRS-design, de π-schatter voor **B** en de OLS-schatter voor β gelijk zijn, de schatters voor hun varianties verschillen. Dit verschil wordt veroorzaakt door de modelveronderstelling dat de varianties van de storingen gelijk zijn (homoscedasticiteit). Als in plaats hiervan verondersteld zou worden dat de storingen ongelijke varianties hebben (maar wel onafhankelijk zijn), dus $E_\xi\varepsilon\varepsilon' = diag(\omega)$ met ω een vector met de varianties σ_i^2 van de storingen ε_i dan is (3.18), afgezien van de eindigheidcorrectie die voor oneindige populaties niet van toepassing is, een consistente schatter voor $\text{var}_\xi(\hat{\beta})$ (v.g.l. (3.7) en zie White, 1980, Royal, 1986. De schatter (3.18) zonder eindigheidcorrectie wordt in de model-benadering dan ook gepropageerd als variantieschatter die robuust is tegen afwijkingen van homoscedasticiteit.

3.4 Gewogen analyse met standaard programmatuur.

De meeste grote statistische pakketten (waaronder SPSS) bieden niet de mogelijkheid om een design based analyse uit te voeren maar wel om gebruik te

maken van gewichten. Bij onderzoeken met ongelijke insluitkansen lijkt het voor de hand te liggen om bij gebruik van bijvoorbeeld SPSS een gewogen analyse uit te voeren. Een gewogen analyse kan op verschillende manieren uitgevoerd worden. Een mogelijkheid is om aan ieder record i een gewicht w_i te geven zodanig dat het record in analyses voor w_i records telt (replicatie-gewichten). Als w_i een geheel getal is, komt het wegen met w_i op hetzelfde neer als het w_i keer meenemen van record i. Bij regressieanalyse kunnen we met de meeste standaardprogrammatuur ook wegen door gebruik te maken van de Weighted Least Squares (WLS) optie. Als we gewichten nemen die gelijk zijn aan (of proportioneel met) de inversen van de insluitkansen, dan leiden beide vormen van gewogen analyse tot de WLS-schatter $\hat{\beta}_\pi$ die ook bij de design based benadering gebruikt werd.

Als gewichten genomen worden die proportioneel zijn met de inversen van de insluitkansen én zodanig geschaald zijn dat ze sommeren tot de steekproefomvang n, wat voor analyse-doeleinden voor de hand ligt, dan is ook de geschatte variantie van $\hat{\beta}_\pi$ voor de beide wegingsmethoden gelijk. Deze variantieschatter kan weergegeven worden als

$$\hat{\text{var}}_{\xi,WLS}(\hat{\beta}_\pi) = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1} s^2_{WLS}\,, \tag{3.19}$$

waarbij $s^2_{WLS} = \mathbf{e}'\mathbf{W}\mathbf{e}/(n-1)$, $\mathbf{e}$ de vector met residuen van het gewogen regressiemodel, (v.gl. 3.9) en $\mathbf{W}$ de diagonaalmatrix met de hierboven beschreven geschaalde gewichten. Het is duidelijk dat deze varianties verschillen van de design-variantie (3.17). Echter, ook vanuit een model standpunt is het moeilijk te verdedigen om deze variantieformules te gebruiken. De model variantie van een gewogen schatter heeft de algemene vorm (v.g.l. (3.7))

$$\text{var}_\xi(\hat{\beta}_\pi) = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}E_\xi(\varepsilon\varepsilon')\mathbf{W}\mathbf{X}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\,. \tag{3.20}$$

In de model based WLS-benadering wordt aangenomen dat de covariantiematrix $E_\xi(\varepsilon\varepsilon')$ van de residuen een diagonaalmatrix is met als diagonaal elementen $\sigma^2 E_\xi \varepsilon_i^2$. Bovendien wordt ervan uitgegaan dat gewichten gekozen worden die omgekeerd evenredig zijn met de varianties van de storingstermen, dus $w_i = a / E_\xi \varepsilon_i^2$ met a een willekeurige factor. Hierdoor reduceert (3.20) tot de WLS-variantie: $\text{var}_{\xi,WLS}(\hat{\beta}_\pi) = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\sigma^2$ en deze variantie kan geschat worden met (3.19).

De variantieschatter (3.19) die door standaardprogrammatuur wordt berekend is dus alleen correct als de gewichten zodanig gekozen zijn dat ze compenseren voor ongelijke varianties van de storingstermen. In dit rapport wordt echter uit gegaan van gewichten die compenseren voor ongelijke insluitkansen (proportioneel met Π^{-1}) en deze gewichten hoeven geen relatie te hebben met de varianties van de storingstermen. Voor de gewichten die we hier beschouwen is (3.19) dus in het algemeen geen correcte variantieschatter voor de gewogen schatter van de regressiecoefficiënt.

4. Illustratie

Ter illustratie laten we in deze paragraaf enkele regressieresultaten zien die berekend zijn op basis van het bestand van de Gezondheidsenquête 1993 van het CBS. Deze enquête is gebaseerd op een tweetrapssteekproef met aselecte trekking van huishoudens in de eerste trap en trekking met ongelijke kansen van personen binnen de geselecteerde huishoudens in de tweede trap. In de gewichten die dit bestand bevat zijn naast de ongelijke insluitkansen ook correcties voor eventuele onder- of oververtegenwoordiging van bepaalde groepen in de steekproef door nonrespons verdisconteerd. Deze gewichten worden in het volgende getallenvoorbeeld gebruikt alsof ze insluitgewichten waren.

We beschouwen een regressiemodel met als afhankelijke variabele *Gezondheid* en als predictoren *Leeftijd*, *Roken* en *Drinken*. De variabele *Gezondheid* is een variabele die de perceptie van de respondenten over hun gezondheid meet. De antwoorden zijn geschaald volgens een vijfpuntsschaal (1 slecht; 2 soms goed en soms slecht; 3 gaat wel; 4 goed; 5 zeer goed. De waarden die de predictoren kunnen aannemen zijn als volgt gedefinieerd

- *Leeftijd*: de leeftijd in jaren
- *Roken*: 1 nee, nooit gerookt; 2 nee, vroeger af en toe; 3 nee, vroeger elke dag; 4 ja, af en toe; 5 ja, elke dag.
- *Drinken*: 0 drinkt nooit; 1 drinkt wel.

Tabel 1 geeft de resultaten van, op verschillende manieren geschatte, lineaire regressies van de variabele *gezondheid* op de verklarende variabelen *leeftijd*, *drinken* en *roken*. Aan de coëfficiënten in de tweede kolom zien we dat het vorderen van de leeftijd een negatief effect heeft op de beleving van de gezondheid. Mensen die drinken beoordelen hun gezondheid positiever dan mensen die niet drinken, bij gelijke score op de andere kenmerken, maar mensen die (meer) roken beoordelen hun gezondheid juist negatiever dan mensen die niet of minder roken. Dit geldt zowel voor de gewogen als de ongewogen schatter voor de regressiecoëfficiënten.

De standaardfouten zijn op verschillende manieren geschat. Bij de ongewogen analyse zijn onder het kopje *model* de standaardfouten weergegeven die berekend zijn volgens de standaard modelveronderstellingen (i.i.d.-storingen), formule (3.9). In de volgende kolom, onder het kopje *design,SRS* zijn de standaardfouten weergegeven die berekend zijn volgens een design based methode waarbij (ten onrechte) aangenomen is dat het steekproefontwerp een enkelvoudige aselecte steekproef is, dus volgens formule (3.18). Onder deze laatste standaardfouten is tussen haakjes de verhouding van de standaardfout volgens het design en de modelveronderstellingen weergegeven. Het verschil tussen deze schatters van de standaardfouten is dat onder de modelveronderstellingen uitgegaan wordt van gelijke varianties van de storingen terwijl in de design based (SRS) schattingen mogelijke heteroscedasticiteit van de storingen verdisconteerd is.

Tabel 1. Regressie van Gezondheid op Leeftijd, Drinken en Roken, tussen haakjes staat de verhouding van de standaardfout volgens het desbetreffende design en het model.

variabele	coëfficiënt	geschatte standaardfouten		
		model	design, SRS	design, 2trap
Ongewogen				
constante	4,309	0,03574	0,03519	0,03745
			(0,9846)	(1,0478)
leeftijd	-0,013	0,00063	0,00064	0,00067
			(1,0159)	(1,0635)
drinken	0,367	0,02880	0,03276	0,03382
			(1,1375)	(1,1743)
roken	-0,038	0,00666	0,00669	0,00690
			(1,0045)	(1,0360)
Gewogen				
constante	4,313		0,03865	0,04109
			(1,0814)	(1,1497)
leeftijd	-0,013		0,00068	0,00071
			(1,0794)	(1,1270)
drinken	0,375		0,03462	0,03561
			(1,2021)	(1,2365)
roken	-0,043		0,00710	0,00735
			(1,0661)	(1,1036)

We zien dat het meenemen van heteroscedasticiteit in de schattingsprocedure over het algemeen leidt tot grotere standaardfouten. In de volgende kolom zijn onder het kopje *design,2trap* de design based standaardfouten weergegeven uitgaande van een tweetrapssteekproef (eerst huishoudens en dan personen). Ook hier is tussen haakjes de verhouding van de standaardfout volgens het design en de modelveronderstellingen weergegeven. Uit deze schattingen blijkt dat de standaardfouten nog wat groter worden als we, behalve met de heteroscedasticiteit ook rekening houden met het feit dat het hier om een tweetrapssteekproef gaat.

Bij de tot nu toe besproken resultaten is nog geen rekening gehouden met de ongelijke insluitkansen. Bij de gewogen analyses in het tweede deel van de tabel is dit wel het geval. In dit deel van de tabel zijn geen standaardfouten voor de model-benadering opgenomen omdat er, onder de i.i.d.-veronderstellingen van het model, geen reden is een gewogen analyse uit te voeren. Onder de design based standaardfouten is wel weer de verhouding weergegeven van de desbetreffende design based standaardfout en de standaardfout volgens de ongewogen model-benadering. Door het meenemen van de gewichten worden de standaardfouten groter

dan bij de ongewogen design based benaderingen en dus wordt ook het verschil met de ongewogen model-benadering groter. De standaardfouten die berekend zijn voor de gewogen analyse, rekening houdend met het tweetrapssteekproefontwerp, zijn standaardfouten die volgens een design based aanpak berekend zouden worden, zij doen het meeste recht aan het steekproefontwerp. Het zijn tevens ook de grootste standaardfouten. Zou men het steekproefontwerp negeren en voor een modelmatige analyse kiezen, dan zouden de standaardfouten onderschat worden en zou te snel tot significantie van de parameters besloten worden.

5. Conclusie

In dit artikel is ingegaan op de gevolgen van een complex steekproefontwerp bij het analyseren van de gegevens. Hierbij zijn de verschillen beschreven tussen model based en design based benaderingen. Door gewogen analyses uit te voeren is het met standaardprogrammatuur wel mogelijk om (design) zuivere schattingen van de parameters te verkrijgen. De model based variantieschattingen die door standaardprogrammatuur berekend worden bij een gewogen analyse zijn echter niet correct als de gewichten de inversen van de insluitkansen zijn. De gewogen model based benadering gaat er namelijk van uit dat de gewichten gekozen zijn om de varianties van de residuen van de regressievergelijking te vergroten/verkleinen. Bij gewichten die de inversen van de insluitkansen zijn is dit niet het geval.

In het algemeen verdient het bij analyses van complexe steekproeven aanbeveling om design based methoden te gebruiken. Hiervoor zijn specialistische software pakketten beschikbaar zoals SUDAAN en PC CARP, maar ook het algemene pakket STATA dat een breed scala aan statistische analyse methoden bevat kent een module waarmee design based (logistische) regressie-analyse uitgevoerd kan worden. Dit laatste pakket is gebruikt voor de illustratie in dit artikel.

6. Literatuur

Bethlehem, J., 2000, De Klassieke Steekproeftheorie - Een Overzicht. Kwantitatieve Methoden, dit nummer.

Cate, A. ten, 1986. Regression analysis using survey data with endogenous design. Survey Methodology, vol. 12, pp. 121-138.

Kish, L. 1965, Survey Sampling. John Wiley & Sons (New York).

Kish, L., 1992. Weighting for unequal P_i. Journal of Official Statistics 8, pp. 183-200.

Nathan, G. & T.M.F. Smith, 1989. The effect of selection on regression analysis. In: Skinner, C.J., D. Holt & T.M.F. Smith, 1989. Analysis of complex surveys. John Wiley & Sons (Chichester, New York).

Royall, R.M., 1986, Model robust confidence intervals using maximum likelihood estimators. International Statistical Review, vol. 54, pp. 221-226.

Särndal, C.E., B. Swensson en J. Wretman, 1991, Model Assisted Survey Sampling. Springer-Verlag (New York).

Skinner, C.J., D. Holt & T.M.F. Smith, 1989. Analysis of complex surveys. John Wiley & Sons (Chichester, New York).

White, H., 1980, A heteroscedasticity-consistent covariance matrix estimator and a direct test for heteroscedasticity. Econometrica, vol. 41, pp. 733-750.

Zeelenberg, C., 1996, A one-line derivation of the linearization of the regression coefficient estimator and the regression estimator. Research Paper no. 9619, Department of Statistical Methods (Statistics Netherlands, Voorburg).