# Handleiding voor de mogelijkheden en het gebruik van paneldata op het LEI

## Het Informatienet en de Landbouwtelling

Stijn Reinhard
Lanie van Staalduinen
Marcel Spijkerman

Januari 2001

Notitie 01.03

LEI, Den Haag

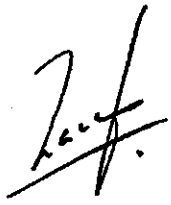
# Inhoud

# Woord vooraf

Dit onderzoek is gefinancierd uit de Strategische Expertise Ontwikkelingsgelden van het LEI. Deze handleiding heeft tot doel de expertise van de medewerkers te verbeteren ten aanzien van de paneldataschattingstechnieken. In deze handleiding moesten we laveren tussen enerzijds het praktisch en eenvoudig aanbieden van de benodigde kennis en anderzijds het correct weergegeven van de econometrische aspecten van paneldataschattingen. Gezien de zeer diverse voorkennis van LEI-medewerkers op dit punt, was dit geen gemakkelijke opgave. Wij hopen dat deze handleiding en de bijbehorende oefenprogramma's het belang van een juiste keuze van de schattingsopzet laten zien, maar ook tonen dat er geen standaard aanpak mogelijk is voor alle onderzoeksvragen. De empirische onderzoeker zal zelf knopen moeten doorhakken over de beste specificatie.

Wij willen graag de LEI-medewerkers bedanken die eerdere versies van deze handleiding en de bijbehorende programma's hebben becommentarieerd: Karel van Bommel, Petra Hellegers, Diti Oudendag en Karel Lodder. Ook zijn wij dank verschuldigd aan Alfons Oude Lansink van de Wageningen Universiteit, hij heeft een kritische blik geworpen op de juiste formulering van de tekst.

De directeur,



Prof.dr.ir. L.C. Zachariasse

# 1. Inleiding

## 1.1 Aanleiding

Op het LEI is veel micro(-economische) informatie (informatie op bedrijfsniveau) aanwezig over de Nederlandse land- en tuinbouw door de aanwezigheid van de Landbouwtellingen (LBT) en het Bedrijven-Informatienet van het LEI (het Informatienet). Het Informatienet bevat een groot aantal variabelen die geschikt zijn om gedegen empirisch financieel, economisch en/of technisch onderzoek te verrichten. Echter de structuur van de dataset (wijze waarop deze tot stand komt en de vorm waarin deze aan de onderzoeker wordt aangeboden) bepaalt de gebruiksmogelijkheden van de dataset voor een bepaald onderzoek. Ervaring leerde dat met dat laatste weinig gedaan werd op het LEI. Het LEI beschikt over prachtige paneldatabestanden zoals het Informatienet en de LBT. Een paneldatabestand is een gegevensbestand met informatie over een aantal waarnemingseenheden of respondenten (bijvoorbeeld boeren) op een aantal tijdstippen (voor het Informatienet worden boeren bijvoorbeeld jaarlijks ondervraagd over een periode van 5 tot 7 jaar). Uit zo'n type bestand kun je hele interessante informatie distileren, mits je de juiste methoden hanteert. Dit was aanleiding om een schriftelijke enquête te houden onder alle LEI-onderzoekers met als hoofddoel zicht te krijgen of wij als LEI efficiënt gebruikmaken van onze paneldatabestanden.

De enquête is uitgevoerd in 1997. In totaal zijn er 123 vragenlijsten verstuurd waarvan er 64 ingevuld teruggekomen zijn. De volgende aspecten zijn in de inventarisatie naar voren gehaald: de mate waarin gebruikgemaakt wordt van data uit het Informatienet en de LBT, de methoden die daarbij gehanteerd worden (regressie, factoranalyse, enzovoort), en de mate waarin onderzoekers in voldoende mate rekening houden met het feit dat zij paneldata onderhanden hebben. Daarnaast is door deze enquête duidelijk geworden hoe de handleiding er volgens de onderzoekers uit dient te zien wil het goed gebruikt kunnen worden door de onderzoekers: theoretisch van opzet of praktisch of juist een mix van beide.

## 1.2 Probleemstelling

Uit de enquête blijkt dat 69% (=44) van alle respondenten (=64) wel eens data uit het Informatienet gebruikt voor onderzoek. Van deze Informatienet-gebruikers schat 64% wel eens een regressievergelijking op basis van het Informatienet. Voor het gebruik van de LBT ligt dit percentage op 61, waarvan 21 wel eens een regressievergelijking schat op basis van de LBT.

Binnen de groep Informatienet-gebruikers blijken 17 personen regelmatig schattingen te doen voor 1 jaar. Schattingen over een langere periode van 2-3 jaren en 4-10 jaren zijn door respectievelijk 11 en 12 mensen wel eens gedaan. Schattingen over een periode van 10 jaar of langer zijn door 8 mensen wel eens uitgevoerd. De meeste respondenten ge-

bruiken de kleinste kwadratenmethode (=OLS) als schattingsmethode. Deze methode is geschikt als de regressieanalyse over 1 jaar gaat, maar is in het algemeen niet geschikt als de analyse meerdere jaren omvat. Reden is dat het gebruik van OLS uitgaat van onafhankelijke waarnemingen. Daar wordt niet aan voldaan bij schattingen over meerdere jaren uit het Informatienet of LBT. In hoofdstuk 2 wordt dit nader uitgelegd. Het blijkt dat in 1997 op het LEI maar 8 responderende hiervan op de hoogte zijn.

Het overgrote deel van de respondenten (77%) geeft aan behoefte te hebben aan een ondersteunende handleiding voor analyses op basis van het Informatienet. Voor de LBT geeft 63% aan dat zij graag een handleiding zouden willen hebben van het schatten van relaties met behulp LBT-data. Het karakter van de handleiding is bijvoorkeur praktisch (55%) of bevat zowel een theoretische verhandeling als een praktische (42%). Tevens is 64% (bijna alle Informatienet-gebruikers) van de respondenten geïnteresseerd in een workshop statistische analyse die is toegespitst op de analyse van gegevens uit het Informatienet en de LBT.

*Concluderend*

Op het LEI wordt gebruikgemaakt van paneldata. Echter, bij multivariate analyse methoden (bijvoorbeeld het schatten van regressievergelijkingen, correlatie bepaling, discriminantenanalyse) op het LEI wordt geen of nauwelijks rekening gehouden met de speciale structuur van de paneldataset zodat een deel van de aanwezige informatie wordt verwaarloosd. De gevolgen zijn mogelijk afwijkende en minder betrouwbare resultaten en conclusies ten opzichte van een situatie waarin wel rekening zou zijn gehouden met de paneldatastructuur. Het is van belang voor de kwaliteit van het onderzoek op LEI dat onderzoekers zich hiervan bewust worden, en dat zij handvatten krijgen aangereikt om in de toekomst hierin verandering te brengen.

## 1.3 Doelstelling

Doelstelling van deze handleiding is te voorzien in de behoefte van LEI-onderzoekers aan een praktische en theoretische handleiding voor het efficiënt gebruik van paneldata op het LEI. Deze handleiding moet het inzicht van LEI-onderzoekers met betrekking tot de mogelijkheden van het gebruik van paneldata in empirisch onderzoek vergroten, zodat er meer, en statistisch efficiënter gebruik kan worden gemaakt van de aanwezige informatie in de bestaande data sets op het LEI.

Een ander doel is om onderzoekers, die nu bij een regressie over meerdere jaren op basis van het Informatienet of de LBT de schattingsmethode OLS gebruiken, op eenvoudige wijze laten kennismaken met paneldata. Hiertoe worden de beginselen van paneldata-analyse beschreven en praktische voorbeelden uitgewerkt. Na bestudering van deze handleiding moet de onderzoeker zelfstandig eenvoudige paneldata-analyses kunnen doen en literatuur kunnen begrijpen waarin ingewikkelder toepassingen worden beschreven.

De handleiding moet leesbaar, begrijpelijk en toepasbaar zijn voor onderzoekers die een basiskennis aan statistiek hebben (HBO, WO-niveau). Dit betekent dat enige voorkennis noodzakelijk is (ze moeten weten of kunnen begrijpen wat een OLS-schatting inhoudt).

## 1.4 Afbakening

Buiten de scope van deze handleiding vallen de volgende onderwerpen:

- de problematiek van het 'wegen'. Deze is groot en zou een aparte rapportage vergen. We geven wel tips hoe hier mee om te gaan maar gaan daar niet diepgaand op in. Als je meer wilt weten over de hele wegingsproblematiek met betrekking tot het Informatienet dan kun je daar Dijk (1989) op na slaan;
- inzichten in de problematiek van het gebruik van het Informatienet met behulp van paneldata-analyses voor voorspellingen in ruimte en tijd, voor dynamische en/of discrete modellen, en voor gebruik in beslissingsondersteunende beleidsmodellen;
- geavanceerde schattingsmethoden die bij een paneldata-analyse gebruikt kunnen worden, zoals SUR, 3SLS en maximum entropy. In paragraaf 2.4.4 wordt de Hausman-Taylor-schattingsmethode beschreven. Deze methode is een versie van 2SLS.

## 1.5 Opzet van de handleiding

Deze handleiding [1] begint bij een uitleg over de theorie van paneldata (hoofdstuk 2). Hierin wordt behandeld hoe data bestanden in elkaar kunnen zitten, hoe het Informatienet en de LBT zijn opgebouwd, wat de voordelen zijn van het gebruik van paneldata, hoe een regressievergelijking er wiskundig uitziet, wat voor schattingsmethoden we kunnen hanteren, enzovoort. Uiteindelijk zullen na lezing van hoofdstuk 2 de toepassingsmogelijkheden, beschreven in hoofdstuk 3, ook beter kunnen worden begrepen en te volgen zijn.

In hoofdstuk 3 is beschreven hoe je zelf een paneldata-analyse kunt uitvoeren. Eerst is aangegeven hoe je met de paneldatabestanden (het Informatienet en LBT) om dient te gaan om een goede paneldataschatting te maken. Daarna wordt aan de hand van een voorbeeld het hele traject van het opvragen van de data tot aan het beoordelen van de schattingsresultaten besproken. In het voorbeeld (paragraaf 3.3, 3.4 en 3.5) wordt de toegediende hoeveelheid stikstof uit kunstmest verklaard. Daarna worden met behulp van een reeds bestaande dataset de schattingen van een klassiek artikel van Mundlak (1961) uitgevoerd en besproken in paragraaf 3.6.

Tenslotte worden conclusies gepresenteerd en aanbevelingen gedaan omtrent het gebruik van paneldata voor onderzoek op het LEI.

In deze handleiding worden voorbeelden uitgewerkt met behulp van LIMDEP7. Dit is een econometrisch software pakket (beschikbaar op het LEI) met standaard procedures voor het analyseren van paneldata. In de LIMDEP-handleiding worden de analyses ook uitgelegd en vaak toegelicht met een voorbeeld.

Naast het voor u liggende specifieke handboek toegespitst op het Informatienet en de LBT, bestaan er in de literatuur standaardwerken over de analyse van paneldata. Een goed begrijpbaar boek (met een basiskennis van econometrie) is geschreven door Badi Baltagi (1995: LEI-informatiecentrum Ab 22) *Econometric analysis of paneldata.* Een wat moeilijker boek is het standaardwerk van Cheng Hsiao (1986) *Analysis of Paneldata* (niet

1 Deze handleiding is samen met de gebruikte programma's in de loop van 2001 ook beschikbaar op Intranet. Eventuele wijzigingen en correcties zijn dan verwerkt in de versie op Intranet.

aanwezig in het LEI-infocentrum). Een ander boek *The econometrics of panel data; a handbook of the theory with applications* van Matyas en Sevestre (1996, LEI-informatiecentrum: Ab 13) geeft een overzicht van wat je met paneldata kan doen, echter het vereist wel een goede kennis van de econometrie (of een wil om je het eigen te maken, uiteraard). Voor degene die hun basiskennis van de econometrie willen opfrissen (betekenis van begrippen, standaard modellen, schattingsmethoden, toetsen en dergelijke) wordt aangeraden een tekstboek econometrie open te slaan. Een aanrader is het boek van Greene (2000) *Econometric Analysis* (aanwezig Ab 27) met daarin speciaal een hoofdstuk gewijd aan paneldata-analyse (hoofdstuk 14, met name paragraaf 14.2 tot en met 14.4). Een andere aanrader is *Estimation and inference in econometrics* van Davidson en MacKinnon (1993; pp. 117, 320-321, 325 voor paneldata-analyse; aanwezig Ab 28). Datzelfde geldt voor het boek van Judge, Hill, Griffiths, Lutkepohl en Lee (1988; niet aanwezig op het LEI), *Introduction to the theory and practice of econometrics* dat de lezer stapsgewijs meeneemt in de leer van de econometrie.

Heb je vragen over paneldata-analyses schroom dan niet om Stijn, Lanie of Marcel aan te schieten voor assistentie.

# 2. Theorie paneldata

## 2.1 Inleiding

Voordat we een beetje de theorie induiken (hoe ziet een regressievergelijking toegepast op paneldata mathematisch eruit) en de voor- en nadelen van het gebruik van paneldata uitleggen, is het noodzakelijk om weet te hebben van de mogelijke structuren van datasets. Ofwel, wanneer spreken we van een paneldatabestand en wanneer van iets anders?

## 2.2 Mogelijke structuren van databestanden

### 2.2.1 Hoofdstructuren: cross-sectie, tijdreeks en paneldata

Er kunnen drie hoofdstructuren van databestanden onderscheiden worden:

1. *cross-sectie;* dwarsdoorsnede door de tijd. Een gegevensbestand met informatie over een aantal waarnemingseenheden (zoals bedrijven, huishoudens, provincies, individuen) op één tijdstip. Bijvoorbeeld je selecteert uit het Informatienet alle gegevens (kunnen bijvoorbeeld ook alleen inkomensgegevens zijn) van alle (of een deel van de) aanwezige land- en tuinbouwbedrijven alleen voor het jaar 1998;
2. *tijdreeks;* Een gegevensbestand met informatie over één waarnemingseenheid op een aantal tijdstippen. Bijvoorbeeld je selecteert 1 bedrijf uit het Informatienet voor de jaren 1993 tot en met 1998 (bijvoorbeeld om te analyseren of er een trend aanwezig is in zijn inkomen gedurende deze jaren);
3. *paneldata;* Combinatie van cross-sectie en tijdreeks. Een gegevensbestand met informatie over een aantal waarnemingseenheden op een aantal tijdstippen. Bijvoorbeeld: je selecteert inkomensgegevens uit het Informatienet van alle (of van melkvee)bedrijven in het Informatienet voor de jaren 1980 tot en met 1998.

Het Informatienet en ook de LBT zijn dus volgens bovenstaande definities te gebruiken als cross-sectie of tijdreeks of als paneldatabestand, afhankelijk van het aantal bedrijven en de hoeveelheid jaren die je selecteert uit het Informatienet of LBT.

### 2.2.2 Structuur van een paneldatabestand: balanced en unbalanced

Het is noodzakelijk dat we nog iets uitleggen over de structuur en opzet van het paneldatabestand zelf. Reden is dat de structuur van je paneldataset invloed heeft op de wijze waarop je je databestand moet opzetten, zodat LIMDEP (het eerder genoemde econometrische pakket) automatisch de juiste schattingsmethode uitvoert. In hoofdstuk 3 komt aan de orde hoe je dit moet doen.

In het marktonderzoek wordt van oudsher al veel gebruikgemaakt van panels. Hierbij gaat het dan bijvoorbeeld om een groep gezinnen of individuen die regelmatig wordt bemonsterd of om een herhaling van precies dezelfde steekproef bij een opinieonderzoek. De verbindende factoren tussen al deze bestanden betreffen het tijdstip en de waarnemingseenheid.

Er wordt onderscheid gemaakt tussen een volledig panelbestand (balanced) en een onvolledig (unbalanced) panelbestand. Balanced wil zeggen dat voor alle waarnemingseenheden (bedrijven, enzovoort) alle variabelen op alle tijdstippen zijn gemeten en unbalanced wil zeggen dat niet alle waarnemingseenheden op alle tijdstippen zijn waargenomen. De een is niet beter als de ander, maar kan voordelen bieden. Er zijn genoeg voorbeelden van grote panelbestanden die met opzet unbalanced gemaakt zijn door het panelonderzoek op roterende basis op te zetten. Een roterend panelbestand wil zeggen dat bij iedere wave, zeg ieder jaar, een vast deel van de respondenten (bedrijven, enzovoort) wordt vervangen door nieuwe respondenten. Het Informatienet is een voorbeeld van een roterend (en dus unbalanced) panelbestand.

Drie algemene redenen om voor een roterend panelbestand te kiezen zijn de volgende:

1. het voorkomt dat personen eeuwig worden gevraagd met als gevolg een hoge uitval gedurende de lange deelname periode;
2. als je weinig vervangt of ververst dan mis je de nieuwe ontwikkelingen in de wereld in je databestand (gedurende de jaren veranderen structuren van bedrijven, worden de bedrijven groter of juist kleiner, en dergelijke). Dit zou je databestand minder representatief maken voor de hele populatie. Je kunt met een roterend panel dus flexibeler reageren op de veranderende buitenwereld;
3. een grote kans op het optreden van het zogenaamde leereffect. Een voorbeeld: boeren en tuinders die 'eindeloos' mee zouden doen aan het Informatienet (bijvoorbeeld van 1960 tot aan 1990) kunnen leren van de inzichten in hun bedrijf, die ze krijgen als beloning van het meedoen aan het Informatienet. Ze kunnen daardoor zaken gaan aanpassen in het bedrijf om de prestaties van het bedrijf te verhogen. Het optreden van leereffecten onder de respondenten heeft in het algemeen een negatieve invloed op de representativiteit van de steekproef. Echter in het geval van het Informatienet is het leereffect niet aangetoond. Een ding is zeker: als er al sprake was van een leereffect in het verleden, dan zal dit effect nu kleiner zijn. Alle agrarische bedrijfshoofden ontvangen nu veel meer informatie over hun bedrijf (via accountantskantoren, stringentere milieubeleid met bijbehorende boekhoudingen/balansen, enzovoort) dan enkele decennia geleden. Het voordeel voor Informatienet-bedrijven is hierdoor veel kleiner geworden.

### 2.2.3 Structuur Landbouwtelling en het Informatienet

*Landbouwtelling*

Ieder jaar op een bepaald tijdstip in april of mei, worden alle land- en tuinbouwbedrijven opgeroepen om zich te laten registreren in de Landbouwtelling. Deze registratie is een momentopname met als gevolg dat bedrijven kunnen ontbreken in de LBT, maar ook dat

het aantal beesten op 1 april verschilt van het aantal beesten op 1 september (in het Informatienet is dan ook het gemiddelde aantal dieren per jaar opgenomen). Daarnaast is er in het algemeen een dalende tendens in het aantal bedrijven. Dit geldt trouwens niet voor alle groepen (type) bedrijven. Zoals eerder beschreven kan de LBT als cross-sectie, tijdreeks of als paneldatabestand gebruikt worden na een specifieke selectie. Als je de LBT als panel wilt gebruiken zul je moeten selecteren op bedrijfsnummer en bekijken voor hoeveel jaren al deze bedrijfsnummers voorkomen. Stel je wilt voor een periode van 5 jaar een panelbestand uit de LBT halen om te kijken of de hoeveelheid aanwezige koeien afhankelijk is van een bepaalde regio. Dan heb je twee keuzes: of je selecteert alleen die bedrijven die gedurende die hele periode van 5 jaar aanwezig zijn in de LBT (=balanced panel) of je selecteert alle opgenomen bedrijven in de LBT in die periode (=unbalanced panel). De laatste selectie zal meer bedrijven bevatten dan de eerste, wegens stoppen van bedrijven, overnames enzovoort.

*Het Informatienet*

Per wave, of jaar in ons geval, maken ongeveer 1.500 land- en tuinbouwbedrijven deel uit van het Informatienet. Deze bedrijven zijn door middel van een gestratificeerde steekproef gekozen uit alle bedrijven in de Landbouwtelling. In het Informatienet wordt in theorie ieder jaar ongeveer 20% van de bedrijven ververst. Bedrijven kunnen in theorie maximaal 5-7 jaar in het Informatienet blijven. Het aantal waarnemingen per meting of wave blijft zo ongeveer gelijk. Het Informatienet is daarom als een roterend panel of een unbalanced panel aan te duiden. Gedurende een wave kunnen bedrijven afvallen door wat voor reden dan ook. Dat betekent dat je in de volgende wave extra nieuwe bedrijven moet bijkiezen, naast de gebruikelijke 20% verversing per wave. Dat dit niet altijd even goed lukt laat zich raden: de non-response (niet mee willen doen voor een eerste deelname aan het Informatienet) onder *geschikte bedrijven* bedroeg in 1999 ongeveer 65%. Uiteraard moet bij de keuze van de bedrijven rekening worden gehouden met de geschiktheid van die bedrijven om de representativiteit van de steekproef niet in gevaar te brengen. Het gevolg voor het Informatienet bijvoorbeeld is dat er nooit precies 1.500 bedrijven per wave of jaar in de steekproef zitten, en dat bedrijven die al in het Informatienet zitten gevraagd worden langer deel te nemen (vandaar dat je soms bedrijven in het Informatienet tegenkomt die al 9 of 10 jaar meedoen).

Als je gebruik wilt maken van de panelstuctuur van het Informatienet bij je onderzoek dan loop je tegen de korte tijdsperiode aan dat bedrijven deelnemen aan het Informatienet (korte tijddimensie; meestal weergegeven door T). De veranderingen die eventueel plaatsvinden binnen het bedrijf kun je niet lang volgen: het gaat dan om veranderingen die op lange termijn spelen (bedrijfsovername, aankoop vaste inputs, en dergelijke). Daarnaast speelt mee dat bij een korte tijdsdimensie de asymptotische argumenten naar oneindigheid, die voor enkele toetsen worden verondersteld (voor een uitleg wordt verwezen naar Greene, 2000) alleen vanuit de N-dimensie gerealiseerd zouden kunnen worden. Dus puur vanuit het onderzoek gezien zouden bedrijven best langer in het Informatienet mogen blijven (zo'n 10 jaar). Dit betekent echter wel een verhoogde kans op uitval (bijvoorbeeld boeren die geen zin meer hebben om mee te doen op een gegeven moment, of verhuizing, enzovoort).

*Representativiteit en weging Informatienet* [1]

In het Informatienet wordt een gedetailleerde administratie bijgehouden van ruim 1.500 land- en tuinbouwbedrijven. Naast financieel-economische gegevens worden ook technisch-economische, milieu-economische en sociaal-economische gegevens van deze bedrijven vastgelegd. Het Informatienet wordt mede bijgehouden voor de Europese Unie (FADN). Daarnaast vormt het Informatienet de basis voor veel onderzoek zoals dat binnen het LEI wordt uitgevoerd. Op basis van de bedrijven in het Informatienet worden uitspraken gedaan over alle land- en tuinbouwbedrijven (of delen daarvan). De vraag die dit wellicht oproept is 'hoe kunnen nu uitspraken worden gedaan over de hele populatie als slechts informatie wordt verzameld bij een deel van de populatie'. Het antwoord ligt in de selectie van bedrijven die in het Informatienet worden opgenomen, de steekproef. Een kok eet immers ook niet de hele pan soep leeg om uitspraken te doen over de kwaliteit. Wel belangrijk is dat voor het proeven goed wordt geroerd, de eetlepel soep die beoordeeld wordt, moet overeenkomen, oftewel moet representatief zijn voor het geheel. Hetzelfde geldt voor het Informatienet. De bedrijven die in het Informatienet zijn opgenomen moeten representatief zijn voor de gehele populatie. Op deze manier kan men zelfs tot betere schattingen komen op basis van slechts een deel van de bedrijven: bij een beperkt aantal bedrijven kan men veel nauwkeuriger en kwalitatief betere gegevens verzamelen dan wanneer men alle bedrijven zou moeten bezoeken en onderzoeken.

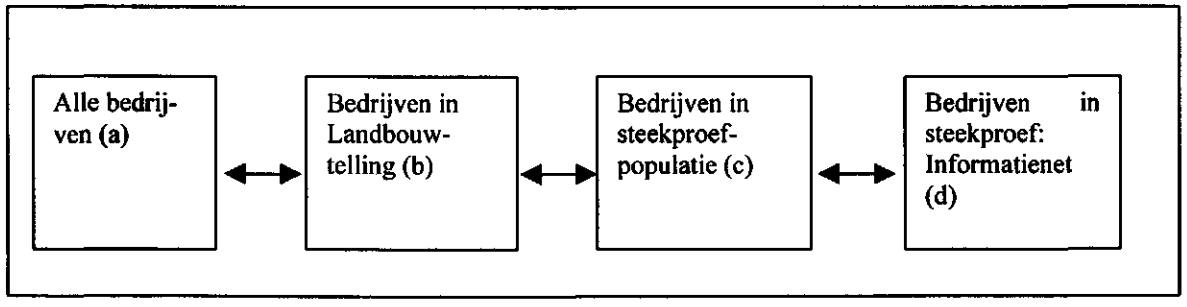
Het is dus belangrijk dat er voor gezorgd wordt dat de bedrijven in het Informatienet representatief zijn voor de bedrijven in de populatie. Hiertoe wordt gebruikgemaakt van een disproportionele gestratificeerde steekproef. Een gestratificeerde steekproef wil zeggen dat de populatie in een aantal groepen wordt opgedeeld en dat er vervolgens bedrijven uit elk van de afzonderlijke groepen worden geselecteerd. De kenmerken op basis waarvan de groepsindeling tot stand komt, moeten belangrijke kenmerken van de populatie zijn zodanig dat bedrijven die in een groep terecht komen veel op elkaar lijken. Door gebruik te maken van deze groepsindeling weet men zeker dat bedrijven uit alle groepen in de steekproef terechtkomen. Disproportioneel wil zeggen dat niet alle bedrijven een even grote kans hebben om in de steekproef terecht te komen. Groepen die heel homogeen zijn, dat wil zeggen dat de bedrijven sterk op elkaar lijken, hebben een lagere trekkingskans. Immers, als alle bedrijven (bijna) identiek zijn kan men op basis van een beperkt aantal waarnemingen een redelijke uitspraak doen (in het extreme geval dat alle bedrijven identiek zijn is één waarneming voldoende om een exacte uitspraak over de hele groep te doen). Bij minder homogene groepen zal men meer bedrijven moeten opnemen om betrouwbare uitspraken te doen. De variabelen of kenmerken op basis waarvan de groepen worden ingedeeld hebben dus een belangrijke invloed op de representativiteit van de steekproef. In het Informatienet worden de groepen ingedeeld op basis van het bedrijfstype, de regio, nge-klassen en meer verfijnd naar de bedrijfsomvang in hectares, de leeftijd van het bedrijfshoofd en een fijnmaziger regio-indeling.

Door op deze manier de bedrijven te selecteren kunnen uitspraken worden gedaan over de hele populatie. Op basis van de bedrijven in een groep kunnen uitspraken worden gedaan voor de hele groep, doordat door de stratificatie bedrijven uit alle groepen zijn op-

[1] Deze sectie is voor een groot deel overgenomen uit Vrolijk (1999) *Agrimonitor.*

genomen kunnen uitspraken worden gedaan over alle groepen. Alle groepen te zamen vormen de gehele populatie. In het Informatienet is dit gerealiseerd door aan elk bedrijf een gewicht toe te kennen (wegingsfactor). Het gewicht wordt berekend door het aantal bedrijven in de populatie (in een bepaalde groep) te delen door het aantal bedrijven in de steekproef (in die zelfde groep). Deze wegingsfactor geeft dus aan hoeveel bedrijven dat ene bedrijf representeert in de totale populatie. Als je een regressievergelijking schat of een gemiddelde uitrekent (of wat voor analyse dan ook doet) op basis van het Informatienet, dan dien je rekening te houden met deze wegingsfactor. Er kan verschil in uitkomsten ontstaan als je wel of niet de wegingsfactor meeneemt in je analyses. Wanneer je bij een paneldataregressie wel of niet moet wegen is een lastig vraagstuk waar geen eenduidig antwoord op te geven is. In de literatuur is men het daar ook niet over eens. De voor- en nadelen van wegen met betrekking tot Informatienet-data bij paneldata-analyses en de gevolgen daarvan, zou al een onderzoek op zich zijn. Wij gaan in deze handleiding dan ook voorbij aan de hele wegingsproblematiek. Het beste is om je schatting zowel met weging te doen als zonder en dan te kijken of daar grote verschillen in zitten. Als dat het geval is dan moet je goed kijken naar de uitkomsten (moet je uiteraard altijd als eerste doen) welke realistisch zijn en welke niet. Als er geen grote verschillen zijn dan bevelen wij aan om vooraf niet te wegen voordat je gaat schatten. In hoofdstuk 3 komen we hierop terug. Voor een diepgaande analyse rond de wegingsproblematiek wordt verwezen naar Dijk (1989).

Op deze manier wordt geprobeerd het Informatienet zo representatief mogelijk te maken voor de gehele populatie. Hierbij moeten twee kanttekeningen worden geplaatst. De eerste is dat de representativiteit is gewaarborgd ten aanzien van de kenmerken op basis waarvan de groepen zijn ingedeeld. Dit wil nog niet zeggen dat de steekproef voor elke willekeurig te bedenken variabele representatief is. Voor een overzicht van onder andere de representativiteit zie Dijk et al. (1999). Ten tweede geldt dat de populatie waarvoor het Informatienet representatief zou moeten zijn, niet betrekking heeft op alle landbouw en tuinbouwbedrijven (a in figuur 2.1). Bedrijven die te klein zijn of te laat zijn geteld maken geen deel uit van de landbouwtelling (b). De steekproefpopulatie (of eigenlijk steekproefkader) (c) wordt gevormd door de bedrijven die in de landbouwtelling zijn opgenomen en een omvang hebben van minimaal 16 nge en maximaal 800 nge. Het percentage bedrijven en het percentage nge dat hiermee wordt gedekt, is vermeld in figuur 2.1. Uit dit steekproefkader wordt de daadwerkelijke steekproef getrokken (d).




*Figuur 2.1 Relatie steekproef en totale populatie*

## 2.3 Voordelen van het gebruik van paneldata

### 2.3.1 Inleiding

Aan het gebruik van paneldata is een aantal substantiële voordelen verbonden in vergelijking met het gebruik van cross-sectie of tijdsreeks data. Er zijn echter ook aandachtspunten bij het gebruik van paneldata, of cross-sectie of tijdreeks. Voordat alle voordelen op een rij gezet worden is het handig om alvast een idee te krijgen van hoe een regressievergelijking op basis van een panelbestand eruit ziet. We gaan uit van het algemene lineaire regressiemodel, waarin een afhankelijke variabele *Y*, zeg de kunstmestgift per ha maïsland, wordt verklaard door onafhankelijke variabelen, zeg de prijs van kunstmest (*X1*), grondsoort (*X2*), en het aantal koeien (*X3*), enzovoort, weergegeven door de matrix X. De index *i* duidt de waarnemingseenheid aan (bedrijf *i*) en *t* het tijdstip. Het totaal aantal bedrijven meegenomen in de regressie bedraagt N, dus i loopt van 1 tot N bedrijven. Het aantal keren dat bedrijf i wordt geënquêteerd bedraagt T keer, dus t loopt van 1 tot T.

Een paneldataregressie verschilt met een reguliere cross-sectie en tijdreeksregressie door het dubbele subscript aan de variabelen (zowel *i* als *t*). Het panel-regressiemodel ziet er voor bedrijf i op tijdstip t als volgt uit:

$$Y_{it} = \alpha + X'_{it}\,\beta + v_{it} \qquad (2.1)$$

met:

$Y_{it}$ = de observatie van bedrijf *i* in jaar *t* van de te verklaren variabele (de kunstmestgift op maïsland):

$\alpha$ = scalair (constante);

$X_{it}$ = de *it*-de observatie van de K verklarende variabelen (veranderen in de tijd en tussen de bedrijven, bijvoorbeeld betaalde kunstmestprijs, aantal koeien, enzovoort);

$\beta$ = K x 1 vector met de te schatten coëfficiënten;

$v_{it}$ = storingsterm (vertegenwoordigt alle niet geobserveerde /niet waargenomen factoren in de vergelijking maar die wel de hoogte van *Y*, de kunstmestgift per hectare maïsland beïnvloeden). v it ~ IID(0, σ 2v) (=onafhankelijk, identiek verdeeld met verwachting 0 en variantie σ2v);

*i* = 1,...,N (bedrijf; cross-sectie dimensie);

*t* = 1,...,Ti (tijdsperiode (jaar) voor betreffende bedrijf *i*; tijdreeks dimensie).

### 2.3.2 Voordelen paneldata

Wat zijn nu de voordelen aan het gebruik van paneldata ten opzichte van cross-sectie en tijdreeks?

*1. Het aantal observaties bij een panelbestand is groter dan bij cross-sectie of tijdreeks*
Het voornaamste doel van een regressieanalyse (al dan niet met behulp van paneldata) is om de elementen van de vector β op een of andere manier goed te schatten. Stel je wilt de kunstmestgift (*Y*) verklaren uit bijvoorbeeld de prijs van kunstmest en het aantal koeien (elementen in de kolomvector *X*). De kunstmestgift op maïs per hectare (*Y*), de prijs van

kunstmest (*X1*) en het aantal koeien (*X2*) weet je uit het Informatienet (zijn exogeen verondersteld). De β vector wordt geschat en geeft de waarde van je coëfficiënten weer (endogeen). De uitkomst ziet er dan als volgt uit: 75 kg kunstmest op maïs per hectare wordt verklaard door bijvoorbeeld -0,2 * prijs van kunstmest + 0,1 * aantal koeien + de storingsterm (restterm).

Een eerste voordeel van paneldata blijkt meteen: vergelijking (2.1) kan worden geschat op grond van N*T waarnemingen, wat betrouwbaarder resultaten oplevert dan een tijdreeks (T waarnemingen) of een cross-sectie (N waarnemingen).

Een voorbeeld: het Informatienet bevat ongeveer 1.500 bedrijven (=N) per wave (jaar). Deze bedrijven worden in theorie maximaal 5 tot 7 jaar achtereenvolgend geënquêteerd, soms wel tot 10 jaar. Als we voor de eenvoud even uitgaan van 7 jaar: door uitval en dergelijke kunnen bedrijven dus minimaal 1 en maximaal 7 (=T) keer worden geënquêteerd. Bij een cross-sectie zou dit bestand bestaan uit 1.500 bedrijven met T=1 (bijvoorbeeld het jaar 1998), zodat het aantal observaties 1.500 bedraagt. Bij een tijdreeks zou het bestand uit één bedrijf bestaan dat 7 keer wordt geënquêteerd, zodat het aantal observaties 7 bedraagt (bijvoorbeeld de jaren 1992-1998). Bij een panelbestand bestaat het aantal observaties (in theorie) uit 1.500 per wave maal 7 jaren = 10.500. Een aanzienlijke toename van het aantal waarnemingen met daardoor een grotere aannemelijkheid om betrouwbaarder coëfficiënt schattingen (van β) te geven. Je kunt met paneldata in feite de veranderingen meten tussen bedrijven maar tegelijkertijd ook binnen de bedrijven zelf door de jaren heen. Op deze manier gebruik je de aanwezige informatie veel beter dan alleen maar te kijken naar verschillen door de tijd heen voor 1 bedrijf (tijdreeks) of kijken wat de verschillen zijn tussen bedrijven voor 1 jaar (cross-sectie).

2. *Minder last van multicollineariteit*

Samenhangend met de toename van het aantal observaties kan het gebruik van paneldata het probleem van multicollineariteit verlichten. Multicollineariteit houdt in dat de verklarende variabelen ofwel regressoren (de variabelen die opgenomen zijn in de $X_{it}$) afhankelijk van elkaar zijn. Dit betekent dat een aantal variabelen die de hoogte van de kunstmestgift op maïs verklaren, (hoog) gecorreleerd zijn met elkaar. Wanneer de verklarende variabelen variëren over twee dimensies (N en T), is het minder aannemelijk of waarschijnlijk, dat ze hoog gecorreleerd zijn omdat de absolute variantie toeneemt in de tijd. Dat wil zeggen de prijs (betaald door de boer) van kunstmest zal meer variëren over 7 jaar dan over 1 jaar. Als je namelijk geen variatie hebt in je gegevens kun je ook geen effecten berekenen (zoals een prijselasticiteit van kunstmest).

3. *Paneldata sets maken het mogelijk om (dynamische) effecten te berekenen die niet te achterhalen zijn in cross-sectie of tijdreeks*

Soms is beargumenteerd dat cross-sectie lange termijn gedrag weergeeft (van een bedrijf of een land of ...) en tijdreeks korte termijn effecten weergeeft. Bij een cross-sectie worden de verschillende stadia van ontwikkeling weergegeven. Combineren van deze twee soorten informatie (tijdreeks en cross-sectie) heeft als gevolg dat je een meer algemeen en veelomvattend dynamische structuur kan formuleren en schatten. Hier gaan wij trouwens in deze handleiding niet verder op in. Dynamische modellen worden hier niet behandeld.

*4. Het gebruik van paneldata kan de schattingsbias elimineren of verkleinen*

Een groot probleem bij het opstellen van een regressievergelijking is het specificeren van je vergelijking: welke variabelen neem ik op in de vergelijking en welke niet, en in welke vorm (logaritmisch of kwadratisch of lineair of anders). Daarbij is ook van belang dat je weet hoe de variabelen die je opneemt in je vergelijking, zich verhouden tot de variabelen die je niet opneemt in je vergelijking maar wel invloed hebben op de uitkomst. Deze niet opgenomen variabelen zitten verwerkt in je storingsterm. Als nu de effecten van je niet-opgenomen variabelen gecorreleerd zijn met de wel opgenomen verklarende variabelen, en als met die correlaties geen rekening wordt gehouden dan zijn je schattingen biased.

Schattingsbias wil zeggen dat de schatting van $\beta$ voor de prijs van kunstmest niet zuiver is (je geschatte $\beta$ wijkt af van de werkelijke waarde van $\beta$), omdat de schatting vertroebeld wordt door het effect van de deskundigheid van de boer op de kunstmestgift op maïs (=*Y*) bijvoorbeeld. Hierbij moet gedacht worden aan het effect van de niet te obverseren variabelen (management capaciteit van de boer) op de hoogte van de kunstmestgift (=*Y*) die uiteindelijk in je $\beta$ voor de kunstmestprijs terechtkomt.

Bij cross-sectie data zullen de schattingen biased zijn omdat het management van de boer niet expliciet opgenomen kan worden in de vergelijking. Met paneldata ben je in staat om deze variabele te controleren door het introduceren van een zogenaamd individueel effect. In dit effect zitten alle variabelen die niet veranderen over de tijd (*t*) maar wel verschillend zijn tussen boeren (*i*). Dit individuele vaste effect zal onder andere in de volgende paragraaf uitgebreid aan de orde komen.

*5. Geen sprake van afwijkingen doordat je niet hoeft te aggregeren over bedrijven*

Paneldata wordt verzameld op micro niveau (bedrijfsniveau, individueel niveau). Veel variabelen kunnen preciezer worden berekend op micro niveau dan op macro niveau. Je hebt dus geen last van aggregatie bias.

Concluderend komt het er op neer dat paneldata de onderzoeker meer mogelijkheden biedt ten aanzien van het specificeren, maken en testen van (gedrags)modellen voor beleidsgericht onderzoek dan met een cross-sectie of tijdreeks alleen, met bovendien een grotere betrouwbaarheid (mits de juiste methoden worden gebruikt).

### 2.3.3 Aandachtspunten bij (het gebruik van) paneldata, cross-sectie en tijdreeks

Zodra je een databestand gebruikt voor je onderzoek (maakt niet uit welk type), moet je je bewust zijn van een aantal zaken die van invloed kunnen zijn op de bruikbaarheid van dat bestand voor jouw vraag en de uitkomst van je onderzoek. Hieronder staan 3 aandachtspunten vermeld:

*1. Optreden van ontwerp en dataverzameling problemen*

Hier vallen allerlei problemen onder die te maken hebben met hoe je nu de juiste steekproef krijgt, non-response, de geënquêteerde die zich de zaken niet goed meer herinnert, fouten van de interviewer, foute antwoorden als gevolg van verkeerde of onduidelijke vragen, bewust fout antwoorden van de respondent, en dergelijke. Geldt voor alle data structuren.

2. *Optreden van selectie problemen*

Non-response: treedt op bij de eerste keer van benaderen doordat respondent niet mee wil doen, niemand thuis, niet te traceren.

Attrition/uitval: treedt op als mensen gedurende de periode dat ze in het panel zitten afhaken doordat ze verhuizen, overlijden of omdat ze er geen zin meer in hebben om verder deel te nemen.

3. *Problemen door korte tijdreeks dimensie*

De meeste panels hebben een relatief korte tijdsdimensie (T). Voor het Informatienet is dat ook zo: 5 tot 7 jaar (in theorie) zitten respondenten in de data set (als ze niet uitvallen). Dit betekent dat de asymptotische argumenten die naar oneindigheid zouden moeten gaan, volledig afhankelijk zijn van de N-dimensie (het aantal boeren dat geënquêteerd wordt, in het Informatienet zijn dat er per wave ongeveer 1.500). Echter de tijdsdimensie vergroten (boeren langer volgen) betekent een kostenstijging en verhoogt tevens de kans op attrition of uitval (zie ook paragraaf 2.2.3).

## 2.4 Schattingsmethoden paneldata

### 2.4.1 Afhankelijke en onafhankelijke waarnemingen in de tijd en de geschikte schattingsmethode bij regressie

Bij multivariate analyse methoden (bijvoorbeeld het schatten van regressievergelijkingen) moet rekening worden gehouden met de speciale structuur van paneldata, aangezien je te maken hebt met meer dan 1 waarneming voor meer dan 1 bedrijf. Dit wil zeggen dat we bedrijf Jansen, Pietersen, Van Dijk enzovoort volgen in het jaar 1996, 1997, t/m 2000. Per jaar verzamelen we dezelfde gegevens, zoals leeftijd en sexe van de boer, type bedrijf, inkomen, grondsoort, enzovoort. Wat blijkt nu: de uitkomsten voor een aantal verzamelde gegevens/variabelen zullen voor boer Jansen onveranderd blijven gedurende zijn deelname aan het Informatienet. Te denken valt aan zijn sexe, grondsoort, bedrijfstype, opleiding, en dergelijke, de zogenaamde individuele bedrijfskenmerken die over deze korte periode van 5 jaar niet veranderen. De uitkomsten van deze variabelen op tijdstip t hangen dus voor 100% samen met de uitkomsten op tijdstip $t$+1. Dit in tegenstelling tot de uitkomsten van de variabelen: inkomen, prijs van kunstmest die een boer heeft betaald, hoeveelheid kunstmest die aangekocht wordt, enzovoort. De uitkomsten op tijdstip t hangen niet samen met die op tijdstip $t$+1. Je kunt ze beschouwen als onafhankelijke waarnemingen door de tijd.

De meest geschikte en efficiënte schattingsmethode onder bepaalde voorwaarde bij een cross-sectie bestand is in het algemeen OLS (ordinary least squares), of in Nederlands wel kleinste kwadraten methode genoemd. Zodra je nu een paneldatabestand hebt (gegevens van boer Jansen, Pietersen enzovoort over meer dan 1 jaar) is OLS niet meer geschikt als schattingsmethode. Voorwaarde voor het gebruik van OLS is dat namelijk uitgegaan wordt van onafhankelijke waarnemingen. Bij een cross-sectie wordt daar in het algemeen aan voldaan. Echter bij een paneldatabestand wordt daar niet aan voldaan omdat je daar te maken hebt met de bedrijfspecifieke kenmerken die niet veranderen door de tijd en daar-

door geen onafhankelijke waarnemingen in je data set vormen. Als zodanig moet je dus proberen in je schatting van de regressievergelijking rekening te houden met de schatting tussen bedrijven (OLS geschikt) *en* met de schatting binnen de bedrijven (OLS niet geschikt). In het algemeen is dan GLS (Generalized Least Squares) het meest efficiënt omdat deze schatter rekening houdt met dit type afhankelijkheid. Voor een uitleg over een OLS- en een GLS-schatter, wordt verwezen naar het boek van Greene (2000).

### 2.4.2 Opbouw storingsterm in een bedrijfseffect en een tijdseffect

We kunnen twee type basis modellen/schattingsmethoden gebruiken om rekening te houden met de panelstructuur (efficiënt schatten tussen en binnen bedrijven): een Random Effects Model (REM) en een Fixed Effects Model (FEM).

Om een idee te krijgen hoe deze modellen eruit zien en wat de belangrijkste verschillen zijn volgt hieronder eerst een korte uitleg over hoe je de storingstermen kunt opbouwen in een One Way Error Component Model en een Two Way Error Component Model.

Het startpunt is ons eerder gepresenteerde lineaire vergelijking (2.1) uit paragraaf 2.3.1

$$Y_{it} = \alpha + X'_{it}\,\beta + v_{it}$$

Vervolgens gaan we het model bruikbaarder maken omdat we allerlei zaken die eerder niet expliciet opgenomen konden worden, (zoals de bedrijfskenmerken die niet veranderen in de tijd) nu wel opnemen in de vergelijking. Dit doen we door de storingsterm op te bouwen uit een bedrijfseffect en een restterm:

$$v_{it} = u_i + \varepsilon_{it} \qquad (2.2)$$

met:

$u_i$ = niet geobserveerde bedrijfsspecifieke grootheid ofwel individuele effect (varieert over de bedrijven maar is constant over de tijd) die relevante variabelen bevat voor de verklaring van $Y$ (denk hierbij aan grondsoort of kwaliteit grondsoort, management, bedrijfstype enzovoort). Hier zitten dus alle verklarende variabelen ($X$) in die je niet op hebt kunnen nemen in je regressievergelijking omdat ze niet bekend zijn of omdat ze niet veranderen door de tijd, zoals management boer, sexe boer, enzovoort);

$\varepsilon_{it}$ = overgebleven stochastische [1] storingsterm ; $\varepsilon_{it} \sim IID(0, \sigma^2 v)$ [met verwachting 0 en variantie $\sigma^2 v$].

Het model ziet er dan als volgt uit:

$$Y_{it} = \alpha + X'_{it}\,\beta + u_i + \varepsilon_{it} \qquad (2.3)$$

[1] Stochast=variabele die waarden kan aannemen welke de uitkomsten zijn van een statistisch experiment.

Dit wordt het zogenaamde One-way Error Compenent Regression Model genoemd, vanwege de opbouw van de storingsterm in een individueel bedrijfseffect [1] en een restterm.

Zoals de naam doet vermoeden bestaat er ook een Two Way Error Compenent Regression Model. Dit model onderscheidt in de storingsterm naast een individueel bedrijfseffect ook een tijdseffect. De storingsterm is dan als volgt opgebouwd:

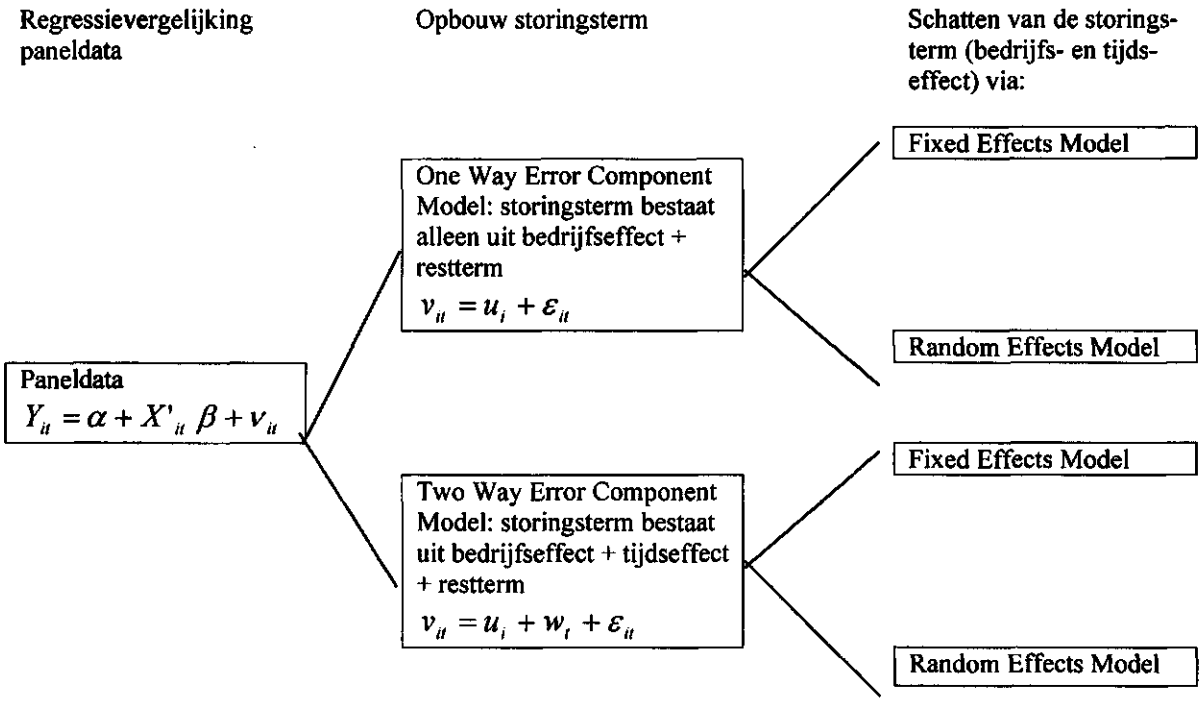
$$v_{it} = u_i + w_t + \varepsilon_{it} \qquad (2.4)$$

$\varepsilon_{it} \sim \text{IID}(0, \sigma^2 v)$

$w_t$ = niet geobserveerde tijdseffect (varieert over de tijd maar niet over de bedrijven): houdt rekening met elk tijd specifiek effect wat niet is meegenomen in de regressie (bijvoorbeeld weersomstandigheden, beleidseffecten, enzovoort).

Het model ziet er dan als volgt uit:

$$Y_{it} = \alpha + X'_{it}\,\beta + u_i + w_t + \varepsilon_{it} \qquad (2.5)$$




*Figuur 2.2 Relatie One Way en Two Way Error Component Model, FEM, REM, en schattingsmethode*

[1] De notatie van het bedrijfseffect, tijdseffect en de statistische ruis verschilt tussen de handboeken. In deze handleiding is de notatie van Greene (2000) en de handleiding van Limdep gebruikt. Bijvoorbeeld Baltagi (1995) gebruikt $u$ voor de gecombineerde storingsterm, $\mu$ voor het bedrijfseffect, $\lambda$ voor het tijdseffect en $\nu$ voor de resterende storingsterm.

Waar nu de voorkeur naar uit zal gaan, is afhankelijk van wat je wilt weten en wat je wilt verklaren met je regressie. Een two-way specificatie heeft als voordeel dat daar impliciet rekening wordt gehouden met een tijdseffect. Een alternatief hiervoor is een one-way specificatie met opname van een jaar trend. Dit laatste heeft als nadeel dat de schatting van deze trend in de vorm van een vaste coëfficiënt (voor ieder jaar gelijk van grootte) naar voren komt. Het is echter in een aantal gevallen niet realistisch om een vaste coëfficiënt te veronderstellen voor technologische ontwikkeling [1], weersomstandigheden, beleidsmaatregelen en dergelijke. In hoofdstuk 3 zullen deze twee type model specificaties op Informatienet data worden toegepast en nader uitgelegd.

Figuur 2.2 geeft bovenstaande nog eens samengevat weer. Daarbij is ook de link gelegd met welk model de storingsterm van een One Way en een Two Way geschat kan worden (beide met zowel een Fixed Effects als een Random Effects Model, afhankelijk van de veronderstellingen die je maakt). Dit onderwerp wordt uitvoerig behandeld in de volgende paragraaf 2.4.3.

### 2.4.3 Fixed Effects Model (FEM) en Random Effects Model (REM)

Zoals uit bovenstaande blijkt maken paneldata het mogelijk om bedrijfseffecten en jaareffecten te isoleren. Er zijn twee mogelijkheden om deze bedrijfs- en jaareffecten te modelleren. (i) De Fixed Effects (FEM) benadering. Hierbij wordt ervan uitgegaan dat ieder bedrijf te karakteriseren is door middel van een specifiek bedrijfseffect dat constant wordt verondersteld voor dat bedrijf. Er zijn geen aannames noodzakelijk over de samenhang van bedrijfseffecten tussen bedrijven en tussen de bedrijfseffecten en de verklarende variabelen. (ii) de Random Effects (REM) benadering. Hierbij wordt ervan uitgegaan dat er 1 bedrijfseffect is voor alle bedrijven en dat verschillen tussen bedrijven (bedrijfseffecten van bedrijven) zijn te modelleren als (stochastische) afwijkingen van het gemiddelde. Het bovenstaande gaat ook op voor de jaareffecten.

*Wat is het verschil tussen een FEM en REM-benadering?*

Het grote verschil tussen de FEM en REM benadering zit in de behandeling van de bedrijfseffecten *ui* 's en tijdseffecten $w_t$ 's bij de schatting. Bij een FEM worden ze als vast behandeld en bij een REM als stochastisch of random. Daarnaast is er een verschil in veronderstellingen die gemaakt moeten worden om zuivere schatters voor $\beta$ te krijgen. Bij een RE benadering wordt onafhankelijkheid verondersteld tussen de $u_i$, $w_t$ en de $\varepsilon_{it}$ onderling, maar ook onafhankelijkheid tussen deze 3 elementen en de $X_{it}$ voor alle *i* en *t*. Bij een FE benadering wordt alleen een onafhankelijkheid verondersteld tussen $X_{it}$ en $\varepsilon_{it}$ voor alle *i* en *t*. Dit betekent dat er bij een FE benadering wel sprake mag zijn van enige correlatie tussen de $X_{it}$ 's en de $u_i$ 's en $w_t$ 's.

In de FEM-benadering wordt het fixed bedrijfseffect (impliciet) gemodelleerd door voor ieder bedrijf een afzonderlijke dummyvariabele op te nemen. Hierdoor stijgt het aantal verklarende variabelen en neemt het aantal vrijheidsgraden af. Als gevolg hiervan

[1] Het is niet realistisch voor een individueel bedrijf, maar vaak wel wanneer 1 coëfficiënt voor de gehele sector wordt geschat.

kunnen geen variabelen worden opgenomen die niet (of incidenteel) veranderen voor een bedrijf (bijvoorbeeld grondsoort, locatie en opleiding van het bedrijfshoofd). De FEM-benadering kent geen aannames over de samenhang tussen de verklarende variabelen en het bedrijfseffect, deze mogen dus een samenhang vertonen. Het hiervoor beschrevene gaat ook op voor een fixed jaareffect. Dat wil zeggen dat bij een two-way fixed effects model benadering (zowel bedrijfseffect als jaareffect) geen trend kan worden opgenomen omdat die onveranderlijk is voor een gegeven jaar. Het individuele bedrijfseffect kan eenvoudig worden afgeleid uit de schattingresultaten. FEM maakt alleen gebruik van de variatie binnen bedrijven (van jaar tot jaar) en niet van de variatie tussen bedrijven, echter bij de schatting wordt wel gebruikgemaakt van de informatie van de andere bedrijven. OLS is hier een geschikte schatter om deze getransformeerde vergelijking te schatten (zie de uitwerking van het voorbeeld in hoofdstuk 3).

In de REM-benadering wordt het random bedrijfseffect gemodelleerd door een voor alle bedrijven gelijk bedrijfseffect plus een random afwijking van dit effect per bedrijf op te nemen. Het aantal vrijheidsgraden is daardoor bij de REM-benadering veel groter dan bij de FEM benadering. Ook kunnen onveranderlijke variabelen zoals grondsoort, locatie enzovoort worden opgenomen. In de REM-benadering wordt verondersteld dat er geen samenhang is tussen de verklarende variabelen en het bedrijfseffect en jaareffect. GLS (Generalized Least Squares) is hier een efficiënte schatter om de vergelijking te schatten.

Uiteraard zijn er naast OLS en GLS nog andere (vaak geavanceerdere) schatters beschikbaar om te gebruiken (zie Davidson en MacKinnon, 1993; Hsiao, 1986). Al naar gelang je wat en hoe wilt schatten zijn andere schatters efficiënt. Te denken valt aan 2SLS (zie ook paragraaf 2.4.4), SUR, 3SLS bij bijvoorbeeld het schatten van een vraagstelsel (je schat dan niet 1 vergelijking maar meerdere tegelijk omdat deze vergelijkingen samenhangen met elkaar), Maximum Likelihood, of via Maximum Entropy. Een andere mogelijke methode is de Hausman-Taylorschatter. Voordeel is dat deze schatter efficiënter is dan de FEM-benadering en deze vooral belangrijk kan zijn voor analyses op Informatienet-data, omdat testen doorgaans uitwijzen dat de REM-benadering wordt verworpen ten gunste van de FEM-benadering. Hier wordt in deze handleiding geen aandacht aan besteed. In hoofdstuk 3 worden twee voorbeelden uitgewerkt met OLS en GLS.

*Wanneer een FEM of REM benadering nemen?*

Het is vaak moeilijk welk model je moet kiezen, hoewel er wel enig houvast is om een keuze te maken. Het is altijd handig om beide schattingen uit te voeren en te kijken wat de verschillen zijn. Dat geeft gevoel voor je data en je specificatie van je vergelijking. Daarnaast kun je onderstaande criteria aflopen om een weloverwogen beslissing te nemen welke te nemen. De criteria op zichzelf zijn niet eenduidig helaas, maar in combinatie met elkaar kunnen ze je goed op weg helpen bij je beslissing over de specificatie van de vergelijking.

1. Als er correlatie bestaat tussen het bedrijfseffect $u_i$ 's en de verklarende variabelen $X_{it}$ 's dan gaat de voorkeur uit naar een FEM.
   Dit kun je testen met de Hausman-test die standaard als output bij je schattingsresultaten door LIMDEP wordt gegeven. Deze teststatistiek geeft aan of het bedrijfseffect

(en eventueel tijdseffect) en de verklarende variabelen samenhangen. Bij een hoge waarde voor de Hausman-test gaat de voorkeur uit naar een FEM-benadering in plaats van een REM-benadering. De Hausman-test heeft een hoge waarde als de *probability value* gelijk is aan 0,000. Dit wordt verder uitgelegd in hoofdstuk 3 (paragraaf 3.4 en 3.5).
Echter, het blijkt uit de praktijk dat een panel met een kleine tijdsdimensie van de deelnemende bedrijven, T, zoals het Informatienet, in theorie geen eenduidige uitspraak kan geven aan de hand van de Hausman-test. Reden is dat de Hausman-test uitgaat van oneindige T.

2. De Breusch and Pagan's Lagrange Multiplier statistic (LM) test of je een FEM of REM benadering moeten nemen boven een klassieke regressie (OLS), ofwel zijn er überhaupt bedrijfs- of tijdseffecten (typerend voor een paneldatamodel)? De LM-test resultaten worden standaard in de output van LIMDEP gepresenteerd.
Wanneer de LM test hoge waarde laat zien (*probability value* kleiner dan 0,05) dan verdient een paneldata-aanpak de voorkeur boven een 'gewone' OLS-regressie (niet expliciet opnemen van bedrijfs- en/of tijdseffecten in de te schatten vergelijking). Als de uitkomst van de test klein is, gaat de voorkeur uit naar een gewone regressie.
Grote waarde voor de LM-test in combinatie met een lage waarde voor de Hausman-test geeft een voorkeur voor een REM-benadering.

3. Een FEM-benadering belemmert de opname van variabelen die niet over de tijd veranderen, bijvoorbeeld bedrijfslocatie en grondsoort. Dit geldt ook voor het tijdseffect uiteraard (bij schatten van een Two Way Error Compenent model).

4. Als laatste kun je ook nog kijken naar waarvoor de schatting dient. Als iets gezegd moet worden over de hele populatie agrarische bedrijven (op basis van een a-selecte steekproef, zoals het Informatienet), bijvoorbeeld over alle bedrijven in Nederland, dan heeft een REM de voorkeur boven een FEM. Een FEM is meer geschikt als iets gezegd moet worden over een *deelpopulatie uit die hele populatie*, dus bijvoorbeeld alleen de melkveehouderij bedrijven in de Veenkoloniën geselecteerd uit het Informatienet.

Nogmaals, schat je vergelijking via beide benaderingen, loop de criteria langs en bedenk vooraf waar je een schatting voor wilt hebben, combineer dit alles en kies er een uit. In hoofdstuk 3 worden bovenstaande zaken verder toegelicht aan de hand van twee voorbeelden op basis van het Informatienet. Als zowel REM als FEM niet geschikt zijn voor je specifieke probleem, moeten alternatieve schattingsmethoden worden overwogen. Deze vergen echter veel meer kennis van econometrie. In de volgende paragraaf wordt de Hausman-Taylormethode als alternatief voor REM en FEM beschreven.

### 2.4.4 De Hausman-Taylor-(HT)schattingsmethode

In de paragraaf 2.4.3 zijn het Fixed Effects Model (FEM) en het Random Effects Model (FEM) voor paneldata besproken. Eveneens zijn criteria besproken voor de keuze van een

FEM- dan wel REM specificatie. Met betrekking tot de hier te bespreken Hausman-Taylormethode (af te korten tot HT) zijn twee van de genoemde criteria van speciaal belang:

1. wanneer het bedrijfseffect ui en de verklarende variabelen $X_{it}$ gecorreleerd zijn gaat de voorkeur uit naar een FEM;
2. wanneer de interesse uitgaat naar variabelen die niet over de tijd veranderen heeft een REM specificatie de voorkeur omdat dergelijke variabelen in een FEM specificatie geëlimineerd worden.

In sommige modellen zijn èn $u_i$ en $X_{it}$ gecorreleerd èn gaat onze de interesse uit naar parameters voor variabelen die niet over de tijd veranderen. In dit geval is noch FEM noch REM geschikt: een REM-specificatie mag niet omdat variabelen met de storingsterm gecorreleerd zijn terwijl in een FEM-specificatie de variabelen die niet over de tijd variëren geëlimineerd worden en de bijbehorende parameters niet geschat kunnen worden. De vraag is wat nu te doen? In Hausman en Taylor (1981) is een methode ontwikkeld die, onder bepaalde voorwaarden, een oplossing biedt voor dit probleem. Deze methode is in de econometrische literatuur bekend geworden onder de naam Hausman-Taylor-schattingsmethode (afgekort HT). Het voordeel van HT is dat de effecten van variabelen die niet over de tijd veranderen geschat kunnen worden terwijl de bedrijfseffecten $u_i$ en de $X_{it}$ gecorreleerd zijn. Een nadeel is dat de theorie achter de methode vrij technisch is en dat de methode niet als voorgeprogrammeerde procedure in econometrische softwarepakketten voorkomt, hetgeen betekent dat er enig programmeerwerk bij komt kijken.

In deze paragraaf wordt de HT besproken. De methode maakt gebruik van enkele econometrische technieken (zoals *two-stage-least-squares*) die niet zijn behandeld in de voorgaande paragraven en die wellicht ook niet behoren tot de parate kennis van de toegepast onderzoeker. Wanneer dit aan de orde is dan wordt of naar bijlage 2 verwezen of naar een aantal referenties van standaard econometrietekstboeken.

Hoewel de HT ook van toepassing is op het Two Way Error Component Model wordt in deze bespreking steeds het One Way Error Component Model als uitgangspunt genomen. Het model in (2.3) kan herschreven worden als:

$$Y_{it} = \beta X_{it} + \gamma Z_i + v_{it} \tag{2.6}$$

met:

$$v_{it} = u_i + \varepsilon_{it}$$

In dit model zijn er naast de variabelen die over de tijd veranderen ($X_{it}$) ook variabelen die niet over de tijd veranderen ($Z_i$). In tegenstelling tot bedrijfseffecten ui die latent aanwezig zijn en eveneens niet over de tijd variëren, zijn de $Z_i$ wel geobserveerd. We kunnen hier denken aan bijvoorbeeld het opleidingsniveau of het aantal jaren scholing van het bedrijfshoofd. De storingsterm $\varepsilon_{it}$ is, zoals gewoonlijk, $IID(0,\sigma^2)$. Stel nu dat we geïnteresseerd zijn in het effect van het opleidingsniveau van het bedrijfshoofd op de productie en dat hierover informatie beschikbaar is. Opleidingsniveau maakt dus deel uit

van $Z_i$. We vermoeden echter ook dat iets als 'aanleg' een effect op de productie heeft terwijl hierover voor de betreffende steekproef geen gegevens in de dataset beschikbaar zijn. Dit betekent dat 'aanleg' een latente variabele is, onderdeel van ui in het model [1]. Problemen doen zich voor als opleidingsniveau en aanleg van een bedrijfshoofd gecorreleerd zijn hetgeen niet op voorhand uit te sluiten valt. Het is bijvoorbeeld denkbaar dat bedrijfshoofden met veel 'aanleg' veel opleiding (gemeten in jaren en/of in niveau) genoten hebben terwijl dit voor bedrijfshoofden met minder 'aanleg' niet het geval is.

Een FEM specificatie biedt geen uitkomst voor dit probleem omdat we dan geen schatting van het effect van opleidingsniveau krijgen terwijl een REM specificatie niet toegepast mag worden omdat $Z_i$ gecorreleerd is met $u_i$. Onder bepaalde omstandigheden kan de HT uitkomst bieden voor dit probleem.

In essentie is HT een gegeneraliseerde versie van *generalised two stage least squares* (af tekorten tot 2SLS) waarbij het probleem van endogeniteit van de regressors wordt aangepakt door instrument variabelen te gebruiken [2]. Nu is het gebruik van instrument variabelen op zichzelf niet bijzonder maar de HT maakt op een speciale manier gebruik van de informatie die in het model zelf aanwezig is. Onder bepaalde voorwaarden kunnen alle structurele parameters in het model geïdentificeerd worden. Zie voor een beknopte uitleg van het begrip *identificatie* bijlage 2.

Essentieel in HT is dat de kolommen van $X_{it}$ en $Z_i$ opgedeeld kunnen worden in variabelen ($X_{1it}$, $Z_{1i}$) die niet gecorreleerd zijn met $u_i$ en variabelen ($X_{2it}$,$Z_{2i}$) die wel gecorreleerd zijn met $u_i$. Speciaal aan de HT is dat de $X_{1it}$ niet alleen gebruikt worden voor schattingen van de eigen parameters ($\beta 1$) maar ook voor die van $Z_{2i}$. Wanneer het aantal variabelen in $X_{1it}$ groter is dan dat in $Z_{2i}$ kunnen alle parameters in het model geïdentificeerd worden. Zonder in de technische details in te gaan worden nu de stappen van de HT beschreven.

Het uiteindelijk te schatten model is:

$$\Omega^{-\frac{1}{2}} Y_{it} = \Omega^{-\frac{1}{2}} \beta X_{it} + \Omega^{-\frac{1}{2}} \gamma Z_i + \Omega^{-\frac{1}{2}} v_{it} \qquad (2.7)$$

Hierin is $\Omega$ de co-variantie matrix (zie bijlage 2). De vermenigvuldiging van, zeg $Y_{it}$ met $\Omega^{-\frac{1}{2}}$ is gelijk aan $Y_{it} - (1-\theta)Y_{i.}$ met $\theta = \left[\frac{\sigma_\varepsilon^2}{\sigma_\varepsilon^2 + T\sigma_u^2}\right]^{\frac{1}{2}}$. Hierin is $Y_{i.}$ het groepsgemiddelde van $Y$ over het aantal jaren $T$.

Dezelfde manipulatie is van toepassing op de overige variabelen en levert de *Generalised Least Squares* (GLS) schatting.

*Het schatten van Ω*

[1] Hausman en Taylor (1981) behandelt een vergelijkbaar model waar het effect van scholing op het salaris van werknemers wordt bepaald.

[2] De begrippen 2SLS, *endogeniteit* en *instrumentvariabelen* worden toegelicht in bijlage 2.

In de praktijk zijn de componenten van $\Omega$ ($\sigma_\varepsilon^2$ en $\sigma_u^2$) niet bekend en moeten geschat worden. Enigszins ironisch zijn hiervoor schattingen voor de parameters $\beta$ en $\gamma$ nodig, de parameters waar het juist om begonnen is. Dit rechtvaardigt de vraag waarom HT dan nog gebruikt moet worden. Het punt is dat er in de econometrie een reeks van schattingsmethoden bestaat die echter niet allemaal even precies (lees: even goed) zijn. Het toepassen van HT is gerechtvaardigd omdat het resulteert in betere schattingen van $\beta$ en $\gamma$ in vergelijking tot de andere schattingsmethoden.

Stap 1: Het schatten van $\sigma_\varepsilon^2$

Een schatting van $\sigma_\varepsilon^2$ is relatief simpel te verkrijgen omdat deze verkregen kunnen worden door de within-group (FEM) regressie uit te voeren:

$$\tilde{Y}_{it} = \tilde{X}_{it} + \varepsilon_{it} \tag{2.8}$$

waarbij $\tilde{Y}_{it}$ en $\tilde{X}_{it}$ afwijkingen van het gemiddelde per bedrijf van $Y_{it}$ respectievelijk $X_{it}$ zijn. In dit model zijn de $Z_i$ èn $u_i$ geëlimineerd zodat de storingsterm alleen nog bestaat uit de componentεit. Nu is

$$\hat{\sigma}_\varepsilon^2 = \frac{1}{N(T-1)} \varepsilon_{it}' Q_v \varepsilon_{it} \quad met \quad Q_v = I_{NT} - \left[ I_N \otimes \frac{1}{T} \iota_T \iota_T' \right]$$

N = aantal bedrijven;
T = aantal jaren dat het bedrijf in het panel aanwezig is;
INT = eenheidsmatrix (NT x NT) met de waarde 1 voor de diagonaalelementen en 0 voor de overige elementen;
ιT = eenheidsvector met lengte T en waarde 1 voor alle elementen.

Hoewel $Q_v$ ingewikkeld lijkt is het niet meer dan een matrix die afwijkingen van groepsgemiddelden berekend: bijvoorbeeld $\tilde{x}_{it} = x_{it} - \bar{x}_{i.}$

Het gebruik van FEM voor de schattingen van $\beta$ is gerechtvaardigd omdat door de eliminatie van $u_i$ ook de correlatie van de verklarende variabelen $X_{it}$ met de storingstermen weggenomen wordt terwijl $Z_i$ in het model helemaal niet meer voorkomen.

Stap 2: Het schatten van $\sigma_u^2$

Helaas is een schatting voor $\sigma_u^2$ is lastiger te verkrijgen omdat hier schattingen van $\beta$ èn $\gamma$ voor nodig zijn. Hausman en Taylor (1981) gebruiken de volgende procedure.

Bereken met behulp van de FEM schattingen voor $\beta$ (deze noemen we voor de duidelijkheid $\beta_W$) de groepsgemiddelden van de residuen:

$\hat{d}_i = Y_{i.} - \hat{\beta}_W X_{i.}$

Merk op dat de geschatte $d_i$, $\hat{d}_i$ informatie bevat over het effecten van de $Z_i$ maar ook de $u_i$ op $Y_{it}$. We zouden nu γ kunnen schatten door de volgende regressie uit te voeren:

$$\hat{d}_i = \gamma Z_i + \eta_i$$

Echter, we hebben opgemerkt dat $\hat{d}_i$ onder andere bepaald wordt door $u_i$ welk effect nu in de storingsterm $\eta_i$ is opgenomen (want niet expliciet in het model opgenomen). Met andere woorden, $Z_i$ is gecorreleerd met $\eta_i$ en OLS mag niet gebruikt worden. In plaats daarvan moeten we 2SLS gebruiken met $[X_{1it}, Z_{1i}]$ als instrumenten (zie voor een beknopte uitleg van 2SLS bijlage 1).

Met de schattingen van γ en β kan de totale variantie van de storingsterm berekend worden. De totale variantie $s^2$ van de componenten $u_i$ en $\varepsilon_{it}$ is gelijk aan $\frac{1}{N}((Y_{i.} - \beta_W X_{i.} - \gamma_W Z_i)'(Y_{i.} - \beta_W X_{i.} - \gamma_W Z_i))$

We weten ook dat $s^2 = \sigma_u^2 + \frac{1}{T}\sigma_\varepsilon^2$ (zie bijlage 2). Met $\sigma_\varepsilon^2$ reeds bekend kan $\sigma_u^2$ berekend worden als:

$$\sigma_u^2 = s^2 - \frac{1}{T}\sigma_\varepsilon^2$$

Met behulp van de schattingen voor $\sigma_u^2$ en $\sigma_\varepsilon^2$ kan nu de transformatie in vergelijking () uit worden gevoerd. Deze transformatie kan worden vereenvoudigd doordat $\Omega^{-\frac{1}{2}} = I_{TN} - (1-\theta)P_V$ met als eindresultaat:

$$Y_{it} - (1-\theta)Y_{i.} = [X_{it} - (1-\theta)X_{i.}]\beta + \theta Z_i\gamma + \theta u_i + [\varepsilon_{it} - (1-\theta)\varepsilon_{i.}]$$

waarbij $\theta = \left[\frac{\sigma_\varepsilon^2}{\sigma_\varepsilon^2 + T\sigma_u^2}\right]^{\frac{1}{2}}$ hetgeen gewoon een getal is.

Vergelijking (2.7) kan nu geschat worden met IV waarbij $Q_v X_1, Q_v X_2, X_1$ *en* $Z_1$ als instrumenten dienen. We zien hier dat de exogene variabelen in $X_1$ op meerdere manieren als instrument kunnen dienen: niet alleen de variabelen zelf maar ook de afwijkingen van de groepsgemiddelden en de gemiddelden zelf kunnen als instrument gebruikt worden. In tegenstelling tot standaard 2SLS methoden wordt in HT alleen informatie (lees: variabelen) gebruikt die in het model zelf aanwezig is.

De praktijk leert dat methoden als HT met name relevant zijn voor microdata (data over individuen, huishoudens of, zoals in het Informatienet, agrarische bedrijven). Vaak zijn we niet in staat om alle bedrijfskarakteristieken te meten terwijl op grond van de theorie verwacht kan worden dat deze karakteristieken gecorreleerd zijn met variabelen die wel gemeten zijn. Dergelijke correlaties leiden tot problemen voor het juist schatten van modelparameters.

In deze paragraaf is gedemonstreerd hoe met behulp van HT parameters geschat kunnen worden van variabelen die constant zijn over de tijd en waarbij er sprake is van correlatie van variabelen met het bedrijfseffect. Met standaardmethoden zoals FEM en REM kan dit niet.

In bijlage 2 is een voorbeeld van HT uitgewerkt waarin alle bovengenoemde stappen uitgevoerd worden.

# 3. Toepassingsmogelijkheden

## 3.1 Inleiding

In dit hoofdstuk wordt beschreven hoe je zelf een paneldata-analyse kunt uitvoeren. Eerst wordt beschreven welke paneldatabestanden het LEI heeft, en hoe je daar mee om kunt (of moet) gaan om een goede paneldataschatting te maken. Daarna wordt aan de hand van een voorbeeld het hele traject van het opvragen van de data tot aan het beoordelen van de schattingsresultaten besproken. In het voorbeeld (paragraaf 3.3, 3.4 en 3.5) wordt de toegediende hoeveelheid stikstof uit kunstmest verklaard; deze schatting is gebruikt voor het stofstromen model (Leneman et al., 2000) [1]. Daarna worden met behulp van een reeds bestaande dataset de schattingen van een klassiek artikel van Mundlak (1961) uitgevoerd en besproken in paragraaf 3.6.

Om de commando's beter van de tekst te kunnen onderscheiden zijn de commando's die ingetypt moeten worden **vet** weergegeven. Opties en andere tekst die op het scherm verschijnt is onderstreept weergegeven.

Om zelf de paneldataschattingen uit te kunnen voeren heb je het software pakket LIMDEP nodig. Als dat nog niet op je PC staat kun je de installatie van Limdep aanvragen bij CIVA Service and Support (mail naar ServiceandSupport@lei.wag-ur.nl). Ook heb je het programmaatje PEBFWIN nodig; je kan een shortcut maken van N:\Appl\Ebfwin\Pebfwin.exe. Alle PC-files die in deze paragraaf als voorbeeld worden besproken zijn te vinden op de n-schijf; onder n:\Afd\Landbouw\Panel de datafiles die worden gebruikt zijn opgeslagen onder n:\Afd\Landbouw\Panel\Data. Alle verwijzingen naar andere files zijn gebaseerd op de volgende procedure; kopieer de folder Panel (inclusief de subfolder Data) naar de e:\Personal directory. De programmafiles staan dan in je eigen subdirectory e:\Personal\Panel. De benodigde data staan dan in de folder Data. Je hoeft dan geen namen en verwijzingen te wijzigen.

Hetzelfde geldt voor de programmafiles op LEI5. Als je die kopieert van [REINHARDA.PANEL] naar je eigen subdirectory [.PANEL] dan hoef je verder ook niets te wijzigen aan de programma's om ze te kunnen draaien.

## 3.2 Het gebruik van de LEI-paneldatabestanden

De meest gebruikte paneldatabestanden van het LEI zijn het Informatienet en de Landbouwtelling (LBT). Beide bestanden zijn beschikbaar via de LEI5 computer. Het Informatienet is onderverdeeld in 4 onderdelen (landbouw, tuinbouw, bosbouw en visserij). Een uitgebreide beschrijving van de beschikbare data is te vinden in de handleiding

[1] Op enkele kleine punten wijkt de hier beschreven aanpak af van die indertijd gevolgd door Leneman et al. (2001). De hier gepresenteerde uitkomsten wijken hierdoor een weinig af van de voorgenoemde publicatie.

'gebruik boekhoudgegevens in BDL' (verkrijgbaar bij CIVA). Op de DATA-schijf van de LEI5 computer staan van zowel ieder onderdeel van het Informatienet als van LBT alle observaties van een jaar in één file. Je kan dat zien door het commando **DIR BKH98L** in te tikken op LEI5. **DIR BKH98T** laat het overeenkomstige bestand zien voor de tuinbouwbedrijven. Voor een paneldatabestand van Informatienet-landbouwbedrijven voor de periode 1990-1999 moeten dus data uit 10 verschillende databestanden worden opgevraagd.

Op dit moment (april 2000) is BDL (Bedrijven Databank LEI) de databank waarin deze bedrijfsgegevens zijn opgeslagen. Via een BDL-programma kunnen gegevens uit de databanken worden geselecteerd en opgevraagd. Informatie over BDL is te vinden in de 'Handleiding voor BDL' (te verkrijgen bij Civa). In dit hoofdstuk worden niet de ins en outs van BDL beschreven, wel wordt een manier beschreven waarmee op een eenvoudige wijze data uit BDL kunnen worden geselecteerd en weggeschreven naar een eigen datafile. Binnen afzienbare termijn worden nieuwe data vastgelegd in ARTIS, ook ARTIS biedt de mogelijkheid om data te selecteren en weg te schrijven naar nieuwe files. (Alleen het gedeelte van deze handleiding waarin het gebruik van BDL wordt beschreven moet dan worden vervangen door een ARTIS-aanpak.)

Via parameters kan exact dezelfde BDL-opvraagfile voor verschillende jaren worden gebruikt. Deze parameters worden met **&&1** aangeduid in de algemene vraagfile voor verschillende jaren en ze worden gedefinieerd in de .BDP file (de BDL programma-file). Een voorbeeld van de opvraagfile voor een jaar is **VRKMST.VAR** (vraagfile voor kunstmest). De parameters worden gedefinieerd in de BDP-file voor de verschillende jaren; bijvoorbeeld **VRKMST97.BDP**. In de file **VRKMST.COM** worden de BDP-files voor de gewenste jaren in een keer opgeroepen.

Via het bedrijfsnummer kunnen de gegevens van verschillende jaren tot hetzelfde bedrijf worden herleid. Het bedrijfsnummer van een bedrijf kan veranderen in de loop der tijd, bijvoorbeeld als gevolg van bedrijfsovername door de zoon. Het is belangrijk om per onderzoek na te gaan of je te maken hebt (of wilt hebben) met 1 bedrijf (met een bijvoorbeeld verschillende bedrijfshoofden) of met 2 verschillende bedrijven. In het laatste geval kun je de benodigde jaren onafhankelijk van elkaar opvragen in BDL. Als je de verschillende bedrijfsnummers in het Informatienet wilt herleiden tot 1 bedrijf moet het bestand in BDL worden gekoppeld aan het voorafgaande jaar. Een bedrijf met twee verschillende nummers komt sporadisch voor in het Informatienet. In de Landbouwtelling ligt dit anders, het komt daar veel voor. Als je een bedrijf met twee verschillende bedrijfsnummers als 1 bedrijf wilt specificeren in de paneldataschatting moet je beide bedrijfsnummers achterhalen. Hiervoor is het nodig om alle jaren op te vragen in BDL om te kunnen achterhalen of het registratienummer is gewijzigd. (Deze optie is in **KOPPEL_BINxx.BDP** [1] en **KOPPEL_LBTxx.BDP** uitgewerkt voor de BDL opvraagfiles. Het SPSSX programma om de nummers te herleiden tot 1 nummer per bedrijf is in **KOPPEL_BIN.SPS** en **KOPPEL_LBT.SPS**.)

Een van de problemen die kunnen opduiken bij het combineren van verschillende jaren is dat enkele variabelen maar voor een beperkte periode in het Informatienet zijn opgenomen (bijvoorbeeld vanwege het invoeren van nieuwe definities). Het niet gedefini-

[1] Voor de 'xx' moeten de laatste twee cijfers van het jaar worden ingevoerd; 95,96 of 97.

eerd zijn van een variabele in een bepaald jaar leidt tot een foutmelding in BDL. Een ander probleem dat voor kan komen in het Informatienet is dat je bij de waarde nul voor een variabele niet zeker weet of er echt nul is waargenomen of dat de variabele niet is opgenomen voor het bedrijf. Dit probleem doet zich regelmatig voor bij de hoeveelheidsgegevens. Van een bedrijf is bijvoorbeeld wel de waarde van de aangekochte stikstofkunstmest bekend maar niet de hoeveelheid in kilogram. In dit geval (waarde groter dan 0, hoeveelheid gelijk aan 0) is het beter om de nulwaarneming van de hoeveelheid kunstmest te vervangen door een missing value in SPSS; **recode var_naam (0=missing)**. Zo voorkom je dat je deelt door 0 (om bijvoorbeeld de gemiddelde prijs uit te rekenen) of dat je een logaritme neemt van 0. Zo kan ook aan de hand van informatie van aansluitende jaren van hetzelfde bedrijf of met behulp van gerelateerde variabelen worden nagegaan wat er aan de hand is (zie ook paragraaf 3.3). Als er echt 0 waargenomen is in het Informatienet en de variabele moet worden gebruikt in een (paneldata) regressieanalyse dan is het nodig de schatting wat anders aan te pakken; zie bijvoorbeeld Battese (1997).

## 3.3 Het generen van een paneldataset (vraagfunctie kunstmest)

Voor het genereren van een paneldataset zijn drie (groepen van) handelingen noodzakelijk (1) het selecteren van variabelen uit het Informatienet of LBT op de LEI5 computer (2) het overzetten van de geselecteerde data naar de PC (3) het combineren van datasets en bewerken van data in SPSS. Eerst wordt de vraag naar kunstmest op maïsland geschat (in paragraaf 3.5). Alle handelingen vanaf het opvragen van de BDL-variabelen worden hierbij behandeld in paragraaf 3.3 en 3.4). Daarna wordt de productiefunctie uit het artikel van Mundlak (1961) behandeld in paragraaf 3.6. In dit laatste geval wordt reeds uitgegaan van een datafile die door LIMDEP kan worden ingelezen.

1. Met behulp van BDL worden de data op de LEI5 computer uit het Informatienet (of Landbouwtelling) gehaald; in dit voorbeeld gaan we uit van Informatienet-gegevens waaraan LBT-gegevens worden gekoppeld [1]. Een praktische optie is om de data binair te laten wegschrijven naar de disk$temp schijf (aan te roepen met het commando **temp$$dir**). Een paneldataset bestaat per definitie uit data van verschillende jaren, deze kunnen het eenvoudigst via het gebruik van een parameter voor het betreffende jaar worden opgevraagd in BDL (zie sectie 3.2 en voor meer details het BDL-handboek). Per jaar wordt dan een dataset weggeschreven naar de disk$temp (via de logical temp$$dir). Dit wegschrijven kan naar een bestand in geformatteerde vorm (je moet dan een format specificeren), of in een alleen voor de computer leesbare vorm (binair). In de voorbeelden worden de data door BDL binair (real*4) weggeschreven.

Met het commando **CREATE/DIR [.PANEL]** wordt een subdirectory [.PANEL] aangemaakt op LEI5. Als je in deze subdirectory staat kun je met het commando **COPY [REINHARDA.PANEL]*.* []** alle relevante files naar je eigen gebied kopiëren.

[1] Het is handig om in BDL alle variabelen waarvan je vermoedt dat je ze nodig hebt in 1 keer op te vragen. Je kunt beter te veel dan te weinig gegevens uit BDL halen, dat voorkomt dat de eerste twee stappen meermalen moeten worden doorlopen.

De file met de BDL-opvraagcodes is na het kopieren aanwezig op je eigen subdirectory **[USERNAME.PANEL]**. (De orginelen zijn te vinden bij [reinharda.panel])
De BDL-file met opvraag codes is: **VRKMST.VAR** [1]
Deze wordt in de BDL programma-file voorzien van de benodigde parameter voor the jaar, bijvoorbeeld:
**VRKMST91.BDP**
De hele set van BDL programmafiles voor de verschillende jaren is samengevoegd in een file voor eventuele batch verwerking, in
**VRKMST.COM**
Na het draaien van deze file **@VRKMST.COM** staan de 7 datafiles voor de verschillende jaren op de temp$$dir-schijf; [tempdir.username].

2. Deze data moeten worden overgezet naar de PC. In het geval van binaire data moeten deze eerst worden geconverteerd op de LEI5 (met behulp van het commando **ebf file.naam/to_pcd**; bijvoorbeeld **ebf vrkmst91.bin/to_pcd**). De commando's om de door BDL aangemaakte files te converteren zijn samengevoegd in de file: **CONVSPS.COM**

De volgende stap is de geconverteerde files met de geselecteerde gegevens van het Informatienet of LBT per jaar te kopieren van LEI5 naar de PC. Voor binaire files is daar het programmaatje **Pefbwin** voor ontwikkeld. De files worden met het PC-programma **Pebfwin** getransporteerd naar de PC (voor paneldata gebruik bij OutPutFormat de optie binair (SPSS, Fortran)).

3. De datafiles voor de verschillende jaren worden ingelezen in SPSS [2], opgeslagen in SPSS-datafiles en met het commando **add files** samengevoegd tot 1 dataset die meer jaren bevat; de betreffende SPSS syntax file heet **VRKMST .SPS** (zie ook bijlage 2). Macro's zijn handig om de data die worden ingelezen te benoemen (je hoeft ze maar 1 keer op te geven, en kan ze dus gemakkelijker wijzigen). Als je teveel data hebt verkregen uit BDL kun je via een macro ook alleen dat deel van de data dat je echt nodig hebt bewaren voor het **add files** commando, zodat de bewerkingen sneller worden uitgevoerd.

Eventueel worden de data eerst nog bewerkt in SPSS.

- In SPSS kun je eenvoudig nagaan of de waarden van variabelen wel geloofwaardig zijn. Je kunt met behulp van het commando **descriptives var=varnaam.** het gemiddelde, minimum, maximum en standaard deviatie van een variabele uit de dataset opvragen. Vooral de minimum en maximum waarde geven aan of een variabele nadere analyse behoeft; bijvoorbeeld een bedrijf waar de arbeidsinzet 0 is.
  Zo worden in het voorbeeld observaties verwijderd waar meer dan 5.000 kg N-kunstmest per hectare wordt toegediend (dat is namelijk een onrealistische hoeveelheid).
- Prijsindices kunnen worden berekend.
- In SPSS kunnen eenvoudig variabelen worden gedefleerd met prijsindices.

---

[1] Van alle files die worden genoemd in de tekst is een uitdraai opgenomen als bijlage. Alle files zijn voorzien van een beschrijving van de gebruikte (specifieke) commando's.
[2] Deze datafiles zijn ook op de N-schijf aanwezig, je hoeft dus niet perse de voorgaande stappen zelf te doorlopen om dit onderdeel uit te kunnen voeren.

Als deze dataset in SPSS gereed is dient deze omgezet te worden in een WK1 (lotusworksheet) bestand format in de SPSS-data editor. Dit WK1-bestand vergemakkelijkt het inlezen van de data in het pakket LIMDEP. Selecteer in de SPSS-data-editor:
File/Save as:
WK1-bestand.
Dit WK1-bestand (**VRKMST .WK1**) kan eenvoudig worden ingelezen in Limdep, en heeft als voordeel dat de variabele namen niet opnieuw hoeven te worden gespecificeerd. (LIMDEP kan niet latere (windows) versies van Lotus-worksheets en Excel-worksheets inlezen).

## 3.4 De paneldataschatting

*Opstarten van LIMDEP*

LIMDEP wordt eerst onder windows opgestart. (Via Start, Programs, Limdep)
Je komt in het LIMDEP scherm. Voor het maken van een nieuw document selecteer je File, New, Text/Command Document. Daarna kom je in een leeg scherm waar je je commando's kan invoeren.

Het gemakkelijkst is echter om de hier gebruikte voorbeeld-file **VRKMST .LIM** te openen (is al gekopieerd naar je E:\Personal\Panel folder) en deze file te wijzigen. Aan de hand van deze **VRKMST .LIM** file worden de belangrijkste commando's uitgelegd.

Een harde beperking van LIMDEP is dat het maar 20.000 bedrijven aan kan en maximaal 100 regressoren. Je kan dus geen paneldataschatting maken van de gehele landbouwtelling. Als LIMDEP crashed of vreemde foutmeldingen geeft, kun je via Stijn Reinhard een nieuwe LIMDEP.EXE file ophalen.

*Help functies in LIMDEP*

Op het LEI zijn enkele handleidingen van LIMDEP aanwezig (CIVA; Service and Support heeft een overzicht van de mensen die een handleiding hebben). Om zonder handleiding de noodzakelijke informatie te kunnen vinden wordt in de volgende alinea de help-functie van LIMDEP uitgebreid beschreven.

Help, Help topics, Index geeft de inhoudsopgave van de handleiding. Bijvoorbeeld paragraaf 17.2 geeft de enkele aandachtpunten voor paneldataschattingen weer. Help, Help topics, Contents geeft de commando's weer in alfabetische volgorde, bijvoorbeeld REGR. Help, Help topics, Find geeft de mogelijkheid om te zoeken op trefwoorden. Help, Limdep website schakelt je meteen door naar de website van Limdep. Onder Documentation, Manual vind je een overzicht van de handleiding.

*Geheugenruimte uitbreiden*

Voor grote datasets en modellen is het nodig de beschikbare geheugenruimte voor LIMDEP te vergroten. (In windows heeft LIMDEP een grens aan de geheugenruimte omdat het anders alle geheugenruimte inpikt). Dit doe je via Project, Settings, Data Area.

Standaard is the number of cells 200.000. Dit kan het makkelijkst worden aangepast tot 2.000.000 door er een nul achter te typen. Er verschijnt een melding dat de dataset opnieuw moet worden ingelezen. Als je **VRKMST .LIM** kopieert en aanpast heb je standaard de 2.000.000 cellen.

*LIMDEP-commando's*

LIMDEP werkt het eenvoudigst door middel van een commandofile die je via selecties of per regel kunt runnen. Run, Run Selection of Run, Run Line. Dit is overeenkomstig het runnen van SPSSX via een syntax-file. De structuur van de LIMDEP-commandoregels is eenvoudig. Een commando wordt beschreven door eerst de commando naam te geven daarna een punt komma ';' en darna een specificatie (afhankelijk van het commando). Alle commando's worden afgesloten met een dollarteken '$'.

*LIMDEP-output*

De LIMDEP-output wordt naar een afzonderlijk window uitgevoerd. De inhoud van dit window kan worden geprint en bewaard als file. In tegenstelling tot SPSS wordt er geen nieuwe outputfile aangemaakt als je de eerste output-file van een sessie verwijdert, LIMDEP voegt de nieuwe output toch gewoon toe aan de output van het eerste deel van de sessie. Een manier om delen van de output-file te printen is het kopieren van de outputfile naar een WORD-bestand. Met lettertype 'courier' en lettergrootte '9'. ziet het er net zo uit als de output file (zie Limdep output in paragraaf 3.5 en 3.6). Via Window kun je van het input scherm naar het output scherm (of springen van een input-file naar een andere).

Extra mogelijkheden:
- je kan commando's en variabelen 'browsen' via het Insert Name balkje juist boven het commando scherm;
- je kunt je werk bewaren in LIMDEP via het commando **SAVE**, via het commando **LOAD** kun je dan snel je werk weer laden in LIMDEP.

*De belangrijkste commando's zijn:*
Het inlezen van een worksheet gebeurt met het volgende commando. Ook de namen van de variabelen worden overgenomen uit het worksheet.
**read ; file=e:\dataext\panel\VRKMST .wk1 ; Format=WKS; names$**
Het inlezen van een WK1-datafile kost veel tijd. Het is daarom handig om deze datafile via het **SAVE** commando door LIMDEP te laten opslaan. Een volgende keer kunnen de data dan snel via **LOAD** worden ingelezen (zie ook **VRKMST .LIM**).
Voor het definiëren van nieuwe variabelen is het create commando nodig.

**create ; newvar=functie(oldvar)$**
**create ; NK_MAISH=NK_MAIS/HAMAIS$**

Indien nodig kunnen de relevante observaties worden geselecteerd. Bijvoorbeeld als er geen waarneming is voor de prijs van kunstmest op een bedrijf (in SPSS gecodeerd als -

999) dan moeten deze waarnemingen worden verwijderd uit de dataset voor er geschat wordt.

**reject; N_MESTP <0$**

Een OLS-schatting gaat als volgt. **lhs** duidt de left hand side variabele (de te verklaren variabele) en **rhs** geeft de right hand side variabelen weer (de onafhankelijke variabelen). Met 'one' wordt het intercept (de constante) gespecificeerd (zonder intercept leg je op dat de functie door de oorsprong gaat).
**regr; lhs=dependent; rhs=one,var1,var2 $**

Het bovenstaande commando kan eenvoudig worden uitgebreid tot een paneldataschatting door het uit te breiden met het woord panel. Standaard worden dan zowel FEM als REM berekend. Als slechts een van beide nodig is, bijvoorbeeld alleen REM omdat er onveranderlijke bedrijfsspecifieke variabelen worden opgenomen dan moet **random** (of **fixed**) worden toegevoegd. We beginnen met een one-way effect model. Limdep moet de verschillende waarnemingen van hetzelfde bedrijf kunnen herkennen, hiertoe moet een voor ieder bedrijf unieke nummer worden toegekend. In ons geval is dat het bedrijfsnummer. Dit wordt aangegeven met **str=num** (de bedrijven moeten zijn gesorteerd op bedrijfsnummer, dit is in SPSS al geschied).

**regr; lhs=dependent; rhs=one,var1,var2; str=num; panel$**

Voor een two-way effect model moet ook de variabele worden aangegeven, die perioden weergeeft. In ons geval is dat de variabele jaar. **;period=jaar** wordt toegevoegd aan het commando.

**regr; lhs=dependent; rhs=one,var1,var2; str=num; period=jaar; panel; random$**

### 3.5 Interpretatie van de resultaten

Aan de hand van de outputfile van het voorbeeld VRKMST .LIM worden de gevonden resultaten besproken. In eerste instantie schatten we een model dat ook met cross-sectiedata kan worden geschat op drie verschillende manieren OLS, FEM en REM (allemaal met 1 Limdep commando). Het te schatten model voor de periode 1991-1997 is:

$$NK\_MAISH_{it} = \alpha + \beta_1 ND\_MAISH_{it} + \beta_2 HATOT_{it} + \beta_3 N\_MSTPI_{it} + u_i + \varepsilon_{it} \qquad (3.1)$$

met:
*NK_MAISH* = hoeveelheid stikstof in kunstmest in kilogram per hectare;
*ND_MAISH* = hoeveelheid stikstof in dierlijke mest in kilogram per hectare;
*HATOT* = totaal areaal van het bedrijf;
*N_MSTPI* = de door het bedrijf betaalde prijs voor stikstofkunstmest.

Eerst geeft LIMDEP de kleinste kwadraten schatting uitkomsten weer (OLS Without Group Dummy Variables). De samenvatting in het kader geeft aan wat de te verklaren variabele is (NK_MAISH), en geeft het aantal waarnemingen (Observations=2965) en het aantal parameters dat is geschat (inclusief de constante). De $R^2$ is gegeven als (R-squared=.131529) [1]. Ook de adjusted $R^2$ is afgedrukt, deze corrigeert de $R^2$ voor het aantal verklarende variabelen [2]. (Door elke verklarende variabele die aan de vergelijking wordt toegevoegd stijgt namelijk de $R^2$ ook als de variabele geen significante bijdrage levert aan het model.) De adjusted $R^2$ is dan ook altijd kleiner dan de standaard $R^2$ score (adjusted R-squared=.13065). De adjusted $R^2$ is niet groot maar ook weer niet dramatisch voor een crosssectiemodel. Onder in het kader wordt de variatie in NK_MAISH uiteengerafeld in variantie tussen de bedrijven; berekend als de som van de kwadraten van het verschil tussen het gemiddelde per bedrijf en het gemiddelde van het hele panel (Between 0.681607E+07). Daarna volgt het residu, meestal trouwens 'within' genoemd, dit is berekend als de som van de kwadraten van het verschil tussen de waarneming en het gemiddelde per bedrijf (Residual 0.192121E+07). Tenslotte volgt de totale variantie; berekend als de som van de kwadraten van het verschil tussen de waarneming en het gemiddelde van het panel. Dit is gelijk aan de som van de 'between' en de 'within'. Een Anova analyse van NK_MAISH met als factor het bedrijfsnummer levert hetzelfde resultaat (kun je zelf in SPSS uitvoeren).

```
+----------------------------------------------------------------------+
| OLS Without Group Dummy Variables                                    |
| Ordinary    least squares regression    Weighting variable = none    |
| Dep. var. = NK_MAISH Mean=   65.73435816    , S.D.=   54.29366748    |
| Model size: Observations =    2965, Parameters =    4, Deg.Fr.=   2961 |
| Residuals:  Sum of squares= 7588077.225     , Std.Dev.=       50.62286 |
| Fit:        R-squared=   .131529, Adjusted R-squared =          .13065 |
| Model test: F[  3,   2961] =  149.48,    Prob value =          .00000 |
| Diagnostic: Log-L = -15841.0070, Restricted(b=0) Log-L =  -16050.0712 |
|             LogAmemiyaPrCrt.=    7.850, Akaike Info. Crt.=     10.688 |
| Panel Data Analysis of NK_MAISH   [ONE way]                          |
|          Unconditional ANOVA (No regressors)                         |
| Source      Variation        Deg. Free.      Mean Square             |
| Between       .681607E+07          995.          6850.32             |
| Residual      .192121E+07         1969.          975.731             |
| Total         .873729E+07         2964.          2947.80             |
+----------------------------------------------------------------------+
+---------+-------------+----------------+--------+---------+----------+
|Variable | Coefficient | Standard Error |b/St.Er.|P[|Z|>z] | Mean of X|
+---------+-------------+----------------+--------+---------+----------+
 ND_MAISH -.1416337546      .93818161E-02  -15.097    .0000  160.90555
 HATOT     .3215585850      .38755166E-01    8.297    .0000  32.765798
 N_MSTPI  -32.09875053      3.8776452       -8.278    .0000  1.0666190
 Constant  112.2250293      5.0492422       22.226    .0000
```

Daaronder staan de parameter schattingen. De coëfficiënt wordt gevolgd door de standard error. Daarna volgt de T-waarde (de coëfficiënt gedeeld door de standaard fout) in Limdep aangeduidt met b/St.Er. en de waarschijnlijkheid (kans p[Z|>z]) dat we deze coëfficiënt zouden aantreffen terwijl de werkelijke waarde 0 is. Als laatste is het gemiddelde

[1] De $R^2$ is een maatstaf voor de juistheid van het model. De $R^2$ geeft aan welk deel van de totale variantie van de te verklaren variabele wordt verklaard door de onafhankelijke variabelen.
[2] Zie Greene (2000:240) voor de definitie en berekening van de adjusted $R^2$.

van de variabele gegeven. Een T-waarde groter dan 1,96 (of kleiner dan -1,96) gaat samen met een waarschijnlijkheid kleiner dan 0.05 en houdt in dat de verklarende variabele significant bijdraagt aan de verklaring van de hoeveelheid stikstof kunstmest toegediend per hectare [1].

De tekens komen overeen met de verwachte resultaten. Een grotere toediening van dierlijke mest leidt tot een kleinere kunstmestgift. Een hogere prijs van kunstmest zal ertoe leiden dat het zuiniger wordt toegediend. Er zijn 2.965 waarnemingen in de dataset, er zijn vier onafhankelijke variabelen (inclusief de constante), dat houdt in dat er 2.961 vrijheidsgraden over zijn. Alle parameters zijn significant afwijkend van 0.

Na de OLS-schattingsresultaten volgt het FEM-kader en de geschatte FEM-coëfficiënten. De fixed effects paneldata-uitkomsten worden aangeduid als Least Squares with Group Dummy Variables. De Group Dummy Variables slaat op de dummy per bedrijf. De samenvatting in het kader geeft aan wat de te verklaren variabele is (NK_MAISH), geeft het aantal waarnemingen (Observations=2965). Het aantal Parameters groter is 999 (is het aantal parameters groter dan staan er sterretjes weergegeven (*Parameters=****). Het aantal parameters is zo groot omdat ieder bedrijf zijn eigen bedrijfsspecifieke parameter heeft (er wordt geen constante gespecificeerd in het FE-model).

```
+----------------------------------------------------------------------+
| Least Squares with Group Dummy Variables                             |
| Ordinary     least squares regression     Weighting variable = none  |
| Dep. var. = NK_MAISH Mean=   65.73435816     , S.D.=   54.29366748   |
| Model size: Observations =    2965, Parameters = 999, Deg.Fr.=   1966 |
| Residuals:  Sum of squares= 1883545.303     , Std.Dev.=      30.95254 |
| Fit:        R-squared=   .784424, Adjusted R-squared =        .67499 |
| Model test: F[998,   1966] =    7.17,    Prob value =          .00000 |
| Diagnostic: Log-L = -13775.2583, Restricted(b=0) Log-L =  -16050.0712 |
|             LogAmemiyaPrCrt.=    7.155, Akaike Info. Crt.=      9.966 |
| Estd. Autocorrelation of e(i,t)    -.110243                          |
+----------------------------------------------------------------------+
+---------+--------------+----------------+--------+---------+----------+
|Variable | Coefficient  | Standard Error |b/St.Er.|P[|Z|>z] | Mean of X|
+---------+--------------+----------------+--------+---------+----------+
 ND_MAISH -.5083873409E-01  .82891270E-02   -6.133    .0000  160.90555
 HATOT     .5380220175E-01  .15390205         .350    .7266  32.765798
 N_MSTPI  2.905161318      3.7494239          .775    .4384  1.0666190
```

Wat opvalt is dat het aantal parameters is toegenomen tot 999 (alle bedrijven hebben een eigen bedrijfsdummy) en daarom is het aantal vrijheidsgraden afgenomen tot 1966. De $R^2$ is gelijk aan 0,78, dit is enorm veel groter dan de $R^2$ van de OLS-schatting (0,13), dit wordt veroorzaakt doordat voor ieder bedrijf een apart bedrijfseffect wordt geschat. Door het grote aantal parameters is de adjusted $R^2$ ook een stuk kleiner dan de standaard $R^2$, namelijk 0,675. Verder is het opvallend dat de parameters van HATOT en N_MSTPI nu niet significant van 0 afwijken. Het teken van N_MSTPI is nu zelfs positief (doch niet signifi-

[1] T-waardes groter dan kritieke T-waarde (of kleiner als de T-waarde negatief is) geven aan de variabele een bijdrage levert aan de verklaring van de te verklaren variabele. De kritieke T-waarde hangt af van het aantal waarnemingen, als er meer dan 100 waarnemingen zijn, is de kritieke T-waarde 1,96, zie ook Greene (2000:248).

cant), terwijl we een negatieve relatie tussen de kunstmestprijs en de toegediende hoeveelheid verwachten.

Het volgende kader geeft een overzicht van de testen die mogelijk zijn aan de hand van de bovenvermelde schattingresultaten.

| | |
|---|---|
| (1) Constant term only | Model met alleen constante |
| (2) Group effects only | Model met bedrijfseffecten |
| (3) X variables only | OLS-schattingen (zonder de constante) |
| (4) X and group effects | Het complete Fixed Effects Model |

Een model met alleen constante heeft natuurlijk een $R^2$=0, omdat er geen variantie wordt verklaard. Alleen de bedrijfsspecifieke dummies leiden al tot een $R^2$=0,780. Er worden verschillende hypotheses getest met behulp van de Likelihood Ratio Test. De testwaarde wordt berekend als 2 keer het verschil tussen de loglikelihood van het gerestricteerde model (het model met de minste verklarende variabelen van de twee) en het ongerestricteerde model [1]. Als de testwaarde groter is dan de kritieke waarde (Chi2 verdeeld) dan wordt het gerestricteerde model verworpen.

De meest wezenlijke test is (4) vs (3), dit is de laatste regel van het volgende kader. Hier worden FEM en OLS met elkaar vergeleken (OLS is het gerestricteerde model met 3 verklarende variabelen; FEM heeft er 998). De probability geeft de kans aan dat de gevonden testwaarde bij toeval wordt bereikt. Deze kans is kleiner dan 0,05, dit duidt op het verwerpen van OLS.

Test Statistics for the Classical Model

| Model | Log-Likelihood | Sum of Squares | R-squared |
|---|---|---|---|
| (1) Constant term only | -16050.07120 | .8737286102D+07 | .0000000 |
| (2) Group effects only | -13804.61426 | .1921214444D+07 | .7801131 |
| (3) X - variables only | -15841.00698 | .7588077225D+07 | .1315293 |
| (4) X and group effects | -13775.25823 | .1883545303D+07 | .7844244 |

Hypothesis Tests

| | Likelihood Ratio Test | | | F Tests | | | |
|---|---|---|---|---|---|---|---|
| | Chi-squared | d.f. | Prob. | F | num. | denom. | Prob value |
| (2) vs (1) | 4490.914 | 995 | .00000 | 7.021 | 995 | 1969 | .00000 |
| (3) vs (1) | 418.128 | 3 | .00000 | 149.480 | 3 | 2961 | .00000 |
| (4) vs (1) | 4549.626 | 998 | .00000 | 7.168 | 998 | 1966 | .00000 |
| (4) vs (2) | 58.712 | 3 | .00000 | 13.106 | 3 | 1966 | .00000 |
| (4) vs (3) | 4131.498 | 995 | .00000 | 5.984 | 995 | 1966 | .00000 |

*Het random effects model*

Eerst wordt de geschatte variante van de standaard storingterm gegeven Var[e] = .958060D+03. Daarna wordt de variantie van het bedrijfseffect gegeven Var[u] = .162687D+04. De daarop volgende correlatie, in tekstboeken ook weergegeven als $\rho$, is berekend als Var[u]/(Var[u]+Var[e]). De storingstermen in verschillende perioden voor een gegeven individu, i, zijn gecorreleerd, omdat deze storingstermen de term ui gemeenschappelijk hebben (zie sectie 2.4.1).

[1] Zie Greene (2000:152 en 392).

Een hoge waarde voor de Lagrange Multiplier test [1] geeft aan dat een paneldatamodel de voorkeur verdient boven het OLS-model. Aan de hand van de probability value kun je snel zien of de Lagrange Multiplier test groot is. Een Prob value kleiner dan 0,05 duidt al op een voorkeur voor paneldata-aanpak. In dit geval geeft het testresultaat aan dat een FEM- of REM-model de voorkeur boven OLS heeft. De Hausman-test geeft aan of FEM dan wel REM-resultaten de voorkeur verdienen. In dit geval is die erg groot 94.64; dit valt af te leiden uit de prob value die gelijk is aan 0,0000. Dit testresultaat wijst op een FEM. We hebben echter in het Informatienet niet te maken met een tijdreeks die naar oneindig gaat, zodat het testresultaat niet alles zegt.

Aan de hand van de FEM-schattingsresultaten (de residuen) worden de varianties berekend (zie het kader hierboven). Deze worden weer gebruikt voor een feasible GLS-schatting van het random effects model. De uitkomsten van de GLS-schatting staan onder in het kader vermeld. Alle parameters zij significant. De REM-uitkomsten zijn veel geloofwaardiger dan de FEM-uitkomsten, omdat alle door REM geschatte parameters significant bijdragen aan de verklaring van de kunstmest gift per hectare dit in tegenstelling tot de FEM uitkomsten. Verder is het teken van de variabele 'kunstmestprijs' (niet significant) positief in het FEM, deze positieve geschatte parameter in niet in overeenstemming met de economische theorie. Daarom leggen we in dit geval de Hausman-test uitkomst naast ons neer en geven we de voorkeur aan het REM (ofschoon de REM resultaten een $R^2$ hebben van 0,13, en die van FEM groter is. Deze grote $R^2$ wordt veroorzaakt door het grote aantal verklarende variabelen; de bedrijfsdummies).

```
+----------------------------------------------------+
| Random Effects Model: v(i,t) = e(i,t) + u(i)       |
| Estimates:  Var[e]                 =    .958060D+03|
|             Var[u]                 =    .162687D+04|
|             Corr[v(i,t),v(i,s)] =    .629367       |
| Lagrange Multiplier Test vs. Model (3) = 1525.29   |
| ( 1 df, prob value =  .000000)                     |
| (High values of LM favor FEM/REM over CR model.)   |
| Fixed vs. Random Effects (Hausman)     =   94.64   |
| ( 3 df, prob value =  .000000)                     |
| (High (low) values of H favor FEM (REM).)          |
| Reestimated using GLS coefficients:                |
| Estimates:  Var[e]                 =    .969839D+03|
|             Var[u]                 =    .175108D+04|
|             Sum of Squares             .779076D+07 |
|             R-squared                  .131529D+00 |
+----------------------------------------------------+
+---------+-------------+----------------+--------+---------+----------+
|Variable | Coefficient | Standard Error |b/St.Er.|P[|Z|>z] | Mean of X|
+---------+-------------+----------------+--------+---------+----------+
 ND_MAISH -.7858856550E-01  .74776291E-02  -10.510   .0000  160.90555
 HATOT     .3944942715      .55309387E-01    7.133   .0000  32.765798
 N_MSTPI  -8.785236116      3.3241179       -2.643   .0082  1.0666190
 Constant  75.10576475      4.6947706       15.998   .0000
```

Als je de relatie tussen de hoeveelheid toegediende dierlijke mest en de hoeveelheid kunstmest wilt weten maakt het nogal uit of je de OLS-uitkomsten gebruikt of de REM-

[1] Limdep presenteert de Breusch and Pagan's Lagrange multiplier statistic. Deze test de $H_0$ dat de variantie van het bedrijfseffect of jaareffect gelijk is aan 0 ($\sigma^2_u$=0), dat is gelijk aan OLS.

uitkomsten. Volgens REM leidt een extra kilogram stikstof uit dierlijke mest tot een reductie van 0,08 kg kunstmest. Volgens OLS is de afname 0,142 kg.

*Intermezzo*

We kunnen natuurlijk ook laten zien dat het FEM-model gelijk is aan een OLS-schatting met bedrijfsdummy variabelen. Daartoe schatten we het bovenstaande model nogmaals voor de eerste honderd bedrijven met FEM en met OLS waarin naast de verklarende variabele ook voor ieder bedrijf een aparte dummy is gespecificeerd (zonder de constante). De SPSS-file en LIMDEP-file evenals de uitkomsten zijn in bijlage 5 opgenomen. Uit de informatie uit het kader (zie bijlage 5) blijkt al dat beide schattingen identiek zijn (de residual sum of squares wijkt een weinig af, dit wordt veroorzaakt door afrondingsverschillen). De OLS met dummies levert voor alle bedrijven meteen een schatting van het bedrijfseffect op.

*Nadere specificatie model*

Nu gaan we het model verder uitbouwen tot een model dat echt gebruik maakt van de paneldatastructuur. We nemen aan dat onder invloed van het beleid de hoeveelheid stikstof die wordt toegediend op maïsland zal afnemen in de loop der tijd.

Voor het stofstromenmodel is het bijvoorbeeld ook van belang dat de relatie tussen kunstmestgift en enkele niet veranderende bedrijfskenmerken (grondsoort, bedrijfstype) wordt bepaald. Deze relatie kan alleen via een REM worden bepaald, vandaar dat we alleen REM schatten. Het model ziet er dan als volgt uit.

$$NK\_MAISH_{it} = \alpha + \beta_1 ND\_MAISH_{it} + \beta_2 DKLEI_{it} + \beta_3 DAKK + \beta_4 HATOT_{it} + \beta_5 N\_MSTPI_{it} + \beta_6 TREND + u_i + \varepsilon_{it} \quad (3.2)$$

met:

| | |
|---|---|
| *DKLEI* | dummy die de belangrijkste grondsoort weergeeft (klei=1, anders=0); |
| *DAKK* | dummy die het bedrijfstype weergeeft (akkerbouw=1, anders=0); |
| *TREND* | trendvariabele (JAAR-1990). |

```
+----------------------------------------------------------------------+
| OLS Without Group Dummy Variables                                    |
| Ordinary     least squares regression    Weighting variable = none   |
| Dep. var. = NK_MAISH Mean=    65.73435816    , S.D.=    54.29366748   |
| Model size: Observations =     2965, Parameters =    7, Deg.Fr.=    2958 |
| Residuals:  Sum of squares= 6723081.579     , Std.Dev.=       47.67439 |
| Fit:        R-squared=   .230530, Adjusted R-squared =          .22897 |
| Model test: F[  6,    2958] =  147.70,    Prob value =          .00000 |
| Diagnostic: Log-L = -15661.5777, Restricted(b=0) Log-L =  -16050.0712 |
|             LogAmemiyaPrCrt.=    7.731, Akaike Info. Crt.=     10.569 |
| Panel Data Analysis of NK_MAISH   [ONE way]                          |
|          Unconditional ANOVA (No regressors)                         |
| Source      Variation          Deg. Free.      Mean Square           |
| Between       .681607E+07          995.          6850.32            |
| Residual      .192121E+07         1969.           975.731           |
| Total         .873729E+07         2964.          2947.80            |
+----------------------------------------------------------------------+
```

Eerst geeft LIMDEP de OLS-uitkomsten weer (OLS Without Group Dummy Variables) De samenvatting in het kader geeft aan wat de te verklaren variable is (NK_MAISH), en geeft het aantal waarnemingen (Observations=2965) en het aantal parameters dat is geschat (inclusief de constante). De $R^2$ is gegeven als (R-squared=.230530). Ook de adjusted $R^2$ is afgedrukt (adjusted R-squared=.22897). Onder in het kader wordt de variatie in NK_MAISH uiteengerafeld in variatie tussen de bedrijven (Between 0.681607E+07) het cijfer 995 duidt op het feit dat er 995 bedrijven in het panel zitten. Daarna volgt het residu (Residual 0.192121E+07), gevolgd door de totale variatie.

Daaronder staan de parameter schattingen. De coëfficiënt wordt gevolgd door de standard error. Daarna volgt de T-waarde en de waarschijnlijkheid (kans) dat we deze coëfficiënt zouden aantreffen terwijl de werkelijke waarde 0 is. Als laatste is het gemiddelde van de variabele gegeven.

```
+---------+--------------+----------------+--------+---------+----------+
|Variable | Coefficient  | Standard Error |b/St.Er.|P[|Z|>z] | Mean of X|
+---------+--------------+----------------+--------+---------+----------+
 ND_MAISH -.1178939219      .92921694E-02  -12.687    .0000   160.90555
 DKLEI     40.88657928     2.1947998       18.629    .0000   .23069140
 DAKK      18.14095550     4.4495906        4.077    .0000   .45868465E-01
 HATOT      .1388332538    .38948748E-01    3.565    .0004   32.765798
 N_MSTPI  -21.21735113     3.7470077       -5.662    .0000   1.0666190
 TREND     -1.975482133    .46224273       -4.274    .0000   4.0563238
 Constant  100.5349074     5.2321972       19.215    .0000
```

De matrix Lastdata bevat de een database met de geschatte coëfficiënten, standaard fouten enzovoort. Als je die binnen Limdep aanklikt, wordt deze via Limdep data-editor geopend. Zo kun je de schattingresultaten gebruiken bij vervolgberekeningen. Limdep kent ook variabelen namen toe aan de laatste schattingsresultaten; die kun je ook gebruiken bij verder rekenwerk in Limdep, of je kunt ze wegschrijven naar een file.

Daarna volgen de random effects paneldata-uitkomsten.

```
+----------------------------------------------------+
| Random Effects Model: v(i,t) = e(i,t) + u(i)       |
| Estimates:  Var[e]                =    .736358D+03 |
|             Var[u]                =    .153649D+04 |
|             Corr[v(i,t),v(i,s)] =    .676020     |
| Lagrange Multiplier Test vs. Model (3) = 1222.14   |
| ( 1 df, prob value =  .000000)                     |
| (High values of LM favor FEM/REM over CR model.)   |
| Reestimated using GLS coefficients:                |
| Estimates:  Var[e]                =    .950350D+03 |
|             Var[u]                =    .149856D+04 |
|             Sum of Squares             .686116D+07 |
|             R-squared                  .230530D+00 |
+----------------------------------------------------+
```

Eerst wordt de geschatte variante van de standaard storingterm gegeven Var[e] = .736358D+03. Daarna wordt de variantie van het bedrijfseffect gegeven Var[u] = .153649D+04. De daarop volgende correlatie, in tekstboeken ook weergegeven als ρ, is berekend als Var[u]/(Var[u]+Var[e]). De storingstermen, vit, in verschillende perioden voor een gegeven individu, i, zijn gecorreleerd, omdat deze storingstermen de term ui gemeenschappelijk hebben (zie sectie 2.4.1).

Een hoge waarde voor de Lagrange Multiplier test [1] geeft aan dat een paneldatamodel de voorkeur verdient boven het OLS-model. Aan de hand van de probability value kun je snel zien of de Lagrange Multiplier test groot is. Een prop value kleiner dan 0,05 duidt al op een voorkeur voor paneldata-aanpak. In dit geval geeft het testresultaat aan dat een REM model de voorkeur boven OLS.

Aan de hand van de OLS-schattingsresultaten (de residuen) worden de varianties berekend (zie het kader hierboven). Deze worden weer gebruikt voor een feasible GLS-schatting van het random effects model. De uitkomsten van de GLS-schatting staan onder in het kader vermeld. De $R^2$ van 0,23 is een best aardig voor een schatting op basis van micro data.

| Variable | Coefficient | Standard Error | b/St.Er. | P[\|Z\|>z] | Mean of X |
|---|---|---|---|---|---|
| ND_MAISH | -.5924878526E-01 | .69052192E-02 | -8.580 | .0000 | 160.90555 |
| DKLEI | 38.73559602 | 3.3081527 | 11.709 | .0000 | .23069140 |
| DAKK | 21.32629041 | 5.3179133 | 4.010 | .0001 | .45868465E-01 |
| HATOT | .2301346572 | .54881347E-01 | 4.193 | .0000 | 32.765798 |
| N_MSTPI | -6.286756417 | 3.0080306 | -2.090 | .0366 | 1.0666190 |
| TREND | -2.427754257 | .37163813 | -6.533 | .0000 | 4.0563238 |
| Constant | 73.95122505 | 4.5822687 | 16.139 | .0000 | |

*Aanvullende opmerkingen*

Als niet expliciet een fixed of random effects model wordt opgevraagd produceert LIMDEP beide. Eerst komt de OLS-schatting, dan de FEM-resultaten en daarna de REM uitkomsten. Nu wordt de varianties benodigd voor de GLS-schatting van het REM bepaald aan de hand van de uitkomsten van het FEM. Als je alleen een random effects schatting uitvoert door **;random** toe te voegen aan de commando regel worden alleen de OLS en

[1] Limdep presenteert de Breusch and Pagan's Lagrange multiplier statistic. Deze test de $H_0$ dat de variantie van het bedrijfseffect of jaareffect gelijk is aan 0 ($\sigma^2_u$=0), dat is gelijk aan OLS.

REM-schatting gedaan. De GLS-schatting kan dan niet meer worden gebaseerd op de FEM-uitkomsten. Als alleen REM wordt berekend, wordt voor de tweede fase gebruikt gemaakt van de OLS-uitkomsten, hierdoor kunnen de REM-schattingsresultaten licht afwijken van het volledige paneldataschatting (als ook FEM wordt berekend).

In dit voorbeeld zijn variabelen opgenomen die niet (per bedrijf) variëren in de tijd; zodat fixed effects schatting niet mogelijk is. Als je een Fixed Effects-model specificeert in LIMDEP waarin variabelen zijn opgenomen die niet variëren voor elk bedrijf (bijvoorbeeld grondsoort, regio of opleiding), dan geeft LIMDEP een foutmelding [variables are collinear].

In het one-way random effects model is een trend opgenomen om de trendmatige afname van de kunstmestgift (bijvoorbeeld door strenge regelgeving) op maïsland te modelleren. Een tijdtrend-term kan niet samen met een two-way error model worden gespecificeerd (Limdep geeft dan een foutmelding na eerst de REM-bedrijfseffecten te hebben geschat). Daarom specificeren we ter vergelijking het two-way REM zonder trend.

Voor een two-way effect paneldataschatting is ook een periode variabele nodig. Je kan gewoon **period=jaar** gebruiken (de variabele hoeft niet bij 1 te beginnen). Als de variabele trend wordt verwijderd van de verklarende variabelenlijst, dan kan een two-way REM worden geschat. Na de OLS-schatting en de eerste ronde REM waarin de bedrijfseffecten worden meegenomen volgt een tweede ronde waarin de jaareffecten worden bepaald.

Deze uitkomsten wijken iets af van de hierboven gepresenteerde resultaten omdat nu de trend-term niet is meegeschat. Zo is nu de Var[e] groter en is de R-squared kleiner (Model 3 is het OLS-model).

```
+----------------------------------------------------------------------+
| OLS Without Group Dummy Variables                                    |
| Ordinary    least squares regression    Weighting variable = none    |
| Dep. var. = NK_MAISH Mean=   65.73435816    , S.D.=   54.29366748    |
| Model size: Observations =    2965, Parameters =   6, Deg.Fr.=   2959|
| Residuals:  Sum of squares= 6764593.826    , Std.Dev.=       47.81326|
| Fit:        R-squared=  .225779, Adjusted R-squared =          .22447|
| Model test: F[  5,   2959] =  172.58,    Prob value =          .00000|
| Diagnostic: Log-L = -15670.7033, Restricted(b=0) Log-L =  -16050.0712|
|             LogAmemiyaPrCrt.=    7.737, Akaike Info. Crt.=     10.575|
| Panel Data Analysis of NK_MAISH   [ONE way]                          |
|          Unconditional ANOVA (No regressors)                         |
| Source      Variation        Deg. Free.     Mean Square              |
| Between       .681607E+07          995.         6850.32              |
| Residual      .192121E+07         1969.          975.731             |
| Total         .873729E+07         2964.          2947.80             |
+----------------------------------------------------------------------+
+---------+-------------+----------------+--------+---------+----------+
|Variable | Coefficient | Standard Error |b/St.Er.|P[|Z|>z] | Mean of X|
+---------+-------------+----------------+--------+---------+----------+
 ND_MAISH -.1295111748      .89115555E-02  -14.533    .0000  160.90555
 DKLEI     40.50568273      2.1993777       18.417    .0000  .23069140
 DAKK      17.75148538      4.4616163        3.979    .0001  .45868465E-01
 HATOT      .1358906687     .39056103E-01    3.479    .0005  32.765798
 N_MSTPI  -19.15399134      3.7265968       -5.140    .0000  1.0666190
 Constant  92.39232402      4.8871602       18.905    .0000
```

```
+----------------------------------------------------+
| Random Effects Model: v(i,t) = e(i,t) + u(i)       |
| Estimates:  Var[e]                =   .751331D+03  |
|             Var[u]                =   .153478D+04  |
|             Corr[v(i,t),v(i,s)] =   .671349        |
| Lagrange Multiplier Test vs. Model (3) = 1195.31   |
| ( 1 df, prob value =  .000000)                     |
| (High values of LM favor FEM/REM over CR model.)   |
| Reestimated using GLS coefficients:                |
| Estimates:  Var[e]                =   .963054D+03  |
|             Var[u]                =   .150580D+04  |
|             Sum of Squares            .690687D+07  |
|             R-squared                 .225779D+00  |
+----------------------------------------------------+
```

| Variable | Coefficient | Standard Error | b/St.Er. | P[\|Z\|>z] | Mean of X |
|---|---|---|---|---|---|
| ND_MAISH | -.7152859303E-01 | .67156461E-02 | -10.651 | .0000 | 160.90555 |
| DKLEI | 38.30400790 | 3.3125624 | 11.563 | .0000 | .23069140 |
| DAKK | 21.44294481 | 5.3401989 | 4.015 | .0001 | .45868465E-01 |
| HATOT | .2040104983 | .54899684E-01 | 3.716 | .0002 | 32.765798 |
| N_MSTPI | -3.619295970 | 3.0038948 | -1.205 | .2283 | 1.0666190 |
| Constant | 64.04656488 | 4.3464719 | 14.735 | .0000 | |

```
+----------------------------------------------------------+
| Random Effects Model: v(i,t) = e(i,t) + u(i) + w(t)      |
| Estimates:  Var[e]                =   .742646D+03        |
|             Var[u]                =   .153910D+04        |
|             Corr[v(i,t),v(i,s)] =   .674528              |
|             Var[w]                =   .435849D+01        |
|             Corr[v(i,t),v(j,t)] =   .005835              |
| Lagrange Multiplier Test vs. Model (3) = 1267.79         |
| ( 2 df, prob value =  .000000)                           |
| (High values of LM favor FEM/REM over CR model.)         |
| Reestimated using GLS coefficients:                      |
| Estimates:  Var[e]                =   .977699D+03        |
|             Var[u]                =   .149923D+04        |
|             Var[w]                =   .192084D+03        |
|             Sum of Squares            .690361D+07        |
|             R-squared                 .225779D+00        |
+----------------------------------------------------------+
```

| Variable | Coefficient | Standard Error | b/St.Er. | P[\|Z\|>z] | Mean of X |
|---|---|---|---|---|---|
| ND_MAISH | -.6923998375E-01 | .80411965E-02 | -8.611 | .0000 | 160.90555 |
| DKLEI | 39.32289860 | 3.3172257 | 11.854 | .0000 | .23069140 |
| DAKK | 21.06954839 | 5.3328107 | 3.951 | .0001 | .45868465E-01 |
| HATOT | .2156228966 | .54948423E-01 | 3.924 | .0001 | 32.765798 |
| N_MSTPI | -5.144663741 | 3.0178786 | -1.705 | .0882 | 1.0666190 |
| Constant | 64.78248513 | 4.4711123 | 14.489 | .0000 | |

De vraag is nu natuurlijk of voor het one-way model moet worden gekozen inclusief trend-term of voor het two-way model (zonder expliciete trend). In het two-way RE-model veronderstel je dat de jaareffecten onafhankelijk zijn en normaal zijn verdeeld. Echter, gezien de significant negatieve coëfficiënt van de trend term in het one-way model neemt de kunstmestgift trendmatig af. Deze trend wordt niet goed (correct) gevangen in het two-way RE-model [1]. De coëfficiënten van beide schattingsresultaten wijken weinig af.

[1] Een two-way FEM-model kent geen veronderstellingen over de verdeling van de jaareffecten, en zou in dit opzicht te prefereren zijn boven een two-way REM. Een mogelijke alternatieve oplossing is het schatten van een one-way REM met jaardummies.

Aangezien deze schatting nodig is voor het stofstromenmodel hebben we een voorkeur voor het one-way model. Je kan dan gebruik maken van de trendmatige afname van de stikstofgift om het stikstofgebruik in een bepaald jaar nauwkeuriger voorspellen in het one-way error model met trend dan in het two-way-error model zonder trend.

Het model dat de voorkeur verdient schatten we nu nog met gebruik making van de wegingsfactoren. Op deze wijze willen we laten zien dat het voor de resultaten wel uitmaakt of je weging gebruikt of niet. Het commando wordt uitgebreid met **;wts=weging.** (de variabele die de wegingsfactoren bevat is weging genoemd in SPSS).

**regress;lhs=NK_maish;rhs=ND_maish,dklei,dakk,hatot,N_mstpi,trend;**
**str=num; panel; random; wts=weging$**

Alleen het REM deel is hieronder weergegeven.

```
+-----------------------------------------------------+
| Random Effects Model: v(i,t) = e(i,t) + u(i)        |
| Estimates:  Var[e]                 =    .178750D+04 |
|             Var[u]                 =    .485267D+03 |
|             Corr[v(i,t),v(i,s)] =    .213514       |
| Lagrange Multiplier Test vs. Model (3) = 4769.04    |
| ( 1 df, prob value =  .000000)                      |
| (High values of LM favor FEM/REM over CR model.)    |
| Reestimated using GLS coefficients:                 |
| Estimates:  Var[e]                 =    .100858D+04 |
|             Var[u]                 =    .111650D+04 |
|             Sum of Squares              .678494D+07 |
|             R-squared                   .215914D+00 |
+-----------------------------------------------------+
```

| Variable | Coefficient | Standard Error | b/St.Er. | P[\|Z\|>z] | Mean of X |
|---|---|---|---|---|---|
| ND_MAISH | -.8937224499E-01 | .94832346E-02 | -9.424 | .0000 | 161.29351 |
| DKLEI | 37.20300339 | 2.8768652 | 12.932 | .0000 | .19991105 |
| DAKK | 28.18828203 | 6.0079102 | 4.692 | .0000 | .29083823E-01 |
| HATOT | .2397892145 | .57584258E-01 | 4.164 | .0000 | 26.706596 |
| N_MSTPI | -14.00839064 | 3.9290159 | -3.565 | .0004 | 1.0789708 |
| TREND | -2.348404378 | .48078622 | -4.885 | .0000 | 4.0247750 |
| Constant | 87.37781051 | 5.6249248 | 15.534 | .0000 | |

Wat blijkt: door de wegingsfactoren zijn alle varianties veranderd. Ook zijn de parameter schattingen beïnvloed; die van de trend is kleiner en die van de dierlijke mestgift is groter dan zonder wegingsfactoren. Het verschil tussen de schattingsresultaten met en zonder wegingsfactor is niet zo groot dat je een duidelijke uitspraak kan doen dat de een goed is en de ander fout.

### 3.6 Schatting van een productiefunctie (FEM versus REM)

Dit voorbeeld is gebaseerd op het artikel van Y. Mundlak (1961). In dat artikel wordt OLS vergeleken met de Fixed Effects schatter (wordt niet zo genoemd door Mundlak).

Om het voorbeeld zo eenvoudig mogelijk te houden wordt een Cobb-Douglas productiefunctie geschat.

$$OUTPH = C^{\alpha 0} ARB^{\alpha 1} KAPH^{\alpha 2} LANDH^{\alpha 3} VARINPH^{\alpha 4} \quad (3.3)$$

*OUTPH* output (in NLG van 1991);
*ARB* gezinsarbeid (in uren);
*KAPH* kapitaalgoederenvoorraad (in NLG van 1991);
*LANDH* land (in hectare);
*VARINPH* hoeveelheid variabele inputs (voer, kunstmest enzovoort) (in NLG van 1991);
$\alpha 1,\ldots, \alpha 4$ te schatten parameters.

$$\ln OUTPH = \alpha_0 + \alpha_1 * \ln ARB + \alpha_2 * \ln KAPH + \alpha_3 * \ln LANDH + \alpha_4 * \ln VARINPH \quad (3.4)$$

Bij OLS neem je aan dat er een productiefunctie is die niet verandert in de tijd en die voor alle bedrijven gelijk is. Bij de FEM- en REM-benadering neem je aan de vorm van de productiefunctie gelijk is voor alle bedrijven maar dat de hoogte kan verschillen tussen de bedrijven. Bij de two-way benadering kan de productiefunctie bovendien ook tussen de jaren verschillen.

In dit geval zijn in SPSS de variabelen die een onderdeel vormen van output (melk, vlees enzovoort) kapitaalgoederen (werktuigen, gebouwen en vee) en variabele inputs (veevoer, kunstmest, enzovoort) geaggregeerd tot een variabele voor output, kapitaalgoederen en een voor variabele inputs.

Eerst worden de data die nodig zijn voor de schatting van de productiefunctie ingelezen in LIMDEP via het worksheet. De dataset bevat 4.310 waarnemingen van 1.153 bedrijven. Dit zijn alle sterk gespecialiseerde melkveehouderijbedrijven die tussen 1985 en 1995 deel hebben genomen aan het Informatienet.

De volende commando's zijn te vinden in de file e:\personal\panel\data\prodfun.lim.

**read ; file=e:\personal\panel\data\prodfun.wk1 ; Format=WKS; names$**

met dit commando wordt de .wk1 file ingelezen en worden ook de namen ingelezen.

Als je wilt controleren of de data goed zijn ingelezen kun je het volgende commando gebruiken (komt overeen met DESCRIPTIVE in SPSSX) voor de melkveehouderijbedrijven uit Informatienet in de periode 1985-1995.

**dstat; rhs=num,jaar,outph,arb,landh,kaph,varinph$**

De data kunnen worden bewerkt, bijvoorbeeld door er logaritmen van te maken nodig voor de schatting van de CobbDouglas productiefunctie (deze bewerking had op zich ook in SPSS plaats kunnen vinden). Bijvoorbeeld:

**create; lo=log(outph)**
**; la=log(arb)**
**; ll=log(landh)**
**; lk=log(kaph)**
**; lv=log(varinph)**
**; lj=jaar-1984$**

De one-way effect paneldataschatting (alleen de bedrijfseffecten worden meegenomen) ziet er nu als volgt uit. De stratificatie variabele om de bedrijven van elkaar te onderscheiden is het bedrijfsnummer **NUM**. Het volgende commando geeft zowel FEM als REM-schatting.

**regr; lhs=lo; rhs=one,la,ll,lk,lv,lj; str=num; panel$**

(* mocht het programma op dit punt crashen dan loop je waarschijnlijk vast op een bug in een oude versie van LIMDEP. Je dient dan een nieuwe .EXE file te gebruiken (en de verwijzing in je shortcut aan te passen.).

Eerst geeft LIMDEP de OLS-uitkomsten weer (OLS Without Group Dummy Variables). In een kader worden de uitkomsten getest enzovoort. Daaronder staan de parameter schattingen.

Na het OLS-kader en de OLS-schattingsresultaten volgt het FEM-kader en de geschatte coëfficiënten. Daarna volgen de fixed effects paneldata-uitkomsten (Least Squares with Group Dummy Variables). De Group Dummy Variables slaat op de dummy per bedrijf. De samenvatting in het kader geeft aan wat de te verklaren variabele is (LO), geeft het aantal waarnemingen (Observations=4310) Als het aantal Parameters groter is dan 999 staan er sterretjes weergegeven (Parameters=***). Het aantal parameters is zo groot omdat ieder bedrijf zijn eigen bedrijfsspecifieke parameter heeft (er wordt geen constante gespecificeerd in het FE-model). Door het grote aantal parameters is de $R^2$ ook erg groot (R-squared=.990564). Ook de adjusted $R^2$ is erg groot (adjusted R-squared=.98710). Dit lijken geweldig goede uitkomsten, maar ieder FEM geeft overeenkomstige resultaten, omdat het aantal te schatten parameters erg groot is.

```
+----------------------------------------------------------------------+
| Least Squares with Group Dummy Variables                             |
| Ordinary    least squares regression    Weighting variable = none    |
| Dep. var. = LO       Mean=   12.66677705    , S.D.=   .6162791939    |
| Model size: Observations =    4310, Parameters = ***, Deg.Fr.=   3152 |
| Residuals:  Sum of squares= 15.44256391    , Std.Dev.=          .06999 |
| Fit:        R-squared=   .990564, Adjusted R-squared =          .98710 |
| Model test: F[***,   3152] =  285.99,    Prob value =          .00000 |
| Diagnostic: Log-L =   6020.3987, Restricted(b=0) Log-L =   -4028.8473 |
|             LogAmemiyaPrCrt.=   -5.081, Akaike Info. Crt.=     -2.256 |
| Estd. Autocorrelation of e(i,t)    -.041723                          |
+----------------------------------------------------------------------+
+---------+-------------+----------------+--------+---------+----------+
|Variable | Coefficient | Standard Error |b/St.Er.|P[|Z|>z] | Mean of X|
+---------+-------------+----------------+--------+---------+----------+
 LA        .8650778246E-01  .11603861E-01    7.455   .0000  8.2585089
 LL        .1990796510      .14940601E-01   13.325   .0000  3.3813826
 LK        .8113945432E-01  .10620232E-01    7.640   .0000  13.321828
 LV        .3060025266      .98641967E-02   31.022   .0000  11.681713
 LJ        .1461509929E-01  .82583329E-03   17.697   .0000  5.8981439
```

Het volgende kader geeft een overzicht van de testen die mogelijk zijn aan de hand van de bovenvermelde schattingresultaten.

(1) Constant term only — Model met alleen constante
(2) Group effects only — Model met bedrijfseffecten

(3) X variables only — OLS-schattingen (zonder de constante)
(4) X and group effects — Het complete Fixed Effects Model

Een model met alleen constante heeft natuurlijk een $R^2$=0, omdat er geen variantie wordt verklaard. Alleen de bedrijfsspecifieke dummies leiden al tot een $R^2$=0,985.

De meest wezenlijke test is (4) vs (3) FEM tegen OLS (zie ook paragraaf 3.5). De probability value geeft de kans aan dat de F-test statistic bij toeval zo groot is; een kans kleiner dan 0,05 duidt op het verwerpen van OLS.

Test Statistics for the Classical Model

| | Model | Log-Likelihood | Sum of Squares | R-squared |
|---|---|---|---|---|
| (1) | Constant term only | -4028.84720 | .1636558393D+04 | .0000000 |
| (2) | Group effects only | 5071.16824 | .2398924953D+02 | .9853416 |
| (3) | X - variables only | 2062.71828 | .9689600167D+02 | .9407928 |
| (4) | X and group effects | 6020.39877 | .1544256391D+02 | .9905640 |

Hypothesis Tests

| | | Likelihood Ratio Test | | | F Tests | | | |
|---|---|---|---|---|---|---|---|---|
| | | Chi-squared | d.f. | Prob. | F | num. | denom. | Prob value |
| (2) | vs (1) | 18200.031 | 1152 | .00000 | 184.214 | 1152 | 3157 | .00000 |
| (3) | vs (1) | 12183.131 | 5 | .00000 | 13677.978 | 5 | 4304 | .00000 |
| (4) | vs (1) | 20098.492 | 1157 | .00000 | 285.988 | 1157 | 3152 | .00000 |
| (4) | vs (2) | 1898.461 | 5 | .00000 | 348.895 | 5 | 3152 | .00000 |
| (4) | vs (3) | 7915.361 | 1152 | .00000 | 14.432 | 1152 | 3152 | .00000 |

Bij een FE-model wordt in feite voor ieder bedrijf een bedrijfsspecifieke factor opgenomen. Het aantal onafhankelijke variabelen stijgt daardoor tot grote hoogte (meer dan N onafhankelijke variabelen). Door dit grote aantal variabelen zal de $R^2$ natuurlijk ook groot zijn, de $R^2$ is dus bij FE-schattingen geen goede maatstaf om te bepalen of de X-variabelen in het model (afgezien van de bedrijfsspecifieke variabelen) een groot deel van de variante van de afhankelijke variabelen verklaren. Daarom geeft LIMDEP bij FE-schattingen ook de $R^2$ van het model met alleen X-variabelen. Op basis van de loglikelihoodwaarden van de 4 onderscheiden modellen (zie het kader) wordt met een loglikelihood test bepaald welk model wordt geprefereerd (er wordt getest of een vereenvoudigd model significant afwijkt van het volledige model; het FEM-model). De Chi-squared geeft de grootte van het testresultaat weer. Een Probability kleiner dan 0,05 geeft aan dat het grote model geen verklaring toevoegt aan het kleine model. In dit geval wordt FEM verkozen boven OLS.

Na de FEM schatting kun je de individuele bedrijfseffecten opvragen (code ;output=2 toevoegen aan de commandoregel. De bedrijfseffecten worden dan weggeschreven naar een matrix Alpha. Deze effecten kun je uitdraaien/kopiëren met het commando
**matrix; list; ALPHA$**

De matrix (matrix:result) die dan in de output-file verschijnt kun je knippen en in SPSS plakken.

*Wat kun je met het bedrijfseffect?*

In de bedrijfseffecten van de FE-schatting zit ook de constante opgenomen. Om deze constante eruit te verwijderen wordt wel het bedrijfseffect bepaald ten opzichte van het beste bedrijf; dit is het bedrijf met het grootste bedrijfseffect (tenminste als het bedrijfeffect wordt verondersteld een positieve relatie te hebben met de te verklaren variabele, zoals in de productiefunctie). In REM kan alleen onder specifieke aannames via enkel berekeningen het bedrijfseffect van een individueel bedrijf worden bepaald. Op zich valt een FE-schatting ook met OLS uit te voeren als je voor alle bedrijven een bedrijfsspecifieke dummy variabele opneemt en dan OLS schat zonder constante (zie een voorbeeld in bijlage 1.3). (In SPSS is het niet mogelijk met veel dummies een OLS te schatten, zodat je dit toch in Limdep zal moeten doen).

Nu wordt de random effects schatting gegeven:

```
+-----------------------------------------------------+
| Random Effects Model: v(i,t) = e(i,t) + u(i)        |
| Estimates:  Var[e]                =    .489929D-02  |
|             Var[u]                =    .175774D-01  |
|             Corr[v(i,t),v(i,s)] =    .782028      |
| Lagrange Multiplier Test vs. Model (3) = 3840.58    |
| ( 1 df, prob value =  .000000)                      |
| (High values of LM favor FEM/REM over CR model.)    |
| Fixed vs. Random Effects (Hausman)      =  905.28   |
| ( 5 df, prob value =  .000000)                      |
| (High (low) values of H favor FEM (REM).)           |
| Reestimated using GLS coefficients:                 |
| Estimates:  Var[e]                =    .587716D-02  |
|             Var[u]                =    .218577D-01  |
|             Sum of Squares             .109509D+03  |
|             R-squared                  .940793D+00  |
+-----------------------------------------------------+
```

In het kader wordt de grootte van de Hausman-test gegeven. In dit geval is die erg groot 905.28; dit valt af te leiden uit de prob value die gelijk is aan 0.0000. Dit testresultaat wijst op een voorkeur voor een FEM. We hebben echter in het Informatienet niet te maken met een tijdreeks die naar oneindig gaat, zodat het testresultaat niet alles zegt.

| Variable | Coefficient | Standard Error | b/St.Er. | P[\|Z\|>z] | Mean of X |
|---|---|---|---|---|---|
| LA | .1273723501 | .90301571E-02 | 14.105 | .0000 | 8.2585089 |
| LL | .2615620163 | .90110376E-02 | 29.027 | .0000 | 3.3813826 |
| LK | .1902917181 | .87612311E-02 | 21.720 | .0000 | 13.321828 |
| LV | .4911279869 | .71079538E-02 | 69.096 | .0000 | 11.681713 |
| LJ | .1851856332E-01 | .67266681E-03 | 27.530 | .0000 | 5.8981439 |
| Constant | 2.338465838 | .10504074 | 22.262 | .0000 | |

Een andere mogelijkheid om te analyseren of FEM of REM de voorkeur geniet is het kijken naar teken en grootte van de parameters. De coëfficiënten van de Cobb Douglas productie functie zijn gelijk aan de partiele elasticiteiten. De som van deze elasticiteiten (uitgezonderd die van de trend variabele) vormt de schaalparameter. Als die groter is dan 1, dan zijn er toenemende schaalopbrengsten en kleiner dan 1 dan afnemende schaalopbrengsten. Met behulp van Limdep kunnen de geschatte parameters worden gesommeerd. Een overzicht van de gevonden schaalparameters is als volgt:

| | Schaalparameter | |
|---|---|---|
| Methode | Totale panel | >= 7 jaar in BIN a) |
| OLS | 1.113 | 1.074 |
| REM oneway | 1.071 | 1.021 |
| REM twoway | 1.077 | 1.034 |
| REM +jaardummy | 1.115 | 1.077 |
| FEM oneway | 0.673 | 0.764 |
| FEM twoway | 0.706 | 0.803 |

a) De observaties zijn geselecteerd op de variabele **NI** (aantal jaren in het paneldatabestand) <7. In dit bestand zijn 954 waarnemingen.

Er wordt over het algemeen verondersteld dat de schaalparameter voor de melkveehouderij iets groter is dan 1. Wat echter bij schattingsresultaten van het Informatienet steeds optreedt is dat de FEM-schatting extreem kleine schaaleffecten te zien geeft. Dit wordt veroorzaakt doordat de bedrijfseffecten een groot deel van de variantie verklaren, zeker als een groep bedrijven slechts korte tijd in het panel is opgenomen. De RE-modellen geven een beeld van de schaalparameter die meer overeenkomst met de verwachte waarde.

Om de invloed van de kleine waarde van T op de schattingsresultaten te laten zien, is de schatting overgedaan voor een dataset waarin alleen de melkveebedrijven uit de eerste dataset zijn opgenomen die 7 jaren of langer deelnamen aan het Informatienet in de periode 1985-1995. De schaalparameter van de FEM-modellen zijn behoorlijk groter geworden, die van de REM-modellen iets kleiner. (Convergeren ze naar elkaar als T naar oneindig gaat.)

Ook bij Mundlak (1961) is de FE-schatter van de schaal veel kleiner dan bij de andere; hij verklaart de kleine schaalparameter uit het feit dat de management input niet is gespecificeerd in het model, deze is wel opgenomen in het bedrijfseffect (als we veronderstellen dat deze input constant is in de tijd). Mundlak heeft een zelfde korte panel als wij dat hebben. OLS is meest gerestricteerde model. Daarom noemt hij two-way fixed effects the unbiased estimator.

# 4. Conclusies en aanbevelingen

In dit hoofdstuk worden conclusies en aanbevelingen beschreven betrekking hebbend op het gebruik van paneldata in het algemeen en het gebruik met behulp van het Bedrijven-Informatienet in het bijzonder.

## 4.1 Conclusies

1. Op het LEI wordt gebruikgemaakt van paneldata. Echter, bij multivariate analyse methoden (bijvoorbeeld het schatten van regressievergelijkingen, correlatie bepaling, discriminantenanalyse) op het LEI wordt geen of nauwelijks rekening gehouden met de speciale structuur van de paneldataset zodat een deel van de aanwezige informatie wordt verwaarloosd. De gevolgen zijn afwijkende en minder betrouwbare resultaten en conclusies ten opzichte van een situatie waarin wel rekening zou zijn gehouden met de paneldatastructuur. Het is van belang voor de kwaliteit van het onderzoek op LEI, dat onderzoekers zich bewust worden van deze tot nog toe gehanteerde inefficiente behandeling van paneldata, en dat zij handvatten krijgen aangereikt om in de toekomst hierin verandering te brengen. Deze handleiding beoogt te voorzien in het aanreiken van handvatten om een paneldataschatting eenvoudiger uit te voeren.
2. Bijna alle Informatienet-gebruikers zijn geïnteresseerd in een workshop statistische analyse, toegespitst op het Informatienet en de LBT (zie paragraaf 1.2).
3. Het aantal observaties bij een panelbestand is groter dan bij cross-sectie of tijdreeks. Vergelijking (1) kan worden geschat op grond van N*T-waarnemingen, wat betrouwbaarder resultaten oplevert dan een tijdreeks (T-waarnemingen) of een cross-sectie (N-waarnemingen). Een aanzienlijke toename van het aantal waarnemingen met daardoor een grotere aannemelijkheid om betrouwbaarder parameter schattingen (van $\beta$) te geven. Je kunt met paneldata in feite de veranderingen meten tussen bedrijven maar tegelijkertijd ook binnen de bedrijven zelf door de jaren heen. Op deze manier gebruik je de aanwezige informatie veel beter dan alleen maar te kijken naar verschillen door de tijd heen voor 1 bedrijf (tijdreeks) of kijken wat de verschillen zijn tussen bedrijven voor 1 jaar (cross-sectie).
4. Met het gebruik van paneldata heb je minder last van multicollineariteit. Samenhangend met de toename van het aantal observaties kan het gebruik van paneldata het probleem van multicollineariteit verlichten. Wanneer de verklarende variabelen varieren over twee dimensies (N en T), is het minder aannemelijk of waarschijnlijk, dat ze hoog gecorreleerd zijn omdat de absolute variantie toeneemt in de tijd. Dat wil zeggen de prijs van kunstmest (die betaald wordt door de boer) zal meer variëren over 7 jaar dan over 1 jaar. Als je geen variatie hebt in je gegevens kun je trouwens ook geen effecten berekenen (zoals een prijselasticiteit van kunstmest).

5. Paneldatasets maken het mogelijk om effecten te bepalen en te berekenen die niet te achterhalen zijn in cross-sectie of tijdreeks. Combineren van cross-sectie en tijdreeks informatie heeft als gevolg dat je een meer algemeen en veelomvattend dynamische structuur kan formuleren en schatten.
6. Het gebruik van paneldata kan de schattingsbias elimineren of verkleinen. Als nu de effecten van je niet-opgenomen variabelen gecorreleerd zijn met de wel opgenomen verklarende variabelen, en als met die correlaties geen rekening wordt gehouden dan zijn je schattingen biased. Bij cross-sectie data zullen de schattingen biased zijn omdat bijvoorbeeld het management van de boer niet expliciet opgenomen kan worden in de vergelijking. Met paneldata ben je in staat om deze variabele te controleren door het introduceren van een zogenaamd individueel effect. In dit effect zitten alle variabelen die niet veranderen over de tijd (t) maar wel verschillend zijn tussen boeren (i).
7. Paneldata wordt verzameld op microniveau (bedrijfsniveau, individueel niveau). Veel variabelen kunnen preciezer worden berekend op microniveau. Als je deze data met paneldataschattingsmethoden analyseert op microniveau, heb je dus geen last van aggregatie bias (afwijkingen door aggregatie).
8. Paneldata bieden de onderzoeker meer mogelijkheden ten aanzien van het specificeren, maken en testen van (gedrags)modellen voor beleidsgericht onderzoek dan met een cross-sectie of tijdreeks alleen, met bovendien een grotere betrouwbaarheid, mits de juiste methoden gebruikt worden.
9. Er is geen standaard aanpak om te komen tot goede keuzes en uitkomsten. Je moet zelf altijd nagaan aan de hand van je theorie en de veronderstellingen en uitkomsten van je schattingen of je nog goed bezig bent.
10. De Hausman-test is niet eenduidig bij een paneldataschatting met behulp van Informatienet-data. Deze test laat je normaliter zien of je een Fixed Effects Model of een Random Effects Model moet gebruiken. Door de relatief kleine tijddimensie van bedrijven in het Informatienet is namelijk lang niet voldaan aan een veronderstelling van de Hausman-test; namelijk dat de tijdshorizon oneindig is.
11. De hele wegingsproblematiek is nog niet uitgekristaliseerd met betrekking tot het gebuik van het Informatienet voor analyse doeleinden.
12. Vanuit onderzoeksoogpunt en paneldata-analyses heeft een roterend panel met een lange tijdsdimensie de voorkeur. Hiermee kunnen zowel de nieuwste ontwikkelingen in de steekproef meegenomen worden (structuurveranderingen, hele grote bedrijven, enzovoort) als de voordelen van de lange tijd dat bedrijven in de steekproef zitten goed worden benut.

## 4.2 Aanbevelingen

1. Op het gebied van paneldataschattingen op basis van Informatienet-gegevens zou het wenselijk zijn dat binnen de kenniseenheid maatschappij meer samengewerkt zou gaan worden, zodat er meer uitwisseling van kennis plaats kan vinden.

2. Vanuit het oogpunt van kwaliteit en onderhouden van statistische kennis is het wenselijk om dit jaar nog een opfris-cursus statistiek met betrekking tot het gebuik van het Informatienet aan de onderzoekers van het LEI aan te bieden.
3. Het is wenselijk om als vervolg op de bovengenoemde basiscursus statistiek, een cursus 'gebruik paneldata op het LEI' aan de onderzoekers van het LEI aan te bieden.
4. Het totaal aan doelstellingen van het huidige Informatienet rechtvaardigen een roterend panel (en een herziening van de steekproefopzet). Gelet op methodologische, organisatorische en budgettaire overwegingen is een herijking van vorm en omvang van de rotatie gewenst. Vanuit het gebruik van het Informatienet voor onderzoeksdoeleinden sec zou het nuttig zijn als de tijdsdimensie (aantal jaren dat respondenten deelnemen aan het Informatienet) vergroot zou worden van 5 à 7 naar 10 jaar.

# Literatuur


Baltagi, B.H., *Econometric analysis of panel data.* Chichester etc. Wiley, 1995.

Battese, G.E., 'On the estimation of Cobb Douglas Production Functions When Some Explanatory Variables Have ZeroValues'. In: *Journal of Agricultural Economics* 48(2):250-252, 1997.

Dijk, J., *De steekproef gewogen.* Onderzoekverslag 53, LEI, Den Haag, 1989.

Dijk, J.P.M. van, J.J.P. Groot, K. Lodder, L.C. van Staalduinen en H.C.J. Vrolijk, *De steekproef voor het Bedrijven-Informatienet van het LEI; Bedrijfskeuze 1998 en selectieplan 1999.* LEI, Den Haag, 1999.

Davidson R. and J.G. Mackinnon, *Estimation and inference in econometrics.* Oxford University Press, 1993.

Greene, W.H., *Econometric Analysis.* Fourth edition, Upper Saddle River, Prentice Hall, 2000.

Gujarati, D.N., *Basic Econometrics.* McGraw-Hill College Division, 1995.

Hausman, J. A. en W.E. Taylor, 'Panel Data and Unobservable Individual Effects'. *Econometrica* Vol. 49, No 6 (November, 1981), 1981.

Hsiao, C., 'Analysis of panel data'. In: *Econometric Society Monographs No. 11*, Cambridge University Press, 1986.

Johnston, J. en J. DiNardo, *Econometric Methods.* McGraw-Hill International Editions, fourth edition, 1997.

Judge, Hill, Griffiths, Lutkepohl en Lee, *Introduction to the theory and practice of econometrics.* Second editon, Wiley, 1988.

Leneman et al., Stofstromenmodel 3. Rapport (nog te verschijnen), LEI, Den Haag, 2001.

Mátyás, L. en P. Sevestre (eds.), *The econometrics of panel data; A handbook of the theory with applications.* Advanced studies in theoretical and applied econometrics, Kluwer Academic Publishers, 1996.

Mundlak, Y., 'Empirical Production Function Free of Management Bias'. In: *Journal of Farm Economics* (February) pp.44-56, 1961.

Pindyck, R.S. and D.L. Rubinfeld, *Econometric Models & Economic Forecasts.* McGraw-Hill International Editions, third edition, 1991.

Vrolijk, H., 'Representativiteit en betrouwbaarheid van het BIN', *Agrimonitor*, jaargang 5, nr. 4, p. 8, 1999.

# Bijlage 1 De programmatuur

## B1.1 De VRKMST.VAR-file; VRKMST97.BDP-file; VRKMST.COM-file; CONVSPS.COM file

---

**! hier staat de VRKMST.VAR file beschreven**

```
bestand b&&1 = 'bkh&&1L'
BESTAND M&&1 ='MEI&&1'
bestand g&&1 ='gewas&&11'

BASIS G&&1

gewas=g&&1.gewas
! voor gras (gewascode 76 en 77) en mais (52 en 79) wordt
! de hoeveelheid stikstof op het gewas uitgedraaid en de werkzame
! zuivere N uit organische mest.
SUBVARG1=g&&1.6.11.2.1.5 ALS [GEWAS=76] ANDERS 0
SUBVARG2=g&&1.6.11.2.1.5 ALS [GEWAS=77] ANDERS 0
SUBVARG3=g&&1.6.11.225 ALS [GEWAS=76] ANDERS 0
SUBVARG4=g&&1.6.11.225 ALS [GEWAS=77] ANDERS 0
! het areaal kunstweide kan uit BIN worden gehaald,
! blijvende weide van pure weide bedrijven zit niet in gewassen codes
! het is consistenter om areaal gegevens uit BIN te halen voor berekening
! kunstemstgift per hectare
SUBVARG5=g&&1.1.4 ALS [GEWAS=77] ANDERS 0

SUBVARM1=g&&1.6.11.2.1.5 ALS [GEWAS=52] ANDERS 0
SUBVARM2=g&&1.6.11.2.1.5 ALS [GEWAS=79] ANDERS 0
SUBVARM3=g&&1.6.11.225 ALS [GEWAS=52] ANDERS 0
SUBVARM4=g&&1.6.11.225 ALS [GEWAS=79] ANDERS 0
SUBVARM5=g&&1.1.4 ALS [GEWAS=52] ANDERS 0
SUBVARM6=g&&1.1.4 ALS [GEWAS=79] ANDERS 0

AGGREGEER gew(sleutel)

BASIS b&&1
KOPPEL gew

varGN=gew.subvarG1+gew.subvarG2
varGND=gew.subvarG3+gew.subvarG4
varMN=gew.subvarM1+gew.subvarM2
varMND=gew.subvarM3+gew.subvarM4

KOPPEL M&&1
```

```
! blijvend grasland en tijdelijk grasland
GRAS=b&&1.1.4.3.1+gew.subvarg5                    !blijvend grasland
! maisland EN OVERIGE VOEDERGEWASSEN
MAIS=gew.subvarM5+gew.subvarm6

selecteer als GRAS > 0 OF MAIS > 0

WEGING=b&&1.1.3.71
selecteer als weging > 0

! UNIEK BEDRIJFSNUMMER
BEDRIJF=bf_dic_real_nummer(sleutel, 'bkh&&1')
JAAR=19&&1
! leeftijd ondernemer
LEEFTIJD=m&&1.1.8.1.03

! 13 LEI-landbouwgebieden
GEB=M&&1.1.1.2.4

NEG=M&&1.1.2.2.2

! cultuurgrond gemeten maat
CULT=M&&1.3

! stikstofgift op  grasland (HOEVEELHEDEN)
! uit boekhoudnet als groter dan 0, anders uit gewassenbestand
! voor 1990 alleen kunstmest in gewassenbestand
! vanaf 1991 inclusief werkzame N uit organische mest
STIK1=b&&1.1.8.102

! kosten stikstofkunstmest/hoeveelheden aangekochte stikstofkunstmest
! op totale bedrijf, dus mogelijk ook voor andere gewassen
STIKWRD=b&&1.6.11.216
STIKHV=b&&1.hv.6.11.216

! melk- en kalfkoeien
MELKKOE=M&&1.4.1.1.2 als M&&1.4.1.1.2>0 ANDERS b&&1.1.8.14

! melkquotum per bedrijf
! het betreft hier het gebruiksquotum
! maar in 87 en 88 referentiequotum
QUOTUM=b&&1.1.8.96 als jaar<89 ANDERS b&&1.1.8.104

! gerealiseerde aan fabriek geleverde melk
MELKHV=b&&1.hv.7.2.1

! gerealiseerde melkopbrengsten (aan fabriek geleverde melk)
! na correctie voor de superheffing
MELKWRD=b&&1.7.2.1

! grondsoort
GRONDS=b&&1.1.6.11.1
```

```
!krachtvoer rundvee paarden schapen en geiten
!kosten en hoeveelheden
KRVWRD=b&&1.5.9.1.1
KRVHV=b&&1.hv.5.9.1.1
!
!ruwvoer kan niet in hoeveelheden worden uitgedrukt (te divers)
!daarom hier alleen kosten en kan later worden gecorrigeerd
!voor de gemiddelde prijs per
RUWWRD=b&&1.5.9.1

rapporteer
file 'scr:vrkmst&&1.bin' REAL*4
```

---

**! hier staat de VRKMST97.BDP file beschreven**

```
@vrkmst.var 97
```

---

**! hier staat de VRKMST.COM file beschreven**

---

```
$ set def [.panel]
$ bdl vrkmst91
$ bdl vrkmst92
$ bdl vrkmst93
$ bdl vrkmst94
$ bdl vrkmst95
$ bdl vrkmst96
$ bdl vrkmst97
```

---

**! hier staat de CONVSPS.COM file beschreven**

```
$ set def temp$$dir
$ ebf vrkmst91.bin/to_pcd
$ ebf vrkmst92.bin/to_pcd
$ ebf vrkmst93.bin/to_pcd
$ ebf vrkmst94.bin/to_pcd
$ ebf vrkmst95.bin/to_pcd
$ ebf vrkmst96.bin/to_pcd
$ ebf vrkmst97.bin/to_pcd
$ set def sys$login
$ set def [.panel]
```

## B1.2 De VRKMST.SPS-file

```
* De macro wordt gedefinieerd die aangeeft welke variabelen worden ingelezen en welke namen deze krijgen
 in de SPSS-file. De SPSS-namen kunnen worden doorgegeven aan het paneldataschattingsprogramma.
 De SPSS namen mogen afwijken vand enamen gebruikt in BDL. Wel is het handig om de .LIS file van BDL
 te gebruiken om na te gaan of de volgorde van de variabelen juist is.

define vrkmdata ()
 N_gras ND_gras N_mais ND_mais hagras hamais weging num jaar leeftijd geb13 bultype hatot n_bkh
 n_mstwrd n_msthv koe# quotum melkhv melkwrd gronds krvwrd krvhv ruwwrd (24rb4).
!enddefine.

* de binaire data files per jaar worden een voor een ingelezen en als SPSS-datafile opgeslagen.
* De hiervoorgedefinieerde makro wordt gebruikt.

file handle vrkmst91 / name='e:\personal\panel\data\vrkmst91.bin' / mode=binary lrecl=96 .
data list file=vrkmst91 notable/ vrkmdata.
xsave outfile='e:\personal\panel\data\vrkmst91.sav' .
execute.

file handle vrkmst92 / name='e:\personal\panel\data\vrkmst92.bin' / mode=binary lrecl=96 .
data list file=vrkmst92 notable/ vrkmdata.
xsave outfile='e:\personal\panel\data\vrkmst92.sav' .
execute.

file handle vrkmst93 / name='e:\personal\panel\data\vrkmst93.bin' / mode=binary lrecl=96 .
data list file=vrkmst93 notable/ vrkmdata.
xsave outfile='e:\personal\panel\data\vrkmst93.sav' .
execute.

file handle vrkmst94 / name='e:\personal\panel\data\vrkmst94.bin' / mode=binary lrecl=96 .
data list file=vrkmst94 notable/ vrkmdata.
xsave outfile='e:\personal\panel\data\vrkmst94.sav' .
execute.

file handle vrkmst95 / name='e:\personal\panel\data\vrkmst95.bin' / mode=binary lrecl=96 .
data list file=vrkmst95 notable/ vrkmdata.
xsave outfile='e:\personal\panel\data\vrkmst95.sav' .
execute.

file handle vrkmst96 / name='e:\personal\panel\data\vrkmst96.bin' / mode=binary lrecl=96 .
data list file=vrkmst96 notable/ vrkmdata.
xsave outfile='e:\personal\panel\data\vrkmst96.sav' .
execute.

file handle vrkmst97 / name='e:\personal\panel\data\vrkmst97.bin' / mode=binary lrecl=96 .
data list file=vrkmst97 notable/ vrkmdata.
xsave outfile='e:\personal\panel\data\vrkmst97.sav' .
execute.

* de hiervoor gesavede SPSS-datafiles worden achterelkaar geplakt.
```

```
add files /file='e:\personal\panel\data\vrkmst91.sav'
 /file='e:\personal\panel\data\vrkmst92.sav'/file='e:\personal\panel\data\vrkmst93.sav'
 /file='e:\personal\panel\data\vrkmst94.sav'/file='e:\personal\panel\data\vrkmst95.sav'
 /file='e:\personal\panel\data\vrkmst96.sav'/file='e:\personal\panel\data\vrkmst97.sav'.

* nu worden enkele controles uitgevoerd om na te gaan of er geen gekke dingen in de dataset voorkomen.
* in wezen overbodig, alleen een check of de goede data uit BIN zijn gehaald; ha ruwvoer moet groter dan 0
zijn.
compute haruwv=hagras+hamais.
descrip var=ha ruwv N_gras ND_gras N_mais ND_mais weging num jaar leeftijd geb13 bultype hatot n_bkh
 n_mstwrd n_msthv koe# quotum melkhv melkwrd gronds krvwrd krvhv ruwwrd.

* check op hoeveelheden per ha.
* het blijkt dat enkele bedrijven geen areaal hebben volgens de landbouwtelling hatot=0.
* hier blijkt dat er verschil kan zijn tussen BIN en LBT, van deze waarnemingen worden jaar en bedrijfs-
nummer gegeven.
do if (hatot eq 0).
print / num jaar hatot haruwv hamais hagras.
end if.
execute.
if (hatot gt 0) N_mstpha=N_msthv/hatot.
descrip var=N_mstpha.
* een waarneming met onwaarschijnlijk veel N per ha.
select if (N_mstpha lt 700).
descrip var=N_mstpha.

* de berekening van de prijzen uit het BIN als waarde gedeeld door de hoeveelheid.
if (N_msthv gt 0) N_mstp=N_mstwrd/N_msthv.
if (melkhv gt 0) melkp=melkwrd/melkhv.
if (krvhv gt 0) krvp=krvwrd/krvhv.
descrip var=N_mstp melkp krvp.

sort cases by jaar num.
* de prijzen worden gedefleerd met de prijsindexcijfers van de gezinsconsumptie basisjaar 1990
(Landbouwcijfers).
* dit prijsindexcijfer is opgeslagen in de SPSS-datafile 'prijsindex.sav'.
match files table='e:\personal\spss95\panel\prijsindex.sav'/file=*/by jaar.
compute N_mstpi=N_mstp*100/consind.
compute melkpi=melkp*100/consind.
compute krvpi=krvp*100/consind.

* de berekening van de gemiddelde prijzen per jaar aan de hand van de BIN-data en het wegschrijven van
deze gemiddelden per jaar.
sort cases by jaar num.
aggregate outfile='e:\personal\panel\data\meanprijs'
 /break =jaar
 /N_mstpm melkpm krvpm N_mstpim melkpim krvpim
 = mean (N_mstp melkp krvp N_mstpi melkpi krvpi ).

* toevoegen van de gemiddelde prijs per jaar aan alle waarnemingen van dat jaar.

match files table='e:\personal\panel\data\meanprijs'/file=*/by jaar.
```

```
descrip var=N_mstpm N_mstp N_mstpim melkpm melkpim melkp krvpm krvp N_mstpi melkpi krvpi gronds
leeftijd.

* Nu worden dummies voor de verschillende grondsoorten gedefinieerd.
* grondsoort 1=zeeklei 2=rivierklei 3=laagveen 4=zand 5=dalgrond 6=loss 7=klei op veen.
do if (gronds eq 1 or gronds eq 2 or gronds eq 7).
compute dklei=1.
else if (gronds eq 3).
compute dveen=1.
else if (gronds eq 4 or gronds eq 5).
compute dzand=1.
else.
compute dloss=1.
end if.
recode dklei dveen dzand dloss (sysmis=0).

* dummies voor de verschillende bedrijfstypen wordne aangemaakt.
do if (bultype ge 1110 and bultype le 1249).
compute dakk=1.
else if (bultype ge 4110 and bultype le 4449).
compute dgraas=1.
else if (bultype ge 5011 and bultype le 5032).
compute dhok=1.
end if.
recode dakk dgraas dhok (sysmis=0).

descrip var=dklei.
sort cases by num jaar.

select if (hamais gt 0).
descrip var=num jaar.

* check of de uitkomsten voor maïsland geloofwaardig zijn.
compute N_mapha=N_mais/hamais.
compute ND_mapha=ND_mais/hamais.
descrip var=N_mapha ND_mapha.
select if (n_mapha le 1000).
descrip var=N_mapha ND_mapha.

compute NK_mais=N_mais-ND_mais.

do if (hagras eq 0 and NK_mais eq 0).
compute codegrnk=1.
else.
compute codegrnk=2.
end if.
descrip var=codegrnk.
select if (codegrnk eq 2).
descrip var=num jaar.
select if (N_mais gt 0).
descrip var=num jaar.
```

```
if (jaar eq 1991) dum91=1.
if (jaar eq 1992) dum92=1.
if (jaar eq 1993) dum93=1.
if (jaar eq 1994) dum94=1.
if (jaar eq 1995) dum95=1.
if (jaar eq 1996) dum96=1.
if (jaar eq 1997) dum97=1.
recode dum91 to dum97 (missing=0).

sort cases by num jaar .

aggregate outfile='e:\personal\panel\data\basagnum'
 /break =num
 /ni = n (num).
match files table='e:\personal\panel\data\basagnum'/file=*/by num.

sort cases by num jaar.
do if ($casenum eq 1).
compute frtnum=1.
compute frtjaar=1.
else if (num ne lag(num)).
compute frtnum=lag(frtnum)+1.
compute frtjaar=1.
else.
compute frtnum=lag(frtnum).
compute frtjaar=lag(frtjaar)+1.
end if.

* door de uitgevoerde selecties kunnen gaten vallen in het aantal observaties van een bedrijf;
 zo kan bijvoorbeeld 1994 eruit vallen zodat er waarnemingen uit 1992 1993 en 1995 overblijven.
* we controleren de dataset op zulke gaten.

if (frtjaar eq 1) startjr=jaar.
if (frtjaar eq ni) eindjr=jaar.

aggregate outfile='e:\personal\panel\data\start.sav'
 /break =num
 /startjr = min (jaar).
aggregate outfile='e:\personal\panel\data\eind.sav'
 /break =num
 /eindjr = max (jaar).
match files table='e:\personal\panel\data\start.sav'/file=*/by num.
match files table='e:\personal\panel\data\eind.sav'/file=*/by num.

compute periode=eindjr-startjr+1.
do if (periode ne ni).
compute gatcode=2.
else.
compute gatcode=1.
end if.
```

```
descrip var=gatcode.
* hoe gaan we om met dit ene bedrijf?.
* hoe gaan we om met bedrijven die maar 1 jaar in de dataset zitten (voor REM maakt dat niet uit).
select if (gatcode eq 1).

recode n_mstp melkp krvp n_mstpi melkpi krvpi (missing=-999).

xsave outfile='e:\personal\panel\data\vrkmst.sav'.
execute.
get file='e:\personal\panel\data\vrkmst.sav'.

DESCRIPTIVES
 VARIABLES=num eindjr startjr ni jaar n_mstpm melkpm krvpm n_mstpim melkpim
 krvpim consind n_gras nd_gras n_mais nd_mais hagras hamais weging leeftijd
 geb13 bultype hatot n_bkh n_mstwrd n_msthv koe# quotum melkhv melkwrd gronds
 krvwrd krvhv ruwwrd haruwv n_mstpha n_mstp melkp krvp n_mstpi melkpi krvpi
 dklei dveen dzand dloss dakk dgraas dhok n_mapha nd_mapha frtnum frtjaar
 periode gatcode dum91 to dum97
 /STATISTICS=MEAN STDDEV MIN MAX .
```

## B1.3 De VRKMST.LIM-file

```
read;file=e:\personal\panel\vrkmst.wk1;names;format=wks$
save;file=e:\personal\panel\data\vrkmlim.sav$

/* eerst de berekeningen voor mais                                    */
load;file=e:\dataext\panel\data\vrkmlim.sav$
dstat; rhs=n_mais,ND_mais,NK_mais$
reject; n_mais <=0$
create; NK_maish=NK_mais/hamais
        ; ND_maish=ND_mais/hamais
        ; ND_mperc=ND_mais/N_mais$
reject; nk_maish > 500$
reject; nk_maish < 0$
dstat; rhs=nk_maish,nd_mperc,N_mstpi$
reject; n_mstp <0$
create; trend=jaar-1990$

regress;lhs=NK_maish;rhs=ND_maish,hatot,N_mstpi;
 str=num; panel$

regress;lhs=NK_maish;rhs=ND_maish,hatot,N_mstpi,trend;
 str=num; panel$

regress;lhs=NK_maish;rhs=ND_maish,dklei,dakk,hatot,N_mstpi,trend;
 str=num; panel$

regress;lhs=NK_maish;rhs=ND_maish,dklei,dakk,hatot,N_mstpi,trend;
 str=num; panel; random$

regress;lhs=NK_maish;rhs=ND_maish,dklei,dakk,hatot,N_mstpi;
 str=num; period=jaar; panel; random$

regress;lhs=NK_maish;rhs=ND_maish,dklei,dakk,hatot,N_mstpi,trend;
 str=num; panel; random; wts=weging$
```

## B1.4 De KOPPEL_BIN; KOPPEL_LBT-files

KOPPEL_BIN.VAR
KOPPEL_BIN97.BDP
KOPPEL_BIN.SPS
KOPPEL_LBT.VAR

---

KOPPEL_LBT97.BDP
KOPPEL_LBT.SPS

### ! hieronder staat de file KOPPEL_BIN.VAR beschreven

```
bestand b&&1 = 'bkh&&1L'
bestand b&&2 = 'bkh&&2L'

BASIS b&&1
voeg toe b&&2 gebruik b&&1.pl

WEGING=b&&1.1.3.71
BEDRIJF=bf_dic_real_nummer(sleutel, 'bkh&&1')
BEDRIJF1=bf_dic_real_nummer(sleutel, 'bkh&&2')
JAAR=19&&1 als [&&1>50] anders 20&&1

rapporteer file 'temp$$dir:koppel_bin&&1.bin' REAL*4
```

---

! hieronder staat de file KOPPEL_BIN97.BDP beschreven

```
@koppel_bin.var 97 96
```

---

**! hieronder staat de file KOPPEL_BIN.SPS beschreven**

```
define koppel ()
 weging num num_j jaar (4rb4).
!enddefine.

file handle kopbin95 / name='e:\personal\panel\data\koppel_bin95.bin' / mode=binary lrecl=16 .
data list file=kopbin95 notable/ koppel.
xsave outfile='e:\personal\panel\data\koppel_bin95.sav' .
execute.

file handle kopbin96 / name='e:\personal\panel\data\koppel_bin96.bin' / mode=binary lrecl=16 .
data list file=kopbin96 notable/ koppel.
xsave outfile='e:\personal\panel\data\koppel_bin96.sav' .
execute.

file handle kopbin97 / name='e:\personal\panel\data\koppel_bin97.bin' / mode=binary lrecl=16 .
data list file=kopbin97 notable/ koppel.
xsave outfile='e:\personal\panel\data\koppel_bin97.sav' .
execute.

add files /file='e:\personal\panel\data\koppel_bin95.sav'/file='e:\personal\panel\data\koppel_bin96.sav'
 /file='e:\personal\panel\data\koppel_bin97.sav'.
execute.
```

```
sort cases by num jaar.
do if (num ne num_j and num_j ne 0).
compute numold=num_j.
print/ num num_j jaar.
else if (num eq num_j and num_j eq lag(num) and lag(numold) gt 0).
compute numold=lag(numold).
print/ num num_j numold jaar.
end if.
execute.

recode numold (missing=-999).
descrip var=numold.

do if (numold gt 0).
compute newnum=numold.
else.
compute newnum=num.
end if.

descrip var=num newnum numold.
```

---

**! hieronder staat de file KOPPEL_LBT.VAR beschreven**

```
bestand m&&1 = 'mei&&1'
bestand m&&2 = 'mei&&2'

BASIS m&&1
KOPPEL m&&2

BEDRIJF=m&&1.1.1.1.4
BEDRIJF1=m&&2.1.1.1.4
JAAR=19&&1

rapporteer
file 'temp$$dir:koppel_lbt&&1.bin' REAL*4
```

---

**! hieronder staat de file KOPPEL_LBT97.BDP beschreven**

```
@koppel_lbt.var 97 96
```

---

**! hieronder staat de file KOPPEL_LBT.SPS beschreven**

```
define koppel ()
 num num_j jaar (3rb4).
!enddefine.

file handle mei95 / name='e:\personal\panel\data\koppel_lbt95.bin' / mode=binary lrecl=12 .
data list file=mei95 notable/ koppel.
xsave outfile='e:\personal\panel\data\koppel_lbt95.sav' .
execute.
```

```
file handle mei96 / name='e:\personal\panel\data\koppel_lbt96.bin' / mode=binary lrecl=12 .
data list file=mei96 notable/ koppel.
xsave outfile='e:\personal\panel\data\koppel_lbt96.sav' .
execute.

file handle mei97 / name='e:\personal\panel\data\koppel_lbt97.bin' / mode=binary lrecl=12 .
data list file=mei97 notable/ koppel.
xsave outfile='e:\personal\panel\data\koppel_lbt97.sav' .
execute.

add files  /file='e:\personal\panel\data\koppel_lbt95.sav'/file='e:\personal\panel\data\koppel_lbt96.sav'
 /file='e:\personal\panel\data\koppel_lbt97.sav'.
execute.

sort cases by num jaar.
do if (num ne 0 and num ne num_j and num_j ne 0).
compute numold=num_j.
print/ num num_j jaar.
else if (num eq num_j and num_j eq lag(num) and lag(numold) gt 0).
compute numold=lag(numold).
print/ num num_j numold jaar.
end if.
execute.

descrip var=numold.

do if (numold gt 0).
compute newnum=numold.
else.
compute newnum=num.
end if.

descrip var=num newnum numold.
```

## B1.5 De VRKMST_DNUM-files

VRKMST_DNUM.SPS
VRKMST_DNUM.LIM
VRKMST_DNUM.OUT

---

**! hieronder staat de file VRKMST_DNUM.SPS beschreven**

```
get file='e:\personal\panel\data\vrkmst.sav'.

select if (n_mais gt 0).
compute NK_maish=NK_mais/hamais.
compute ND_maish=ND_mais/hamais.
compute ND_mperc=ND_mais/N_mais.
select if (nk_maish le 500).
select if (nk_maish ge 0).
select if (n_mstp gt 0).

sort cases by num jaar.
do if ($casenum eq 1).
compute frtnum=1.
compute frtjaar=1.
else if (num ne lag(num)).
compute frtnum=lag(frtnum)+1.
compute frtjaar=1.
else.
compute frtnum=lag(frtnum).
compute frtjaar=lag(frtjaar)+1.
end if.

descrip var=frtnum.

select if (frtnum le 100).

vector dnum(100).
compute dnum(frtnum)=1.
recode dnum1 to dnum100 (missing=0).
execute.

REGRESSION
 /MISSING LISTWISE
 /STATISTICS COEFF OUTS R ANOVA
 /CRITERIA=PIN(.05) POUT(.10)
 /ORIGIN
 /DEPENDENT nk_maish
 /METHOD=ENTER nd_maish hatot n_mstpi dnum1 dnum2 dnum3 dnum4 dnum5 dnum6 dnum7 dnum8
dnum9
 dnum10 dnum11 dnum12 dnum13 dnum14 dnum15 dnum16 dnum17 dnum18 dnum19 dnum20
 dnum21 dnum22 dnum23 dnum24 dnum25 dnum26 dnum27 dnum28 dnum29 dnum30 dnum31
 dnum32 dnum33 dnum34 dnum35 dnum36 dnum37 dnum38 dnum39 dnum40 dnum41 dnum42
 dnum43 dnum44 dnum45 dnum46 dnum47 dnum48 dnum49 dnum50 dnum51 dnum52 dnum53
 dnum54 dnum55 dnum56 dnum57 dnum58 dnum59 dnum60 dnum61 dnum62 dnum63 dnum64
 dnum65 dnum66 dnum67 dnum68 dnum69 dnum70 dnum71 dnum72 dnum73 dnum74 dnum75
 dnum76 dnum77 dnum78 dnum79 dnum80 dnum81 dnum82 dnum83 dnum84 dnum85 dnum86
```

dnum87 dnum88 dnum89 dnum90 dnum91 dnum92 dnum93 dnum94 dnum95 dnum96 dnum97 dnum98 dnum99 dnum100 .

---

**! hieronder staat de file KOPPEL_LBT.LIM beschreven**

```
/* de volgende commando's berekenen het LIMDEP FEM-model en het overeenkomstige      */
/* model maar dan gespecificeerd met expliciete bedrijfsdummies. De datafile         */
/* wordt ingelezen (bevat de eerste 100 bedrijven uit de dataset, aangemaakt met     */
/* de file vrkmst_dnum.sps                                                           */

reset$
read;file=e:\persoanl\panel\data\vrkmst_dnum.wk1;names;format=wks$

regress;lhs=NK_maish;rhs=ND_maish,hatot,N_mstpi;
 str=num; panel; fixed$

regress;lhs=NK_maish;rhs=ND_maish,hatot,N_mstpi,
 dnum1,dnum2,dnum3,dnum4,dnum5,dnum6,dnum7,dnum8,dnum9,dnum10,
dnum11,dnum12,dnum13,dnum14,dnum15,dnum16,dnum17,dnum18,dnum19,dnum20,
 dnum21,dnum22,dnum23,dnum24,dnum25,dnum26,dnum27,dnum28,dnum29,dnum30,
 dnum31,dnum32,dnum33,dnum34,dnum35,dnum36,dnum37,dnum38,dnum39,dnum40,
 dnum41,dnum42,dnum43,dnum44,dnum45,dnum46,dnum47,dnum48,dnum49,dnum50,
 dnum51,dnum52,dnum53,dnum54,dnum55,dnum56,dnum57,dnum58,dnum59,dnum60,
 dnum61,dnum62,dnum63,dnum64,dnum65,dnum66,dnum67,dnum68,dnum69,dnum70,
 dnum71,dnum72,dnum73,dnum74,dnum75,dnum76,dnum77,dnum78,dnum79,dnum80,
 dnum81,dnum82,dnum83,dnum84,dnum85,dnum86,dnum87,dnum88,dnum89,dnum90,
 dnum91,dnum92,dnum93,dnum94,dnum95,dnum96,dnum97,dnum98,dnum99,dnum100$
```

---

**! hieronder staat de file VRKMST_DNUM.OUT beschreven**

```
--> reset$
--> read;file=e:\dataext\panel\vrkmst_dnum.wk1;names;format=wks$
--> regress;lhs=NK_maish;rhs=ND_maish,hatot,N_mstpi;
 str=num; panel; fixed$

+----------------------------------------------------------------------+
| OLS Without Group Dummy Variables                                    |
| Ordinary    least squares regression    Weighting variable = none    |
| Dep. var. = NK_MAISH Mean=   92.23407337    , S.D.=   54.23056810    |
| Model size: Observations =     281, Parameters =   4, Deg.Fr.=    277 |
| Residuals:  Sum of squares= 716717.1889     , Std.Dev.=       50.86675 |
| Fit:        R-squared=   .129635, Adjusted R-squared =          .12021 |
| Model test: F[  3,    277] =   13.75,    Prob value =          .00000 |
| Diagnostic: Log-L =  -1500.8152, Restricted(b=0) Log-L =   -1520.3226 |
|             LogAmemiyaPrCrt.=    7.873, Akaike Info. Crt.=     10.710 |
| Panel Data Analysis of NK_MAISH   [ONE way]                          |
|          Unconditional ANOVA (No regressors)                         |
| Source      Variation        Deg. Free.      Mean Square             |
| Between       642441.             99.          6489.30              |
| Residual      181026.            181.          1000.15              |
| Total         823467.            280.          2940.95              |
+----------------------------------------------------------------------+
+---------+--------------+----------------+--------+---------+----------+
|Variable | Coefficient  | Standard Error |t-ratio |P[|T|>t] | Mean of X|
+---------+--------------+----------------+--------+---------+----------+
 ND_MAISH -.1957432500      .35274263E-01   -5.549    .0000  128.70150
 HATOT    -.3766314334E-01  .10336436        -.364    .7159  48.703808
 N_MSTPI   .5606554811E-01  .23175422E-01    2.419    .0162 -16.784205
 Constant  120.2018776      7.7797036       15.451    .0000
```

Least Squares with Group Dummy Variables
Ordinary least squares regression Weighting variable = none
Dep. var. = NK_MAISH Mean= 92.23407337 , S.D.= 54.23056810
Model size: Observations = 281, Parameters = 103, Deg.Fr.= 178
Residuals: Sum of squares= 160811.6528 , Std.Dev.= 30.05722
Fit: R-squared= .804714, Adjusted R-squared = .69281
Model test: F[102, 178] = 7.19, Prob value = .00000
Diagnostic: Log-L = -1290.8454, Restricted(b=0) Log-L = -1520.3226
LogAmemiyaPrCrt.= 7.118, Akaike Info. Crt.= 9.921
Estd. Autocorrelation of e(i,t) -.107374

| Variable | Coefficient | Standard Error | t-ratio | P[\|T\|>t] | Mean of X |
|---|---|---|---|---|---|
| ND_MAISH | -.1200798881 | .29952001E-01 | -4.009 | .0001 | 128.70150 |
| HATOT | -.4397500241 | .40461695 | -1.087 | .2781 | 48.703808 |
| N_MSTPI | .4888396770E-01 | .20584436E-01 | 2.375 | .0182 | -16.784205 |

Test Statistics for the Classical Model

| Model | | Log-Likelihood | Sum of Squares | R-squared |
|---|---|---|---|---|
| (1) | Constant term only | -1520.32261 | .8234672646D+06 | .0000000 |
| (2) | Group effects only | -1307.48187 | .1810264352D+06 | .7801656 |
| (3) | X - variables only | -1500.81524 | .7167171889D+06 | .1296349 |
| (4) | X and group effects | -1290.84537 | .1608116528D+06 | .8047140 |

Hypothesis Tests

| | Likelihood Ratio Test | | | F Tests | | | |
|---|---|---|---|---|---|---|---|
| | Chi-squared | d.f. | Prob. | F | num. | denom. | Prob value |
| (2) vs (1) | 425.681 | 99 | .00000 | 6.488 | 99 | 181 | .00000 |
| (3) vs (1) | 39.015 | 3 | .00000 | 13.752 | 3 | 277 | .00000 |
| (4) vs (1) | 458.954 | 102 | .00000 | 7.191 | 102 | 178 | .00000 |
| (4) vs (2) | 33.273 | 3 | .00000 | 7.458 | 3 | 178 | .00010 |
| (4) vs (3) | 419.940 | 99 | .00000 | 6.215 | 99 | 178 | .00000 |

Ordinary least squares regression Weighting variable = none
Dep. var. = NK_MAISH Mean= 92.23407337 , S.D.= 54.23056810
Model size: Observations = 281, Parameters = 103, Deg.Fr.= 178
Residuals: Sum of squares= 160811.6233 , Std.Dev.= 30.05721
Fit: R-squared= .804714, Adjusted R-squared = .69281
Model test: F[102, 178] = 7.19, Prob value = .00000
Diagnostic: Log-L = -1290.8453, Restricted(b=0) Log-L = -1520.3226
LogAmemiyaPrCrt.= 7.118, Akaike Info. Crt.= 9.921
Autocorrel: Durbin-Watson Statistic = 2.18259, Rho = -.09129

| Variable | Coefficient | Standard Error | t-ratio | P[\|T\|>t] | Mean of X |
|---|---|---|---|---|---|
| ND_MAISH | -.1200799049 | .29951998E-01 | -4.009 | .0001 | 128.70150 |
| HATOT | -.4397512206 | .40461700 | -1.087 | .2786 | 48.703808 |
| N_MSTPI | .4888396504E-01 | .20584435E-01 | 2.375 | .0186 | -16.784205 |
| DNUM1 | 165.2005204 | 33.166439 | 4.981 | .0000 | .35587189E-02 |
| DNUM2 | 182.2112796 | 31.301738 | 5.821 | .0000 | .14234875E-01 |
| DNUM3 | 208.4985250 | 36.629083 | 5.692 | .0000 | .35587189E-02 |
| DNUM4 | 112.9685246 | 28.041319 | 4.029 | .0001 | .17793594E-01 |
| DNUM5 | 172.5781489 | 22.570736 | 7.646 | .0000 | .71174377E-02 |
| DNUM6 | 112.6609782 | 20.064335 | 5.615 | .0000 | .17793594E-01 |
| DNUM7 | 217.0219398 | 42.130000 | 5.151 | .0000 | .71174377E-02 |
| DNUM8 | 119.6005121 | 32.416058 | 3.690 | .0003 | .35587189E-02 |

| | | | | | |
|---|---|---|---|---|---|
| DNUM9 | 169.6237032 | 19.336890 | 8.772 | .0000 | .10676157E-01 |
| DNUM10 | 95.39097512 | 33.808084 | 2.822 | .0053 | .35587189E-02 |
| DNUM11 | 193.2096461 | 33.245750 | 5.812 | .0000 | .71174377E-02 |
| DNUM12 | 108.6723984 | 19.368987 | 5.611 | .0000 | .10676157E-01 |
| DNUM13 | 155.9451783 | 27.792303 | 5.611 | .0000 | .71174377E-02 |
| DNUM14 | 111.1019981 | 25.120613 | 4.423 | .0000 | .17793594E-01 |
| DNUM15 | 219.4796982 | 35.942082 | 6.106 | .0000 | .71174377E-02 |
| DNUM16 | 176.3632482 | 16.965901 | 10.395 | .0000 | .14234875E-01 |
| DNUM17 | 173.1463034 | 31.283089 | 5.535 | .0000 | .10676157E-01 |
| DNUM18 | 72.59936068 | 17.614395 | 4.122 | .0001 | .14234875E-01 |
| DNUM19 | 146.8818264 | 18.795184 | 7.815 | .0000 | .10676157E-01 |
| DNUM20 | 217.3707261 | 25.603835 | 8.490 | .0000 | .71174377E-02 |
| DNUM21 | 68.19475942 | 18.987722 | 3.592 | .0004 | .14234875E-01 |
| DNUM22 | 77.36895084 | 26.588841 | 2.910 | .0041 | .14234875E-01 |
| DNUM23 | 47.82351077 | 16.192835 | 2.953 | .0036 | .14234875E-01 |
| DNUM24 | 143.0922222 | 34.726627 | 4.121 | .0001 | .17793594E-01 |
| DNUM25 | 123.5501343 | 35.709458 | 3.460 | .0007 | .71174377E-02 |
| DNUM26 | 114.3025117 | 52.992872 | 2.157 | .0324 | .14234875E-01 |
| DNUM27 | 109.7762281 | 21.464731 | 5.114 | .0000 | .14234875E-01 |
| DNUM28 | 164.6381266 | 40.612829 | 4.054 | .0001 | .35587189E-02 |
| DNUM29 | 160.5162502 | 20.877494 | 7.688 | .0000 | .24911032E-01 |
| DNUM30 | 149.4778218 | 15.668611 | 9.540 | .0000 | .24911032E-01 |
| DNUM31 | 103.8477878 | 17.844605 | 5.820 | .0000 | .24911032E-01 |
| DNUM32 | 53.87388035 | 18.366091 | 2.933 | .0038 | .10676157E-01 |
| DNUM33 | 110.8977729 | 24.125243 | 4.597 | .0000 | .24911032E-01 |
| DNUM34 | 148.5645875 | 34.181341 | 4.346 | .0000 | .21352313E-01 |
| DNUM35 | 94.59453101 | 34.014087 | 2.781 | .0060 | .35587189E-02 |
| DNUM36 | 121.5088388 | 19.815040 | 6.132 | .0000 | .21352313E-01 |
| DNUM37 | 168.8431102 | 16.411729 | 10.288 | .0000 | .17793594E-01 |
| DNUM38 | 69.88875972 | 17.466641 | 4.001 | .0001 | .17793594E-01 |
| DNUM39 | 85.88041850 | 41.610759 | 2.064 | .0405 | .35587189E-02 |
| DNUM40 | 112.2467440 | 20.016451 | 5.608 | .0000 | .14234875E-01 |
| DNUM41 | 149.4098466 | 16.461564 | 9.076 | .0000 | .17793594E-01 |
| DNUM42 | 82.44065104 | 19.411369 | 4.247 | .0000 | .17793594E-01 |
| DNUM43 | 243.7531301 | 26.108839 | 9.336 | .0000 | .17793594E-01 |
| DNUM44 | 102.3273962 | 41.123378 | 2.488 | .0138 | .35587189E-02 |
| DNUM45 | 146.6655831 | 51.275332 | 2.860 | .0047 | .35587189E-02 |
| DNUM46 | 128.4381316 | 22.414657 | 5.730 | .0000 | .14234875E-01 |
| DNUM47 | 122.4084801 | 21.782706 | 5.620 | .0000 | .14234875E-01 |
| DNUM48 | 60.98234600 | 33.924004 | 1.798 | .0739 | .35587189E-02 |
| DNUM49 | 100.6124738 | 27.647905 | 3.639 | .0004 | .10676157E-01 |
| DNUM50 | 229.8737433 | 54.670992 | 4.205 | .0000 | .14234875E-01 |
| DNUM51 | 129.7587366 | 23.505628 | 5.520 | .0000 | .14234875E-01 |
| DNUM52 | 212.4263581 | 32.823444 | 6.472 | .0000 | .14234875E-01 |
| DNUM53 | 46.32512212 | 23.807711 | 1.946 | .0533 | .10676157E-01 |
| DNUM54 | 188.1640398 | 41.678219 | 4.515 | .0000 | .10676157E-01 |
| DNUM55 | 222.8129518 | 39.947772 | 5.578 | .0000 | .10676157E-01 |
| DNUM56 | 179.3434457 | 27.575990 | 6.504 | .0000 | .10676157E-01 |
| DNUM57 | 75.68802274 | 36.443712 | 2.077 | .0393 | .71174377E-02 |
| DNUM58 | 68.61087406 | 27.420424 | 2.502 | .0132 | .71174377E-02 |
| DNUM59 | 68.51688299 | 28.545594 | 2.400 | .0174 | .71174377E-02 |
| DNUM60 | 74.26339614 | 32.760074 | 2.267 | .0246 | .71174377E-02 |
| DNUM61 | 118.9830848 | 28.718020 | 4.143 | .0001 | .71174377E-02 |
| DNUM62 | 93.12439170 | 58.942037 | 1.580 | .1159 | .35587189E-02 |
| DNUM63 | 155.6291597 | 30.817360 | 5.050 | .0000 | .35587189E-02 |
| DNUM64 | 162.8322440 | 31.525540 | 5.165 | .0000 | .71174377E-02 |
| DNUM65 | 165.0401695 | 28.802106 | 5.730 | .0000 | .71174377E-02 |
| DNUM66 | 60.96842667 | 24.483865 | 2.490 | .0137 | .71174377E-02 |
| DNUM67 | 146.4130519 | 22.701328 | 6.450 | .0000 | .71174377E-02 |
| DNUM68 | 55.61385176 | 23.186887 | 2.399 | .0175 | .71174377E-02 |
| DNUM69 | 176.3016063 | 34.821409 | 5.063 | .0000 | .71174377E-02 |
| DNUM70 | 208.4763079 | 50.469308 | 4.131 | .0001 | .35587189E-02 |
| DNUM81 | 48.87234977 | 31.454271 | 1.554 | .1220 | .35587189E-02 |
| DNUM72 | 161.6842332 | 32.829996 | 4.925 | .0000 | .35587189E-02 |
| DNUM73 | 100.1391630 | 70.936672 | 1.412 | .1598 | .35587189E-02 |
| DNUM74 | 165.0442658 | 48.268392 | 3.419 | .0008 | .35587189E-02 |

| | | | | | |
|---|---|---|---|---|---|
| DNUM71 | 158.2934688 | 34.158088 | 4.634 | .0000 | .35587189E-02 |
| DNUM82 | 93.77243958 | 22.454145 | 4.176 | .0000 | .10676157E-01 |
| DNUM83 | 175.0729894 | 32.527544 | 5.382 | .0000 | .71174377E-02 |
| DNUM84 | 168.5063430 | 43.082149 | 3.911 | .0001 | .35587189E-02 |
| DNUM85 | 80.40649812 | 46.909346 | 1.714 | .0883 | .35587189E-02 |
| DNUM86 | 162.2819682 | 39.903295 | 4.067 | .0001 | .35587189E-02 |
| DNUM87 | 210.0966236 | 30.479963 | 6.893 | .0000 | .35587189E-02 |
| DNUM88 | 101.8302654 | 27.956233 | 3.642 | .0004 | .71174377E-02 |
| DNUM89 | 56.51281820 | 24.056124 | 2.349 | .0199 | .71174377E-02 |
| DNUM90 | 147.1008568 | 25.981974 | 5.662 | .0000 | .71174377E-02 |
| DNUM91 | 97.26824949 | 20.277027 | 4.797 | .0000 | .10676157E-01 |
| DNUM92 | 156.4335291 | 31.465170 | 4.972 | .0000 | .35587189E-02 |
| DNUM93 | 82.97903448 | 22.181312 | 3.741 | .0002 | .71174377E-02 |
| DNUM94 | 147.8127452 | 20.108443 | 7.351 | .0000 | .10676157E-01 |
| DNUM95 | 106.0958866 | 23.940094 | 4.432 | .0000 | .14234875E-01 |
| DNUM96 | 100.4619126 | 29.651558 | 3.388 | .0009 | .10676157E-01 |
| DNUM97 | 150.3106438 | 29.946512 | 5.019 | .0000 | .10676157E-01 |
| DNUM98 | 96.65366185 | 58.157491 | 1.662 | .0983 | .14234875E-01 |
| DNUM99 | 141.1977546 | 21.510433 | 6.564 | .0000 | .10676157E-01 |
| DNUM100 | 90.74915174 | 28.045697 | 3.236 | .0014 | .17793594E-01 |

## B1.6 De PRODFUN.LIM-file

```
reset$
read ; file=e:\PERSONAL\panel\DATA\cost_panel.wk1 ; Format=WKS; names$

dstat; rhs=num,jaar,outph,arb,landh,kaph,varinph$

create; lo=log(outph)
      ; la=log(arb)
      ; ll=log(landh)
      ; lk=log(kaph)
      ; lv=log(varinph)
      ; lj=jaar-1984$

regr; lhs=lo; rhs=one,la,ll,lk,lv,lj$

save; file=e:\dataext\panel\prodfun.sav$
load; file=e:\dataext\panel\prodfun.sav$

dstat; rhs=ind,lout,la,ll,lk,lv,lj$

regr; lhs=lo; rhs=la,ll,lk,lv,lj; str=num ; panel$

regr; lhs=lo; rhs=la,ll,lk,lv,lj; str=num ; panel; random$
create; b1=b(1)
      ; b2=b(2)
      ; b3=b(3)
      ; b4=b(4)
      ; b5=b(5)
      ; schaal=b1+b2+b3+b4+b5$
dstat; rhs=b1,b2,b3,b4,b5,schaal$

regr; lhs=lo; rhs=la,ll,lk,lv; str=num ; period=jaar; panel; fixed; output=2$
matrix;list;alpha$
regr; lhs=lo; rhs=la,ll,lk,lv; str=num ; period=jaar; panel; random$

create; b1=b(1)
      ; b2=b(2)
      ; b3=b(3)
      ; b4=b(4)
      ; b5=b(5)
      ; schaal=b1+b2+b3+b4$
dstat; rhs=b1,b2,b3,b4,schaal$

reject; ni<7$

regr; lhs=lo; rhs=la,ll,lk,lv; str=num ; panel; fixed$
regr; lhs=lo; rhs=la,ll,lk,lv,lj; str=num ; panel; random$
regr; lhs=lo; rhs=la,ll,lk,lv,lj; str=num ; panel; fixed$
regr; lhs=lo; rhs=la,ll,lk,lv,lj$
regr; lhs=lo; rhs=la,ll,lk,lv,lj; str=num ; panel$
regr; lhs=lo; rhs=la,ll,lk,lv,lj; str=num; period=lj ; panel$
```

# Bijlage 2 Methodologische achtergronden

## B2.1 Toelichting op de begrippen 2SLS, endogeniteit en instrument variabelen

Bij de bespreking van de Hausman-Taylor-schattingsmethode (HT) in paragraaf 2.4 zijn de begrippen 2SLS, endogeniteit en instrumentvariabelen genoemd. Deze begrippen behoren wellicht niet (meer) tot de parate kennis van LEI-medewerkers en daarom wordt in deze bijlage hierover een beknopte uitleg gegeven. Deze uitleg is niet uitputtend en voor een fundamentelere aanpak wordt verwezen naar standaard econometrie tekstboeken zoals Pindyck en Rubinfeld (1991), Gujarati (1995), Johnston en DiNardo (1997) en Greene (2000).

*Simultane vergelijkingen en identificatie*

Een standaard aanname in econometrische methoden als OLS is dat de verklarende slechts de te verklaren variabele *endogeen* is terwijl de verklarende variabelen alle *exogeen* zijn. De term exogeen houdt in dit verband in dat de waarden van deze variabelen vooraf gedetermineerd zijn en dus niet door het model bepaald worden. Bijvoorbeeld de hoeveelheid regenval is in economische modellen exogeen omdat het gemodelleerde proces (bijvoorbeeld een productiefunctie) op geen enkele manier de regenval zal beïnvloeden. In veel economische modellen komen meerdere endogene variabelen voor. Neem het bekende voorbeeld van vraag en aanbod van een bepaald goed. We weten dat de prijs van een goed wordt bepaald door het marktevenwicht dat weer het resultaat is van vraag- en aanbodrelaties. Een vraag- en aanbodmodel bestaat dan ook uit twee (of eigenlijk drie) vergelijkingen [1]:

$$\begin{aligned} &\textit{Vraagvergelijking}: && Q_t^d = \alpha_0 + \alpha_1 P_t + u_{1t} \quad \alpha_1 < 0 \\ &\textit{Aanbodvergelijking}: && Q_t^s = \beta_0 + \beta_1 P_t + u_{2t} \quad \beta_1 > 0 \\ &\textit{Evenwichtconditie}: && Q_t^d = Q_t^s \end{aligned} \tag{B1}$$

Dit systeem van vergelijkingen bepaalt de vraag, het aanbod en de prijs. $Q_t^d, Q_t^s$ en $P_t$ zijn dus endogeen in dit systeem. De vergelijkingen in het bovengenoemde systeem worden structurele vergelijkingen genoemd: ze beschrijven de relaties tussen economische variabelen volgens de economische theorie. In structurele vergelijkingen kunnen zowel endogene als exogene variabelen voorkomen en dus kunnen we ze niet met OLS schatten.

Wanneer we geïnteresseerd zijn in alle structurele parameters van het systeem (dus $\alpha_1$ en $\beta_1$) dan moeten we de structurele vergelijkingen (door middel van substitutie) omschrijven naar *gereduceerde* vergelijkingen. In gereduceerde vergelijkingen zijn alle

[1] De in deze bijlage gebruikte voorbeelden van simultane vergelijkingen zijn afkomstig van Gujarati (1995).

verklarende variabelen exogeen. In hoeverre structurele parameters geïdentificeerd kunnen worden hangt af van de hoeveelheid exogene informatie in het model.

In het bovenstaande model zijn alle variabelen endogeen en is er dus geen exogene informatie aanwezig. Herschrijving van het model met behulp van de evenwichtconditie levert de volgende gereduceerde vergelijkingen:

$$P_t = \Pi_0 + v_t \qquad \textit{waarin}$$
$$\Pi_0 = \frac{\beta_0 - \alpha_0}{\alpha_1 - \beta_1} \quad en \quad v_t = \frac{u_{2t} - u_{1t}}{\alpha_1 - \beta_1} \tag{B2}$$

Substitutie van de uitdrukking voor *Pt* in één van de structurele vergelijkingen levert

$$Q_t = \Pi_1 + w_t \qquad \textit{waarin}$$
$$\Pi_1 = \frac{\alpha_1\beta_0 - \alpha_0\beta_1}{\alpha_1 - \beta_1} \quad en \quad w_t = \frac{\alpha_1 u_{2t} - \beta_1 u_{1t}}{\alpha_1 - \beta_1} \tag{B3}$$

Er zijn nu twee gereduceerde vergelijkingen (één voor $P_t$ en één voor $Q_t$) terwijl alle vier de parameters in elk van deze vergelijkingen voorkomen. In een dergelijk systeem is het onmogelijk om unieke schattingen voor om het even welke van deze parameters te verkrijgen. In de econometrische literatuur wordt dit *under-identification* genoemd.

In veel gevallen is er wel exogene informatie aanwezig. Neem het volgende vraag- en aanbodmodel.

$$\begin{aligned} &\textit{Vraagvergelijking}: && Q_t^d = \alpha_0 + \alpha_1 P_t + \alpha_2 I_t + u_{1t} && \alpha_1 < 0, \alpha_2 > 0 \\ &\textit{Aanbodvergelijking}: && Q_t^s = \beta_0 + \beta_1 P_t + u_{2t} && \beta_1 > 0 \\ &\textit{Evenwichtconditie}: && Q_t^d = Q_t^s \end{aligned} \tag{B4}$$

De vraag naar een bepaald goed hangt nu niet alleen af van de prijs maar ook van het inkomen dat exogeen verondersteld wordt. De gereduceerde vergelijkingen zijn nu als volgt.

$$P_t = \Pi_0 + \Pi_1 I_t + v_t \qquad \textit{waarin}$$
$$\Pi_0 = \frac{\beta_0 - \alpha_0}{\alpha_1 - \beta_1}, \quad \Pi_1 = -\frac{\alpha_2}{\alpha_1 - \beta_1} \; en \; v_t = \frac{u_{2t} - u_{1t}}{\alpha_1 - \beta_1} \tag{B5}$$

en voor Qt

$$Q_t = \Pi_2 + \Pi_3 I_t + w_t \qquad \textit{waarin}$$
$$\Pi_2 = \frac{\alpha_1\beta_0 - \alpha_0\beta_1}{\alpha_1 - \beta_1}, \quad \Pi_3 = -\frac{\alpha_2\beta_1}{\alpha_1 - \beta_1} en \quad w_t = \frac{\alpha_1 u_{2t} - \beta_1 u_{1t}}{\alpha_1 - \beta_1} \tag{B6}$$

In dit systeem zijn er vijf parameters terwijl er vier vergelijkingen zijn. Dit doet vermoeden dat (nog steeds) niet alle structurele parameters uniek bepaald kunnen worden. Na substitutie kan echter worden afgeleid dat

$$\beta_0 = \Pi_2 - \beta_1 \Pi_0$$

$$\beta_1 = \frac{\Pi_3}{\Pi_1}$$

Hieruit blijkt dat de parameters van de aanbodfunctie wel uniek bepaald kunnen worden terwijl die van de vraagfunctie *under-identified* blijven.

Tenslotte bestaat de mogelijkheid dat vergelijkingen over-geïdentificeerd zijn. Dit wordt geïllustreerd in het volgende voorbeeld.

$$\begin{aligned}
&\textit{Vraagvergelijking}: && Q_t^d = \alpha_0 + \alpha_1 P_t + \alpha_2 I_t + \alpha_3 R_t + u_{1t} && \alpha_1 < 0, \alpha_2 > 0, \alpha_3 > 0 \\
&\textit{Aanbodvergelijking}: && Q_t^s = \beta_0 + \beta_1 P_t + \beta_2 P_{t-1} + u_{2t} && \beta_1 > 0, \beta_2 > 0 \\
&\textit{Evenwichtconditie}: && Q_t^d = Q_t^s
\end{aligned} \qquad \text{(B7)}$$

$$\begin{aligned}
&P_t = \Pi_0 + \Pi_1 I_t + \Pi_2 R_t + \Pi_3 P_{t-1} + v_t \qquad \textit{waarin} \\
&\Pi_0 = \frac{\beta_0 - \alpha_0}{\alpha_1 - \beta_1}, \quad \Pi_1 = -\frac{\alpha_2}{\alpha_1 - \beta_1}, \\
&\Pi_2 = -\frac{\alpha_3}{\alpha_1 - \beta_1}, \quad \Pi_3 = \frac{\beta_2}{\alpha_1 - \beta_1} \quad \textit{en} \quad v_t = \frac{u_{2t} - u_{1t}}{\alpha_1 - \beta_1}
\end{aligned} \qquad \text{(B8)}$$

$$\begin{aligned}
&Q_t = \Pi_4 + \Pi_5 I_t + \Pi_6 R_t + \Pi_7 P_{t-1} + w_t \qquad \textit{waarin} \\
&\Pi_4 = \frac{\alpha_1 \beta_0 - \alpha_0 \beta_1}{\alpha_1 - \beta_1}, \quad \Pi_5 = -\frac{\alpha_2 \beta_1}{\alpha_1 - \beta_1}, \\
&\Pi_6 = -\frac{\alpha_3 \beta_1}{\alpha_1 - \beta_1}, \quad \Pi_7 = \frac{\alpha_1 \beta_2}{\alpha_1 - \beta_1} \quad \textit{en} \quad w_t = \frac{\alpha_1 u_{2t} - \beta_1 u_{1t}}{\alpha_1 - \beta_1}
\end{aligned} \qquad \text{(B9)}$$

In dit model zijn er 7 structurele parameters terwijl de reduceerde vergelijkingen (B8) geschatte parameters opleveren. Als gevolg hiervan kunnen niet alle structurele parameters uniek bepaald kunnen worden. Dit kan eenvoudig gedemonstreerd worden.

$$\beta_1 = \frac{\Pi_6}{\Pi_2} \quad \textit{maar ook} \quad \beta_1 = \frac{\Pi_5}{\Pi_1}$$

Er is geen enkele garantie dat deze twee uitkomsten hetzelfde zijn. Omdat $\beta_1$ in alle parameters van de gereduceerde vergelijkingen voorkomt, wordt deze ambiguïteit overgedragen op alle andere structurele parameters.

Het herschrijven van structurele vergelijkingen naar gereduceerde vergelijkingen en deze met OLS te schatten staat bekend onder de naam *Indirect Least Squares* (ILS).

In de praktijk worden meestal zogenaamde *single-equation* technieken toegepast omdat we niet in alle structurele parameters geïnteresseerd zijn. Dus in plaats van het hele systeem van vergelijkingen te beschouwen beperken we ons tot slecht één vergelijking uit het systeem. Of we voor de desbetreffende vergelijking alle parameters kunnen identificeren hangt af van de exogene informatie die in het model aanwezig is: is het aantal exogene variabelen minder dan het aantal endogene dan lukt dit niet (*under-identification*). Zijn er meer exogene dan endogene variabelen dan leidt ILS niet tot unieke schattingen voor de structurele parameters (*over-identification*). In dergelijke gevallen kan *Two-Stage-Least-Squares* (2SLS) uitkomst bieden.

Neem het volgende voorbeeld.

$$\begin{aligned} &\textit{Vraagvergelijking}: && Q_t^d = \alpha_0 + \alpha_1 P_t + \alpha_2 I_t + \alpha_3 R_t + u_{1t} && \alpha_1 < 0, \alpha_2 > 0, \alpha_3 > 0 \\ &\textit{Aanbodvergelijking}: && Q_t^s = \beta_0 + \beta_1 P_t + u_{2t} && \beta_1 > 0 \\ &\textit{Evenwichtconditie}: && Q_t^d = Q_t^s \end{aligned} \qquad \text{(B10)}$$

Stel dat onze belangstelling uitgaat naar de structurele parameters van de aanbodvergelijking. Uit het voorgaande weten we dat de aanbodvergelijking over-geïdentificeerd is en ILS niet leidt tot unieke parameterschattingen. Stel nu dat we een proxy $\hat{Q}_t^d$ vinden voor $Q_t^d$ die niet gecorreleerd is met de storingsterm $u_{2t}$. Nu zijn er in het model twee kandidaten voor deze proxy: $I_t$ en $R_t$. Deze zijn immers exogeen en dus per definitie niet gecorreleerd met de storingsterm $u_{2t}$. Bovendien kunnen we op grond van hun aanwezigheid in het model ook verwachten dat ze gecorreleerd zijn met de endogene variabelen. In plaats van te kiezen voor of de één of de ander wordt in de praktijk een composiet gebruikt die gemaakt wordt door middel van de volgende regressie:

$$Q_d' = \Pi_0 + \Pi_1 I_t + \Pi_2 R_t + u_t \qquad \text{(B11)}$$

Het kan aangetoond worden dat deze composiet een hogere correlatie heeft met de endogene variabele dan elk van de instrumenten afzonderlijk. De kwaliteit van een instrument neemt toe als deze sterker gecorreleerd is met de endogene variabele.

De geschatte waarden voor $Q_t^d$ zijn nu gegeven als

$$\hat{Q}_d' = \hat{\Pi}_0 + \hat{\Pi}_1 I_t + \hat{\Pi}_2 R_t$$

ofwel

$$Q_d' = \hat{Q}_d' + \hat{u}_t$$

Omdat It en Rt strikt exogene variabelen zijn is $\hat{Q}_d^t$ niet gecorreleerd met $\hat{u}_t$. Substitutie van $\hat{Q}_d^t$ in (B10) levert

$$Q_s^t = \beta_0 + \beta_1(\hat{Q}_d^t + \hat{u}_t) + u_{2t}$$
$$Q_s^t = \beta_0 + \beta_2\hat{Q}_d^t + u_t^*$$
$$waarin\ u_t^* = u_{2t} + \beta_2\hat{u}_t$$

Omdat $\hat{Q}_d^t$ niet gecorreleerd is met $u_t^*$ kan deze aanbodvergelijking gewoon met OLS geschat worden. Het kan aangetoond kan worden dat in exact geïdentificeerde modellen ILS en 2SLS exact dezelfde parameterschattingen opleveren. Het voordeel van 2SLS ten opzichte van ILS is dat ook in het geval van over-identificatie unieke schattingen van structurele parameters verkregen kunnen worden. De 'kwaliteit' van de schattingen wordt in sterke mate bepaald door de fit ($R^2$) van de instrumentvariabele: wanneer de exogene variabelen in het model een groot deel van de variantie van $Q_d^t$ verklaren ($R^2$>0.80) dan zullen de 2SLS schattingen dicht bij de werkelijke waarden van de structurele parameters liggen.

*Relevante maar niet geciteerde literatuur*

Cornwell, C., P. Schmidt and D. Wyhowski, Simultaneous Equations and Panel Data. *Journal of Econometrics* 51 (1992) pp. 151-181, 1992.

Breusch, T.S., G.E. Mizon and P. Schmidt, Efficient Estimation Using Panel Data. *Econometrica*, Vol 57, No. 3 (May, 1989), 1989.

### B2.2 De structuur van de co-variantiemartrix in een error-component model

Een belangrijke aanname voor OLS is dat de storingsterm $\varepsilon_{it}$ normaal verdeeld is met verwachtingswaarde nul en met constante variantie $\sigma_\varepsilon^2$. Bovendien wordt aangenomen dat de storingstermen onderling niet gecorreleerd zijn. Componenten van storingstermen worden vaak weergegeven in een co-variantiematrix. Als wordt voldaan aan deze aannames voor OLS ziet deze er als volgt uit:

$$\Omega = \begin{bmatrix} \sigma_i^2 & 0 & 0 & 0 \\ 0 & \sigma_i^2 & 0 & 0 \\ 0 & 0 & \sigma_i^2 & 0 \\ 0 & 0 & 0 & \sigma_i^2 \end{bmatrix}$$

Vaak worden aannames met betrekking tot de storingsterm genoteerd als $E(\varepsilon\varepsilon') = \sigma^2 I$. Wanneer niet aan deze aannames wordt voldaan heeft dit consequenties voor de eigenschappen (lees: kwaliteit) van de OLS-schatter.

Neem het volgende model:

$$Y_{it} = \beta X_{it} + \gamma Z_i + u_i + \varepsilon_{it} \qquad \text{(B12)}$$

In een REM specificatie wordt $u_i$ opgenomen in de storingsterm die er dan als volgt uit ziet:

$$v_{it} = u_i + \varepsilon_{it}$$

De co-variantiematrix $\Omega$ voor een dergelijk model heeft een speciale structuur omdat de storingstermen in de tijd gecorreleerd zijn. Deze correlatie wordt veroorzaakt door het feit dat ui over de tijd constant blijft. De co-variantiematrix voor *bedrijf i* heeft dan de volgende structuur:

$$\Omega_i = \begin{bmatrix} \sigma_\varepsilon^2 + \sigma_u^2 & \sigma_u^2 & \sigma_u^2 \\ \sigma_u^2 & \sigma_\varepsilon^2 + \sigma_u^2 & \sigma_u^2 \\ \sigma_u^2 & \sigma_u^2 & \sigma_\varepsilon^2 + \sigma_u^2 \end{bmatrix} \qquad \text{terwijl}$$

$$\Omega = \begin{pmatrix} \Omega_i & 0 & 0 \\ 0 & \Omega_i & 0 \\ 0 & 0 & \Omega_i \end{pmatrix} \quad \because \forall\, i = 1.....N$$

Het is duidelijk dat aan de aannames van OLS niet langer voldaan wordt. In plaats daarvan kan *generalised-least-squares* (GLS) gebruikt worden. In de praktijk betekent dit dat het originele model getransformeerd wordt (net als bij HT) en vervolgens op dit model OLS toegepast wordt. De transformatie dient ervoor om weer aan de aannames voor de storingsterm te voldoen. Welke transformatie er nodig is om dit te bereiken hangt af van de structuur van de storingstermen. In het geval van FEM wordt de volgende transformatie uitgevoerd:

$$\Omega_i^{-\frac{1}{2}} y_i = \begin{bmatrix} y_{i1} - \theta y_{i.} \\ y_{i2} - \theta y_{i.} \\ y_{iT} - \theta y_{i.} \end{bmatrix} \quad \text{waarbij } \theta = 1 - \frac{\sigma_\varepsilon}{\sqrt{T\sigma_u^2 + \sigma_\varepsilon^2}}$$

Met behulp van enige matrixalgebra kan worden aangetoond dat een deze transformatie inderdaad leidt tot een model met storingstermen die voldoen aan de aannames voor OLS (dus $E(\varepsilon\varepsilon') = \sigma^2 I$). Zie bijvoorbeeld Pindyck en Rubinfeld (1991) en Greene (2000).