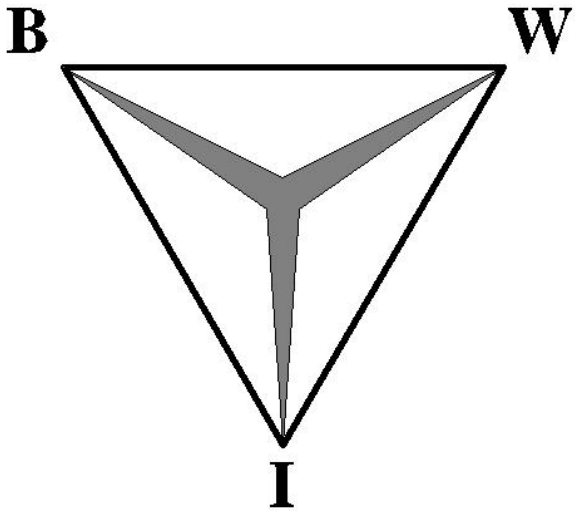
Statistische analyse van kleine datasets
Een praktijkstudie



J.F. Fokkens

BWI-werkstuk

vrije Universiteit Amsterdam
Faculteit der Exacte Wetenschappen
Divisie Wiskunde en Informatica
De Boelelaan 1081a
1081 HV Amsterdam

juni 2003

Inhoudsopgave

Voorwoord

Een onderdeel van de opleiding Bedrijfswiskunde en Informatica (BWI) aan de *vrije* Universiteit is het BWI-werkstuk. Naar aanleiding van een probleemstelling verricht de student een (literatuur-)onderzoek, waarvan de resultaten zowel schriftelijk als mondeling gepresenteerd dienen te worden. De student dient tijdens het onderzoek voldoende aandacht te besteden aan het bedrijfskundige aspect van de studie, naast de wiskunde- en informatica-aspecten.

Het onderwerp van dit werkstuk, de statistische analyse van kleine datasets, komt voort uit een actuele vraag uit de praktijk. Binnen een organisatie, die verantwoordelijk is voor het beheer van informatiesystemen, heerst de vraag of het mogelijk is om met behulp van een kleine dataset met software metrieken uitspraken te doen omtrent de onderhouds-inspanningen van deze informatiesystemen. In dit werkstuk zal getracht worden om inzicht te verschaffen in verscheidene wiskundige technieken die gebruikt kunnen worden voor de statistische analyse van (kleine) datasets. Tevens zal een analyse worden uitgevoerd op data uit de praktijk.

Tot slot wens ik nog een woord van dank uit spreken. Graag wil ik Frank Niessink danken voor zijn uiteenzetting van de praktijksituatie en zijn begeleiding bij het opzetten van het raamwerk welke ten grondslag ligt aan dit werkstuk. Tevens dank ik Geurt Jongbloed voor zijn kritische beschouwing van de statistische aspecten van dit werkstuk, zodat de aanwezigheid van theoretische onderbouwing gewaarborgd is. Beiden hebben bijgedragen tot een hogere kwaliteit van dit werkstuk.

Juni 2003,
Jan Fokkens

Samenvatting

De doelstelling van dit werkstuk is om na te gaan welke wiskundige techniek het meest geschikt is voor het analyseren van een specifieke multivariate dataset. Aan de hand van een dataset met *software metrieken* wenst de beheerder van informatiesystemen uitspraken te doen omtrent de benodigde onderhoudsinspanningen als gevolg van wijzigingsverzoeken van klanten. Deze wijzigingsverzoeken kunnen zowel betrekking hebben op problemen met het product als op gewenste nieuwe functionaliteit.

Het voornaamste kenmerk van de dataset in praktijksituatie die beschouwd is, is de beperkte hoeveelheid gegevens die beschikbaar is; slechts 19 waarnemingen. Een ander belangrijk kenmerk is dat de software metrieken in verschillende grootheden en verschillende schaaltypen uitgedrukt zijn. Deze kenmerken in ogenschouw genomen hebben geleid tot de conclusie dat een *gegeneraliseerd lineair model* het meest geschikt is vanwege het feit dat het model een goede theoretische basis heeft, rekening houdt met eventuele meetfouten (en modelfouten), geen aanname maakt omtrent de verdeling van de meetfouten en een grote verscheidenheid van mogelijke structuren voor de onderliggende relatie tussen de variabelen toelaat.

Een analyse van de beschikbare data is uitgevoerd met behulp van een gegeneraliseerd lineair model. Hierbij zijn eerst de variabelen van een wijzigingsverzoek (lees: software metrieken) die van invloed zijn op de benodigde onderhoudsinspanning bepaald. Vervolgens is de omvang van de invloed bepaald door middel van parameterschattingen.

De resultaten van het gegeneraliseerde model zijn vergeleken met de “natte vinger” schattingen die voorheen gebruikt werden. Aan de hand van de resultaten is het echter nog steeds niet duidelijk of het model ook daadwerkelijk geschikt is voor het voorspellen van onderhoudsinspanningen, hiertoe zou het model met behulp van nieuwe waarnemingen getoetst kunnen worden. Vervolgens is het aan de potentiële gebruiker van het model om te bepalen of de resultaten bevredigend genoeg zijn voor de acceptatie van het model als voorspellingsmethode.

Hoofdstuk 1

Inleiding

Deze paragraaf is grotendeels gebaseerd op [Möl'93].

Het verbeteren van de kwaliteit en prestaties van software producten, en het vergroten van de productiviteit van medewerkers aan het ontwikkelingsproces van software, is van primair belang voor vrijwel alle organisaties die afhankelijk zijn van computers. Terwijl de prestatie van computerhardware ongeveer elke drie jaar verdubbelt, zijn de verbeteringen in de prestaties van software zeer bescheiden.

De toename in de mogelijkheden van de hardware doet de behoefte aan krachtigere nieuwe software ontstaan, tevens neemt de benodigde onderhoudsinspanning voor software toe met de toename van de totale hoeveelheid software. Het ontwikkelingsproces van nieuwe software en het onderhoudsproces van oude systemen zijn echter in vele gevallen slecht geïmplementeerd. Het resultaat hiervan is dat schattingen omtrent doorlooptijd en kosten van softwareprojecten sterk afwijken van de realiteit. Vrijwel alle grote softwareprojecten zijn later gereed dan gepland, budgetten worden overschreden, en de kwaliteit van de producten is vaak onder de maat.

Het mag duidelijk zijn dat er een sterke behoefte bestaat om het softwareproces te beheersen, om dit te bereiken is het noodzakelijk een goed inzicht in het proces te verkrijgen. Om dit inzicht te verkrijgen kan gebruik gemaakt worden van kwantitatieve methoden. Deze methoden zijn gebaseerd op het argument van [DeM'82]; 'wat men niet kan meten is niet te beheersen'. De kwantitatieve gegevens die benodigd zijn voor deze methoden worden door middel van metingen aan het softwareproces onttrokken, de grootheden die kunnen worden gemeten noemen we *software metrieken*.

Wanneer de gegevens verzameld zijn wil men, door middel van analyse, voorspellingen doen omtrent bepaalde aspecten van het softwareproces. De vraag is nu welke analytische methoden voor dit doel het meest geschikt zijn. Het antwoord op deze vraag is echter afhankelijk van de situatie. Van belang zijn onder andere de hoeveelheid en de vorm van de beschikbare gegevens. In dit werkstuk zal ik trachten antwoord te geven op bovengenoemde vraag voor een specifieke praktijksituatie.

Het voornaamste kenmerk van de praktijksituatie die in ogenschouw genomen zal worden is de beperkte hoeveelheid gegevens die beschikbaar is. De gegevens hebben betrekking op wijzigingsverzoeken van een gebruiker van een specifiek software product aan de beheerder van het product. Deze wijzigingsverzoeken kunnen zowel betrekking hebben op problemen met het product als op gewenste nieuwe functionaliteit. De beheerder zou graag

aan de hand van de gegevens van een wijzigingsverzoek een voorspelling willen doen over de benodigde onderhoudsinspanningen als gevolg van het wijzigingsverzoek.

De probleemstelling van dit werkstuk kan als volgt geformuleerd worden:

> *"Is het mogelijk om, aan de hand van een kleine dataset met software metrieken, uitspraken te doen omtrent de benodigde inspanning voor software onderhoud, en welke wiskundige technieken zijn daarvoor het meest geschikt?"*

In het tweede hoofdstuk zal, voor een beter begrip van de probleemstelling, wat dieper ingegaan worden op het nut en gebruik van software metrieken.

In hoofdstuk 3 zullen een aantal analytische methoden besproken worden die gebruikt kunnen worden voor de analyse van software metrieken. De methoden zullen geëvalueerd worden op basis van de voor- en nadelen. Tevens zullen de verschillende methoden met elkaar worden vergeleken, waarbij speciale aandacht besteed wordt aan de bruikbaarheid van de techniek in geval van een beperkte hoeveelheid gegevens.

In het vierde hoofdstuk wordt de eerder genoemde praktijksituatie uitvoerig beschreven. Vervolgens zal, aan de hand van de beschikbare data, de meest geschikte methode bepaald worden waarmee de analyse wordt uitgevoerd. De bevindingen en resultaten van de analyse zullen ten slotte besproken worden.

Hoofdstuk 2

Software metrieken

Dit hoofdstuk is grotendeels gebaseerd op [Fen'95].

Voor een beter begrip van de probleemstelling van dit werkstuk zal in dit hoofdstuk dieper ingegaan worden op het begrip *software metrieken*. In het eerste hoofdstuk is al gezegd dat software metrieken kwantitatieve gegevens zijn die door middel van metingen aan een softwareproces of –product worden onttrokken. In paragraaf 2.1 wordt besproken waarom het zinvol is om deze metingen te verrichten, en wat er zoal gemeten kan worden. In paragraaf 2.2 zal aandacht besteed worden aan de eigenschappen van metrieken.

2.1. Metingen in software engineering

Software engineering is de term die gebruikt wordt om het geheel van methoden en technieken te beschrijven dat toegepast wordt bij het op economische en effectieve wijze construeren en in stand houden van grote programma's [vVl'93]. Het gaat hierbij om engineering technieken zoals het managen, budgetteren, plannen, modelleren, analyseren, ontwerpen, implementeren, testen en onderhouden van softwareproducten. Deze activiteiten werden lang gezien als de oplossing voor de in hoofdstuk 1 genoemde problematiek: slechte kwaliteit systemen die te laat en tegen te hoge kosten worden opgeleverd. Deze activiteiten zijn echter niet voldoende gebleken.

Om toch tot een oplossing te komen van het probleem hebben de engineering technieken methoden nodig die ondersteund worden door modellen en theorieën. Voor de ontwikkeling van de theorieën worden metingen gebruikt om ze toepasbaar te maken voor specifieke situaties.

2.1.1. Waarom metingen verrichten?

Er zijn, volgens [Fen'95], twee hoofdredenen waarom men metingen zou willen verrichten:

1. *assessment:* De metingen worden verricht om de voortgang van een specifiek project te beoordelen. De voortgang van een proces kan op verschillende eigenschappen beoordeeld worden, zoals: tijd, geld en kwaliteit. De metingen worden gebruikt om na te gaan of een project verloopt volgens van tevoren bepaalde planning en budget. Tevens kunnen de metingen gebruikt worden om na te gaan of gestelde kwaliteitseisen worden gehaald. Zoals duidelijk mag zijn zullen deze metingen tijdens het gehele proces verricht worden.

2. *prediction:* Hierbij zullen de metingen gebruikt worden om voorspellingen te doen over belangrijke karakteristieken van projecten. Enkele voorbeelden van karakteristieken waar men voorspellingen over zou willen doen zijn kosten, inspanning, doorlooptijd en productiviteit.

[Möl'93] specificeert de redenen waarom men metingen zou willen verrichten aan de hand van de toepassing van de metingen, er worden zes primaire toepassingsmogelijkheden genoemd:

1. *goal-setting:* Metingen kunnen gebruikt worden voor het kwantificeren van, door het management gestelde, (kwalitatieve) doelen. Het management zou bijvoorbeeld als doel kunnen hebben, het verdubbelen van de productiviteit in de komende drie jaar. Om na te gaan of dit doel daadwerkelijk gerealiseerd wordt is het wenselijk om een metriek te definiëren die de productiviteit meet. Deze metriek kan gebruikt worden om de huidige productiviteit te meten, en vervolgens kunnen er periodiek metingen verricht worden om na te gaan of de gewenste productiviteit gerealiseerd wordt.

2. *improving quality:* Metingen kunnen gebruikt worden als ondersteuning van een programma voor de verbetering van kwaliteit, bijvoorbeeld Total Quality Management (TQM). Met behulp van metingen kan nagegaan worden of aan gestelde kwaliteitseisen wordt voldaan. Indien dit niet het geval is zouden er maatregelen getroffen kunnen worden om toch tot het gewenste kwaliteitsniveau te komen.

3. *improving productivity:* Metingen kunnen ook gebruikt worden als ondersteuning van een programma voor de verbetering van productiviteit, op een zelfde manier als hierboven is beschreven voor een programma voor de verbetering van kwaliteit.

4. *project planning:* Het verzamelen van meetgegevens zal ertoe bijdragen dat de vaardigheid van het plannen van projecten zal toenemen. Historische meetgegevens zullen beschikbaar zijn voor projectmanagers. Een nieuw project kan worden vergeleken met een soortgelijk project uit het verleden zodat er gedegen voorspellingen gemaakt kunnen worden omtrent complexiteit, kosten, doorlooptijd, inspanning etc. Dit leidt ertoe dat een project beter gepland kan worden.

5. *managing:* Metrieken zijn een handig hulpmiddel bij het managen van software projecten, de voortgang van een project kan nauwkeurig in de gaten worden gehouden. Indien meetgegevens op frequente basis beschikbaar zijn kan een manager correctieve handelingen verrichten zodat de kans op een succesvolle oplevering van het project toeneemt.

6. *improving costumer confidence:* Het gebruik van metrieken, zoals in de vorige vijf toepassingen is besproken, zal (als het goed is) resulteren in een hogere kwaliteit softwaresystemen, met als gevolg een grotere tevredenheid van de klant en een toenemende mate van vertouwen in de onderneming.

Wanneer we de twee verschillende onderverdelingen van redenen om te meten vergelijken zien we dat het geheel van goal-setting, improving quality, improving productivity en managing ongeveer overeenkomt met assessment, en dat project planning ongeveer overeenkomt met prediction. Improving costumer confidence is het algemenere doel van het gebruik van metrieken, en vloeit voort uit de andere toepassingen van metrieken.

Software metrieken worden dus gebruikt om de kwaliteit van softwaresystemen te verhogen, en betere planningen te kunnen maken betreffende doorlooptijd en kosten van de ontwikkeling (en onderhoud) van softwaresystemen.

2.1.2. Metingen en metrieken

De begrippen 'metingen' en 'metrieken' zijn tot dusverre in dit werkstuk door elkaar heen gebruikt. Dit is gezien de definitie van het begrip *metriek*, die in dit werkstuk wordt gebruikt, niet geheel correct. Deze definitie luidt [Fen'95]:

> *Een getal dat is afgeleid van een product, proces of hulpbron.*

In de literatuur die over dit onderwerp verschenen is komen echter vele verschillende definities voor. Wanneer we de bovenstaande definitie hanteren, kunnen we een meting zien als een waargenomen (ofwel: gemeten) waarde van een metriek. Meten is dus het onttrekken van kwantitatieve gegevens aan een product, proces of hulpbron.

Een voorbeeld van een metriek is 'Lines of code (LOC)', het aantal regels code van een software product. Als voor een specifiek product het aantal regels code 10.000 is, noemen we deze waarde een meting.

2.1.3. Wat kunnen we meten?

Als het gaat om software zijn er drie klassen van entiteiten waarvan men attributen zou willen meten:

- *Processen* zijn software gerelateerde activiteiten, normaal gesproken met een tijdsfactor.
- *Producten* zijn programma's en documenten voortvloeiend uit processen.
- *Hulpbronnen* zijn de benodigdheden voor een proces.

Alles wat we zouden willen meten of voorspellen is een attribuut van een entiteit uit één van deze drie klassen. Een attribuut kan *direct* of *indirect* gemeten worden. Directe metingen hangen niet af van andere metingen, bijvoorbeeld 'aantal bugs gevonden tijdens testen'. Indirecte metingen hangen af van metingen van één of meerdere andere attributen, bijvoorbeeld 'aantal bugs per 1000 regels code'.

Bij attributen wordt onderscheid gemaakt tussen interne en externe attributen:

- *Interne attributen* van een product, proces of hulpbron zijn attributen die strikt afhankelijk zijn van het product, proces of hulpbron zelf.
- *Externe attributen* van een product, proces of hulpbron zijn die attributen die voortkomen uit de relatie van het product, proces of hulpbron met zijn omgeving.

In paragraaf 2.1.1 zijn er verschillende redenen besproken waarom men metingen zou willen verrichten. Welke attributen gemeten gaan worden is afhankelijk van de reden van de metingen. Aangezien het verkrijgen van meetgegevens een kostbaar proces is, is het van groot belang om de te meten attributen goed te bepalen, een handig hulpmiddel hierbij is het Goal/Question/Metric-paradigma (GQM) van Basili en Rombach (1988). In dit werkstuk zal niet ingegaan worden op de keuze van de metrieken of op het verkrijgen van meetgegevens.

2.2. Eigenschappen van metrieken

Volgens [Möl'93] zijn er een aantal technische eisen waaraan goede metrieken moeten voldoen:

1. *limited number:* Het aantal metrieken dat een individu moet bijhouden zou gelimiteerd moeten worden tot maximaal vijf. Dit zal de kwaliteit van de metingen ten goede komen.

2. *easily calculated:* De metrieken die gekozen zijn moeten gemakkelijk te berekenen zijn.

3. *readily available data:* De data die gebruikt wordt als bron voor de berekening van metrieken dient direct beschikbaar te zijn.

4. *precisely defined:* De definitie van metrieken, en berekeningswijze van de metrieken, dienen zeer precies gedefinieerd en eenduidig te zijn.

5. *tools support:* Het is belangrijk dat metrieken zo gekozen worden dat ze de al aanwezige software tools goed ondersteunen.

Het is de ontwikkelaars van software metrieken aan te raden om langere tijd zeer nauwkeurig te experimenteren met metrieken, voordat een keuze gemaakt wordt. Het is mogelijk dat metrieken die op het eerste oog significant lijken te zijn voor een proces dat eigenlijk niet zijn.

2.2.1. De representatie conditie

Software metrieken dienen behalve aan eerder genoemde eisen ook nog aan de zogenoemde *representatie conditie* te voldoen. Dit houdt in dat de toewijzing van getallen aan attributen zodanig moet geschieden dat de bestaande relaties behouden blijven. Het attribuut 'temperatuur', maakt het bijvoorbeeld wenselijk om een relatie 'warmer dan' te introduceren. Bij de toewijzing van getallen aan het attribuut 'temperatuur' van verschillende entiteiten dient de relatie 'warmer dan' in stand te blijven.

2.2.2. Schaaltypen

Software metrieken kunnen in verschillende schaaltypen uitgedrukt worden:

- *nominaal:* Denk hierbij aan het labelen of classificeren van entiteiten.
- *ordinaal:* Het aanbrengen van een rangorde in de entiteiten.
- *interval:* Reëelwaardige attributen waarop relaties zoals "twee keer zo …" niet toepasbaar zijn, maar waar wel een rangorde/volgorde in aan te brengen is. Denk hierbij aan tijd ("later dan …") of temperatuur ("warmer dan …").
- *ratio:* Reëelwaardige attributen waarop relaties zoals "twee keer zo …" wel toepasbaar zijn. Denk hierbij aan lengte of leeftijd.
- *absoluut:* Een attribuut dat iets telt, bijvoorbeeld 'het aantal gepubliceerde artikelen' van een academicus.

De keuze van het schaaltype hangt af van de metriek zelf, maar ook van de meetinstrumenten die beschikbaar zijn. Het is aan te raden om een zo hoog mogelijk schaaltype te gebruiken (ordinaal is hoger dan nominaal etc.) zonder dat de representatie conditie wordt geschonden, zodat er zoveel mogelijk relaties kunnen worden uitgedrukt.

Hoofdstuk 3

Analytische methoden

In dit hoofdstuk zullen enkele analytische methoden besproken worden die gebruikt kunnen worden voor de analyse van gegevens. Per paragraaf zal een methode worden beschreven en zal er een bespreking zijn over de voor- en nadelen van de methode. Achtereenvolgens worden principale componenten analyse, factor analyse, lineaire modellen, niet-lineaire modellen en gegeneraliseerde lineaire modellen behandeld. De eerste twee methoden zijn gekozen omdat zij in de praktijk veel gebruikt worden voor de analyse van software metrieken. De laatste drie methoden zijn gekozen op basis van het feit dat zij goed gebruikt kunnen worden als voorspellingsmodellen. Voorspellingsmodellen trachten met behulp van een aantal grootheden (verklarende variabelen) voorspellingen te doen over de waarde van een (onbekende) grootheid waarin we geïnteresseerd zijn (respons variabele).

Andere methoden zouden in theorie misschien ook in aanmerking komen, maar deze methoden hebben vaak andere kwaliteiten die ze meer geschikt maakt voor bijvoorbeeld het voorspellen van een grootheid door de tijd heen (tijdreeksen), het classificeren van gegevens (cluster analyse) of het geschikt maken van gegevens voor een visuele analyse (correspondentie analyse). Deze laatste methode zou eventueel als aanvulling op één van de bovengenoemde methoden gebruikt kunnen worden.

In de laatste paragraaf van dit hoofdstuk zullen de besproken methoden met elkaar vergeleken worden, waarbij met name aandacht besteed zal worden aan de bruikbaarheid van de methoden in geval van een beperkte hoeveelheid gegevens.

3.1. Principale componenten analyse (PCA)

Deze paragraaf is grotendeels gebaseerd op [Eve'01].

Principale componenten analyse behoort tot de oudste en meest gebruikte methoden voor de analyse van multivariate data. De methode is geïntroduceerd door Pearson (1901) en onafhankelijk daarvan door Hotelling (1933). Het idee achter de methode is om het aantal verklarende variabelen te verkleinen door middel van het introduceren van een aantal nieuwe variabelen (componenten genoemd). Deze nieuwe variabelen zijn lineaire combinaties van de oorspronkelijke variabelen, en zijn ongecorreleerd ten opzichte van elkaar. De eerste principale component wordt zo gekozen dat deze zoveel mogelijk van de variantie in de oorspronkelijke data verklaart. De tweede component wordt zo gekozen dat deze zoveel mogelijk van de resterende variantie verklaart, enz. Het doel van deze analytische methode is

om met de eerste paar componenten de oorspronkelijke data samen te vatten zonder al teveel verlies aan informatie.

3.1.1. Het model

De principale componenten zijn van de volgende vorm:

(3.1) $$Y_i = a_{i1}X_1 + a_{i2}X_2 + \ldots + a_{ip}X_p.$$

Hierbij is Y_i de i^e principale component, is $X^T = [X_1,\ldots,X_p]$ de p-dimensionale stochastische vector met de originele verklarende variabelen en is a_{ij} de coëfficiënt die bij de j^e variabele van de i^e principale component hoort.

Aangezien de variantie van Y_i simpelweg kan worden verhoogd, zonder limiet, door middel van het verhogen van de coëfficiënten $a_{i1}, a_{i2},\ldots,a_{ip}$ zal er een extra conditie moeten worden vastgesteld. Als conditie zullen we stellen dat de som der kwadraten van de coëfficiënten ($a_i^T a_i$), gelijk moet zijn aan één. Om de ongecorreleerdheid van de componenten te waarborgen moeten de sommen $a_j^T a_i$, voor $i < j$, gelijk zijn aan nul.

Om de coëfficiënten te vinden van de lineaire combinatie van de eerste component dienen we de elementen van de vector a_1 zo te kiezen dat de variantie van Y_1 maximaal is onder de gestelde conditie. De variantie van Y_1 wordt gegeven door:

(3.2) $$Var(Y_1) = Var(a_1^T X) = a_1^T \Sigma a_1,$$

waarbij Σ de covariantiematrix van X voorstelt. Om deze functie te maximaliseren kan de methode van Lagrange multipliers gebruikt worden, in welk geval de oplossing voor a_1 de eigenvector van Σ behorende bij de grootste eigenwaarde is. De andere componenten worden op gelijke wijze verkregen, waarbij a_j gegeven wordt door de eigenvector behorende bij de j^e grootste eigenwaarde. De eigenwaarden van Σ worden verkregen door de wortels te bepalen van $|\Sigma - \lambda I| = 0$. Als $\lambda_1,\lambda_2,\ldots,\lambda_p$ de eigenwaarden van Σ zijn, dan is de variantie van de i^e component gelijk aan λ_i, aangezien $a_i^T a_i = 1$. Nu kan de bijdrage van component i aan de variantie bepaald worden door:

(3.3) $$P_i = \frac{\lambda_i}{\sum_{j=1}^{p} \lambda_j}.$$

De bijdrage van de eerste q componenten aan de variantie wordt gegeven door:

(3.4) $$P_q^* = \frac{\sum_{i=1}^{q} \lambda_i}{\sum_{j=1}^{p} \lambda_j}.$$

Laat $\gamma_1, \gamma_2, ..., \gamma_p$ de genormaliseerde eigenvectoren van Σ zijn, dan definiëren we:

(3.5) $$Y_i = \gamma_i^T X,$$

zodat de Y_i's ongecorreleerd zijn en de variantie wordt gegeven door λ_i. De genormaliseerde eigenvectoren $\gamma_1, \gamma_2, ..., \gamma_p$ geven nu de vectoren met de coëfficiënten $a_1, a_2, ..., a_p$ voor de principale componenten $Y_1, Y_2, ..., Y_p$ weer ($\gamma_i = a_i$, *voor* $i = 1, ..., p$).

Wanneer we de p originele verklarende variabelen vervangen worden door de eerste q principale componenten $Y_1, ..., Y_q$, met $q < p$, dan wordt het aantal variabelen verkleind zonder al te veel verlies aan informatie. Het aantal componenten dat gekozen dient te worden in het model kan op verschillende manieren bepaald worden, bijvoorbeeld het minimum aantal componenten waarbij 80 % van de variantie verklaard wordt. Er zijn ook een aantal formelere methoden ontwikkeld voor het bepalen van het aantal toe te voegen componenten, zie onder andere [Jol'86].

Principale componenten analyse is een techniek die gebruikt als voorbereiding op vervolg analyse, zoals regressie analyse. PCA kan handig zijn als:

- er relatief veel verklarende variabelen zijn in vergelijking tot het aantal waarnemingen,
- de verklarende variabelen sterk gecorreleerd zijn.

Beide situaties leiden tot problemen bij het gebruik van regressie technieken, problemen die kunnen worden opgelost door het reduceren van het aantal verklarende variabelen tot een kleiner aantal ongecorreleerde principale componenten.

3.1.2. Bespreking model

Principale componenten analyse dient volgens [Fen'95] om de volgende redenen gebruikt te worden:

- voor het bepalen van de onderliggende dimensionaliteit van een verzameling gecorreleerde variabelen.
- voor het vervangen van een verzameling gecorreleerde variabelen door een verzameling ongecorreleerde variabelen.
- als hulp bij uitbijteranalyse.

Bij het bepalen van de onderliggende dimensionaliteit wordt verondersteld dat er aan de principale componenten een betekenis is toe te kennen. Het aantal componenten dat een relevante bijdrage levert aan het verklaren van de variantie bepaalt de dimensionaliteit. Het bepalen welke componenten relevant zijn is echter arbitrair [Eve'01], verschillende methoden leveren verschillende uitkomsten. Bij het bepalen van het aantal componenten dient met name rekening gehouden te worden met het aantal originele variabelen en het aantal waarnemingen. Wanneer het aantal componenten bijna net zo groot is als het aantal originele variabelen heeft PCA weinig zin omdat het dan nauwelijks toegevoegde waarde heeft. Het aantal componenten dient ook substantieel kleiner te zijn dan het aantal waarnemingen omdat anders het gevaar bestaat dat er “teveel waarde wordt toegekend aan” individuele waarnemingen (met name slecht in geval van uitbijters).

Het vervangen van gecorreleerde variabelen door ongecorreleerde variabelen is een goede toepassing, en maakt PCA zeer geschikt als voorbereiding op vervolganalyse. Indien de verklarende variabelen bijna ongecorreleerd zijn heeft het echter weinig zin om PCA toe te passen, omdat er slechts componenten gevonden zullen worden die weinig verschillen van de originele variabelen, slechts gerangschikt op variantie [Eve'01].

PCA is een goed hulpmiddel bij het bepalen van uitbijters. Wanneer de eerste twee componenten tegen elkaar geplot worden zal er geen relatie zijn; echter, punten die erg ver van het zwaartepunt van de punten liggen zouden als uitzonderingen beschouwd dienen te worden. Er wordt voor gekozen om de eerste twee componenten tegen elkaar te plotten omdat deze de waarnemingen voor het grootste deel verklaren.

[Eve'01] geeft twee redenen voor het gebruik van principale componenten analyse:

- er zijn relatief veel verklarende variabelen in vergelijking tot het aantal waarnemingen,
- de verklarende variabelen zijn sterk gecorreleerd.

Beide situaties leiden tot problemen bij het gebruik van regressie technieken, problemen die kunnen worden opgelost door het reduceren van het aantal verklarende variabelen tot een kleiner aantal principale componenten.

De redenen die [Eve'01] geeft voor het gebruik van PCA komen grotendeels overeen met de eerste twee redenen die [Fen'95] geeft. De eerste reden die [Eve'01] geeft komt voort uit het feit dat, wanneer het aantal verklarende variabelen relatief groot is in vergelijking met het aantal waarnemingen, het model te specifiek zal zijn voor de desbetreffende data en waarschijnlijk niet geschikt is voor voorspellingen betreffende toekomstige waarnemingen. PCA zou dus een goede methode kunnen zijn voor het analyseren van kleinere datasets. In het geval van kleine datasets zal immers vaak het aantal verklarende variabelen in vergelijking met het aantal waarnemingen groot zijn. Het gevaar is echter dat er te weinig componenten overblijven om een gedegen analyse uit te voeren; de componenten wordt te veel belang toebedacht. De kans is ook groot dat, wanneer het aantal originele variabelen klein is, het aantal componenten hier nauwelijks van zal verschillen, waardoor het toepassen van PCA weinig nut zal hebben. Het enige voordeel is dat de componenten ongecorreleerd zijn, waar de originele variabelen dat niet waren.

Een groot nadeel van PCA is dat de verklarende variabelen geen nominale of ordinale schaal mogen hebben, de opdeling in waarden die een variabele kan aannemen is immers arbitrair. Zo zal bij verschillende opdelingen de uitkomst van de analyse verschillend zijn. Ook is het bepalen van gemiddelden en varianties niet mogelijk. De verklarende variabelen mogen wel een interval-, ratio- of absolute schaal hebben. Indien de variabelen niet in dezelfde eenheid gemeten zijn kunnen de variabelen herschaald worden, of er kan gebruik worden gemaakt van de correlatiematrix in plaats van de covariantiematrix, zie [Eve'01].

Een ander groot nadeel van PCA is dat het niet gebaseerd is op een onderliggend model; er wordt geen rekening gehouden met mogelijke meetfouten c.q. residuele variantie [Eve'01]. Het gevolg is dat de componenten ook de optredende meetfouten trachten te verklaren.

We kunnen concluderen dat principale componenten analyse met name geschikt is als voorbereiding op vervolganalyse. PCA wordt dan gebruikt voor het vervangen van een verzameling gecorreleerde variabelen door een (kleinere) verzameling ongecorreleerde variabelen. Tevens kan er een uitbijteranalyse worden uitgevoerd met behulp van plots.

Voorwaarde voor het gebruik van PCA is echter wel dat de originele variabelen geen ordinale of nominale schaal hebben.

3.2. Factor analyse (FA)

In 1904 publiceerde Charles Spearman een beroemd artikel welke gezien wordt als het startpunt van *factor analyse*. Spearman onderzocht behaalde resultaten voor examens, en ontdekte bepaalde systematische effecten in de correlatiematrix tussen de behaalde resultaten van verschillende vakken. De specifieke bevindingen van dat onderzoek zullen hier verder niet besproken worden. Spearman concludeerde dat de variabelen X_i, $i = 1,...,p$ (met p het aantal vakken), bestonden uit twee delen:

(3.6) $$X_i = \lambda_i^T f + \varepsilon_i,$$

met X_i de resultaten voor examen van vak i, λ_i de vector met "factor loadings" behorende bij variabele i, f een vector van niet waargenomen random variabelen, gemeenschappelijk voor alle X_i's, en ε_i specifiek voor alle X_i's [Ken'80]. In paragraaf 3.2.1. zal de methode in zijn algemeenheid besproken worden.

Factor analyse wordt veelal gebruikt wanneer het niet mogelijk is om de variabelen waarin we geïnteresseerd zijn te meten. Voorbeelden zijn intelligentie en sociale klasse. In zulke gevallen zijn we gedwongen om informatie te ontlenen aan variabelen die wel meetbaar zijn, en indirect gerelateerd zijn aan de variabele van onze interesse.

Het factor analyse model is in essentie hetzelfde als een lineair regressie model, dat in paragraaf 3.3 besproken wordt, behalve dan dat de meetbare (verklarende) variabelen beschreven worden door een lineaire combinatie van de onbekende variabelen (ook wel factoren genoemd).

3.2.1. Het model

Deze paragraaf is grotendeels gebaseerd op [Eve'01] en [Cha'80].

Stel we hebben een stochastische vector van p variabelen $X^T = [X_1,...,X_p]$ met gemiddelden vector μ en covariantiematrix Σ. Het factor analyse model veronderstelt dat er k niet waargenomen factoren $f_1,f_2,...,f_k$, $(k < p)$ zijn, welke middels een regressie model gekoppeld zijn aan de verklarende variabelen:

(3.7) $$X_i = \lambda_{i1} f_1 + \lambda_{i2} f_2 + \ldots + \lambda_{ik} f_k + \varepsilon_i.$$

In matrix notatie:

(3.8) $$X = \Lambda f + \varepsilon.$$

We nemen aan dat de 'residuele' termen $\varepsilon_1, \ldots, \varepsilon_p$ onderling onafhankelijk zijn en ook ten opzichte van de factoren $f_1, \ldots, f_k$. Dit impliceert dat, gegeven de waarden van de factoren, de variabelen $X_1, \ldots, X_p$ onafhankelijk zijn; ze worden volledig bepaald door de factoren en de residuele termen. Bij factor analyse worden de regressie coëfficiënten in Λ ook wel *factor loadings* genoemd.

Aangezien de factoren niet zijn waargenomen kunnen we de locatie en de schaal zelf vastleggen. We zullen aannemen dat ze gemiddeld nul zijn en standaarddeviatie één hebben. Tevens zullen we aannemen dat de factoren onderling onafhankelijk zijn. Met deze aannamen geldt dat de varianties van de X_i's gegeven worden door:

$$\sigma_i^2 = \sum_j \lambda_{ij}^2 + \phi_i, \tag{3.9}$$

met ϕ_i de variantie van ε_i.

Het factor analyse model impliceert dus dat de variantie van iedere geobserveerde variabele in twee delen kan worden opgesplitst. Het eerste deel, h_i^2, dat gegeven wordt door:

$$h_i^2 = \sum_j \lambda_{ij}^2, \tag{3.10}$$

representeert de variantie die met alle variabelen gedeeld wordt door middel van de factoren. Het tweede deel, ϕ_i, is de specifieke variantie van X_i die niet gedeeld wordt met de andere variabelen. Tevens geldt als covariantie voor de variabelen X_i en X_j:

$$\sigma_{ij} = \sum_l \lambda_{il}\lambda_{jl}. \tag{3.11}$$

De covariantiematrix van X wordt dus gegeven door:

$$\Sigma = \Lambda\Lambda^T + \Phi, \tag{3.12}$$

waarbij:

$$\Phi = diag(\phi_i). \tag{3.13}$$

De factor loadings en varianties van de residuele termen zijn vrijwel altijd onbekend, en dienen geschat te worden met behulp van de beschikbare data. Methoden voor het schatten van de parameters worden beschreven in [Eve'01]. De parameters maken het mogelijk om de factor scores te berekenen, gegeven een vector waarnemingen x. Als we aannemen dat de factoren en residuele termen normaal verdeeld zijn (impliciet aannemend dat de variabelen $X_1, \ldots, X_p$ normaal verdeeld zijn) worden de factor scores gegeven door:

$$\hat{f} = \hat{\Lambda}^T \Sigma^{-1}(x - \mu). \tag{3.14}$$

Andere methoden voor het bepalen van factor scores worden beschreven in [Ren'95].

3.2.2. Bespreking model

Deze paragraaf is grotendeels gebaseerd op [Eve'01].

Factor analyse en principale componenten analyse proberen beide door middel van het verkleinen van de dimensionaliteit een verzameling gegevens te verklaren, de manieren waarop ze dit trachten te bereiken verschillen echter nogal. Enkele verschillen zijn:

- Factor analyse is gebaseerd op een onderliggend model, PCA is dat niet.
- Factor analyse tracht de covarianties of correlaties van de waarnemingen te verklaren aan de hand van een aantal gemeenschappelijke factoren. PCA is primair gebaseerd op het verklaren van de variantie van de waarnemingen.
- Wanneer het aantal componenten toeneemt, zeg van q naar $q + 1$, dan blijven de eerste q componenten ongewijzigd. Dit is niet het geval bij factor analyse, waarbij substantiële veranderingen kunnen plaatsvinden in *alle* factoren.
- Het bepalen van de parameters bij PCA is recht toe recht aan, terwijl het bepalen van de factorscores complexer is.
- Er is normaal gesproken geen relatie tussen de principale componenten bepaald aan de hand van de covariantiematrix en die bepaald aan de hand van de correlatiematrix. In geval van factor analyse zijn de resultaten bij analyse met beide matrices equivalent.

Zoals het eerste verschil al aangeeft is factor analyse gebaseerd op een onderliggend model, wat wil zeggen dat het rekening houdt met meetfouten en modelfouten, wat een pluspunt is van factor analyse ten opzichte van PCA.

Ondanks de verschillen zijn de resultaten van de beide analytische methoden vaak gelijk. Met name wanneer de varianties erg klein zijn verwachten we gelijke resultaten, omdat de residuele termen $\varepsilon_1, \ldots, \varepsilon_p$ dan niet significant zijn en het PCA model en het FA model vrijwel identiek zijn. Er dient opgemerkt te worden dat zowel factor analyse als PCA geen nut hebben indien de verklarende variabelen vrijwel ongecorreleerd zijn. In dat geval heeft factor analyse niets te verklaren, en zal PCA componenten opleveren die vrijwel gelijk zijn aan de originele variabelen.

Factor analyse heeft waarschijnlijk meer kritiek geoogst dan enige andere statistische techniek. Enkele van deze kritieken zijn de volgende:

- Zoals de kritische lezer misschien al is opgevallen bij de beschrijving van het model worden er zeer veel aannamen gemaakt aangaande de residuele termen en de factoren. Deze aannamen zijn echter in veel gevallen niet realistisch.
- Vanwege het feit dat de *factor loadings* niet uniek vastgelegd worden door het factor model kan door rotatie van de factoren een gewenst antwoord verkregen worden. De analyse kan dus beïnvloed worden door de uitvoerder van de analyse, wat in strijd is met de uitgangspositie dat de data de waarnemingen verklaart (meetfouten en modelfouten uitgezonderd).
- In veel gevallen kan men zich afvragen of het concept van onderliggende ongeobserveerde factoren acceptabel is.

Concluderend kan gezegd worden dat er nogal wat nadelen kleven aan het gebruik van factor analyse. In vergelijking met PCA zijn de te verrichten berekeningen nogal ingewikkeld. Tevens worden er een groot aantal aannamen gemaakt waarvan men zich kan afvragen of deze wel allemaal correct zijn. Een ander punt van kritiek is dat de uitvoerder van de analyse

het model kan manipuleren. Op basis van het bovengenoemde zijn er zelfs critici die suggereren dat het gebruik van factor analyse de moeite niet waard is, het feit dat factor analyse gebaseerd is op een onderliggend model dat rekening houdt met meetfouten doet hier niets aan af. Persoonlijk denk ik dat factor analyse in sommige gevallen een nuttige methode kan zijn voor het onderzoeken van karakteristieken of structuren van een multivariate dataset, er dient echter wel gekeken te worden of de aannamen die gemaakt worden realistisch zijn

3.3. Lineaire modellen

In het geval van lineaire modellen wordt verondersteld dat de waarde van de responsvariabele lineair afhangt van de waarden van een aantal verklarende variabelen, ook wel factoren of effecten genoemd. De respons wordt waargenomen met een meetfout: de bijbehorende verklarende variabelen worden meestal exact bekend verondersteld. De relatie tussen de twee soorten variabelen hangt af van de waarden van een stel onbekende parameters die geschat moeten worden [deG'98].

3.3.1. Het model

Deze paragraaf is grotendeels gebaseerd op [deG'99].

Een lineair model Ω is een statistisch model waarbij random waarnemingen $Y_1,\ldots,Y_n$ worden beschreven door een lineaire combinatie van $p + 1$ onbekende parameters $\beta_0,\ldots,\beta_p$ plus niet waarneembare random fouten $e_1,\ldots,e_n$,

$$\Omega:\begin{cases} Y_i = \beta_0 + x_{i1}\beta_1 + \ldots + x_{ip}\beta_p + e_{i,} \\ Cov(e_i, e_j) = \begin{cases} \sigma^2, & i = j \\ 0, & i \neq j \end{cases}, \\ Ee_i = 0, \end{cases} \tag{3.15}$$

voor $i,j = 1,\ldots,n$, en met de $\{x_{ij}\}$ de bekende waarden van de verklarende variabelen. Soms is het praktisch om de matrixnotatie voor Ω te gebruiken,

$$\Omega:\begin{cases} Y_i = X\beta + e \\ Cov(e) = \sigma^2 I_{n\times n}, \\ Ee = 0, \end{cases} \tag{3.16}$$

met $Y = (Y_1,\ldots,Y_n)^T$ de vector van waarnemingen, X de $n\times(p+1)$-matrix met als i^{e} rij $x_i^T = (1, x_{i1},\ldots,x_{ip}), \beta = (\beta_0,\ldots,\beta_p)^T$ de vector onbekende parameters, $I_{n\times n}$ de $n\times n$ identiteitsmatrix, en $e = (e_1,\ldots,e_n)^T$ de vector met random fouten.

Indien *rang(X) = p + 1*, dus X is van maximale rang, dan hebben we te maken met regressieanalyse. Indien *rang(X) < p + 1* en alle velden van de matrix X zijn 0 of 1, waarmee de afwezigheid of aanwezigheid van een effect wordt aangegeven, dan *kan* er gebruik gemaakt worden van variantieanalyse (ANOVA). Indien een variabele een nominale of ordinale schaal heeft kan de variabele vervangen worden door een aantal "dummy"-variabelen waardoor het gebruik van variantieanalyse toch mogelijk is. Voorbeeld: zij X een ordinale

variabele met 3 mogelijke waarden, dan kan variabele X vervangen worden door 3 "dummy"-variabelen X_1, X_2, X_3 met $X_i = 1$ ($i = 1, 2, 3$) indien X waarde i heeft, en 0 anders. Een combinatie van regressieanalyse en variantieanalyse noemen we covariantieanalyse.

[Sch'59] geeft een minder wiskundige uitleg voor het bepalen van de analytische methode. Variantieanalyse kan gebruikt worden indien alle factoren kwalitatief behandeld worden (nominale of ordinale schaal), regressieanalyse kan gebruikt worden indien alle factoren kwantitatief zijn en ook kwantitatief behandeld worden (interval-, ratio- of absolute schaal) en covariantieanalyse kan in de overige gevallen gebruikt worden, dus als er zowel kwalitatieve als kwantitatieve factoren aanwezig zijn.

Het is mogelijk om kwantitatieve factoren om te zetten naar kwalitatieve factoren. Bijvoorbeeld lichaamstemperatuur in graden Celsius zou vervangen kunnen worden door een aantal binaire variabelen, die aangeven of een persoon koorts heeft, verhoging heeft, een normale lichaamstemperatuur heeft of onderkoeld is. Op deze manier kan het model aangepast worden zodat er andere analytische methoden gebruikt zouden kunnen worden. Het omzetten van kwalitatieve factoren naar kwantitatieve factoren is niet mogelijk. Gezien de aard van de beschikbare data in mijn analyse (zowel kwalitatieve als kwantitatieve factoren) zal lineaire regressie niet bruikbaar zijn. Lineaire regressie zal in dit werkstuk dan ook niet verder behandeld worden, de geïnteresseerde lezer verwijs ik naar [deG'98]. Variantieanalyse en covariantieanalyse komen wel in aanmerking voor de mijn analyse. In het vervolg van deze paragraaf zal variantieanalyse nader besproken worden, voor behandeling van covariantieanalyse zie [Sch'59].

Beschouw het model (3.15). In dit model wordt de parametervector bepaald door $EY = X\beta$. Om de parametervector te schatten hebben we een aanname gemaakt ten aanzien van de meetfouten $e_1,\ldots,e_n$. We veronderstellen dat de meetfouten onderling onafhankelijk zijn en identiek $N(0,\sigma^2)$-verdeeld. De parametervector β kan geschat worden met behulp van de kleinste kwadraten schatter $\hat{\beta}$, die verkregen wordt door $S(\beta)$ te minimaliseren:

$$
\begin{aligned}
S(\beta) &= \sum_{i=1}^{n} (Y_i - EY_i)^2 \\
&= \sum_{i=1}^{n} (Y_i - \sum_{j=0}^{p} x_{ij}\beta_j)^2 \\
&= (Y - X\beta)^T (Y - X\beta).
\end{aligned}
\tag{3.17}
$$

Het minimaliseren van $S(\beta)$ zal hier niet verder besproken worden, zie [deG'99]. Als de matrix X niet van maximale rang is zal de oplossing niet uniek zijn. Om dit probleem op te lossen is het gebruikelijk om extra aannamen te doen ten aanzien van de parameters. Deze extra aannamen hebben geen nadelige gevolgen voor de kwaliteit van de voorspellingen die uiteindelijk met het model gedaan kunnen worden ten aanzien van de responsvariabele.

Een mooie eigenschap van variantieanalyse is dat er makkelijk rekening gehouden kan worden met interacties tussen verschillende factoren. Het gelijktijdig "optreden van twee of meerdere verschillende factoren" kan een bepaald gecombineerd effect hebben, dat eenvoudig in het model kan worden opgenomen door middel van het invoeren van een binaire variabelen die de af- en aanwezigheid van de combinatie van deze factoren aangeeft. Een model met meer dan twee factoren en zonder interacties wordt additief genoemd. Additieve modellen zijn vaak makkelijker te interpreteren, maar voordat overgegaan wordt op het gebruik van een additief model moet eerst worden nagegaan of een model met interacties niet beter is. Voor

een beschrijving van de toetsen die gebruikt kunnen worden om te bepalen welke factoren en interacties in het model moeten worden opgenomen, zie [deG'99].
In deze paragraaf hebben we aangenomen dat de waarnemingen 'random' zijn en de effecten niet. Een analyse waarbij de effecten eveneens 'random' zijn is ingewikkelder, maar kan op soortgelijke wijze uitgevoerd worden. Tevens is verondersteld dat alle combinaties van effecten tenminste één keer is waargenomen. Als dit niet het geval is zal de analyse minder eenvoudig zijn.

3.3.2. Bespreking model

Deze paragraaf is grotendeels gebaseerd op [deG'99].

Zoals in de vorige paragraaf al is opgemerkt komt regressieanalyse niet in aanmerking voor de analyse van de praktijksituatie die in hoofdstuk 4 zal worden uitgevoerd, bij deze bespreking van lineaire modellen zullen dan ook alleen variantieanalyse en covariantieanalyse beschouwd worden.

Het voornaamste verschil tussen variantieanalyse en covariantieanalyse zit in het feit dat er bij variantieanalyse slechts binaire verklarende variabelen mogen zijn, bij covariantieanalyse is dit geen vereiste. Wanneer er echter niet alleen binaire variabelen voorkomen kunnen de niet-binaire variabelen eenvoudig worden omgezet in binaire variabelen, de keuze van de domeinen van de nieuwe variabelen is echter arbitrair en kan van invloed zijn op de kwaliteit van het model. Het is aan te raden dat wanneer het aantal niet-binaire variabelen groot is (in vergelijking met het totaal aantal verklarende variabelen) te kiezen voor covariantieanalyse in plaats van de variabelen om te zetten.

Door het gebruik van binaire variabelen zal het aantal verklarende variabelen relatief groot zijn, variabelen met een ordinale of nominale schaal zullen immers uitgedrukt worden middels meerdere binaire variabelen. Tevens zal bij het omzetten van interval-, ratio- of absolute variabelen naar binaire variabelen het aantal variabelen toenemen. In het geval van een kleine dataset lijkt dit misschien problemen op te leveren omdat het aantal verklarende variabelen in vergelijking met het aantal waarnemingen groot zal zijn. Echter, de dimensionaliteit verandert niet; het aantal variabelen dat informatie bevat per waarneming is gelijk aan het aantal originele verklarende variabelen.

Het grote voordeel van variantieanalyse is dat er vrij eenvoudig rekening kan worden gehouden met gecombineerde effecten van factoren, dit kan bij covariantieanalyse alleen voor de binaire variabelen. Deze gecombineerde effecten worden interacties genoemd.

Met behulp van verschillende toetsen kan bepaald worden welke factoren en interacties van invloed zijn op de responsvariabele, en dus moeten worden meegenomen in het model. Doordat per variabele, op theoretische wijze, wordt bepaald of deze wordt meegenomen in het definitieve model hopen we dat het aantal verklarende variabelen kleiner zal zijn dan wanneer PCA of factor analyse wordt gebruikt. Met name in geval van een kleine dataset is dit een gewenst resultaat.

Bij variantieanalyse wordt aangenomen dat de responsvariabelen normaal verdeeld zijn. Normaliteit van de responsvariabelen is in veel gevallen echter niet aannemelijk. De toetsen die gebruikt worden om te bepalen welke variabelen dienen te worden opgenomen in het model gaan ook uit van deze veronderstelling, en zouden dus verkeerde resultaten kunnen opleveren.

De aanname, zoals in de vorige paragraaf vermeldt, dat iedere combinatie van effecten tenminste één keer moet zijn waargenomen zal in geval van een kleine dataset zelden

voorkomen, variantieanalyse behoort dan nog wel tot de mogelijkheden maar zal minder eenvoudig uit te voeren zijn.

3.4. Niet-lineaire modellen

Deze paragraaf is grotendeels gebaseerd op [deG'99].

In deze paragraaf zullen niet-lineaire regressie modellen besproken worden. De structuur van niet-lineaire modellen verschilt niet zo veel van de lineaire modellen die in de vorige paragraaf besproken zijn. Het grootste verschil is, zoals de naam al aangeeft, dat de verwachte waarde van een responsvariabele op niet-lineaire wijze van de verklarende variabelen afhangt.

3.4.1. Het model

Er zal voor niet-lineaire modellen een gelijksoortige notatie gebruikt worden als voor lineaire modellen: er zijn n random waarnemingen Y_i, $i = 1,\ldots,n$ en een k-dimensionale vector x_i geeft de vector van verklarende variabelen voor Y_i weer. Alle waarnemingen worden onafhankelijk verondersteld. Aangenomen wordt dat de regressie relatie tussen Y_i en x_i de som is van een systematisch deel, dat beschreven wordt door de functie f van x_i, en een random deel ε_i. De functie f hangt meestal af van een aantal onbekende parameters. We definiëren:

$$Y_i = f(x_i, \theta) + \varepsilon_i, \quad i = 1, \ldots, n, \tag{3.18}$$

met $\theta = (\theta_1, \ldots, \theta_p)^T$ de onbekende parametervector. De functie f wordt de regressiefunctie genoemd. De random variabele ε_i wordt de meetfout genoemd, en heeft verwachting 0 en variantie σ_i^2. In het geval de structuur van de onderliggende functie f bekend is, bijvoorbeeld door een fysische wet, dan zijn de ε_i's echte meetfouten, in andere gevallen geven de ε_i's de discrepantie weer tussen Y_i en $f(x_i, \theta)$ inclusief de meetfout. Het doel is om de geschikte functie f te vinden en de bijbehorende parameters te schatten. Indien de onderliggende structuur van f niet bekend is, is het verstandig om f zo te kiezen dat, door het variëren van de parameters, een brede variëteit aan mogelijke functies ontstaat.

Het schatten van de parameters wordt, net als bij lineaire modellen, veelal gedaan met de kleinste kwadraten schatter, die wordt verkregen door het minimaliseren van $S(\theta)$,

$$S(\theta) = \sum_{i=1}^{n} (Y_i - f(x_i, \theta))^2. \tag{3.19}$$

In tegenstelling tot lineaire modellen is het echter mogelijk dat $S(\theta)$ meerdere locale minima heeft naast het absolute minimum $S(\hat{\theta})$. Het minimaliseren van $S(\theta)$ zal hier niet verder worden besproken, zie [deG'99].

Modellen kunnen met elkaar vergeleken worden door middel van grafieken en/of toetsen, zie [deG'99]. Indien meerdere modellen de data even goed verklaren dan wordt het eenvoudigste model geprefereerd. Het eenvoudigste model is het model met het kleinste aantal verklarende variabelen.

3.4.2. Bespreking model

Niet-lineaire modellen zijn met name nuttig wanneer de onderliggende structuur van de relatie tussen de verklarende variabelen en de responsvariabelen bekend is, dit kan door een fysische wet of om een andere reden. Indien de structuur niet bekend is kan door een geschikte keuze van de functie *f* een zo breed mogelijk scala aan mogelijke regressiefuncties beschreven worden, waarna alleen de parameters nog bepaald moeten worden. In het geval van niet-lineaire modellen hoeven de verklarende variabelen niet dezelfde schaal te zijn. Verder gelden voor niet-lineaire modellen praktisch dezelfde voor- en nadelen als voor lineaire modellen.

3.5. Gegeneraliseerde lineaire modellen

Deze paragraaf is grotendeels gebaseerd op [deG'99].

De modellen die in de vorige twee paragrafen behandeld zijn kunnen gezien worden als uitbreidingen op lineaire regressie modellen. In 1972 hebben Nelder en Wedderburn een algemene en toegankelijke theorie ontwikkeld voor een hele klasse van soortgelijke modellen, genaamd *gegeneraliseerde lineaire modellen.*

3.5.1. Het model

Evenals in voorgaande paragrafen hebben we n onafhankelijke waarnemingen $Y_1,\ldots,Y_n$ van een responsvariabele. Het doel is om de relatie te onderzoeken tussen de responsvariabelen en de p verklarende variabelen, die bekend verondersteld worden. Een voorbeeld van een gegeneraliseerd lineair model voor het standaard lineaire model ziet er als volgt uit:

$$\begin{aligned} &(i) \quad Y_i = N(\mu_i, \sigma^2), \\ &(ii) \quad \eta_i = x_i^T \beta, \\ &(iii) \quad \eta_i = g(\mu_i) = \mu_i, \end{aligned} \tag{3.20}$$

voor $i = 1,\ldots,n$, met $\beta = (\beta_0,\ldots,\beta_p)^T$ de vector onbekende parameters. In (3.17) is het model voor $Y_1,\ldots,Y_n$ opgesplitst in een random component *(i)*, een systematische component *(ii)* en een zogenaamde linkfunctie *(iii)*. De random component van het model specificeert de verdeling van Y_i. De systematische component geeft het lineaire verband tussen de responsvariabele en de verklarende variabelen weer. De linkfunctie g specificeert het verband tussen de random- en de systematische component. Preciezer; g drukt η_i uit als een functie van EY_i, ofwel $g(EY_i) = \eta_i$. Een algemene notatie voor gegeneraliseerde lineaire modellen luidt als volgt:

$$\begin{aligned} &(i) \quad Y_i \text{ heeft kansdichtheidsfunctie } f(\cdot;\theta_i)_i, \\ &(ii) \quad \eta_i = x_i^T \beta, \\ &(iii) \quad \eta_i = g(\theta_i). \end{aligned} \tag{3.21}$$

De componenten *(i)* en *(iii)* kunnen gegeneraliseerd worden, component *(ii)* kan niet gegeneraliseerd worden, vandaar de naam *gegeneraliseerd lineair model*. De kansdichtheidsfunctie f_i kan gekozen worden uit een één-parameter exponentiele familie, deze zijn van de volgende vorm:

$$f_i(y,\theta_i,\phi) = \exp\left(\frac{y\theta_i - b(\theta_i)}{\phi} + c(y,\phi)\right). \tag{3.22}$$

De specifieke vorm van f_i wordt bepaald door de functies b en c, zo kunnen door geschikte keuze van b en c de normale, binomiale en exponentiële verdeling verkregen worden [McC'89]. Aangezien ϕ een schaalparameter is, gelijk voor alle Y_i, zijn we niet geïnteresseerd in het schatten van deze parameter. De keuze van de exponentiele familie en de linkfunctie leggen het gegeneraliseerde model vast.

De uitdrukking (3.22) kan de suggestie wekken dat er n onbekende parameters $\theta_1,\ldots,\theta_n$ geschat dienen te worden. Het is echter zo dat er slechts $p + 1$ parameters geschat dienen te worden, namelijk $\beta_0,\ldots,\beta_p$. Omdat de parameters $\beta_0,\ldots,\beta_p$ gelinkt zijn aan de parameters $\theta_1,\ldots,\theta_n$ door de vergelijkingen *(ii)* en *(iii)* en (3.22), de kansdichtheidsfuncties $f_1,\ldots,f_n$ hangen impliciet af van $\beta_0,\ldots,\beta_p$. Dit is waarom een natuurlijke methode voor het schatten van de parameters de meest aannemelijke schatters-methode is. De schatters worden bepaald door het maximaliseren van de aannemelijkheidsfunctie of de log-aannemelijkheidsfunctie, zie [Oos'97]. In veel gevallen is een expliciete uitdrukking van de aannemelijkheidsfunctie niet aanwezig, dan dienen de parameters numeriek te worden bepaald, dit kan met behulp van Fisher scoring, zie [McC'89].

3.5.2. Bespreking model

Evenals voor niet-lineaire modellen gelden voor gegeneraliseerde lineaire modellen dezelfde voordelen als voor lineaire modellen, doch enkele nadelen worden weggenomen. De twee voornaamste verschillen zijn:

- er worden, buiten de normale verdeling, andere verdelingen toegelaten voor de respons-variabele
- door middel van het toelaten van andere linkfuncties kunnen verschillende relaties tussen de systematische- en de random component beschreven worden.

De verschillen zijn tevens de grote voordelen van gegeneraliseerde lineaire modellen ten opzichte van lineaire modellen. Data die niet geschikt is voor een analyse met behulp van een lineair model kan met behulp van een gegeneraliseerd lineair model eventueel wel worden beschreven.

Enkelen eigenschappen van gegeneraliseerde lineaire modellen zijn:

- ze zijn ook bruikbaar wanneer de schaaltypen niet gelijk zijn.
- biedt de mogelijkheid om te bepalen welke variabelen van invloed zijn op de respons-variabelen, inclusief interacties.

Het laatste kan bereikt worden door toetsen uit te voeren die twee of meerdere 'geneste' modellen met elkaar vergelijken. Tevens kan, vanwege de statistische basis van het model, de betrouwbaarheid van de parameterschattingen bepaald worden, bijvoorbeeld door middel van betrouwbaarheidsintervallen. Achteraf kan een model ook nog getoetst worden op zijn 'goodness of fit'; er wordt nagegaan of de gemaakte modelaannamen correct zijn geweest.

De aanname van onafhankelijke waarnemingen kan een nadeel zijn in specifieke situaties. Indien er wel afhankelijkheid bestaat tussen de waarnemingen, bijvoorbeeld bij tijdreeksen, kunnen gegeneraliseerde lineaire modellen niet gebruikt worden.

3.6. Conclusies

In deze paragraaf zullen de voornaamste voor- en nadelen van analytische methoden die in dit hoofdstuk besproken zijn naast elkaar gezet worden, met speciale aandacht voor te toepassingsmogelijkheden van de methoden indien de hoeveelheid beschikbare gegevens beperkt is.

In paragraaf 3.1 werd principale componenten analyse besproken. Het uitgangspunt van PCA is het verkleinen van de dimensionaliteit, door de originele verklarende variabelen te vervangen door een kleiner aantal nieuwe ongecorreleerde variabelen (zonder verlies van informatie). In het geval van een kleine dataset is de kans groot dat het aantal verklarende variabelen relatief groot is ten opzichte van het aantal waarnemingen. Om deze reden lijkt PCA dus een geschikte methode omdat het tracht het aantal verklarende variabelen te reduceren; we willen immers voorkomen dat een klein aantal waarnemingen met een groot aantal variabelen verklaart wordt omdat de uitkomsten te specifiek zullen zijn voor de data en niet bruikbaar zijn voor voorspellingen over toekomstige waarnemingen.

Doordat PCA geen onderliggend model heeft en geen rekening houdt met meetfouten zal het model de meetfouten trachten te verklaren door de componenten. In het geval van een kleine dataset zullen de componenten dus zeer specifiek zijn voor de waarnemingen waarop het model is gebaseerd en zeer gevoelig voor meetfouten in de waarnemingen. PCA zal dan niet erg bruikbaar zijn wanneer we voorspellingen zouden willen doen over toekomstige waarnemingen. Vanwege het ontbreken van een onderliggend model is het tevens niet mogelijk om uitspraken te doen omtrent te betrouwbaarheid van het model. Indien het model gebaseerd is op een klein aantal waarnemingen zal deze betrouwbaarheid niet groot zijn.

PCA kan, mits de originele variabelen geen nominale- of ordinale schaal hebben, gebruikt worden als voorbereiding op vervolganalyse voor het reduceren van het aantal variabelen en het bepalen van uitbijters. In geval van een kleine dataset is PCA echter niet geschikt.

In paragraaf 3.2 is factor analyse beschouwd. Een groot voordeel van factor analyse ten opzichte van PCA is het feit dat het gebaseerd is op een onderliggend model, met name met kleine datasets is dit wel een vereiste. Factor analyse heeft echter nogal wat nadelen. Het feit dat uitkomsten gemanipuleerd kunnen worden en de grote hoeveelheid aannamen die gemaakt worden zonder dat deze verklaard worden zorgen ervoor dat factor analyse in de meeste gevallen beter niet gebruikt kan worden, zeker als er alternatieven zijn.

In paragraaf 3.3 zijn verschillende lineaire modellen behandeld, waaronder variantieanalyse en covariantieanalyse. Lineaire modellen veronderstellen dat de meetfouten normaal verdeeld zijn, wat in veel gevallen niet aannemelijk is. Ook de toetsen die bepalen welke variabelen in het model moeten worden opgenomen gaan uit van deze veronderstelling, en zullen in bepaalde gevallen dus niet al te betrouwbaar zijn. In geval van een kleine dataset is de keuze van de variabelen extra belangrijk omdat het er maar zo weinig zijn. Lineaire modellen zijn goed te gebruiken in de gevallen dat het aannemelijk is dat de meetfouten normaal verdeeld zijn gezien de goede theoretische grondslag, wanneer de meetfouten niet normaal verdeeld zijn dienen alternatieven gezocht te worden.

In paragraaf 3.4 kwamen niet-lineaire modellen aan bod. Grote voordelen ten opzichte van lineaire modellen zijn; de onderliggende structuur van de relatie tussen de verklarende variabelen en de responsvariabelen hoeft niet lineair te zijn en er wordt geen normaliteit van de meetfouten verondersteld. Het gebruik van niet-lineaire modellen heeft echter alleen zin als er een vermoeden is omtrent de onderliggende structuur. Bij kleine datasets geldt dit natuurlijk ook, maar indien er geen vermoeden is omtrent de structuur zijn niet-lineaire modellen niet geschikt.

Als laatste zijn in paragraaf 3.5 gegeneraliseerde lineaire modellen besproken. Deze modellen nemen alle belangrijke nadelen van de andere methoden weg, en zal in veel gevallen de meest geschikte methode zijn voor analyse. Het voornaamste kenmerk van de praktijksituatie zoals die in het volgende hoofdstuk aan de orde zal komen is de beperkte hoeveelheid gegevens. Vanwege het feit dat het model een goede theoretische basis heeft, rekening houdt met eventuele meetfouten (en modelfouten), geen aanname maakt omtrent de verdeling van de meetfouten en een grote verscheidenheid van mogelijke structuren voor de onderliggende relatie tussen de variabelen toelaat doen deze conclusie tot stand komen.

Hoofdstuk 4

Een praktijkstudie

In dit hoofdstuk zal een analyse worden uitgevoerd aan de hand van gegevens uit de praktijk. De praktijksituatie en de beschikbare data zullen in paragraaf 4.1 besproken worden. In paragraaf 4.2 zal aan de hand van de beschikbare data, met behulp van de bevindingen uit hoofdstuk 3, de meest geschikte methode worden gekozen voor de analyse. De uitvoering van de analyse zal vervolgens in paragraaf 4.3 behandeld worden. Tot slot zullen in paragraaf 4.4 de bevindingen en resultaten van de analyse gepresenteerd worden.

4.1. Beschrijving praktijksituatie

De gegevens die in dit hoofdstuk bekeken zullen worden zijn afkomstig van een organisatie die verantwoordelijk is voor het beheer van een aantal informatiesystemen die gebruikt worden in de Nederlandse sociale zekerheden sector. De gegevens hebben betrekking op wijzigingverzoeken van een gebruiker van een specifiek softwareproduct aan de beheerder van het product. Deze wijzigingsverzoeken kunnen zowel betrekking hebben op problemen met het product als op gewenste nieuwe functionaliteit. De beheerder zou graag aan de hand van de gegevens van een wijzigingsverzoek een voorspelling willen doen omtrent de benodigde onderhoudsinspanningen als gevolg van het wijzigingsverzoek.

De gegevens die ter beschikking zijn hebben betrekking op 19 wijzigingsverzoeken. Per wijzigingsverzoek zijn de volgende attributen bekend:

- *kwaliteit van de specificatie van de klant:* ordinale schaal met 4 voorkomende waarden; goed, voldoende, slecht, nvt
- *herkomst wijziging:* ordinale schaal met 3 voorkomende waarden; gebruikersprobleem of wens, wettelijke maatregelen, interne of externe verantwoording
- *aantal betrokken externe systemen:* absolute schaal
- *database gerelateerde wijziging:* ordinale schaal met 3 voorkomende waarden; geen, alleen raadpleging, structuur wijziging
- *aantal te wijzigen componenten:* absolute schaal

Per te wijzigen component zijn de volgende attributen bekend:

- *complexiteit van de component:* ordinale schaal met 4 voorkomende waarden; zeer hoog, hoog, gemiddeld, laag
- *omvang van de component:* ordinale schaal met 4 voorkomende waarden; zeer groot, groot, gemiddeld, klein
- *omvang van de wijziging:* ordinale schaal met 3 voorkomende waarden; een klein gedeelte, een flink gedeelte, bijna de gehele component

In appendix A zijn als voorbeeld de gegevens van enkele wijzigingsverzoeken opgenomen.

Een moeilijkheid die we tegenkomen wanneer we de data bekijken is de gelaagdheid. De hoeveelheid gegevens per wijzigingsverzoek is als gevolg hiervan niet gelijk, het aantal te wijzigen componenten is immers niet gelijk en per component worden een aantal attributen gegeven.

Een mogelijke oplossing voor dit probleem is per wijzigingsverzoek een aantal nieuwe variabelen in te voeren $[x_{111}, x_{112}, x_{113}, \ldots, x_{443}]$. Hierbij duidt de variabele x_{ijk} aan hoeveel componenten er zijn met complexiteit i, omvang j en omvang van de wijziging k, met:

- $i \in \{1 = \text{zeer hoog}, 2 = \text{hoog}, 3 = \text{gemiddeld}, 4 = \text{laag}\}$,
- $j \in \{1 = \text{zeer groot}, 2 = \text{groot}, 3 = \text{gemiddeld}, 4 = \text{klein}\}$ en
- $k \in \{1 = \text{een klein gedeelte}, 2 = \text{een flink gedeelte}, 3 = \text{bijna de gehele component}\}$.

Indien er 3 grote componenten zijn met een lage complexiteit die voor een klein gedeelte gewijzigd dienen te worden zal de variabele x_{421} de waarde 3 hebben. Door de toevoeging van de nieuwe variabelen kan de originele variabele "aantal te wijzigen componenten" met de bijbehorende attributen van de componenten worden weggelaten. Het grote nadeel van deze aanpak is dat er een zeer groot aantal nieuwe variabelen ontstaat (4 x 4 x 3 = 48) waarvan een groot aantal niet is waargenomen in de praktijk. Indien in de toekomst variabelen worden waargenomen die nog niet eerder zijn waargenomen, dan zal het model slechte voorspellingen doen. Gezien het feit dat we een beperkte hoeveelheid gegevens ter beschikking hebben voor het opstellen van het model zal de kans hierop groot zijn. Deze oplossing zal hier dus niet gehanteerd worden, maar kan in de toekomst wanneer er meer gegevens beschikbaar zijn wel worden toegepast.

Een andere oplossing is om per wijzigingsverzoek de gemiddelde waarden van de attributen van de te wijzigen componenten te gebruiken. Er zullen dan per wijzigingsverzoek 8 attributen zijn. Een argument tegen deze aanpak kan zijn dat het nemen van het gemiddelde van een variabele met ordinale schaal niet goed mogelijk is omdat dan een bepaald (lineair) verband tussen de mogelijk waarden gesuggereerd wordt. Aangezien er geen geschikt alternatief is en we de gegevens van de componenten toch mee willen nemen in het model kiezen we toch voor deze oplossing. In appendix A is als voorbeeld de aangepaste data van het eerste wijzigingsverzoek gegeven.

Een andere moeilijkheid is het feit dat sommige waarden van een attribuut slechts één keer voorkomen, dit is een typische eigenschap van een kleine dataset en kan niet voorkomen worden. Een analyse kan gewoon uitgevoerd worden maar de betrouwbaarheid van de analyse zal hierdoor afnemen, dit als gevolg van het feit dat mogelijk teveel waarde wordt gehecht aan specifieke waarnemingen.

De waarde “bijna gehele component” van het attribuut “gemiddelde omvang wijziging” komt zelfs niet één keer voor, dit zal voor de analyse geen problemen opleveren. Wanneer er echter nieuwe waarnemingen zijn die wel deze waarde hebben voor het betreffende attribuut dan zal de waarde niet kunnen worden ingevoerd in het model. Het is aan te raden om de analyse opnieuw uit te voeren met de nieuwe waarnemingen erbij.

Het is sowieso verstandig om na verloop van tijd de analyse opnieuw uit te voeren omdat de analyse dan op een groter aantal waarnemingen zal zijn gebaseerd, wat de betrouwbaarheid zal doen toenemen.

4.2. Keuze analytische methode

In deze paragraaf zal bepaald worden welke analytische methode zal worden gebruikt voor de analyse van de data zoals die in de vorige paragraaf beschreven is. Bij de keuze van de methode zal gekeken worden naar de kwaliteit van de verschillende methoden en de toepasbaarheid in geval van een beperkte hoeveelheid gegevens. Tevens zal rekening gehouden worden met de specifieke structuur van de gegevens.

In hoofdstuk 3 is al opgemerkt dat principale componenten analyse geen geschikte methode voor het analyseren van een kleine dataset is, mede doordat de methode geen onderliggend model heeft.

Lineaire regressie komt niet in aanmerking omdat er variabelen met een ordinale schaal voorkomen. Variantieanalyse zou gebruikt kunnen worden indien de variabele met een absolute schaal zouden worden omgezet in variabelen met een ordinale schaal. De keuze van de domeinen van de variabelen is echter arbitrair en kan van invloed zijn op de kwaliteit van het model. Covariantieanalyse heeft geen beperkingen met betrekking tot de schalen van de variabelen, er wordt echter verondersteld (zoals bij alle lineaire modellen) dat de waarnemingen normaal verdeeld zijn, in de besproken praktijksituatie is het echter niet aannemelijk dat de onderhoudsinspanningen normaal verdeeld zijn. Lineaire modellen zijn dus niet geschikt.

Zoals in het vorige hoofdstuk besproken is worden er bij factor analyse nogal veel aannamen, die niet verklaard worden en veelal niet realistisch zijn, gemaakt. Ook kan het model door de uitvoerder van de analyse gemanipuleerd worden zodat gewenste resultaten verkregen zouden kunnen worden. Factor analyse zou alleen gebruikt moeten worden wanneer de andere methoden niet in aanmerking komen.

Blijven niet-lineaire modellen en gegeneraliseerde lineaire modellen over. Aangezien de onderliggende structuur van de relatie tussen de verklarende variabelen en de respons variabele niet bekend is, en er ook geen vermoede is wat de structuur zou kunnen zijn, is een niet-lineair model geen voor de hand liggende keus als analyse methode. Indien er wel een geschikte regressiefunctie gekozen kan worden zou een niet-lineair model gebruikt kunnen worden. Toch lijkt het beter om een gegeneraliseerd lineair model te gebruiken.

In onze specifieke situatie is het niet aannemelijk dat de respons variabele een normale verdeling heeft, met behulp van gegeneraliseerde modellen kunnen ook andere verdelingen gebruikt worden. Tevens is de onderliggende structuur betreffende de relatie tussen de verklarende variabelen en respons variabele niet bekend, door middel van linkfuncties kan de onderliggende structuur gevonden en beschreven worden.

Nu is besloten tot het gebruik van een gegeneraliseerd lineair model voor de analyse, dienen de verdelingsfunctie f_i en de linkfunctie bepaald te worden. De onderhoudsinspanningen zijn in halve uren gegeven, en kunnen dus als discrete waarnemingen gezien worden. Aangezien alleen positieve waarden voorkomen is de poisson verdeling een geschikte kandidaat voor de

verdelingsfunctie *f*. De linkfunctie dient nu zo gekozen te worden dat het model alleen positieve waarden aan kan nemen, we besluiten om de *log* als linkfunctie te gebruiken. De keuze van de *log* voor de linkfunctie levert een aantal wenselijke eigenschappen, welke hier niet nader besproken zullen worden, zie hiervoor [McC'89]. Het model met deze keuze voor de verdelingsfunctie en linkfunctie, heet een log-lineair model, en ziet er als volgt uit:

$$\begin{aligned} &(i) \quad Y_i \sim Poisson(\mu_i), \\ &(ii) \quad \eta_i = x_i^T \beta, \\ &(iii) \quad \eta_i = g(\mu_i) = \log(\mu_i). \end{aligned} \tag{4.1}$$

4.3. Uitwerking analyse

In deze paragraaf zal beschreven worden hoe de analyse van de praktijkgegevens verricht is. Bij het uitvoeren van de analyse is gebruik gemaakt van het statistische softwarepakket "Splus".

Voordat er kan worden aangevangen met de analyse dienen de gegevens eerst aangepast te worden zodat de analyse met behulp van "Splus" mogelijk wordt. Voor alle variabelen met een ordinale schaal zullen een aantal, zogenoemde, "dummy"-variabelen worden ingevoerd. Voor elke waarde die de originele variabele aan kan nemen wordt een binaire variabele geïntroduceerd die aangeeft of de originele variabele die waarde heeft (1) of niet heeft (0). Zo heeft altijd precies één "dummy" variabele de waarde 1, en de rest 0.

Het is zelfs zo dat we per ordinale variabele met één "dummy" variabele minder af zouden kunnen zonder verlies van informatie. Wanneer één van de waarden optreedt waarvoor een "dummy" variabele bestaat zal deze variabele de waarde 1 hebben en de rest 0, wanneer de waarde optreedt waarvoor geen "dummy" variabele bestaat hebben alle "dummy" variabelen de waarde 0. Alle waarden kunnen dus op een unieke manier worden gerepresenteerd. Aangezien we zo min mogelijk variabelen in ons uiteindelijke model willen, mits er geen verlies van informatie optreedt, zullen we ordinale variabelen op deze manier aanpassen.

Zoals in de vorige paragraaf al aangegeven is zullen de onderhoudsinspanning als discrete variabelen worden beschouwd, voor de analyse zullen we de onderhoudsinspanningen met 2 vermenigvuldigen zodat het een geheelwaardige variabele wordt. Bij beschouwing van de resultaten van de analyse zal er rekening mee gehouden moeten worden dat de onderhoudsinspanning in halve uren gegeven wordt. De variabelen met een absolute schaal kunnen onveranderd blijven.

4.3.1. Keuze van de variabelen

De eerste stap van de analyse zal zijn om te bepalen welke van de mogelijke verklarende variabelen dienen te worden opgenomen in het model. Er zijn twee verschillende methoden om dit te doen, door middel van een t-toets of deviantie analyse. Laatstgenoemde methode is echter bij een beperkt aantal waarnemingen niet erg betrouwbaar, zie [deG'99]. Er zal dus een t-toets worden uitgevoerd met een betrouwbaarheidsdrempel $\alpha = 0.05$. Normaalgesproken zal in deze fase van de analyse een hoger betrouwbaarheidsdrempel gebruikt worden, maar aangezien het aantal mogelijke variabelen ten opzichte van het aantal waarnemingen groot is, en we graag het aantal verklarende variabelen enigszins zouden willen terugbrengen, zal er een laag betrouwbaarheidsdrempel gehanteerd worden. In tabel 4.1 zijn de t-ratio's voor alle variabelen gegeven.

Variabele	t-ratio
(Intercept)	5.124
Kwaliteit (1): goed	4.920
Kwaliteit (2): voldoende	6.523
Kwaliteit (3): slecht	-3.065
Probleem (1): gebruikers probleem	-6.946
Probleem (2): wettelijke maatregelen	-3.259
Aantal systemen	4.330
Wijzigingen (1): geen	-2.159
Wijzigingen (2): alleen raadpleging	1.538
Aantal componenten	4.740
Complexiteit (1): zeer hoog	3.797
Complexiteit (2): hoog	-1.375
Complexiteit (3): gemiddeld	-0.498
Omvang (1): zeer groot	NVT
Omvang (2): groot	3.116
Omvang (3): gemiddeld	1.805
Omvang wijziging (1): klein	0.741

Tabel 4.1: t-ratio's voor alle mogelijke variabelen

Bij een betrouwbaarheidsdrempel $\alpha = 0.05$ zullen de variabelen met een t-ratio groter dan 4.30 worden opgenomen in het model en de variabelen met een t-ratio kleiner dan 4.30 niet, zie appendix B4. Indien een variabele is weergegeven door middel van "dummy" variabelen zullen alle "dummy" variabelen, behorende bij deze variabele, worden opgenomen in het model indien minimaal één een t-ratio groter dan 4.30 heeft. Er kan dus geconcludeerd worden dat de variabelen "Kwaliteit", "Probleem", "Aantal systemen" en "Aantal componenten" in het model worden opgenomen. De variabele "Omvang" heeft één "dummy" variabele waarvoor geen t-ratio bepaald kan worden, dit komt doordat deze variabele als een lineaire combinatie van andere variabelen geschreven kan worden en daardoor geen extra invloed op het model uitoefent, zie [Str'88].

De volgende stap is om na te gaan of er eventueel gecombineerde effecten van de variabelen in het model moeten worden opgenomen. Er is voor gekozen om alleen te kijken naar de eerste orde interacties tussen de variabele "Aantal componenten" met de variabelen "Complexiteit", "Omvang" en "Omvang wijziging". Het ligt voor de hand om deze interacties nader te bekijken omdat de laatst genoemde drie variabelen informatie bevatten over de componenten, overige interacties zijn achterwege gelaten omdat het aantal verklarende variabelen niet te groot dient te worden. Er is gekozen voor het uitvoeren van t-toetsen, waarbij telkens één eerste orde interactie wordt toegevoegd, met een betrouwbaarheids-drempel $\alpha = 0.05$. In tabel 4.2 zijn de t-ratio's weergegeven.

Interactie "Aantal componenten" met:	t-ratio
Complexiteit: zeer hoog	4.468
Complexiteit: hoog	2.084
Complexiteit: gemiddeld	2.202
Omvang: zeer groot	5.676
Omvang: groot	4.633
Omvang: gemiddeld	3.535
Omvang wijziging: klein	-0.524

Tabel 4.2: t-ratio's voor alle beschouwde interacties

Bij de interacties “Aantal componenten” met “Complexiteit” en “Omvang” zal de interactie worden toegevoegd indien de t-ratio groter is dan 2.31, voor de interactie met “Omvang wijziging” is dat 2.23. Uit tabel 4.2 blijkt dus dat de interacties “Aantal componenten” met “Complexiteit” en “Omvang” zullen worden toegevoegd aan het model. Door het toevoegen van de interacties aan het model ontstaat het vermoeden dat de variabele “Aantal componenten” wellicht overbodig is geworden omdat de informatie die deze variabele bevat misschien al via de interacties tot uitdrukking komt. We voeren nog een laatste t-toets uit om dit na te gaan. De betreffende t-ratio’s zijn in tabel 4.3 opgenomen.

Variabele	t-ratio
(Intercept)	25.218
Kwaliteit: goed	1.366
Kwaliteit: voldoende	8.983
Kwaliteit: slecht	-6.884
Probleem: gebruikers probleem	-24.296
Probleem: wettelijke maatregelen	-4.150
Aantal systemen	12.376
Aantal componenten	-2.051
Interactie "Aantal componenten" met:	t-ratio
Complexiteit: zeer hoog	5.186
Complexiteit: hoog	-5.277
Complexiteit: gemiddeld	-3.489
Omvang: zeer groot	NVT
Omvang: groot	7.262
Omvang: gemiddeld	5.244

Tabel 4.3: t-ratio’s van alle opgenomen variabelen

Wanneer er een betrouwbaarheidsdrempel van $\alpha = 0.05$ wordt gehanteerd, zal de variabele “Aantal componenten” uit het model worden verwijderd indien de t-ratio lager dan 2.57 is. Dat is hier inderdaad het geval, het vermoeden dat we hadden is dus juist gebleken. We merken tevens op dat de t-ratio van één van de interacties niet bepaald kan worden, dit heeft dezelfde oorzaak als we eerder zijn tegengekomen voor een “Dummy” variabele. Deze interactie kan zonder problemen buiten het model gelaten worden.

4.3.2. Parameterschattingen

De parameters worden geschat met behulp van de meest aannemelijke schatter-methode, zie appendix B4. Indien er geen expliciete uitdrukking is voor de aannemelijkheidsfunctie zullen de parameters numeriek bepaald worden met behulp van *Fisher scoring,* zie [McC’89]. In tabel 4.4 zijn de parameterschattingen van het model weergegeven en zijn tevens voor alle variabelen de 95 % betrouwbaarheidsintervallen opgenomen.

Variabele	Schatting	95 % bti
(Intercept)	4.957	[4.492, 5.423]
Kwaliteit: goed	0.164	[-0.253, 0.582]
Kwaliteit: voldoende	1.561	[1.145, 1.977]
Kwaliteit: slecht	-3.812	[-5.080, -2.544]
Probleem: gebruikers probleem	-3.276	[-3.563, -2.989]
Probleem: wettelijke maatregelen	-6.106	[-8.606, -3.606]
Aantal systemen	0.657	[0.546, 0.769]
Interactie "Aantal componenten" met:	Schatting	95 % bti
Complexiteit: zeer hoog	0.687	[0.562, 0.813]
Complexiteit: hoog	-1.501	[-2.036, -0.966]
Complexiteit: gemiddeld	-1.321	[-1.990, -0.652]
Omvang: groot	1.868	[1.293, 2.443]
Omvang: gemiddeld	1.923	[1.081, 2.765]

Tabel 4.4: parameterschattingen en 95 % betrouwbaarheidsintervallen

4.3.3. Kwaliteit van het model

De deviantie D is een globale grootheid voor het bepalen hoe goed het model de data beschrijft, zie appendix B4., die geïnterpreteerd kan worden zoals de residuele kwadratensom in het klassieke lineaire regressie model. Als uitdrukking voor de hoeveelheid verklaarde variantie kan de volgende grootheid gebruikt worden:

(4.2) $$\frac{D_0 - D}{D}.$$

De hoeveelheid verklaarde variantie met het model zoals dat in de vorige paragraaf is bepaald is 95.7 %. Dit wil zeggen dat het model de data zeer goed beschrijft, dit wil echter nog niet zeggen dat het model ook zeer goede voorspellingen zal doen. Er dient gerealiseerd te worden dat het aantal verklarende variabelen ten opzichte van het aantal waarnemingen uiteindelijk nog steeds behoorlijk groot was.

4.4. Resultaten analyse

In onderstaande tabel zijn de voorspellingen van de onderhoudsinspanningen gegeven zoals deze bepaald zijn met het gegeneraliseerde lineaire model. Tevens zijn de werkelijke onderhoudsinspanningen gegeven en de afwijking van de schattingen ten opzichte van de werkelijke waarden. In de laatste kolom treft u de verbetering in schatting aan ten opzichte van de “natte vinger” schatting die tot dus verre gehanteerd wordt.

Verzoek	Voorspelling GLM	Werkelijk	Verschil	Verbetering
1	6,0	1,0	5,0	-3,0
2	6,0	17,5	-11,5	-10,0
3	154,5	189,0	-34,5	-5,5
4	49,5	49,5	0,0	6,5
5	19,0	19,0	0,0	7,0
6	6,0	5,0	1,0	4,0
7	23,5	7,0	16,5	-11,5
8	20,5	20,5	0,0	7,5
9	6,0	3,0	3,0	18,0
10	191,5	172,0	19,5	4,5
11	59,5	43,5	16,0	20,5
12	42,5	37,5	5,0	-4,5
13	6,0	12,0	-6,0	2,0
14	45,0	59,5	-14,5	13,0
15	20,5	20,5	0,0	0,5
16	310,5	311,5	-1,0	207,5
17	153,0	153,0	0,0	39,0
18	18,0	25,0	-7,0	13,0
19	19,5	12,5	7,0	0,5

Tabel 4.5: schattingen en werkelijke waarden onderhoudsinspanningen in uren

Wanneer we de schattingen vergelijken met de werkelijk waargenomen waarden zien we dat 5 van de 19 schattingen precies goed zijn, 8 van de 19 verschillen minder dan 10 uur en de grootste afwijking is 34.5 uur. In vergelijking met de “natte vinger” schattingen geeft het model in 5 van de 19 gevallen een minder goede schatting, met als grootste waarde 11.5 uur. In de overige 14 gevallen geeft het model betere schattingen, met als uitschieter een verbetering van 207.5 uur.

Aan de hand van deze resultaten is het nog steeds niet duidelijk of het model ook daadwerkelijk geschikt is voor het voorspellen van onderhoudsinspanningen, hiertoe zou het model met behulp van nieuwe waarnemingen getoetst kunnen worden. Vervolgens is het aan de potentiële gebruiker van het model om te bepalen of de resultaten bevredigend genoeg zijn voor de acceptatie van het model als voorspellingsmethode.

Appendix A

Data

In deze appendix zullen voorbeelden gegeven worden van de originele data, zoals deze beschikbaar is gesteld, van de beschreven praktijksituatie uit hoofdstuk 4 en de bewerking van de originele data die gebruikt is voor de analyse.

A1. De originele data

Enkele voorbeelden van de gegevens per wijzigingsverzoek:

Attribuut	Wijzigingsverzoek 1	Wijzigingsverzoek 2
Wijzigingsverzoek ID	995	996
Aantal te wijzigen componenten	1	1
Database gerelateerde wijziging?	geen	alleen raadpleging
Aantal betrokken externe systemen	0	0
Herkomst wijziging	gebruikersprobleem of wens	gebruikersprobleem of wens
Kwaliteit van de specificatie door de klant	goed	goed
Werkelijke inspanning ontwerp wijzigingsverzoek	0.5	3
Werkelijke inspanning realisatie wijzigingsverzoek	0.5	14.5
Werkelijke totale inspanning wijzigingsverzoek	1	17.5

Tabel A1: voorbeelden van wijzigingsverzoeken

Enkele voorbeelden van gegevens per “te wijzigen component”:

Attribuut	Component 1	Component 2	Component 3
Wijzigingsverzoek ID	995	996	998
Complexiteit van de component	gemiddeld	gemiddeld	zeer hoog
Omvang van de component	gemiddeld	gemiddeld	zeer groot
Omvang van de wijziging	een klein gedeelte	een klein gedeelte	een klein gedeelte

Tabel A2: voorbeelden van componenten

De componenten zijn aan de wijzigingsverzoeken verbonden door middel van het “Wijzigingsverzoek ID”.

A2. De bewerkte data

Enkele voorbeelden van de bewerkte gegevens per wijzigingsverzoek:

Attribuut	Wijzigingsverzoek 3	Wijzigingsverzoek 4
Wijzigingsverzoek ID	998	1002
Aantal te wijzigen componenten	5	4
Database gerelateerde wijziging?	alleen raadpleging	alleen raadpleging
Aantal betrokken externe systemen	1	0
Herkomst wijziging	gebruikersprobleem of wens	Gebruikersprobleem of wens
Kwaliteit van de specificatie door de klant	voldoende	goed
Gemiddelde complexiteit van de componenten	hoog	zeer hoog
Gemiddelde omvang van de componenten	groot	zeer groot
Gemiddelde omvang van de wijzigingen	een klein gedeelte	een klein gedeelte
Werkelijke inspanning ontwerp wijzigingsverzoek	36	12
Werkelijke inspanning realisatie wijzigingsverzoek	153	37.5
Werkelijke totale inspanning wijzigingsverzoek	189	49.5

Tabel A3: voorbeelden van wijzigingsverzoeken na bewerking

Appendix B

Wiskundige terminologie

In deze appendix zullen enkele wiskundige begrippen, die in dit werkstuk genoemd zijn, uitgelegd worden.

B1. Kansverdelingen

Grotendeels gebaseerd op [Har'95].

De normale verdeling

Een variabele X heeft een normale verdeling indien X absoluut continu verdeeld is met kansdichtheidsfunctie:

$$f(x) = P(X = x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma} \right)^2 \right), \qquad [x \in R].$$

We zeggen in zo'n geval dat X een normale verdeling heeft met parameters $\mu \in R$ en $\sigma^2 > 0$, notatie $X \sim N(\mu, \sigma^2)$.

Enkele eigenschappen van de normale verdeling zijn:

(verwachting) $EX = \int_{-\infty}^{\infty} x f(x) dx = \mu.$

(variantie) $VarX = EX^2 - (EX)^2 = \int_{-\infty}^{\infty} x^2 f(x) dx - \mu^2 = \sigma^2.$

De poisson verdeling

Een variabele X heeft een poisson verdeling met indien X (positief) discreet verdeeld is met kansdichtheidsfunctie:

$$f(x) = \frac{\lambda^x}{x!} e^{-\lambda}, \qquad [x \in Z_+].$$

We zeggen in zo'n geval dat X een normale verdeling heeft met parameters $\lambda > 0$, notatie $X \sim Poisson(\lambda)$.

Enkele eigenschappen zijn van de poisson verdeling zijn;

(verwachting) $EX = \lambda$.

(variantie) $VarX = \lambda$.

B2. Kansrekening

Grotendeels gebaseerd op [Har'95].

Standaardafwijking of standaarddeviatie

De standaardafwijking is gedefinieerd als de wortel van de variantie.

Covariantie

De definitie van de covariantie van de variabelen X en Y is:

$$Cov(X,Y) = EXY - (EX)(EY).$$

Een eigenschap van covariantie is:

$$Cov(X,Y) = Cov(Y,X).$$

en in geval $Y = X$ geldt:

$$Cov(X,X) = VarX.$$

Een covariantiematrix is de matrix met als $(i,j)^e$ element (i^e rij, j^e kolom) de covariantie van het paar (X_i,X_j).

Correlatie

De correlatiecoëfficiënt ρ_{ij} wordt verkregen door de covariantie te delen door het product van de standaardafwijkingen:

$$\rho_{ij} = \frac{Cov(X_i, X_j)}{\sigma_i \sigma_j}.$$

De correlatiematrix is de matrix met als $(i,j)^e$ element (i^e rij, j^e kolom) de correlatiecoëfficiënt van het paar (X_i,X_j), ofwel ρ_{ij}.

B3. Matrices

Grotendeels gebaseerd op [Str'88].

Rang

De rang van een matrix is het aantal onafhankelijke kolommen van een matrix. Een matrix is van maximale rang indien de rang gelijk is aan het aantal kolommen van de matrix. Voor een uitgebreide uitleg van het begrip rang, en onafhankelijke kolommen, wordt de lezer verwezen naar [Str'88].

Som der kwadraten

Zij a een vector met p elementen $a_1,\ldots,a_p$ dan is de som der kwadraten $a'a$ als volgt gedefinieerd:

$$a^T a = \sum_{i=1}^{p} a_i^2.$$

Eigenwaarde

De eigenwaarden van een matrix A worden verkregen door de wortels $\lambda_1,\ldots,\ \lambda_p$ te bepalen van:

$$\left|A - \lambda I\right| = 0,$$

met I de identiteitsmatrix (matrix met één-en op de diagonaal en de rest nullen).

Eigenvector

Elke eigenwaarde heeft een bijbehorende eigenvector x, waarvoor geldt:

$$Ax = \lambda x.$$

Een genormaliseerde eigenvector is de vector die verkregen wordt door alle elementen van een eigenvector x te vermenigvuldigen met een factor α zodat de som der kwadraten van de nieuwe vector 1 is.

B4. Gegeneraliseerde lineaire modellen

Grotendeels gebaseerd op [Oos'97] en [deG'99].

Meest aannemelijke schatter

Onder de kansdichtheid p_θ van een stochastische vector X verstaan we de functie $x \mapsto P_\theta(X = x)$ als X discreet verdeeld is en de (echte) kansdichtheid als X continu verdeeld is. Zij X een stochastische vector met een kansdichtheid $p_\theta(x)$ die van een parameter $\theta \in \Theta$ afhangt. De functie

$$\theta \mapsto L(\theta; x) := p_\theta(x),$$

opgevat als functie van $\theta \in \Theta$ voor vaste x heet de aannemelijkheidsfunctie. De log-aannemelijkheidsfunctie wordt gegeven door $log(\, p_\theta(x))$.

Vaak is $X = (X_1, \ldots, X_n)$ een vector met onderling onafhankelijke identiek *discreet* verdeelde coördinaten X_i, die van een onbekende parameter $\theta \in \Theta$ afhangt. Dan is de dichtheid van X het product van de dichtheden van de X_i, en de log-aannemelijkheidsfunctie wordt

$$\log L(\theta; x_1, \ldots, x_n) = \sum_{i=1}^{n} \log p_\theta(x_i),$$

waarin p_θ nu de (marginale) dichtheid van een enkele X_i weergeeft. De meest aannemelijke schatter voor θ is die waarde $\theta \in \Theta$ die de log-aannemelijkheidsfunctie maximaliseert.

t-toets

Om na te gaan of een variabele in het model dient te worden opgenomen als verklarende variabele wordt de bijbehorende t-ratio bepaald. De t-ratio wordt verkregen door de schatting te delen door de geschatte standaard deviatie. Vervolgens wordt de t-ratio vergeleken met (+ of -) het $(1-\alpha/2)$ kwantiel van de Students-t verdeling met $(n-p-1)$ vrijheidsgraden, met n het aantal waarnemingen en p het aantal variabelen. Indien de t-ratio niet significant is (kleiner dan in absolute waarde), is dit een indicatie dat de corresponderende verklarende variabele misschien wel uit het model weggelaten kan worden.

Betrouwbaarheidsinterval

Gebaseerd op de asymptotische normaliteit van de meest aannemelijke schatter kunnen we het $(1 - \alpha)100\%$ betrouwbaarheidsinterval voor β_j schatten met behulp van de volgende uitdrukking:

$$\hat{\beta}_j \pm t_{(n-p-1);(1-\alpha/2)} \sqrt{\hat{\phi}((X^T \hat{W} X)^{-1})_{jj}}$$

waarbij $t_{(n-p-1);(1-\alpha/2)}$ het $(1-\alpha/2)$ kwantiel is van de Students-t verdeling met $(1-n-p)$ vrijheidsgraden, en de geschatte standaardafwijking van β_j.

Deviantie

Een kwantitatieve grootheid waarmee het verschil tussen twee modellen gemeten kan worden heet deviantie, wat is gedefinieerd als de geschaalde log-aannemelijkheids ratio

$$D = 2\phi[l(\tilde{\theta};\phi) - l(\hat{\theta};\phi)]$$
$$= 2\sum_{i=1}^{n}\left\{Y_i(\tilde{\theta}_i - \hat{\theta}_i) - b(\tilde{\theta}_i) + b(\hat{\theta}_i)\right\}$$

Met $l(\theta;\phi)$ de log-aannemelijkheidsfunctie is, $b(\theta)$ de functie b uit de gedefinieerde exponentiele familie. De nuldeviantie D_0 is de deviantie behorende bij een model zonder verklarende variabelen, dus alleen met een intercept.

Bibliografie

[Cha'80] C. Chatfield and A.J. Collins. *Introduction to multivariate analysis.* Chapman & Hall 1980.

[deG'98] M.C.M. de Gunst en A.W. van der Vaart. *Statistische data analyse.* *vrije* Universiteit 1998.

[deG'99] M.C.M. de Gunst. *Statistische modellen.* *vrije* Universiteit 1999.

[DeM'82] T. DeMarco. *Controlling software projects.* Prentice Hall, New York 1982.

[Eve'01] B.S. Everitt and G. Dunn. *Applied multivariate data analysis.* Arnold 2001.

[Fen'95] N.E. Fenton. *Software metrics, A rigorous approach.* International Thompson Computer Press 1995.

[Jol'86] I.T. Jolliffe. *Principal component analysis.* Springer-Verlag, New York 1986.

[Har'95] K. van Harn en P.J. Holewijn. *Inleiding in de waarschijnlijkheidsrekening.* *vrije* Universiteit 1995.

[Ken'80] Sir M. Kendall. *Multivariate analysis.* Charles Griffin & Company LTD, London and High Wycombe 1980.

[Law'63] D.N. Lawley and A.E. Maxwell. *Factor analysis as a statistical method.* Butterworth & Co. Ltd. 1963.

[McC'89] P. McCullagh and J.A. Nelder. *Generalized linear models.* Chapman & Hall 1989.

[Möl'93] K.H. Möller and D.J. Paulish. *Software metrics, A practitioner's guide to improved product development.* Chapman & Hall 1993.

[Mor'90] D.F. Morrison. *Multivariate statistical methods.* McGraw-Hill, New York 1990.

[Oos'97] J. Oosterhoff en A.W. van der Vaart. *Algemene statistiek.* *vrije* Universiteit 1997.

[Ren'95] A.C. Rencher. *Methods for multivariate analysis.* John Wiley & Sons 1995.

[Sch'59] H. Scheffé. *The analysis of variance.* John Wiley & Sons 1959.

[Str'88] G. Strang. *Linear algebra and its applications.* Harcourt Brace & Company 1988.

[vVl'93] H. van Vliet. *Software principals, Principles and practice.* John Wiley & Sons 1993.