Drs. J.P. Elhorst

Onderzoekverslag 36

REGRESSIEANALYSE OP BASIS VAN EEN GESTRATIFICEERDE STEEKPROEF

LANDBOUW-ECONOMISCH INSTITUUT
DEN HAAG
BIBLIOTHEEK
SIGN: L 28-36
EX. NO: C
MLV:

Februari 1988

Landbouw-Economisch Instituut
Afdeling Landbouw

REFERAAT

REGRESSIEANALYSE OP BASIS VAN EEN GESTRATIFICEERDE STEEKPROEF
Elhorst, J.P.
Den Haag, Landbouw-Economisch Instituut, 1988
Onderzoekverslag 36
47 p., tab., fig.

Dit verslag gaat in op de vraag of bij regressieanalyse rekening moet worden gehouden 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. De beste oplossing uit een oogpunt van modelspecificatie is die vergelijking waarin geen significant verschil optreedt in de schatting van de regressiecoëfficiënten 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, die aangeeft of het verschil tussen beide significant is, wordt in dit verslag 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.

Stratificatie/Trekkingskans/Wegingsfactoren/Econometrie/Schattingstechnieken

Overname van de inhoud toegestaan, mits met duidelijke bronvermelding.

Inhoud

Woord vooraf

Om de ontwikkelingen in de Nederlandse landbouw te kunnen volgen houdt het LEI boekhoudingen bij van ruim duizend landbouwbedrijven. Deze bedrijven worden gekozen op basis van een gestratificeerde steekproef. Daar de steekproefpercentages in de onderscheiden strata verschillend zijn, kan bij het berekenen van bepaalde resultaten niet worden volstaan met statistische procedures afgeleid voor een enkelvoudige steekproef. Zo kan een gemiddelde niet berekend worden door een eenvoudig optellen en middelen van de bedrijfsgegevens, maar is weging noodzakelijk. Evenzo ligt het voor de hand de bedrijfsgegevens te wegen bij het schatten van een regressievergelijking. Toch blijkt hier verschillend over te worden gedacht: in de literatuur treft men zowel voorstanders van ongewogen als van gewogen kleinste kwadraten, beide met redelijke argumenten. Het thans voor U liggende onderzoekverslag baant zich een weg door deze literatuur, zet de argumenten op een rij en geeft uiteindelijk antwoord op de klemmende vraag of nu wel of niet moet worden gewogen. Een woord van dank gaat daarbij uit naar verschillende collega's van het LEI voor hun medewerking en kritische commentaar.

De Directeur,

J. de Veer

Den Haag, februari 1988

Samenvatting

Doel van dit onderzoek is de beantwoording van de vraag hoe te handelen als een regressieanalyse wordt uitgevoerd op basis van een gestratificeerde steekproef. Gebleken is dat de te volgen handelwijze in twee delen uiteenvalt, te weten de bepaling van de wegingsfactoren en de toets op het gebruik van deze wegingsfactoren.

De bepaling van de wegingsfactoren is afhankelijk van de steekproef en de populatie waarop zij betrekking heeft. Op het LEI, waar een gestratificeerde steekproef wordt getrokken uit alle in Nederland geregistreerde landbouwbedrijven boven een bepaalde miniumomvang, zijn drie factoren van invloed op de hoogte van de wegingsfactoren. Ten eerste de aard van de gegevens die men wil analyseren, waarbij niet alleen een onderscheid mogelijk is tussen bedrijfsuitkomsten en financieringsgegevens, maar ook bijvoorbeeld naar het aantal jaren dat een bedrijf achtereen in administratie is gehouden. Ten tweede het moment waarop de wegingsfactoren worden bepaald, in verband met het aantal bedrijven dat is uitgewerkt, en ten derde het deel van de populatie waarover men een uitspraak wil doen. Afhankelijk van deze factoren nemen de wegingsfactoren andere waarden aan, zodat zij in het algemeen niet als een vaststaand gegeven kunnen worden beschouwd.

Nadat de wegingsfactoren zijn bepaald volgt de toets op het gebruik. Deze toets is bedoeld om te kunnen kiezen tussen de OLS-schattingsmethode en de gewogen schattingsmethode. Het lineair regressiemodel geschat volgens de OLS-schattingsmethode geeft als resultaat

$$\hat{b} = (X'X)^{-1} X'Y$$

$$E(\hat{b}) = b$$

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

$$\hat{\sigma}^2 = \frac{1}{n-k} \sum_{i=1}^{n} (Y_i - b'X_i)^2$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (Y_i - b'X_i)^2}{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}$$

Het lineair regressiemodel geschat volgens de gewogen schattingsmethode als resultaat

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

$$E(\hat{b}_w) = b$$

$$\mathrm{var}(\hat{b}_w) = \sigma^2 (X'WX)^{-1} X'W^2 X (X'WX)^{-1}$$

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

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

met H het aantal onderscheiden strata. Het antwoord op de vraag of rekening moet worden gehouden met de hoogte van de wegingsfactoren is tweeledig. De beste oplossing uit een oogpunt van modelspecificatie is die vergelijking waarin geen significant verschil optreedt in de schatting van de regressiecoëfficiënten volgens de OLS-schattingsmethode en de gewogen schattingsmethode. Onder deze voorwaarde wordt het gedrag, dat men met het model wil verklaren, namelijk beter beschreven. De toets die is bedoeld om te kunnen kiezen tussen de OLS-schattingsmethode en de gewogen schattingsmethode, geeft aan of het verschil tussen beide significant is en kan worden uitgevoerd zonder dat de gewogen schatting van de regressiecoëfficiënten is bepaald. Deze toets is in dit verslag 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. In dat geval geeft alleen regressieanalyse inclusief de hoogte van de wegingsfactoren een representatief beeld voor alle bedrijven.

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

1. Inleiding

1.1 Probleemstelling

Om de ontwikkelingen in de Nederlandse landbouw te kunnen volgen houdt het LEI boekhoudingen bij van ruim duizend landbouwbedrijven. Deze bedrijven worden gekozen op basis van een steekproef uit de in de Landbouwtelling geregistreerde bedrijven.

Om tot een zo getrouw mogelijke afspiegeling te komen van de Nederlandse landbouw in zijn volle verscheidenheid, is het nodig dat de samenstelling van de steekproef zo veel mogelijk overeenkomt met die van de gehele populatie. Afwijkingen zouden kunnen ontstaan als door toevallige factoren bij het trekken bepaalde groepen (bedrijfstypen, grootteklassen, intensiteitsklassen, etc.) onder- of oververtegenwoordigd zouden zijn. Door een tweetal maatregelen wordt getracht de kans op zulke afwijkingen te verkleinen.

In de eerste plaats wordt de steekproef gestratificeerd: de samenstelling ervan wordt gereguleerd via een onderverdeling in strata en het trekken daaruit van afzonderlijke steekproeven. Ten tweede wordt binnen de zo ontstane strata rekening gehouden met de grootte van de onderlinge verschillen tussen de bedrijven door het steekproefpercentage te variëren. Naarmate die verschillen groter zijn neemt namelijk ook de kans toe op relatief grote toevallige afwijkingen tussen steekproef en werkelijkheid en deze kans kan worden beperkt door het steekproefpercentage van het betreffende stratum te verhogen.

Naast het verkleinen van de kans op toevallige afwijkingen helpt stratificatie ook om een bepaalde systematische afwijking te voorkomen: in bepaalde strata kan de bereidheid of de praktische mogelijkheid tot deelname groter zijn dan in een andere. Door uit de afzonderlijke strata steekproeven te trekken en de bedrijven die door weigering uitvallen te vervangen door bedrijven uit datzelfde stratum wordt in zo'n geval bereikt dat de verhoudingen tussen de strata niet verstoord worden.

Daar de steekproefpercentages in de onderscheiden strata verschillend zijn, kan bij het berekenen van bepaalde resultaten niet worden volstaan met statistische procedures afgeleid voor een enkelvoudige steekproef. Een zuivere schatting van een gemiddelde bijvoorbeeld kan alleen worden verkregen door middel van weging, waarbij de gewichten worden bepaald als de verhouding tussen het aantal bedrijven dat in de populatie en het aantal bedrijven dat in de steekproef is verdeeld over de strata.

Dit verslag gaat in op de vraag hoe te handelen als men niet een gemiddelde wil berekenen, maar een regressieanalyse wil uitvoeren op het op deze wijze tot stand gekomen databestand en spitst zich toe op de vraag of bij regressieanalyse de gegevens

eveneens moeten worden gewogen. Hierbij zullen wij ons beperken tot het klassieke lineaire schattingsmodel bestaande uit één vergelijking.

1.2 De stand van zaken

De literatuur op het terrein van de steekproeftrekking richt zich op de schatting van Y, geschreven als $\hat{Y}$, en haar standaardfout, $stf(\hat{Y})$. 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. $Stf(\hat{Y}) = \sqrt{var(Y)}$ 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 $\hat{Y} \pm t_p \cdot stf(\hat{Y})$.

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 $\hat{Y} \pm t_p \cdot stf(\hat{Y})$, where $\hat{Y}$ is some complex statistic, and $stf(\hat{Y})$ 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 steekproeftechniek 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 figuur 1.1. 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 verslag 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.

Figuur 1.1 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	STANDAARD		
Gestratificeerde steekproef		beschikbaar	in ontwikkeling

Bron : Kish en Frankel (1974)

Rij 1 heeft betrekking op het bepalen van betrouwbaarheidsintervallen voor statistische grootheden bij een 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 verslag tracht een uitspraak te doen over regressiecoëfficiënten door recent verschenen literatuur te bundelen.

1.3 Opbouw van het verslag

Twee hoofdstukken vormen de kern van dit verslag. Eén hoofdstuk waarin wordt ingegaan op de wijze waarop wegingsfactoren in

de LEI-steekproef tot stand komen, als ook factoren die van invloed zijn op de hoogte van deze wegingsfactoren, en één hoofdstuk waarin wordt ingegaan op de vraag of het nodig is om bij regressieanalyse rekening te houden met de hoogte van de wegingsfactoren. Onderwerpen die hierin aan de orde komen zijn standaard regressiemethoden, de analytische uitwerking van het lineair regressiemodel inclusief en exclusief de hoogte van wegingsfactoren en een toets om te bepalen of het gebruik van wegingsfactoren noodzakelijk is. Eén en ander wordt geïllustreerd met schattingen van de produktiefunctie - de Cobb-Douglas functie en de translogfunctie - voor de Nederlandse landbouw.

Centraal in dit verslag staat het boekjaar 1984/85: de illustratie van de wijze waarop de wegingsfactoren in de LEI-steekproef tot stand komen, als ook de schattingen van de produktiefunctie zijn gebaseerd op data van dit boekjaar. Dat is geen beperking, omdat dit jaar representatief is voor een achterliggende periode die teruggaat tot 1975/76.

2. Wegingsfactoren

2.1 Wegingsfactoren in de LEI-steekproef

Dit hoofdstuk behandelt de wijze waarop wegingsfactoren in de LEI-steekproef tot stand komen. Het is toegevoegd, omdat de vraag of bij regressieanalyse rekening moet worden gehouden met wegingsfactoren niet beantwoord kan worden voordat überhaupt is nagegaan hoe ze tot stand komen en ook welke factoren van invloed zijn op de hoogte van deze wegingsfactoren. Echter, de lezer die uitsluitend is geïnteresseerd in de vraag of het nodig is om bij regressieanalyse rekening te houden met deze wegingsfactoren kan dit hoofdstuk overslaan. De lezer die zich meer inzicht wil verschaffen in de opzet van de LEI-steekproef dan in dit hoofdstuk wordt geboden, wordt verwezen naar een LEI-mededeling van Lodder (1987).

Op grond van de in de Landbouwtelling geregistreerde gegevens wordt een indeling gemaakt van alle bedrijven in 32 basisstrata. Dit geschiedt aan de hand van twee criteria:

A De bedrijfstypering waarbij acht typen worden onderscheiden:
   1. A : akkerbouw
   2. R : rundveehouderij
   3. Va : varkenshouderij
   4. Pl : pluimveehouderij
   5. A+ : gemengd akkerbouw
   6. RA : gemengd rundveehouderij/akkerbouw
   7. RV : gemengd rundveehouderij/intensieve veehouderij
   8. V+ : gemengd intensieve veehouderij

   Deze indeling wordt bepaald aan de hand van het aantal standaardbedrijfseenheden (sbe) opgegeven in de Landbouwtelling, waarbij standaardbedrijfseenheden zijn gedefinieërd als een maatstaf voor de bedrijfsomvang en/of voor de afzonderlijke produktierichtingen binnen een bedrijf. Een standaardbedrijfseenheid komt overeen met een bepaald bedrag aan toegerekende kosten in een basisperiode bij een doelmatige bedrijfsvoering onder normale produktieomstandigheden (zie ook Cleveringa, 1972).

B Een onderverdeling van de bedrijfstypen in vier grootteklassen wederom gebaseerd op het aantal standaardbedrijfseenheden.

Een totaaloverzicht van de indeling die is aangehouden in het boekjaar 1984/85 is weergegeven in tabel 2.1.

Uit deze tabel blijkt dat de LEI-steekproef zich richt op landbouwbedrijven boven een bepaalde miniumomvang, welke in het boekjaar 1984/85 79 standaardbedrijfseenheden bedroeg, en beneden een bepaalde maximumomvang van 2000 standaardbedrijfseenheden. Het aanhouden van een miniumbedrijfsomvang berust op twee overwegingen:

Tabel 2.1 De indeling in 32 basisstrata in het boekjaar 1984/85

Bedrijfs-type	Sbe-klasse											
	1			2			3			4		
	cel no.	klasse-grenzen	trk. kans	cel no.	klasse-grenzen	trk. kans	cel no.	klasse-grenzen	trk. kans	cel no.	klasse-grenzen	trk. kans
A	1	79-160	1.33	2	160-251	1.44	3	251-412	2.11	4	412 ≤ 2000	7.19
R	5	79-154	1.01	6	154-232	1.06	7	232-347	1.37	8	347 ≤ 2000	3.76
Va	9	79-154	1.21	10	154-237	1.31	11	237-413	2.20	12	413 ≤ 2000	8.40
P1	13	79-166	1.31	14	166-289	1.70	15	289-538	3.12	16	538 ≤ 2000	9.26
A+	17	79-146	1.19	18	146-230	1.35	19	230-392	2.11	20	392 ≤ 2000	7.14
RA	21	79-142	1.13	22	142-224	1.32	23	224-358	1.91	24	358 ≤ 2000	4.39
RV	25	79-153	1.18	26	153-229	1.19	27	229-348	1.64	28	348 ≤ 2000	3.91
V+	29	79-144	1.11	30	144-218	1.22	31	218-351	1.83	32	351 ≤ 2000	6.54

Noot: De trekkingskansen zijn uitgedrukt in procenten.

a. Ondanks dat het aantal bedrijven kleiner dan 79 standaardbedrijfseenheden groot is, is hun aandeel in de agrarische produktie klein. Ze liggen ver beneden het niveau dat nodig is voor een bestaan in de landbouw en worden voor het grootste deel als nevenbedrijf geëxploiteerd, of door oudere boeren als aflopend bedrijf aangehouden.

b. De ervaring heeft geleerd dat op deze bedrijven de bereidheid en de mogelijkheid tot deelname klein zijn en dat het praktisch niet goed mogelijk is om van deze bedrijven een betrouwbaar beeld te verkrijgen.

Door het aanhouden van een bovengrens wordt vermeden dat zeer grote bedrijven, zoals het landbouwbedrijf van de Rijksdienst voor de IJsselmeerpolders, in de LEI-steekproef worden opgenomen.

De steekproefbasis, dat wil zeggen het adresmateriaal dat op grond van deze indeling in aanmerking komt voor de bedrijfskeuze, is omschreven in tabel 2.2.

Tabel 2.2 Aantal landbouwbedrijven van 79 tot 2000 sbe per bedrijfstype en sbe-klasse in de Landbouwtelling van 1984

Bedrijfstype	Sbe-klasse				
	1	2	3	4	Totaal
Akkerbouw	4147	3826	2610	764	11347
Rundveehouderij	14081	13424	10390	3770	41665
Varkenshouderij	2307	2135	1272	332	6046
Pluimveehouderij	689	530	289	97	1605
Akkerbouw/gemengd	589	517	331	98	1535
Rundvee/akkerbouw	529	454	314	137	1434
Rundvee/intens.veeh.	1016	1005	733	307	3061
Gemengd/intens.veeh.	1439	1316	875	244	3874
Totaal	24797	23207	16814	5749	70567

Binnen de 32 basisstrata wordt daar waar mogelijk nog verder gestratificeerd naar oppervlakte, leeftijd en regio. Dit gebeurt vooral met het oog op de bij de bedrijfskeuze optredende non-respons: bij de vervanging van bedrijven die in eerste instantie niet in de steekproef terecht komen kunnen de onderlinge verhoudingen tussen oppervlakteklassen, leeftijdsklassen en regio's op deze manier in takt worden gehouden, hetgeen de representativiteit ten goede komt. De trekkingskans wordt bij deze verdere stratificatie echter gelijk gehouden, zodat zij voor de verdere bepaling van de wegingsfactoren niet van belang is.

Van de bedrijven die in de steekproef worden gekozen verzamelt het LEI gegevens over de bedrijfsuitkomsten en daar waar mogelijk gegevens over de vermogenspositie van de ondernemer en over de inkomensvorming en -besteding van de ondernemer en zijn gezinsleden. Het aantal bedrijven in het boekjaar 1984/85 waarvoor gegevens over de bedrijfsuitkomsten zijn verzameld staat vermeld in tabel 2.3.

Tabel 2.3 Het aantal bedrijven in het boekjaar 1984/85 waarvoor gegevens over de bedrijfsuitkomsten zijn verzameld

Bedrijfstype	Sbe-klasse				
	1	2	3	4	Totaal
Akkerbouw	63	68	73	59	263
Rundveehouderij	117	140	134	123	514
Varkenshouderij	18	26	25	18	87
Pluimveehouderij	6	14	6	7	33
Akkerbouw/gemengd	5	8	6	4	23
Rundvee/akkerbouw	4	4	7	5	20
Rundvee/intens.veeh.	12	14	11	9	46
Gemengd/intens.veeh.	11	17	13	8	49
Totaal	236	291	275	233	1035

Het aantal bedrijven waarvoor gegevens over de bedrijfsuitkomsten zijn verzameld behoeft niet precies overeen te komen met de trekkingskans. Redenen dat het aantal bedrijven soms kleiner is dan men op grond van de trekkingskans zou mogen verwachten, zijn dat onvoldoende bedrijven meedoen in de steekproef, bijvoorbeeld omdat de bereidheid tot deelname of de praktische mogelijkheid tot deelname klein zijn, of ook dat de gegevens niet volledig zijn uitgewerkt, bijvoorbeeld door:

- Spontane opzegging door de deelnemer.
- Het niet insturen van de gegevens.
- Ziekte of ongevallen van het bedrijfshoofd of gezinsleden.
- Opheffing van het bedrijf.
- Andere redenen zoals verhuizing, overdracht, reorganisatie, struktuurverandering van het bedrijf, etc.

Een reden dat het aantal bedrijven soms groter is dan men op grond van de trekkingskans zou mogen verwachten is een aanvulling op de steekproef die voorkomt uit de doelstellling om niet alleen gegevens te kunnen verstrekken die representatief zijn voor de landbouw in zijn geheel, maar ook voor bepaalde groepen binnen de populatie, met name voor de akkerbouw en de rundveehouderij in bepaalde regio's.

De wegingsfactoren behorend bij het aantal bedrijven in de steekproefbasis en het aantal bedrijven in de steekproef waarvoor de bedrijfsuitkomsten zijn uitgewerkt, staan vermeld in tabel 2.4.

Tabel 2.4 Aantal bedrijven dat door één bedrijf in de steekproef wordt vertegenwoordigd

Bedrijfstype	Sbe-klasse			
	1	2	3	4
Akkerbouw	65.83	56.26	35.75	12.95
Rundveehouderij	120.35	95.89	77.54	30.65
Varkenshouderij	128.17	82.12	50.88	18.44
Pluimveehouderij	114.83	37.86	48.17	13.86
Akkerbouw/gemengd	117.80	64.63	55.17	24.50
Rundvee/akkerbouw	132.25	113.50	44.86	27.40
Rundvee/intens.veeh.	84.67	71.79	66.64	34.11
Gemengd/intens.veeh.	130.82	77.42	67.31	30.50

2.2 Factoren van invloed op de hoogte van de wegingsfactoren

Het is goed zich te realiseren dat de wegingsfactoren geen vaststaand gegeven zijn, maar afhankelijk van een aantal factoren, waarvan wij onderscheiden:

1) De aard van de gegevens die men wil analyseren.
2) Het moment waarop men de wegingsfactoren bepaalt.
3) Het deel van de landbouw waarover men een uitspraak wil doen.

ad 1.

Hierboven is opgemerkt dat het LEI daar waar mogelijk gegevens verzamelt over de vermogenspositie van de ondernemer en over de inkomensvorming en -besteding van de ondernemer en zijn gezinsleden. Dit om tevens inzicht te kunnen verschaffen in de financiële positie van landbouwbedrijven. Het aantal bedrijven dat meewerkt aan de opstelling van deze uitgebreide boekhouding is kleiner dan het aantal bedrijven waarvoor bedrijfsuitkomsten worden verzameld (circa 80-85%). Bij de presentatie van de financiële positie van de landbouw in het boekjaar 1984/85 (zie Aukema en Overgaauw, 1986) worden dan ook andere wegingsfactoren gebruikt dan bij de presentatie van de bedrijfsuitkomsten (zie LEI, 1986). Op het LEI spreekt men wel over wegingsfactoren met betrekking tot de financiering en over wegingsfactoren met betrekking tot de bedrijfsuitkomsten. De wegingsfactoren met betrekking

tot de financiering zijn groter, omdat het aantal deelnemende bedrijven aan deze vorm van administratie kleiner is en zodoende het aantal bedrijven dat ieder bedrijf in de populatie vertegenwoordigt groter.

Eén en ander betekent dat de wegingsfactoren die in een onderzoek worden gebruikt afhankelijk zijn van de gegevens die men wil analyseren. Analyseert men gegevens die op alle bedrijven zijn verzameld, dan komen de wegingsfactoren met betrekking tot de bedrijfsuitkomsten in aanmerking. Analyseert men gegevens die alleen zijn verzameld op bedrijven met een financieringsboekhouding, dan komen de wegingsfactoren met betrekking tot de financiering in aanmerking.

Een ander geval treedt op als men gegevens wil analyseren van bedrijven die tenminste twee jaar achtereen in administratie zijn gehouden. Bijvoorbeeld als men het aanbod van produkten of de vraag naar produktiefactoren wil verklaren uit variabelen die één jaar zijn vertraagd. Nu is het zo dat jaarlijks een kwart van de steekproef wordt vervangen. Dit wordt gedaan om de door het jaar heen opgetreden uitval aan te vullen en om de samenstelling van de steekproef zodanig aan te passen dat de jaarlijks optredende structuurveranderingen er afdoende door worden weerspiegeld. Bedrijven die "hun tijd uitzitten", dat wil zeggen bedrijven die niet onvoorzien uitvallen, worden in de regel niet langer dan zes jaar in administratie gehouden. Wil men nu in een voorkomend geval het aanbod van produkten of de vraag naar produktiefactoren verklaren uit variabelen die één jaar zijn vertraagd, dan is een herberekening van de wegingsfactoren naar die bedrijven die tenminste twee jaar achtereen in administratie zijn gehouden noodzakelijk. Soortgelijke herberekeningen zijn ook noodzakelijk als men bedrijven elimineert, bijvoorbeeld omdat bepaalde gegevens ontbreken.

ad 2.

Het moment waarop men de wegingsfactoren bepaalt kan van belang zijn voor het aantal bedrijven waarvoor de gegevens zijn uitgewerkt. Toen over de bedrijfsuitkomsten in het boekjaar 1984/85 werd gerapporteerd (LEI, 1986) waren 1008 van de 1035, dit is 97.4%, van de bedrijven uitgewerkt. Dit betekent dat de wegingsfactoren, zoals die bij de presentatie van de bedrijfsuitkomsten werden gebruikt, waren gebaseerd op "slechts" een deel (97.4%) van de bedrijven. Daarna is het aantal bedrijven dat is uitgewerkt nog iets toegenomen. Het verdient daarom aanbeveling om bij onderzoekingen die nadien worden opgestart tevens de bedrijven op te nemen die niet bij de presentatie van de bedrijfsuitkomsten zijn gebruikt, omdat men anders onnodig informatie verloren laat gaan. Dit geldt tevens voor bedrijven met een financieringsboekhouding.

ad 3.

De bepaling van de wegingsfactoren kan ook een verandering ondergaan als men een uitspraak wil doen over een subgroep in de landbouw. Als deze subgroep overeenkomt met de wijze waarop de steekproef is gestratificeerd, bijvoorbeeld één van de bedrijfstypen gedefinieërd volgens de bedrijfstypering onder A, dan veranderen de wegingsfactoren niet. Maar als de subgroep dwars door de strata heen loopt, bijvoorbeeld de akkerbouw op klei, de akkerbouw in de IJsselmeerpolders, etc., dan is het raadzaam een herberekening uit te voeren van de wegingsfactoren door middel van poststratificatie. Hieronder wordt deze methode besproken.

Uitgangspunt hierbij is een subgroep met kenmerk S die gedekt is door de steekproef. Bepaal per stratum het aantal bedrijven in de populatie en het aantal bedrijven in de steekproef, die voldoen aan het kenmerk S. Is geen van de strata leeg, bereken dan de wegingsfactoren door deze op elkaar te delen. Is een stratum leeg met betrekking tot het aantal bedrijven in de steekproef, voeg deze dan bij een aanverwant stratum dat niet leeg is en bereken voor deze strata tezamen de wegingsfactor. Is een stratum leeg met betrekking tot het aantal bedrijven in de steekproef en het aantal bedrijven in de populatie, voer dan het stratum af. Een voorbeeld van deze gang van zaken is opgenomen in tabel 2.5.

Opgemerkt moet worden dat poststratificatie alleen mogelijk is als de bedrijven in de populatie, waaruit de steekproef wordt getrokken, ook naar dit kenmerk zijn in te delen. Is dat niet zo - bijvoorbeeld bij een indeling in eigendoms- en pachtbedrijven - dan rest alleen het gebruik van de oorspronkelijke wegingsfactoren.

2.3 Conclusie

De wegingsfactoren zijn geen vaststaand gegeven, dat men kan opvragen uit het databestand van het LEI. Zij zijn afhankelijk van de aard van de gegevens die men wil analyseren, waarbij niet alleen een onderscheid mogelijk is tussen bedrijfsuitkomsten en financieringsgegevens, maar ook bijvoorbeeld naar het aantal jaren dat een bedrijf achtereen in administratie is gehouden. Van het moment waarop de wegingsfactoren worden bepaald, in verband met het aantal bedrijven dat is uitgewerkt, en van het deel van de populatie waarover men een uitspraak wil doen. Afhankelijk van deze factoren nemen de wegingsfactoren andere waarden aan, zodat in het algemeen niet kan worden volstaan met die wegingsfactoren die worden gebruikt voor het opstellen van de bedrijfsuitkomsten en de financiële positie.

Tabel 2.5 Bepaling van de wegingsfactoren voor een subgroep met kenmerk S, die gedekt is door de steekproef

Stratum no.	Aantal bedrijven in de populatie met kenmerk S	Aantal bedrijven in de steekproef met kenmerk S	Wegingsfactor
1	3000	40	75
2	2100	50	42
3	300	4	75
4	10	2	5
5	0	0	
.	.	.	
.	.	.	afvoeren
.	.	.	
16	0	0	
17	200	4	50
18	300	4	75
19	50	2	30
20	10	0	
21	100	1	100
22	20	3	
23	10	0	12
24	6	0	
25	0	0	
.	.	.	
.	.	.	afvoeren
.	.	.	
32	0	0	

3. Regressieanalyse: wegen ja of nee?

3.1 Standaard regressiemethoden

Een onderzoeker die een regressieanalyse 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 afgeleid uit de steekproef.

Methode A is het lineair regressiemodel dat als volgt kan worden beschreven

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

met Y_i de te verklaren variabele, X_i een (k*1) vector van verklarende variabelen, u_i de storingsterm, b een (k*1) vector van te schatten parameters en n het aantal waarnemingen. Als uitkomsten heeft dit model

$$\hat{b} = (X'X)^{-1} X'Y$$

$$E(\hat{b}) = b$$

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

$$\hat{\sigma}^2 = \frac{1}{n-k} \sum_{i=1}^{n} (Y_i - b'X_i)^2$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (Y_i - b'X_i)^2}{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}$$

Methode B, zoals deze in het algemeen wordt toegepast in standaard regressieprogrammatuur, bijvoorbeeld SPSS, BMDP en Genstat, hanteert een regressiemodel analoog aan het lineair 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 waarneming i. Als uitkomsten heeft dit model

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

$$E(\hat{b}) = b$$

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

$$\hat{\sigma}^2 = \frac{1}{n-k} \sum_{i=1}^{n} W_i (Y_i - b'X_i)^2$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n} W_i (Y_i - b'X_i)^2}{\sum_{i=1}^{n} W_i (Y_i - \bar{Y})^2}$$

met W de diagonaal matrix van de wegingsfactoren.

Een derde methode die veelvuldig wordt toegepast is kleinste kwadraten op basis van gemiddelden per groep van bedrijven. Een reden voor het groeperen en middelen wordt vaak niet gegeven, maar in het algemeen betreft het een reductie van het databestand tot een naar eigen zeggen aanvaardbaar niveau. Ondanks dat regressieanalyse toegepast op gemiddelden per groep van bedrijven los staat van de wegingsproblematiek - immers het middelen van gegevens kan op ieder databestand van dwarsdoorsnedegegevens worden toegepast - is besloten hier toch aandacht aan te besteden, juist omdat het veelvuldig wordt toegepast en de LEI-steekproef hierin geen uitzondering vormt. Voor voorbeelden zie Burger (1983) alsmede Douma en Poppe (1987). Op deze plaats echter beperken wij ons tot de belangrijkste bevindingen en verwijzen wij naar de bijlage voor meer achtergrondinformatie. Uit de betreffende bijlage blijkt dat het schatten op basis van groepsgemiddelden sterk moet worden afgeraden, omdat

- de schatting van de regressiecoëfficiënten onzuiver is. Deze bevinding, die afwijkt van de tot nu toe gangbare literatuur, is gebaseerd op een recent artikel van Deaton (1985) en zou men kunnen opvatten als een nieuw gezichtspunt op het werken met groepsgemiddelden;
- de standaardfouten toe- en zodoende de T-waarden afnemen; en
- de determinatiecoëfficiënt (R^2) toeneemt door een reductie van het aantal waarnemingen.

Deze punten zijn het gevolg van de aggregatiefout, die voortkomt uit het ten onrechte als volkomen identiek beschouwen van bedrijven die aan het gemiddelde bedrijf ten grondslag liggen. Alleen door de aggregatiefout in de schatting te betrekken volgens het

regressiemodel met errors-in-variables, kan een schatting worden verkregen die consistent is. Deze methode echter is omslachtig en nodeloos ingewikkeld. In dit hoofdstuk zullen wij ons daarom verder toeleggen op de vraag of het nodig is om rekening te houden met de hoogte van de wegingsfactoren indien regressieanalyse wordt toegepast op bedrijfsgegevens die niet zijn gegroepeerd en gemiddeld.

3.2 Regressieanalyse zonder en met wegingsfactoren

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

1) $E(u_i \mid X_i) = 0$ voor alle i.

2) homoskedasticiteit: $var(u_i \mid X_i) = \sigma^2$ voor alle i

3) onafhankelijkheid van de waarnemingen:

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

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 aselector 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 paragraaf 1.2), 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 al-

leen dan de betrouwbaarheid belangrijk als deze twee hoog gecorreleerd zijn (zie bijvoorbeeld Moors en Muilwijk, 1975, blz 63). Op het LEI, waar een gestratificeerde steekproef wordt getrokken uit alle in Nederland geregistreerde landbouwbedrijven boven een bepaalde miniumomvang, 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 worden verzameld, wordt beinvloed 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 lineair 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 regressieanalyse rekening te houden met de hoogte van de wegingsfactoren.

Afhankelijk van het model dat men specificeert en gegeven de steekproefpercentages p_i $(i=1,\ldots,n)$ die over de strata kunnen variëren zijn drie aanpassingen te onderscheiden:

1) de stratificatievariabele is tevens de te verklaren variabele

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

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).

2) de stratificatievariabele is tevens een verklarende variabele

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

Volgens Holt et al. (1980) kan in dit geval de OLS-schattingsmethode worden toegepast, omdat voldaan is aan de eis $E(u_i|S_i, X_i)=0$. Volgens ons echter kan de OLS-schattingsmethode alleen dan 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 stra-

tificatievariabele 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

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

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

$$b_w = (X'WX)^{-1} X'WY.$$

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

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

$$\mathrm{var}(\hat{b}_w) = E\,[(X'WX)^{-1} X'W\,(Xb+u)-b]\,[(X'WX)^{-1} X'W\,(Xb+u)-b]' =$$

$$= E\,[(X'WX)^{-1} X'Wu]\,[(X'WX)^{-1} X'Wu] =$$

$$= E\,((X'WX)^{-1} X'Wuu'WX(X'WX)^{-1}) =$$

$$= (X'WX)^{-1} X'W\,E(uu')\,WX(X'WX)^{-1} =$$

$$= (X'WX)^{-1} X'W\,\sigma^2 I_n\,WX(X'WX)^{-1} =$$

$$= \sigma^2 (X'WX)^{-1} X'W^2 X\,(X'WX)^{-1}$$

Het blijkt dat de $\mathrm{var}(\hat{b}_w)$ 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 daarmede 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 gedeeltelijk is ontleend aan Cochran (1977, hoofdstuk 7). Dit levert als uitkomsten

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

$$E(\hat{b}_w) = b$$

$$\text{var}(\hat{b}_w) = \sigma^2 (X'WX)^{-1} X'W^2 X (X'WX)^{-1}$$

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

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

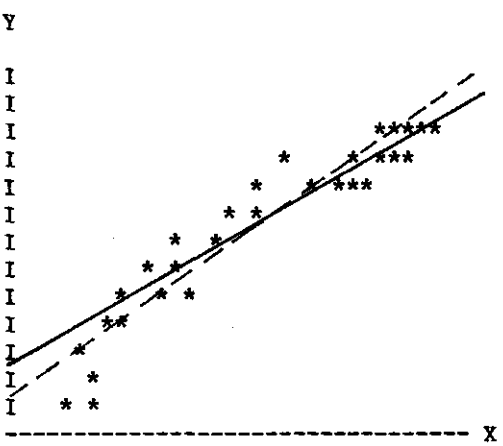
met H het aantal onderscheiden strata.

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, 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 van bedrijven in de steekproef zijn onder- of oververtegenwoordigd.

Daarnaast wordt tegen de argumenten die de voorstanders van de OLS-schattingsmethode hanteren 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 parameters, 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 3.1.

Figuur 3.1 Gewogen en ongewogen regressie van Y op X



Y = de te verklaren variabele
X = de verklarende variabele
_____ het lineair verband tussen Y en X geschat volgens de OLS-schattingsmethode
----- het lineair verband tussen Y en X geschat volgens de gewogen schattingsmethode
* waarnemingen

Deze figuur toont een aantal waarnemingen tussen de te verklaren variabele Y en de verklarende variabele X, waarbij is ver-

ondersteld dat het steekproefpercentage en daarmede het aantal waarnemingen groter is naarmate de waarden van Y en X toenemen. Stel dat een onderzoeker op grond van dit waarnemingspatroon besluit een lineair verband tussen Y en X te specificeren. Schat deze nu volgens de OLS-schattingsmethode, dan wordt de ononderbroken lijn verkregen. Schat deze volgens de gewogen schattingsmethode, dan wordt de onderbroken lijn verkregen. Duidelijk is dat de veronderstelling van een lineair verband slechts een benadering is voor het daadwerkelijke verband tussen Y en X, dat bij de onderzoeker niet bekend is. Naar aanleiding van dit veronderstelde verband zou men kunnen opmerken dat de parameterwaarden niet voor alle bedrijven gelijk zijn. Immers splitst men de bedrijven op in twee groepen, zeg groot en klein, dan wordt voor beide groepen een verschillende hellingscoëfficiënt verkregen. Maar - en dit is een belangrijk punt - dit wordt niet veroorzaakt doordat groepen van bedrijven verschillend reageren, maar uitsluitend omdat de veronderstelde schattingsvergelijking voor deze bedrijven onjuist is gespecificeerd. Men kan dan ook zeggen dat een 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. De beste oplossing uit een oogpunt van modelspecificatie is dan ook die vergelijking waarin geen significant verschil optreedt tussen de onderbroken en de ononderbroken regressielijn, ofwel tussen de schatting van b volgens de OLS-schattingsmethode en volgens de gewogen schattingsmethode. Is men daartoe niet in staat, dan moet de OLS-schattingsmethode worden verworpen ten gunste van de gewogen schattingsmethode. Want alleen regressieanalyse inclusief de hoogte van de wegingsfactoren geeft in dat geval een representatief beeld voor alle bedrijven.

3.3 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 tussen deze twee, dat is gedefinieërd als

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

Het te toetsen schattingsmodel is

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

terwijl als alternatief schattingsmodel wordt onderscheiden

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

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 $\hat{D}=0$ kan worden vervangen door de bekende F-toets op $\hat{c}=0$ als zou het te toetsen schattingsmodel zijn gespecificeerd als

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

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 figuur 3.2.

Figuur 3.2 Toets op het gebruik van wegingsfactoren

Schattingsmodel	Aantal vrijheidsgraden	Residuele kwadratensom	Gemiddelde residuele kwadratensom
Regressie zonder Z variabelen	n-k	SS_{ols}	
Regressie met Z variabelen	n-2k	SS_w	$SS_w/(n-2k)$
Verschil toe te schrijven aan de wegingsfactoren	k	$SS_{ols}-SS_w$	$(SS_{ols}-SS_w)/k$

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.

3.4 Voorbeelden

Om de uitkomsten van regressieanalyse zonder en met wegingsfactoren alsmede de werking van de toets te illustreren is een schattting gemaakt van de produktiefunctie voor de Nederlandse landbouw. Hierbij is gekozen voor een vorm die bekend staat als de Cobb-Douglas produktiefunctie (CD-functie) en de translog-produktiefunctie (translog-functie), omdat de kennis van deze functies het grootst is (zie Elhorst, 1986). De specificatie van de CD-functie is van de vorm

$$\ln PY_{b} = a_{0} + a_{1} \ln H_{b} + a_{2} \ln L_{b} + a_{3} \ln K_{b} + a_{4} \ln M_{b},$$

PY = bruto-bedrijfsopbrengst.
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 exclusief meststieren, mestkalveren, fokvarkens, mestvarkens en slachtpluimvee.
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.

a_0 t/m a_4 zijn de te schatten parameters van het model met a_0 de efficiency parameter. De CD-functie en de wijze waarop de inzet van de produktiefactoren is berekend staat in dit verslag niet ter discussie. Het gaat ons uitsluitend om een vergelijking van de uitkomsten. Voor de discussie zij verwezen naar Elhorst (1986).

De specificatie van de translog-functie is van de vorm

$$\ln PY = a_0 + [\, a_1 \; a_2 \; a_3 \; a_4 \,] \begin{bmatrix} \ln H \\ \ln L \\ \ln K \\ \ln M \end{bmatrix}_b +$$

$$+ 1/2 \begin{bmatrix} \ln H \\ \ln L \\ \ln K \\ \ln M \end{bmatrix}_b' \begin{bmatrix} A_{11} & A_{12} & A_{13} & A_{14} \\ A_{21} & A_{22} & A_{23} & A_{24} \\ A_{31} & A_{32} & A_{33} & A_{34} \\ A_{41} & A_{42} & A_{43} & A_{44} \end{bmatrix} \begin{bmatrix} \ln H \\ \ln L \\ \ln K \\ \ln M \end{bmatrix}_b$$

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

De translog-functie wordt gekenmerkt door flexibiliteit. De functie is namelijk een 2e orde benadering van de produktiefunctie uitgedrukt in logarithmen. Tevens is het een generalisatie van de CD-functie, omdat de translog-functie in de CD-functie overgaat als de parameters A_{ij} allen gelijk zijn aan nul. Dit opent de mogelijkheid om te toetsen of de produktiefunctie moet worden beschreven volgens een translog-functie of dat mag worden volstaan met een functie die eenvoudiger is, de CD-functie.

De resultaten verkregen voor de CD-functie staan vermeld in tabel 3.1a en de uitwerking van de toets in tabel 3.1b. Bovendien zijn in tabel 3.1a de verschillen berekend tussen regressieanalyse zonder en met wegingsfactoren. De resultaten tonen aan dat het verschil significant is en dat regressieanalyse zonder wegingsfactoren moet worden verworpen. Ofwel dat de CD-functie onjuist is gespecificeerd en geen goede beschrijving geeft van het produktieproces op het landbouwbedrijf. Dit sluit aan op het onderzoek van Elhorst (1986), waarin wordt geconstateerd dat de CD-functie uit statistisch oogpunt moet worden verworpen, omdat het verschil tussen de verklaringsgraad van de translog-functie en de CD-functie significant is.

De resultaten verkregen voor de translog-functie staan vermeld in tabel 3.2a en 3.2b.

Tabel 3.1a De CD-functie geschat voor de landbouw in zijn geheel *)

Verklarende variabelen	Regressieanalyse zonder wegingsfactoren		Regressieanalyse met wegingsfactoren		Het verschil tussen beide	
	par.	T-waarde	par.	T-waarde	par.	T-waarde
constante	2.2446	24.59	2.2261	23.87	.0185	.42
grond	.0781	11.70	.0641	9.08	.0141	3.76
arbeid	.2219	7.97	.1888	6.45	.0331	2.16
kapitaal	.1327	11.36	.1545	12.94	-.0219	-3.90
non-factor inputs	.7168	71.63	.7044	68.98	.0124	2.61
R^2	.95		.81			

*) gebaseerd op 1035 waarnemingen over het boekjaar 1984/85.

Tabel 3.1b Toets op het gebruik van wegingsfactoren

	Aantal vrijheidsgraden	Residuele kwadratensom	Gemiddelde residuele kwadratensom
Regressie zonder Z variabelen	1030	26.112	
Regressie met Z variabelen	1025	22.043	.022
Verschil toe te schrijven aan de wegingsfactoren	5	4.069	.814

Uitkomst toetsgrootheid .814/.022=37.84.
Kritische grens 4.37 (betrouwbaarheidsdrempel 95%).

Tabel 3.2a De translog-functie geschat voor de landbouw in zijn geheel *)

Verklarende variabelen	Regressieanalyse zonder wegingsfactoren		Regressieanalyse met wegingsfactoren		Het verschil tussen beide	
	par.	T-waarde	par.	T-waarde	par.	T-waarde
Constante	6.1199	5.46	7.3085	6.11	-1.1886	-2.04
H	.2812	2.36	.1391	1.09	.1421	2.23
L	2.1071	4.55	1.7092	3.57	.3979	1.95
K	.1396	.71	.2617	1.26	-.1221	-1.24
M	-.1398	-.83	-.3857	-2.15	.2459	2.81
H * H	.1153	10.72	.1287	10.78	-.0134	-2.04
H * L	-.0063	-.19	.0112	.31	-.0175	-.87
H * K	-.0695	-5.60	-.0715	-5.36	.0020	.30
H * M	.0160	1.68	.0255	2.53	-.0095	-1.98
L * L	.1938	1.13	.1420	.75	.0518	.51
L * K	.0381	.72	.0196	.35	.0185	.72
L * M	-.2056	-4.67	-.1613	-3.52	-.0444	-2.21
K * K	.0960	3.59	.1174	4.14	-.0215	-1.59
K * M	-.0624	-3.23	-.0894	-4.39	.0270	2.87
M * M	.1313	6.38	.1701	7.85	-.0388	-3.87
R^2	.96		.84			

*) gebaseerd op 1035 waarnemingen over het boekjaar 1984/85.

Tabel 3.2b Toets op het gebruik van wegingsfactoren

	Aantal vrijheidsgraden	Residuele kwadratensom	Gemiddelde residuele kwadratensom
Regressie zonder Z variabelen	1020	20.581	
Regressie met Z variabelen	1005	18.198	.018
Verschil toe te schrijven aan de wegingsfactoren	15	2.383	.159

uitkomst toetsgrootheid .159/.018=8.77
kritische grens 2.07 (betrouwbaarheidsdrempel 95%)

Tabel 3.3a De translog-functie geschat voor de akkerbouw *)

Verklarende variabelen	Regressieanalyse zonder wegingsfactoren		Regressieanalyse 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 3.3b 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
Verschil toe te schrijven aan de wegingsfactoren	15	1.433	.096

uitkomst toetsgrootheid .096/.024=4.06
kritische grens 2.07 (betrouwbaarheidsdrempel 95%)

Tabel 3.4a De translog-functie geschat voor de rundveehouderij*)

Verklarende variabelen	Regressieanalyse zonder wegingsfactoren		Regressieanalyse 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 3.4b 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
Verschil toe te schrijven aan de wegingsfactoren	15	.333	.022

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

Hieruit blijkt dat ook de translog-functie onjuist is gespecificeerd en geen goede beschrijving geeft van het produktieproces op het landbouwbedrijf. 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 logarithmen - 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 landbouw in zijn geheel vormt dan ook een goed voorbeeld van een onderzoek, waarin wordt volstaan met de translog-functie geschat volgens de gewogen schattingsmethode.

Om te kunnen beoordelen hoe groot de invloed is van de groep of populatie waarop de schattingen betrekking hebben, zijn ook schattingen verricht voor de akkerbouw en de rundveehouderij. De resultaten van deze exercitie, waarbij wij ons hebben beperkt tot de translog-functie, staan vermeld in tabel 3.3a t/m 3.4b. Uit deze schattingsresultaten blijkt dat alleen de translog-functie voor de rundveehouderij leidt tot het gewenste resultaat. Er bestaat geen significant verschil, bij een betrouwbaarheidsdrempel van 95%, tussen de schatting van de regressiecoëfficiënten volgens de OLS-schattingsmethode en de gewogen schattingsmethode, waardoor de toepassing van de OLS-schattingsmethode is gerechtvaardigd. Wij zouden ook kunnen zeggen dat de translog-functie een goede beschrijving geeft van het produktieproces op het rundveehouderijbedrijf ongeacht de grootte van het bedrijf.

3.5 Conclusie

Het lineair regressiemodel geschat volgens de OLS-schattingsmethode geeft als resultaat

$$\hat{b} = (X'X)^{-1} X'Y$$

$$E(\hat{b}) = b$$

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

$$\hat{\sigma}^2 = \frac{1}{n-k} \sum_{i=1}^{n} (Y_i - b'X_i)^2$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (Y_i - b'X_i)^2}{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}$$

Het lineair regressiemodel geschat volgens de gewogen schattingsmethode geeft, indien de stratificatievariabele niet gelijk is aan de te verklaren variabele, als resultaat

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

$$E(\hat{b}_w) = b$$

$$var(\hat{b}_w) = \sigma^2 (X'WX)^{-1} X'W X (X'WX)^{-1}$$

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

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

Het antwoord op de vraag of men rekening moet houden met de hoogte van de wegingsfactoren is tweeledig. De beste oplossing uit een oogpunt van modelspecificatie is die vergelijking 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 wordt beschreven. Een toets, die aangeeft 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. In dat geval geeft alleen regressieanalyse inclusief de hoogte van de wegingsfactoren een representatief beeld voor alle bedrijven.

Tot slot is gebleken dat de standaard regressieprogrammatuur geen goede uitkomsten aflevert voor de variantie-covariantie matrix en daarmede voor de standaardfouten en de T-waarden als de gewogen schattingsmethode wordt toegepast.

4. Nabeschouwing

Doel van dit onderzoek was de beantwoording van de vraag hoe te handelen als een regressieanalyse wordt uitgevoerd op basis van een gestratificeerde steekproef. Gebleken is dat te volgen handelwijze in twee delen uiteenvalt, te weten de bepaling van de wegingsfactoren en de toets op het gebruik van deze wegingsfactoren.

De bepaling van de wegingsfactoren is afhankelijk van de steekproef en de populatie waarop zij betrekking heeft. Op het LEI, waar een gestratificeerde steekproef wordt getrokken uit alle in Nederland geregistreerde landbouwbedrijven boven een bepaalde miniumomvang, zijn drie factoren van invloed op de hoogte van de wegingsfactoren. Ten eerste de aard van de gegevens die men wil analyseren, waarbij niet alleen een onderscheid mogelijk is tussen bedrijfsuitkomsten en financieringsgegevens, maar ook bijvoorbeeld naar het aantal jaren dat een bedrijf achtereen in administratie is gehouden. Ten tweede het moment waarop de wegingsfactoren worden bepaald, in verband met het aantal bedrijven dat is uitgewerkt, en ten derde het deel van de populatie waarover men een uitspraak wil doen. Afhankelijk van deze factoren nemen de wegingsfactoren andere waarden aan, zodat zij in het algemeen niet als een vaststaand gegeven kunnen worden beschouwd.

Nadat de wegingsfactoren zijn bepaald volgt de toets op het gebruik. Deze toets is bedoeld om te kunnen kiezen tussen de OLS-schattingsmethode en de gewogen schattingsmethode. De beste oplossing uit een oogpunt van modelspecificatie is die vergelijking waarin geen significant verschil optreedt in de schatting van de regressiecoëfficiënten volgens de OLS-schattingsmethode en de gewogen schattingsmethode. Onder deze voorwaarde wordt het gedrag, dat men met het model wil verklaren, namelijk beter beschreven. De toets die is bedoeld om te kunnen kiezen tussen de OLS-schattingsmethode en de gewogen schattingsmethode, geeft aan of het verschil significant is. 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.

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

De vraag die dit resultaat tenslotte oproept is hoe het is gesteld met andere multivariate analysetechnieken, zoals clusteranalyse, factoranalyse, discriminantieanalyse en variantie-covariantie analyse. Verschillende statistische standaard paketten 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

Aukema, S. en J.G.A. Overgaauw,
De financiele positie van de landbouw boekjaar 1984/85.
Den Haag, LEI, 1986. Periodieke rapportage 12-84/85.

Burger, C.P.J.,
Investeringen in de akkerbouw. In: Herfst, A.C.C. et al., Financiering en belegging; stand van zaken anno 1983.
Rotterdam (Erasmus Universiteit) 1983.

Bekker, P.A.,
Essays on identification in linear models with latent variables.
Helmond (Wibro) 1986. Proefschrift.

Chow, G.C.,
Econometrics.
Tokyo (McGraw-Hill) 1983. Economics handbook series.

Cleveringa, C.J.,
Standaardbedrijfseenheden (SBE) als criterium voor bedrijfsgrootte en bedrijfstype.
Den Haag, LEI, 1972. Mededelingen en Overdrukken 94.

Cochran, W.G.,
Sampling techniques.
New York (John Wiley and Sons) 1977.

Cramer, J.S.,
"Efficient grouping, regression and correlation in Engel curve analysis".
Journal of the American statistical association 59 (1964) 233-250.

Deaton, A.,
"Panel data from time series of cross-sections".
Journal of econometrics 30 (1985) 109-126.

Douma, B.E., en K.J. Poppe,
Akkerbouw 1985.
Den Haag, LEI, 1987. Periodieke rapportage 5-85.

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.

LITERATUUR (le vervolg)

Dijk, J. van,
De bedrijfskeuze voor het boekhoudnet 1986/87.
Den Haag, LEI, 1986. Interne notitie.

Elhorst, J.P.,
Een schatting van de produktiefunctie en de winstfunctie voor de landbouw in Nederland.
Den Haag, LEI, 1986. Onderzoekverslag 25.

Holt, D. et al.,
"Regression analysis of data from complex surveys".
Journal of the royal statistical society, A, 143 (1980) 474-487.

Johnston, J.,
Econometric methods.
Tokyo (McGraw-Hill) 1972.

Judge, G.G. et al.,
The theory and practice of econometrics.
New York (John Wiley and Sons) 1980.

Ketellapper, R.H.,
The impact of observational errors on parameter estimation in econometrics.
Groningen (Veenstra-Offset) 1982. Proefschrift.

Kish, L. en M.R. Frankel,
"Inference from complex samples (with discussion)".
Journal of the royal statistical society, B, 36 (1974) 1-37.

LEI,
Bedrijfsuitkomsten in de landbouw boekjaar 1984/85.
Den Haag, 1986. Periodieke rapportage 11-84/85.

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.

LITERATUUR (2e vervolg)

Nathan, G. en D. Holt,
"The effect of survey design on regression analysis".
Journal of the royal statistical society, B, 42 (1980) 377-386.

Rijken van Olst, H.,
Algemene statistiek.
Assen (van Gorcum) 1974.

Bijlage

Bijlage : Regressieanalyse op basis van groepsgemiddelden

Deze bijlage behandelt de gevolgen van regressieanalyse toegepast op gemiddelden per groep van bedrijven. Het groeperen en middelen van dwarsdoorsnedegegevens wordt regelmatig toegepast met als argument dat zodoende een reductie van het databestand optreedt tot een naar eigen zeggen aanvaardbaar niveau. In woorden van de econometrist Johnston (1972): "it is sometimes the case that an investigator faced with very large numbers of observations will undertake some prior grouping of the data in order to reduce the sheer bulk of the calculations".

De literatuur (Cramer, 1964; Johnston, 1972) noemt een drietal eigenschappen van het groeperen en middelen van data:

(1) De schatting van de coëfficiënten, gebaseerd op gemiddelden per groep van bedrijven, is zuiver.
(2) De standaardfouten van de regressiecoëfficiënten nemen toe en zodoende de T-waarden af.
(3) De determinatiecoëfficiënt (R^2) neemt toe.

Ondanks dat het groeperen en middelen van data los staat van de wijze van steekproeftrekking - immers het middelen van gegevens kan op ieder databestand van dwarsdoorsnedegegevens worden toegepast - is besloten hier toch aandacht aan te besteden, juist omdat het veelvuldig wordt toegepast en de LEI-steekproef hierin geen uitzondering vormt. Voor voorbeelden zie Burger (1983) alsmede Douma en Poppe (1987). Sterker, het groeperen en middelen in een gestratificeerde steekproef ligt voor de hand, omdat een indeling in groepen reeds voorhanden is.

Tabel 1 toont de schatting voor de CD-functie (zie paragraaf 3.4) op basis van die groepsgemiddelden, die overeenkomen met de 32 basisstrata (zie hoofdstuk 2). Uitgegaan is van het schattingsmodel

$$\bar{Y}_h = b'\bar{X}_h + u_h, \quad h = 1,\ldots,H, \quad u_h \sim N(0,\sigma_u^2/n_h)$$

$$\text{met } \bar{Y}_h = \frac{1}{n_h}\sum_{i=1}^{n} Y_i, \quad \bar{X}_{hj} = \frac{1}{n_h}\sum_{i=1}^{n} X_{ij}, \qquad (1)$$

en n_h het aantal waarnemingen in stratum h. Dit aantal staat vermeld in tabel 2.3.

Tabel 1 Schattingsresultaten voor de CD-functie op basis van groepsgemiddelden en op basis van individuele bedrijven (zie paragraaf 3.4)

Verklarende variabelen	Groepsgemiddelden		Bedrijfsgegevens	
	coëf.	T-waarde	coëf.	T-waarde
Constante	1.6075	4.11	2.2446	24.59
Grond	.1176	5.31	.0781	11.70
Arbeid	.1584	.86	.2219	7.97
Kapitaal	.0293	.50	.1327	11.36
Non-factor inputs	.8518	20.88	.7168	71.63
R^2	.99		.95	

Een aantal eigenschappen die hierboven zijn genoemd kunnen op grond van deze resultaten worden bevestigd. De T-waarden, die worden afgeleid van de standaardfouten, nemen inderdaad af en de determinatiecoëfficiënt neemt toe. Daarnaast blijkt de onderzoeker die de CD-functie schat op basis van groepsgemiddelden, in tegenstelling tot de onderzoeker die schat op basis van individuele data, te constateren dat de coëfficiënten van arbeid en kapitaal niet significant van nul verschillend zijn.

Het is zeer de vraag of in dit geval is voldaan aan de eis van zuiverheid. Niet alleen door het verschil in uitkomsten, maar bovenal naar aanleiding van een recent artikel van Deaton (1985). Deze werd geconfronteerd met het probleem dat voor een bepaald onderzoek, dat hij wilde verrichten, geen panel-data, maar alleen cross-sectie data over verschillende jaren beschikbaar waren. Om toch optimaal gebruik te maken van de beschikbare data, besloot hij tot de constructie van een databestand met een zogenaamde panel-structuur. Hiertoe benoemde hij "cohorts", die zijn gedefinieërd als groepen met een vast kenmerk. Voorbeelden hiervan zijn leeftijdsklassen, regio's, bedrijven met en zonder een ligboxenstal, e.d. Vervolgens berekende hij voor de verschillende cohorts over de verschillende jaren gemiddelden, zodat ieder cohort kon worden gevolgd in de loop van de tijd. Tot slot beargumenteerde Deaton dat niet de OLS-schattingsmethode, maar de schattingsmethode die bekend staat als de schattingsmethode met errors-in-variables, moet worden toegepast, omdat de aggregatiefout, die voortkomt uit het ten onrechte als volkomen identiek beschouwen van actoren die aan de gemiddelde actor ten grondslag liggen, bij de schatting moet worden betrokken. In de woorden van Deaton: "I consider economic relationships that are linear in the parameters. Corresponding to these individual relationships, there will exist averaged versions of the same form for the cohort population, but with unobservable data points. The sample cohort means from the surveys are consistent but error-ridden estimates of the unobservable cohort population means. Since the micro data are used to construct the means, they can also be used to construct estimates of the variances and covariances of the sample means. It is therefore possible to use errors-in-variable estimators to estimate consistently the population relationships".

De gevolgen van deze aanname worden duidelijk bij de specificatie van het schattingsmodel

$$Y_h = b'X_h + u_h, \quad h = 1,\ldots,H, \quad u_h \sim N(0,\sigma_u^2/n_h),$$

$$\bar{Y}_h = Y_h + v_h, \quad h = 1,\ldots,H,$$

$$\bar{X}_h = X_h + w_h, \quad h = 1,\ldots,H.$$

Dit schattingsmodel vertoont dezelfde eigenschappen als het lineaire schattingsmodel, met dit verschil dat thans is aangenomen dat Y_h en X_h niet waarneembaar zijn. In plaats daarvan worden $\bar{Y}_h$ en $\bar{X}_h$ waargenomen met respectievelijk een waarnemingsfout van v_h en w_h. Verondersteld wordt dat v_h en w_h ongecorreleerd zijn met u_h en de werkelijke waarden van Y_h en X_h, en dat

$$E(v_h) = 0, \quad E(w_h) = 0,$$

$$E\begin{bmatrix} v_h \\ w_h \end{bmatrix}\begin{bmatrix} v_h & w_h' \end{bmatrix} = \begin{bmatrix} \sigma_{00} & \sigma' \\ \sigma & \Sigma \end{bmatrix},$$

voor alle h en bovendien dat (v_h, w_h') stochastisch onafhankelijk is van (v_k, w_k) voor $h \neq k$. Dit is het standaard regressiemodel met errors-in-variables, gedeeltelijk beschreven door Ketellapper (1982) en Bekker (1986).

Van het standaard regressiemodel met errors-in-variables is bekend, dat de OLS-schatting van de parameter b

$$\hat{b} = (\bar{X}'\bar{X})^{-1}\bar{X}'\bar{Y},$$

waarbij $\bar{Y}$ en $\bar{X}$ de waargenomen waarden van Y en X, niet zuiver is. Dit wordt behandeld in bijna ieder standaardwerk op het terrein van de econometrie (zie onder meer Johnston, 1972; Judge et al., 1980; en Chow, 1983). De richting en de grootte van deze onzuiverheid is afhankelijk van de variabelen behept met waarnemingsfouten en van de grootte van deze variabelen in verhouding tot de waarnemingsfouten. De conclusie is dan ook dat bij regressieanalyse toegepast op gemiddelden per groep van bedrijven niet voldaan is aan de eis van zuiverheid $(E\hat{b}=b)$.

Een consistente schatting, beter bekend als de "corrected least squares" schatting, voor b wordt verkregen door (ontleend aan Deaton, 1985; en Bekker, 1986)

$$\hat{b} = (X'X - n\Sigma)^{-1}(X'Y - n\sigma),$$

$$\hat{\sigma}_u^2 = (Y'Y - n\sigma_{00} - \hat{b}'(X'X - n\Sigma)\hat{b}) / n,$$

$$\text{var}(\hat{b}) = (X'X - n\Sigma)^{-1}\,[1/n\,(X'X)e'e + X'ee'X]\,(X'X - n\Sigma)^{-1},$$

met $e = Y - X\hat{b}$

en n het aantal waarnemingen, dat wil zeggen het aantal groepsgemiddelden vermenigvuldigt met het aantal jaren waarover waarnemingen beschikbaar zijn.

Een schatting van de elementen in de matrix $\begin{bmatrix} \sigma_{00} & \sigma' \\ \sigma & \Sigma \end{bmatrix}$

wordt gevonden door

$$S_{ht} = \sqrt{1/n \sum_{i=1}^{n}\sum_{j=1}^{n} (Z_i - \bar{Z})(Z_j - \bar{Z})} \quad \text{met } Z = Y_{ht}, X_{ht}$$

per stratum (h) en per jaartal (t) te bepalen en vervolgens een gemiddelde te bepalen over alle groepsgemiddelden en jaartallen. S_{ht} staat hierbij voor een spreidingsmaatstaf, die aangeeft de mate waarin de waarnemingen binnen een bepaald stratum en binnen een bepaald jaar zijn gespreid rondom het bijbehorende gemiddelde, ofwel een indikator voor de mate waarin het gemiddelde het geheel van de waarnemingen binnen dat stratum en dat jaar representeert (zie ook Rijken van Olst, 1974, blz 36).

De uitwerking van deze schattingsmethode leek ons om twee redenen niet nodig. Ten eerste omdat het eerst berekenen van groepsgemiddelden en spreidings-

maatstaven omslachtig is, en ten tweede omdat het op die manier afbreuk doet aan het doel waarvoor het oorspronkelijk is opgezet, namelijk een reductie van het databestand ten einde het aantal uit te voeren berekeningen te verminderen. Met deze conclusie sluiten wij de bijlage af.