

REGRESSIE-ANALYSE OP BASIS VAN EEN GESTRATIFICEERDE STEEKPROEF

J.P. Elhorst

Samenvatting

Dit artikel gaat in op de vraag of men bij regressie-analyse rekening dient te houden met wegingsfactoren in die gevallen waarin de opbouw van de steekproef wordt gereguleerd via een onderverdeling in strata en het trekken daaruit van afzonderlijke steekproeven met ongelijke steekproefpercentages.

Het antwoord op deze vraag is tweeledig. Het advies is om te streven naar een specificatie van de schattingsvergelijking waarin geen significant verschil optreedt in de regressie-coëfficiënten volgens de methoden van gewogen en ongewogen kleinste kwadraten. Onder deze voorwaarde wordt het gedrag, dat men met het model wil verklaren, namelijk beter beschreven. Een toets om te bepalen of het verschil tussen beide significant is wordt in dit artikel besproken. Is men niet in staat om een specificatie op te sporen die aan deze voorwaarde voldoet, dan moet de ongewogen schattingsmethode worden verworpen ten gunste van de gewogen schattingsmethode.

J.P. Elhorst
Landbouw-Economisch Instituut
Conradkade 175
2517 CL s'-Gravenhage
tel. 070-614161.

1. Inleiding

De literatuur op het terrein van de steekproeftrekking richt zich op de schatting van Y , geschreven als $\hat{Y}$, en haar standaardfout, $\text{stf}(\hat{Y})$. De schatting van Y behoeft niet betrekking te hebben op slechts één element. Ook totalen $\sum Y_i$, gemiddelden $\sum Y_i / n$, gewogen gemiddelden $\sum W_i Y_i / \sum W_i$, ratio- of verschilschatters van gemiddelden en totalen, correlatie- en regressiecoëfficiënten komen hiervoor in aanmerking. $\text{Stf}(\hat{Y}) = \sqrt{\text{var}(\hat{Y})}$ is de geschatte standaardfout, berekend uit de elementen in de steekproef en als het goed is in overeenstemming met de wijze van steekproeftrekking. Het doel van deze statistische grootheden is om informatie te verschaffen over de waarde van Y in de populatie middels betrouwbaarheidsintervallen van de vorm $\hat{Y} \pm t_p \cdot \text{stf}(\hat{Y})$.

Met opzet is hiervoor gezegd "als het goed is", want de praktijk is dat meer en meer gebruik wordt gemaakt van complexe steekproeftechnieken zonder dat de berekening van betrouwbaarheidsintervallen hier op aansluit. In de woorden van Kish en Frankel (1974): "We think it imperative and urgent to extend to more complex statistics. More and more researchers are able to obtain data from complex samples, and write computer programs for complex analytical statistics. We need methods for dealing properly with complex statistics from complex samples. We need statistics for probability statements. Such statements are symbolized with $\hat{Y} \pm t_p \cdot \text{stf}(\hat{Y})$, where $\hat{Y}$ is some complex statistic, and $\text{stf}(\hat{Y})$ its computed standard error. Standard errors should be computed in accord with the complexity of the sample designs; neglect of that complexity is a common source of serious mistakes". Dat de berekening van betrouwbaarheidsintervallen niet altijd aansluit op de wijze van steekproeftrekking is verklaarbaar, daar vooral de bepaling van de standaardfouten bijzonder gecompliceerd is en in vele gevallen nog niet uitgezocht.

Een goed overzicht van de stand van zaken met betrekking tot het bepalen van betrouwbaarheidsintervallen voor statistische grootheden in relatie tot de wijze van steekproeftrekking is gegeven door dezelfde Kish en Frankel, opgenomen in tabel 1. Terzijde kan worden opgemerkt dat dit overzicht niet volledig is met betrekking tot de wijze van steekproeftrekking en arbitrair met betrekking tot de statistische grootheden. Niet volledig, omdat meer methoden van steekproeftrekking bekend zijn. In dit artikel beperken wij ons echter tot gestratificeerde steekproeven. Arbitrair, omdat ook een andere indeling van statistische grootheden mogelijk is. Volgens Kish en Frankel echter geeft deze indeling het best de stand van zaken weer.

Rij 1 heeft betrekking op het bepalen van betrouwbaarheidsintervallen voor statistische grootheden bij een

Tabel 1 Stand van zaken met betrekking tot het bepalen van betrouwbaarheidsintervallen voor statistische grootheden in relatie tot de wijze van steekproeftrekking.

wijze van steekproef- trekking	statistische grootheid									
	gemiddelden en totalen van de populatie			gemiddelden en verschillen tussen gemiddelden van subgroepen				complexe statistische grootheden, bijvoorbeeld regressie- coëfficiënten		
enkelvoudige steekproef	S	T	A	N	D	A	A	R	D	
	T									
	A									
	N									
	D									
gestrati- ficeerde steekproef	A					beschik- baar			in ontwikkeling	
	A									
	R									
	D									

Bron : Kish en Frankel (1974)

enkelvoudige steekproef. Dit is standaard theorie, waarover boekenkasten zijn vol geschreven. De ontwikkeling in deze rij staat niet stil, maar blijft voortdurend in beweging. De literatuur handelend over steekproeven is in hoofdzaak toegelegd op kolom 1.

Hoe in een gestratificeerde steekproef betrouwbaarheidsintervallen te berekenen voor gemiddelden van subgroepen of verschillen hiertussen (rij 2, kolom 2) is bekend - zie Moors en Muilwijk (1975) alsmede Cochran (1977) -, maar nog geen gemeengoed. Vandaar dat niet de kwalificatie "standaard" is toegekend. Hoe in een gestratificeerde steekproef betrouwbaarheidsintervallen te berekenen voor complexe statistische grootheden (rij 2, kolom 3) is in vele gevallen nog niet bekend, maar bevindt zich in ontwikkeling. Dit artikel tracht een uitspraak te doen over regressiecoëfficiënten, en spitst zich toe op de vraag of men rekening dient te houden met wegingsfactoren omgekeerd evenredig aan de steekproefpercentages binnen de verschillende strata.

2. Standaard regressiemethoden

Een onderzoeker die een regressie-analyse wil uitvoeren op

basis van een gestratificeerde steekproef maakt in het algemeen de keuze uit onderstaande methoden:

A. Kleinste kwadraten (OLS).

B. Gewogen kleinste kwadraten (WLS) op basis van de wegingsfactoren.

Methode A is gebaseerd op het lineaire regressiemodel dat als volgt kan worden beschreven

$$Y_i = b_i' X_i + u_i, \quad i = 1, \dots, n, \quad u_i \sim N(0, \sigma^2)$$

met Y_i de te verklaren variabele, X_i een $(k \times 1)$ vector van verklarende variabelen, u_i de storingsterm, b een $(k \times 1)$ vector van te schatten parameters en n het aantal waarnemingen. Als uitkomsten levert deze methode

$$\hat{b} = (X'X)^{-1} X'Y \quad \text{en} \quad \text{var}(\hat{b}) = \sigma^2 (X'X)^{-1}$$

Methode B, zoals deze in het algemeen wordt toegepast in standaard regressieprogrammatuur, bijvoorbeeld SPSS, BMDP en Genstat, is gebaseerd op een regressiemodel analoog aan het lineaire regressiemodel, met dit verschil dat wordt uitgegaan van de veronderstelling $u_i \sim N(0, \sigma^2/W_i)$, waarbij W_i de wegingsfactor van element i . Als uitkomsten levert deze methode

$$\hat{b} = (X'WX)^{-1} X'WY \quad \text{en} \quad \text{var}(\hat{b}) = \sigma^2 (X'WX)^{-1}$$

met W de diagonaal matrix van de wegingsfactoren. De formulering voor de bepaling van de variantie van de storingsterm (σ^2) en de determinatiecoëfficiënt (R^2) wordt in beide gevallen overgelaten aan de lezer.

3. Regressie-analyse zonder en met wegingsfactoren

De theorie die aan het lineaire regressiemodel ten grondslag ligt berust op drie veronderstellingen:

- 1) $E(u_i | X_i) = 0$ voor alle i .
- 2) homoskedasticiteit: $\text{var}(u_i | X_i) = \sigma^2$ voor alle i
- 3) onafhankelijkheid van de waarnemingen:

$$\text{cov}(u_i, u_j | X_i, X_j) = 0 \quad \text{voor alle } i \neq j$$

De veronderstelling onder 1) is fundamenteel voor het

regressiemodel en wil zeggen dat de verwaarloosde termen die in u_i zijn opgenomen onafhankelijk zijn van de waarden die X_i kan aannemen, of ook dat men de storingen in een gedachtenexperiment door middel van een aselektor over de waarnemingen zou kunnen verdelen zonder dat dit kan worden opgemerkt. De veronderstellingen onder 2) en 3) zijn niet noodzakelijk en kunnen worden verzwakt.

De eigenschappen die toebehoren aan de kleinste kwadraten schatter van b zijn in de literatuur uitgebreid besproken: lineair, zuiver en met minimale variantie. Te weinig echter wordt ingegaan op de wijze waarop de steekproefelementen uit de populatie zijn getrokken. Dit is een tekortkoming, juist in die gevallen waarin de opbouw van de steekproef wordt gereguleerd via een onderverdeling in strata en het trekken daaruit van afzonderlijke steekproeven met ongelijke steekproefpercentages. Dit zal worden toegelicht.

Het doel van stratificatie is om de betrouwbaarheid van de steekproefuitkomsten, in het bijzonder van gemiddelden en totalen (zie tabel 1, kolom 1) te vergroten. Door te stratificeren wordt de standaardfout van de steekproefuitkomsten teruggebracht ten opzichte van de standaardfout welke uit een enkelvoudige steekproef zou resulteren. Essentieel voor de reductie die optreedt in de standaardfout is de mate van correlatie van de onderzoeksvariabele(n) en de stratificatievariabele(n): stratificatie verhoogt alleen dan de betrouwbaarheid belangrijk als deze twee hoog gecorreleerd zijn (zie bijvoorbeeld Moors en Muilwijk, 1975, blz 63). Op het Landbouw-Economisch Instituut, waar een gestratificeerde steekproef wordt getrokken uit alle in Nederland geregistreerde landbouwbedrijven boven een bepaalde minimumomvang, gelden als onderzoeksvariabelen "het netto-overschot per bedrijf" en "de arbeidsopbrengst van de ondernemer". Hierbij wordt uitgegaan van de veronderstelling dat deze twee variabelen maatgevend zijn voor een breed scala van achterliggende variabelen (Lodder, 1987, blz 15), zoals opbrengsten- en kostenpatronen alsmede de inkomensvorming en -besteding van de ondernemer en zijn gezinsleden. Men kan dan ook met zekerheid zeggen dat een groot deel van de variabelen, die in deze steekproef zijn verzameld, wordt beïnvloed door de indeling in afzonderlijke strata, of, anders gezegd, dat deze indeling informatie verschaft over de hoogte van de onderzoeksvariabelen. Het gevolg hiervan is dat niet voldaan is aan de veronderstelling $E(u_i | X_i) = 0$, waarop de theorie van het lineaire regressiemodel berust. Dit betekent dat de schattingsmethode aanpassing behoeft en het is deze aanpassing die belangrijk is voor de vraag of het nodig is om bij regressie-analyse rekening te houden met wegingsfactoren.

Gegeven het steekproefpercentage p_i behorend bij ieder element in de steekproef - veronderstel dat de populatie is

verdeeld in H strata en dat de omvang van stratum h ($h=1, \dots, H$) gelijk is aan N_h . Als uit stratum h n_h steekproefelementen zijn getrokken, dan is het steekproefpercentage in dat stratum n_h/N_h (*100%). Daar ieder steekproefelement tot precies één stratum behoort, kan het steekproefpercentage alleen variëren tussen elementen die uit verschillende strata zijn getrokken - en afhankelijk van het model dat men specificiert zijn drie aanpassingen te onderscheiden:

- 1) de stratificatievariabele (S geheten) is tevens de te verklaren variabele

$$S_i = b_i' X_i + u_i, \quad u_i \sim N(0, \sigma^2), \quad W_i = 1/p_i, \quad i=1, \dots, n.$$

Dit is een uitzonderlijk geval, omdat de stratificatievariabele in het algemeen geen onderzoeksvariabele is. Wil men niettemin deze variabele verklaren, dan dient te worden uitgegaan van een zogenaamd truncated regressiemodel, omdat de te verklaren variabele binnen ieder stratum bepaalde waarden niet kan aannemen. Voor een bespreking van deze klasse van modellen zij verwezen naar Maddala (1983, blz 170-174).

- 2) de stratificatievariabele is tevens een verklarende variabele

$$Y_i = a_i S_i + b_i' X_i + u_i, \quad u_i \sim N(0, \sigma^2), \quad W_i = 1/p_i, \quad i=1, \dots, n.$$

Volgens Holt et al. (1980) kan in dit geval de OLS-schattingmethode worden toegepast, omdat voldaan is aan de eis $E(u_i | S_i, X_i) = 0$. Volgens ons echter kan de OLS-schattingmethode alleen worden toegepast als men veronderstelt dat geen interactie bestaat tussen de verklarende variabelen en de stratificatievariabele alsmede als men veronderstelt dat het verband tussen de verklarende variabele en de stratificatievariabele lineair is. Aangezien dit als vrij uitzonderlijk geldt, kan men dit geval beter scharen onder het nu volgende.

- 3) de stratificatievariabele is geen onderdeel van de regressievergelijking

$$Y_i = b_i' X_i + u_i, \quad u_i \sim N(0, \sigma^2), \quad W_i = 1/p_i, \quad i=1, \dots, n.$$

De schatting van de regressiecoëfficiënten b die in dit geval door verschillende auteurs (Kish en Frankel, 1974; Holt et al., 1980, gebaseerd op vier referenties; en DuMouchel en Duncan, 1983) wordt voorgesteld is

$$\hat{b}_w = (X'WX)^{-1} X'WY.$$

Merk op dat de schatting van b overeenkomt met de gewogen schatting van methode B, maar dat een belangrijk verschil ontstaat bij de berekening van de variantie-covariantie matrix.

$$\left. \begin{aligned} \text{var}(\hat{b}_w) &= E(\hat{b}_w - b)(\hat{b}_w - b)' \\ Y &= Xb + u \end{aligned} \right\}$$

$$\begin{aligned} \text{var}(\hat{b}_w) &= E[(X'WX)^{-1} X'W(Xb+u) - b][(X'WX)^{-1} X'W(Xb+u) - b]' = \\ &= E[(X'WX)^{-1} X'Wuu'WX(X'WX)^{-1}] = \\ &= (X'WX)^{-1} X'W E(uu') WX(X'WX)^{-1} = \\ &= (X'WX)^{-1} X'W \sigma^2 I_n WX(X'WX)^{-1} = \\ &= \sigma^2 (X'WX)^{-1} X'W X (X'WX)^{-1} \end{aligned}$$

Het blijkt dat de $\text{var}(\hat{b}_w)$ niet gelijk is aan $\sigma^2 (X'WX)^{-1}$, hetgeen het geval zou zijn onder de veronderstelling $u_i \sim N(0, \sigma^2/W_i)$. Anders gezegd, de standaard regressieprogrammatuur voldoet niet als de wegingsfactoren voortkomen uit een gestratificeerde steekproef. De schatting van b is juist, maar de variantie-covariantie matrix en daarmee de standaardfouten en de T-waarden, die uit deze matrix worden afgeleid, niet.

Hoe in dit specifieke geval de variantie van de storingsterm (σ^2) en de determinatiecoëfficiënt (R^2) te bepalen wordt door de verschillende auteurs niet behandeld en toont nog eens aan hoezeer de bepaling van statistische grootheden op basis van een gestratificeerde steekproef in ontwikkeling is. Om toch een compleet beeld te verkrijgen, stellen wij zelf een formulering voor die mede is ontleend aan Cochran (1977, hoofdstuk 7)

$$\hat{\sigma}^2 = \frac{1}{\sum_{i=1}^n W_i} \sum_{h=1}^H \sum_{i \in h} W_i (Y_i - \bar{Y}_h - b'(X_i - \bar{X}_h))^2,$$

$$R^2 = 1 - \frac{\sum_{h=1}^H \sum_{i \in h} W_i (Y_i - \bar{Y}_h - b'(X_i - \bar{X}_h))^2}{\sum_{h=1}^H \sum_{i \in h} W_i (Y_i - \bar{Y})^2}.$$

De aanpassing die is voorgesteld als de stratificatievariabele geen onderdeel is van de regressievergelijking is het meest voorkomende geval en wijkt af van de standaard regressiemethoden beschreven onder A en B. Wellicht is dit de reden dat de aanpassing door verschillende personen wordt betwist, hetgeen is uiteengezet door DuMouchel en Duncan (1983). Elementen van deze controverse zijn ook terug te vinden in de verschillende discussiebijdragen toegevoegd aan het artikel van Kish en Frankel (1974). In het kort komen de meningen hier op neer. De voorstanders van de OLS-schattingsmethode beargumenteren dat de parameters in een regressievergelijking onafhankelijk zijn van de wijze van stratificatie. Als de landbouwbedrijven bijvoorbeeld worden ingedeeld naar de kleur van dakpannen, zeg rood en zwart, dan is er geen reden om in een regressievergelijking rekening te houden met de steekproefpercentages binnen deze twee typen van landbouwbedrijven. De voorstanders van de gewogen schattingsmethode beargumenteren dat de wegingsfactoren moeten worden gebruikt om redenen, analoog aan de bepaling van gemiddelden en totalen, namelijk dat bepaalde groepen in de steekproef zijn onder- of oververtegenwoordigd.

Daarnaast wordt tegen de argumenten van de voorstanders van de OLS-schattingsmethode ingebracht, dat stratificatie is gebaseerd op variabelen die hoog zijn gecorreleerd met de onderzoeksvariabelen. Het indelen van landbouwbedrijven naar de kleur van dakpannen is dus zeker niet maatgevend voor de wijze waarop wordt gestratificeerd. Alsmede dat de parameterwaarden, die in de regressievergelijking zijn opgenomen, alleen dan onafhankelijk zijn van de wijze van stratificatie als de regressievergelijking juist is gespecificeerd. Dat wil zeggen als geen verklarende variabelen in de vergelijking ontbreken en ook als het functioneel verband tussen de te verklaren variabele Y en de vector van verklarende variabelen X in overeenstemming is met de werkelijkheid. Om dit te illustreren zie figuur 1. Deze

schattingsvergelijking onjuist is gespecificeerd, zolang de wegingsfactoren afgeleid van de steekproefpercentages binnen de verschillende strata de ligging van de regressielijn significant beïnvloeden. Zolang ook zal men op zoek moeten gaan naar mogelijkheden om de specificatie te verbeteren, althans om de OLS-schattingsmethode te kunnen rechtvaardigen. In woorden van DuMouchel en Duncan (1983): "the rationale for preferring unweighted to weighted regression is rejected unless some other variables can be found that lead one to accept an extended model". Wij zouden deze zinsnede willen uitbreiden tot tevens een verbetering van het functioneel verband tussen Y en X . Het advies is dan ook pas dan de OLS-schattingsmethode toe te passen als geen significant verschil bestaat tussen de onderbroken en de ononderbroken regressielijn, ofwel tussen de schatting van b volgens de OLS-schattingsmethode en volgens de gewogen schattingsmethode. Is men daar niet toe in staat, bijvoorbeeld omdat verschillende specificatievormen niet hebben geleid tot dit gewenste resultaat, dan moet de OLS-schattingsmethode worden verworpen ten gunste van de gewogen schattingsmethode. De best mogelijke schatting van de regressiecoëfficiënten, uitgaande van de veronderstelling dat de bijbehorende specificatie tevens langs een aantal andere wegen is getest, wordt in dat geval verkregen door rekening te houden met de hoogte van de wegingsfactoren.

4. Toets op het gebruik van wegingsfactoren

Om te bepalen of een significant verschil bestaat tussen de OLS-schattingsmethode en de gewogen schattingsmethode wordt aangesloten op een toets die is beschreven door DuMouchel en Duncan (1983). Centraal in deze toets staat het verschil in de schatting van b volgens de gewogen schattingsmethode en de OLS-schattingsmethode, dat is gedefinieerd als

$$\hat{D} = \hat{b} - \hat{b}^w \quad \text{met} \quad \text{var}(\hat{D}) = \sigma^2 A A' \quad \text{en} \quad A = (X'WX)^{-1} X'W - (X'X)^{-1} X'.$$

Het te toetsen schattingsmodel is

$$Y_i = b'X_i + u_i, \quad i=1, \dots, n, \quad u_i \sim N(0, \sigma^2),$$

terwijl als alternatief schattingsmodel wordt onderscheiden

$$Y_i = b'X_i + c'Z_i + u_i, \quad i=1, \dots, n, \quad u_i \sim N(0, \sigma^2),$$

waarbij Z_i een vector van verklarende variabelen die in het te toetsen schattingsmodel ontbreken. Dit kunnen ook

tussenprodukten zijn van variabelen onder de vector X_i .

De toets op weging beschreven door DuMouchel en Duncan komt hier op neer dat de F-toets op $D=0$ kan worden vervangen door de bekende F-toets op $c=0$ als zou het te toetsen schattingsmodel zijn gespecificeerd als

$$Y_i = b_i' X_i + c_i' W_i + u_i, \quad i=1, \dots, n, \quad u_i \sim N(0, \sigma^2),$$

en geschat volgens de OLS-schattingsmethode. Met andere woorden: bepaal de variabelen $Z=WX$, waaronder ook de constante is begrepen, en toets of de invloed van Z , weergegeven door de vector van parameters c , significant van nul verschillend is. Dit leidt tot een analyse van de variantie toe te schrijven aan de wegingsfactoren zoals weergegeven in tabel 2.

Tabel 2 Toets op het gebruik van wegingsfactoren

	aantal vrijheidsgraden	residuele kwadratensom	gemiddelde residuele kwadratensom
Regressie zonder Z variabelen	$n-k$	SS ols	
Regressie met Z variabelen	$n-2k$	SS w	SS / $(n-2k)$ w
Verschil toe te schrijven aan de wegingsfactoren	k	SS - SS ols w	(SS - SS) / k ols w

SS_{ols} en SS_w kunnen aan de berekeningen worden ontleend, n is het aantal waarnemingen en k het aantal variabelen in het te toetsen schattingsmodel.

De toetsgrootheid met een $F(k, n-2k)$ verdeling wordt verkregen door de twee grootheden, die helemaal rechts staan, op elkaar te delen

$$\frac{(SS_{ols} - SS_w) / k}{SS_w / (n-2k)}.$$

Het voordeel van deze toets is dat zij is uit te voeren zonder

dat de gewogen schatting van b bepaalt behoeft te worden. Zij is opgebouwd uit een drietal stappen. Voer een regressie uit van Y op de variabelen X en bereken SS_{ols} . Deze wordt in het algemeen geleverd door de standaard regressieprogrammatuur. Voer vervolgens een regressie uit van Y op de variabelen X en Z met $Z=WX$ en bereken SS_w . Vul tot slot de tabel in en bereken de toetsgrootheid.

5. Voorbeelden

Om de uitkomsten van regressie-analyse zonder en met wegingsfactoren alsmede de werking van de toets te illustreren is een schatting gemaakt van de productiefunctie voor de rundveehouderij en de akkerbouw, waarbij gekozen is voor een vorm die bekend staat als de translog-productiefunctie. Deze functie is als volgt gespecificeerd

$$\ln PY_b = a_0 + [a_1 \ a_2 \ a_3 \ a_4] \begin{bmatrix} \ln H \\ \ln L \\ \ln K \\ \ln M \end{bmatrix}_b + \frac{1}{2} \begin{bmatrix} \ln H \\ \ln L \\ \ln K \\ \ln M \end{bmatrix}_b' \begin{bmatrix} A_{11} & A_{12} & A_{13} & A_{14} \\ A_{21} & A_{22} & A_{23} & A_{24} \\ A_{31} & A_{32} & A_{33} & A_{34} \\ A_{41} & A_{42} & A_{43} & A_{44} \end{bmatrix} \begin{bmatrix} \ln H \\ \ln L \\ \ln K \\ \ln M \end{bmatrix}_b$$

$$A_{ij} = A_{ji}, \quad i, j = 1, \dots, 4.$$

PY = bruto-bedrijfsopbrengst (gld).

H = oppervlakte in hectaren kadastraal (ha).

L = aantal gezins- en vreemde arbeidskrachten (vak).

K = de kapitaalgoederenvoorraad berekend als de rente die aan de kapitaalgoederenvoorraad in rekening wordt gebracht (gld).

Tot de kapitaalgoederenvoorraad wordt gerekend gebouwen inclusief pachtersinvesteringen, werktuigen en vee. De laatste alleen als sprake is van de rundveehouderij.

M = de inzet van non-factor inputs opgebouwd uit en berekend als de kosten van loonwerk, bestrijdingsmiddelen, zaai-, plant- en pootgoederen, energie, onderhoud aan gebouwen en werktuigen, veevoer en meststoffen (gld).

b = bedrijfsindex.

De translog-functie en de wijze waarop de inzet van de produktiefactoren is berekend staat in dit artikel niet ter discussie. Het gaat ons uitsluitend om een vergelijking van de uitkomsten. Voor de discussie zij verwezen naar Elhorst (1986). De data die bij de schattingen zijn gebruikt zijn afkomstig van bedrijven, die het Landbouw-Economisch Instituut gebruikt voor het opstellen van bedrijfsuitkomsten in de landbouw. Zij bestaan uit de bedrijfseconomische administratie van bedrijfsgegevens over het boekjaar 1984/85. Daar in de steekproef wordt gestratificeerd naar bedrijfstype en bedrijfsgrootte, maar onder de bedrijfstypering ook de rundveehouderij en de akkerbouw zijn begrepen, bleef de bepaling van wegingsfactoren beperkt tot de verschillende bedrijfsgrootteklassen. Voor meer achtergrondinformatie zij verwezen naar Elhorst (1988).

Uit de schattingsresultaten blijkt dat de toepassing van de OLS-schattingsmethode niet kan worden verworpen voor de rundveehouderij. Er bestaat geen significant verschil, bij een betrouwbaarheidsdrempel van 95%, tussen de schatting van de parameters volgens de OLS-schattingsmethode en de gewogen schattingsmethode. Wij zouden dan ook kunnen zeggen dat de translog-functie een goede beschrijving geeft van het productieproces op het rundveehouderijbedrijf ongeacht de grootte van het bedrijf.

De toepassing van de OLS-schattingsmethode kan wel worden verworpen voor de akkerbouw. Dit doet ons belanden in een stadium dat een functie die wordt gekenmerkt door flexibiliteit - de translog-functie is een 2e orde benadering voor de produktiefunctie uitgedrukt in logaritmen - niet leidt tot het gewenste resultaat, in die zin dat wij geen mogelijkheden zien om de specificatie zodanig te verbeteren, dat de OLS-schattingsmethode kan worden gerechtvaardigd. Natuurlijk bestaan naast het functioneel verband mogelijkheden om de specificatie te verbeteren, door stil te staan bij:

- a) de bepaling van de produktiefactoren, dat wil zeggen het aantal produktiefactoren dat wordt onderscheiden en de wijze waarop ze worden gemeten; en
- b) de groep of populatie waarop de schattingen betrekking hebben,

maar of deze mogelijkheden leiden tot het gewenste resultaat is zeer de vraag. De schatting van de produktiefunctie voor de akkerbouw vormt dan ook een goed voorbeeld van een onderzoek, waarin men volstaat met de translog-functie geschat volgens de gewogen schattingsmethode.

Tabel 3a Schattingsresultaten verkregen voor de rundveehouderij bij regressie-analyse zonder en met wegingsfactoren *).

Verklarende variabelen	regressie-analyse zonder wegingsfactoren		regressie-analyse met wegingsfactoren		het verschil tussen beide	
	par.	T-waarde	par.	T-waarde	par.	T-waarde
Constante	2.1657	1.33	3.3281	1.87	-1.1624	-1.97
H	1.2325	4.36	1.0505	3.46	.1819	2.02
L	-1.3073	-1.94	-1.2392	-1.72	-.0681	-.33
K	1.0248	2.86	1.0744	2.75	-.0496	-.37
M	-.2222	-.69	-.4197	-1.22	.1976	2.00
H * H	.0174	.30	.0221	.35	-.0047	-.24
H * L	.1835	1.98	.1933	1.93	-.0098	-.31
H * K	.1193	2.04	.1293	2.06	-.0100	-.53
H * M	-.2082	-4.22	-.2035	-3.82	-.0046	-.28
L * L	-.5316	-1.90	-.7166	-2.30	.1851	1.58
L * K	.0492	.41	.0843	.66	-.0351	-.92
L * M	.0667	.64	.0412	.37	.0252	.71
K * K	-.2109	-2.37	-.2146	-2.28	.0038	.16
K * M	.0725	1.05	.0660	.90	.0066	.36
M * M	.0625	1.01	.0859	1.31	-.0234	-1.41
R ²	.97		.85			

*) gebaseerd op 514 waarnemingen over het boekjaar 1984/85.

Tabel 3b Toets op het gebruik van wegingsfactoren

	aantal vrijheidsgraden	residuele kwadratensom	gemiddelde residuele kwadratensom
Regressie zonder Z variabelen	499	5.756	
Regressie met Z variabelen	484	5.423	.011
Verskil toe te schrijven aan de wegingsfactoren	15	.333	.022

uitkomst toetsgrootheid .022/.011=1.98
kritische grens 2.07 (betrouwbaarheidsdrempel 95%)

Tabel 4a Schattingsresultaten verkregen voor de akkerbouw bij regressie-analyse zonder en met wegingsfactoren *).

Verklarende variabelen	regressie-analyse zonder wegingsfactoren		regressie-analyse met wegingsfactoren		het verschil tussen beide	
	par.	T-waarde	par.	T-waarde	par.	T-waarde
Constante	4.9419	1.28	6.4285	1.63	-1.4867	-1.24
H	1.5699	2.39	1.7360	2.48	-.1661	-.59
L	2.6993	1.92	2.8938	1.99	-.1944	-.40
K	-1.7099	-2.37	-1.9450	-2.56	.2351	.82
M	1.1710	1.68	1.0481	1.47	.1229	.56
H * H	.1789	1.14	.1028	.59	.0761	.90
H * L	-.2228	-1.39	-.1452	-.86	-.0776	-1.19
H * K	-.0520	-.57	-.0442	-.46	-.0078	-.22
H * M	-.1148	-1.41	-.1181	-1.36	.0033	.09
L * L	.6972	2.23	1.0500	3.10	-.3529	-2.37
L * K	-.1188	-1.31	-.1565	-1.68	.0376	1.27
L * M	-.0812	-.56	-.1171	-.77	.0359	.69
K * K	.1173	2.53	.1369	2.78	-.0195	-1.01
K * M	.0827	1.01	.0868	1.02	-.0040	-.13
M * M	-.0702	-.63	-.0594	-.51	-.0108	-.27
R ²	.93		.74			

*) gebaseerd op 263 waarnemingen over het boekjaar 1984/85.

Tabel 4b Toets op het gebruik van wegingsfactoren

	aantal vrijheidsgraden	residuele kwadratensom	gemiddelde residuele kwadratensom
Regressie zonder Z variabelen	248	6.918	
Regressie met Z variabelen	233	5.485	.024
Vershil toe te schrijven aan de wegingsfactoren	15	1.433	.096

uitkomst toetsgrootheid .096/.024=4.06

kritische grens 2.07 (betrouwbaarheidsdrempel 95%)

6. Conclusie

Dit artikel heeft zich toegespitst op de vraag of men bij regressie-analyse rekening dient te houden met wegingsfactoren omgekeerd evenredig aan het steekproefpercentage indien sprake is van een gestratificeerde steekproef.

Het antwoord op deze vraag is tweeledig. Het advies is om te streven naar een specificatie waarin geen significant verschil optreedt in de schatting van b volgens de OLS-schattingsmethode en de gewogen schattingsmethode. Onder deze voorwaarde wordt het gedrag, dat men met het model wil verklaren, namelijk beter beschreven. Een toets om te bepalen of het verschil tussen beide significant is kan worden uitgevoerd zonder dat de gewogen schatting van b , $\hat{b}_w$, bepaald behoeft te worden. Deze toets is hierboven besproken. Is men niet in staat om een specificatie op te sporen die aan deze voorwaarde voldoet, dan moet de OLS-schattingsmethode worden verworpen ten gunste van de gewogen schattingsmethode. De best mogelijke schatting van de regressiecoëfficiënten, uitgaande van de veronderstelling dat de bijbehorende specificatie tevens langs een aantal andere wegen is getest, wordt in dat geval verkregen door rekening te houden met de hoogte van de wegingsfactoren.

Past men de gewogen schattingsmethode toe, dan is voorzichtigheid geboden, omdat de variantie-covariantie matrix en daarmee de standaardfouten en de T-waarden berekend volgens de standaard regressieprogrammatuur niet voldoen.

De vraag die dit resultaat tenslotte oproept is hoe het is gesteld met andere multivariate analysetechnieken, zoals clusteranalyse, factoranalyse, discriminantie-analyse en variantie-covariantie analyse. Verschillende statistische standaard pakketten bieden de mogelijkheid om rekening te houden met wegingsfactoren, maar, zo is de vraag, voldoen zij in dat geval waarin sprake is van wegingsfactoren die voortkomen uit een gestratificeerde steekproef? Berekenen zij net als de standaard regressieprogrammatuur alleen de puntschattingen goed, en niet de lengte van de betrouwbaarheidsintervallen, of ligt dit anders? Duidelijk is dat dit in de toekomst nader onderzocht dient te worden.

LITERATUUR

- Cochran, W.G., Sampling techniques. New York (John Wiley and Sons) 1977.
- DuMouchel, W.H. en G.J. Duncan, "Using sample survey weights in multiple regression analysis of stratified samples". Journal of the American statistical association 78 (1983) 535-543.
- Elhorst, J.P., Een schatting van de produktiefunctie en de winstfunctie voor de landbouw in Nederland. Den Haag, LEI, 1986. Onderzoekverslag 25.
- Elhorst, J.P., Regressie-analyse op basis van een gestratificeerde steekproef. Den Haag, LEI, 1988. Onderzoekverslag 36.
- Holt, D. et al., "Regression analysis of data from complex surveys". Journal of the royal statistical society, A, 143 (1980) 474-487.
- Kish, L. en M.R. Frankel, "Inference from complex samples (with discussion)". Journal of the royal statistical society, B, 36 (1974) 1-37.
- Lodder, K., Het boekhoudnet landbouwbedrijven; een statistische verantwoording. Den Haag, LEI, 1986. Mededeling 358.
- Maddala, G.S., Limited-dependent and quantitative variables in econometrics. Cambridge (University Press) 1983.
- Moors, J.J.A. en J. Muilwijk, Steekproeven; een inleiding tot de praktijk. Amsterdam/Brussel: (Agon Elsevier) 1975.
- Nathan, G. en D. Holt, "The effect of survey design on regression analysis". Journal of the royal statistical society, B, 42 (1980) 377-386.

Ontvangen: 22-10-1987

Geaccepteerd: 15-02-1988