Introductie op meervoudige regressie en verwante procedures

Een handboek voor onderzoekers in de gedragswetenschappen en de sociale wetenschappen

Voeten en Van den Bercken

Lineaire regressieanalyse

Noordhoff Uitgevers

Lineaire regressieanalyse

Lineaire regressieanalyse

Marinus J.M. Voeten

John H.L. van den Bercken

Noordhoff Uitgevers Groningen | Houten

Ontwerp omslag: Total Identity, Amsterdam

Eventuele op- en aanmerkingen over deze of andere uitgaven kunt u richten aan: Noordhoff Uitgevers bv, Afdeling Hoger Onderwijs, Antwoordnummer 13, 9700 VB Groningen, e-mail: info@noordhoff.nl

2 3 4 5 / 14 13 12 11 10

ISBN (ebook) 978-018-4944-3
ISBN 978-207-3231-3
NUR 173

Woord vooraf

Dit boek is het resultaat van een jarenlange praktijk in het onderwijzen van de technieken van regressie- en variantieanalyse en in het begeleiden van onderzoeksprojecten waarin regressie- en variantieanalyse in diverse vormen zijn toegepast. Zoals dat vaak gaat in dergelijke gevallen, bleken onze aantekeningen en aanvullende teksten bij bestaande leerboeken op den duur zo omvangrijk dat er langzamerhand een eigen en onafhankelijk te gebruiken cursusboek ontstond, met een min of meer stabiele inhoud. In 2002 publiceerden we vanuit dit materiaal een leerboek over variantieanalyse. Dit wordt nu vervolgd met een boek over regressieanalyse. Er zijn al veel (goede) boeken over regressieanalyse geschreven. Het feit echter dat er over regressieanalyse nog geen uitgebreid Nederlandstalig leerboek bestaat en de overtuiging dat onze didactische uitgangspunten de bestudering van regressieanalyse voor een meerderheid van onze studenten gemakkelijker maken, bracht ons ertoe onze tekst uit te willen geven.

Wij beschouwen regressie- en variantieanalyse als de basis voor de meeste data-analytische procedures die nodig zijn en toegepast worden in de gedrags- en de sociale wetenschappen. Regressie- en variantieanalyse leggen het fundament voor verdergaande bestudering van onderzoeksmethoden en methoden van data-analyse. Daarnaast en vooral zijn regressie- en variantieanalyse een noodzakelijke basis voor het kritisch kunnen lezen van een groot deel van de gedragswetenschappelijke en sociaal-wetenschappelijke onderzoeksliteratuur.

Wij danken allen die op enigerlei wijze hebben meegewerkt aan het totstandkomen van dit boek. Op de eerste plaats zijn dat de personen die op verzoek van de uitgever het manuscript hebben beoordeeld: Martijn Berger, Gerhard van de Bunt, Wilfried de Corte, Arie van Peet en Yves Rosseel. Hun soms zeer gedetailleerde commentaren hebben wezenlijk bijgedragen aan de definitieve tekst. Verder danken wij onze collega in de data-analyse Chris Michels voor zijn bruikbare suggesties. Marieke Voeten danken wij voor haar bijdragen aan het verhogen van de leesbaarheid van de tekst. Wij danken ook de medewerkers van Noordhoff Uitgevers voor het in ons gestelde vertrouwen en de begeleiding op afstand bij het schrijfproces. Ten slotte en bovenal zijn we dank verschuldigd aan de generaties van studenten wier kritische houding ten aanzien van de successieve 'verbeterde' versies van de tekst een stimulerende factor was voor de evolutie ervan.

Bij deze tekst is aanvullend materiaal beschikbaar. Wij verwijzen daarvoor naar de inleiding.

Voor opmerkingen naar aanleiding van de tekst en voor suggesties ten behoeve van de onderwijspraktijk houden wij ons aanbevolen.

Nijmegen, juni 2003
Marinus J.M. Voeten en John H.L. van den Bercken

Inhoud

Samenvattingen in vragen en antwoorden

Inleiding

Dit boek is bestemd voor universitaire studenten in de sociale en de gedragswetenschappen, met name pedagogiek en onderwijskunde, psychologie en sociologie. Regressieanalyse is, samen met variantieanalyse, de meest gebruikte statistische techniek bij onderzoek in deze disciplines en vrijwel elke universitaire opleiding besteedt er aandacht aan, hetzij als onderdeel van een meer algemene cursus over statistische technieken van data-analyse, hetzij in de vorm van een gespecialiseerde cursus die volgt op zo'n algemene cursus. Het boek is geschikt voor beide cursusvarianten.

Wat de statistiek betreft veronderstellen we de volgende voorkennis. We gaan ervan uit dat de lezer vertrouwd is met de basisbegrippen van de beschrijvende statistiek, zoals gemiddelde, standaarddeviatie, variantie, correlatie en covariantie. Daarnaast moet de lezer de fundamentele begrippen van de toetsende statistiek kennen, zoals kansverdeling, steekproevenverdeling, toetsingsgrootheid, significantieniveau, type-I- en type-II-fouten. Van de specifieke (parametrische) toetsingsprocedures worden met name de t- en de F-toets gebruikt. We besteden uitgebreid aandacht aan de principes waarop deze toetsen berusten, zodat de behandeling van de stof een op zichzelf staand geheel blijft.

Wij willen de methode van regressieanalyse begrijpelijk maken. Dat wil zeggen: na bestudering van het boek weet men wat deze methode inhoudt, waarvoor en wanneer ze kan worden gebruikt (bij welke onderzoeksopzet en bij welke vraagstelling), welke voorwaarden gelden voor de toepassing ervan (aard en kwaliteit van de data), wat de relevante statistische resultaten zijn en hoe men die moet interpreteren. Met deze kennis kan men in wetenschappelijke publicaties waarin een vorm van regressieanalyse toegepast wordt, de gerapporteerde resultaten begrijpen en op hun waarde beoordelen.

Deze doelstellingen willen we realiseren met een tekst waarbij de nadruk ligt op de centrale begrippen en principes van regressieanalyse. We hanteren daarbij een algemeen conceptueel kader (dat van het vergelijken van modellen) en illustreren de bewerkingen die de data tijdens analyse ondergaan bij voorkeur aan de hand van procedures gebaseerd op algemene definities en principes, ook al zijn deze vaak minder efficiënt dan de specifieke rekenformules die men in veel leerboeken aantreft. Voor zover we vergelijkingen en formules gebruiken, is dat vrijwel uitsluitend om de relaties tussen fundamentele begrippen tot uitdrukking te brengen. We willen laten zien wat er bij regressieanalyse gebeurt met de data en waarom. Voor het feitelijke rekenwerk in concrete toepassingen zijn uitstekende computerprogramma's beschikbaar. Ook van die programma's hoeft men niet de technische ins en outs te kennen om ze op een verstandige manier te kunnen gebruiken. Waar het om gaat is dat men weet wat de regressieanalyse conceptueel gesproken inhoudt en dat men van de geëigende programma's de functionaliteit en de gebruiksvoorwaarden kent en de output begrijpt.

We kunnen deze doelstellingen alleen realiseren door ons te committeren aan concrete computerprogramma's. Dat zijn met name de procedures REGRESSION en in mindere mate GLM in het pakket SPSS. We gebruiken deze procedures om de relevante resultaten van een regressieanalyse te berekenen en om ze aan de hand van de output te kunnen bespreken. Maar we vinden het ook van belang dat studenten zelf met deze programma's kunnen werken. Tegelijkertijd willen we voorkomen dat de behandeling van de algemene principes van regressieanalyse doorspekt wordt met details over de manier waarop een programma aangestuurd moet worden, informatie die bovendien in de opeenvolgende versies van een programmapakket sterk aan wijziging onderhevig kan zijn. Onze oplossing voor dit probleem is de volgende. In de tekst bespreken we wel de output die de programma's leveren, maar we zeggen (vrijwel) niets over de manier waarop die output verkregen wordt. Op de cd-rom bij dit boek en op de bijbehorende website vindt men echter uitgebreide informatie over hoe men de data-analyse met behulp van SPSS-procedures concreet kan uitvoeren. Deze informatie bestaat uit een algemene handleiding voor het werken met de procedures REGRESSION en GLM en uit tekstbestanden met de syntax van alle aansturingen voor de in dit boek getoonde output.

Er zijn in het boek geen vragen, oefeningen en practicumopdrachten opgenomen. Niet omdat deze zaken van ondergeschikt belang zouden zijn. Integendeel, we beseffen ten volle dat ze onmisbare hulpmiddelen zijn voor het consolideren van theoretische kennis en het ontwikkelen van praktische vaardigheid. In onze onderwijspraktijk maken we er dan ook wel degelijk gebruik van. Het betreffende materiaal hebben we echter ondergebracht op de cd-rom en op de website bij het boek.
Wij willen sterk benadrukken dat het leren van data-analytische procedures het beste kan gebeuren door data-analyses uit te voeren met een computerprogramma en de output ervan grondig te bestuderen en te interpreteren. We raden de student daarom aan zelf de voorbeelden in de tekst met een computerprogramma te analyseren, evenals de oefeningen op de cd-rom en de website die het gebruik van een computer vereisen. Een goede manier om vertrouwd te raken met de praktische kanten van regressieanalyse is het uitvoeren van de in het boek geïllustreerde analyses. Het voordeel daarvan is dat men de context van het probleem kent en weet hoe de resultaten er uit moeten zien.

Data-analyse is een fase in een onderzoeksproces, ingebed in de inhoudelijke context van onderzoeksvragen, theoretische achtergrond, conclusies en praktische aanbevelingen. Data-analyse is niet een doel op zich maar een middel, een instrument in het proces van beantwoorden van belangrijke onderzoeksvragen op basis van empirische informatie. Regressieanalyse is een van de hulpmiddelen om patronen in data te ontwarren en samen te vatten, zodat conclusies kunnen worden bereikt die ons verder helpen in de studie van gedrag van mensen in sociale contexten. Het gaat niet om de getallen en de statistiek, maar om wat je uit die getallen kunt afleiden. Dit boek is dan ook geschreven vanuit het perspectief van toepassing van regressieanalyse in gedrags- en sociale wetenschappen. We maken veel gebruik van voorbeelden uit gepubliceerd onderzoek om de relatie van de statistische techniek met de inhoudelijke context waar het eigenlijk om gaat, zichtbaar te maken. Bestudering van statistische procedures schiet zijn doel voorbij als men deze procedures niet in verband kan brengen met de wetenschappelijke vragen en processen waar het om gaat.

De basisprincipes van regressieanalyse en de in de uitvoering van data-analyse te volgen stappen worden uitgelegd aan de hand van voorbeelden. Hierbij gebruiken we zowel artificiële voorbeelden als aan de onderzoekspraktijk ontleende voorbeelden. Wij zijn ervan overtuigd dat het van groot belang is data-analyse te bestuderen aan de hand van reële databestanden en vanuit de wetenschappelijke context waarin die bestanden een functie hebben. Van de andere kant vraagt een ordelijke uiteenzetting van de procedures van regressieanalyse dat niet problemen van allerlei aard tegelijkertijd aan de orde komen. Daarom gebruiken we voor onze uiteenzettingen ook vaak onrealistisch kleine, artificiële databestanden die het toelaten bepaalde specifieke punten uit te leggen.

De behandeling van de stof wordt langzaam opgebouwd vanuit elementaire principes. Dat kan voor sommige lezers betekenen dat met name in de eerste vier hoofdstukken veel bekende zaken uitvoerig uitgelegd worden. We maken ook veel gebruik van herhalingen en samenvattingen. In de meeste hoofdstukken zijn tekstboxen te vinden waarin een bepaalde vraag gesteld en beantwoord wordt. Deze tekstboxen dienen vooral om hoofdzaken samen te vatten. Aan het eind van het boek is een lijst van symbolen en formules en een lijst van kernbegrippen te vinden. Ook deze hebben een samenvattende rol en beperken zich tot hoofdzaken. De tekstblokken tussen lijnen betreffen uitweidingen en wetenswaardigheden die, zeker bij eerste lezing, kunnen worden overgeslagen.

Naast de basisprincipes van regressieanalyse komen in het boek ook verschillende uitbreidingen en inleidingen op meer geavanceerde vormen van regressieanalyse aan de orde. De hoofdstukken laten zich als volgt groeperen. Hoofdstuk 1 is voorbereidend. Aan de hand van onderzoeksvoorbeelden wordt duidelijk gemaakt welke vraagstellingen en onderzoeksopzetten zich lenen voor regressieanalyse en met name welke rollen de variabelen spelen in de analyse. Verder worden begrippen herhaald uit de elementaire statistiek en methodologie die van belang zijn voor regressieanalyse. De hoofdstukken 2 tot en met 4 behandelen de basisprincipes van enkelvoudige en meervoudige regressie. Hoofdstuk 2 handelt over het regressiemodel en de interpretatie van de parameters, de regressiecoëfficiënten, terwijl in hoofdstuk 3 de schatting en toetsing van deze parameters aan de orde zijn, evenals de interpretatie van de resultaten. Hierbij zijn de statistische principes van betrouwbaarheidsintervallen en hypothesetoetsing van belang, met name de t-toets. Hoofdstuk 4 completeert de basisprincipes door te kijken naar de kwaliteit van het regressiemodel als geheel en naar de bijdrage die het model levert aan het voorspellen of verklaren van de afhankelijke variabele.

We volgen dus, net als in ons boek *Variantieanalyse* (Van den Bercken & Voeten, 2002) de methodologie van het formuleren en uitwerken van een model voor waargenomen data, waarbij we de afzonderlijke stappen behandelen, met name de specificatie van een structureel en statistisch model, het schatten van de modelparameters en het toetsen van het model tegen de data. In hoofdstuk 4 en 5 komen daarbij de variantieanalysetabel en de F-toets aan de orde die ook uitgebreid in het boek over variantieanalyse worden behandeld. Hoofdstuk 5 is deels herhaling van de hypothesetoetsingen die al in hoofdstuk 3 en 4 aan de orde waren, maar die nu op een andere manier worden behandeld. Hoofdstuk 5 handelt over het vergelijken

van regressiemodellen, hetgeen een algemeen toepasbare benadering geeft van het toetsen van modellen in regressieanalyse (en variantieanalyse). Die benadering komt overeen met het algemene conceptuele kader van het vergelijken van modellen dat we ook in *Variantieanalyse* hanteren. Naast herhaling biedt hoofdstuk 5 dus veralgemening naar een toetsingsprocedure, de algemene F-toets, die gebruikt kan worden voor het toetsen van hypothesen over een regressiemodel als geheel, over afzonderlijke variabelen in het model en over subsets van variabelen.

Een belangrijk principe van regressieanalyse is dat je er onderzoek mee kunt doen naar de onderlinge relaties van een veelheid van variabelen. De relatie van de ene variabele met een andere variabele kan er verschillend gaan uitzien als ook nog een derde of een vierde variabele mee in de analyse wordt betrokken. In een regressieanalyse controleren als het ware de variabelen elkaar. Hoe dat in zijn werk gaat en welke consequenties dat heeft voor de interpretatie van resultaten van een regressieanalyse, is aan de orde in hoofdstuk 6. Daarbij maken we gebruik van het vergelijken van modellen dat geïntroduceerd is in hoofdstuk 5.

In hoofdstuk 7 en 8 vragen we aandacht voor de regressiediagnostiek. Met het formuleren van een regressiemodel gaan assumpties gepaard. In hoofdstuk 2 en 7 maken we duidelijk wat die assumpties zijn. Hoofdstuk 7 gaat verder over manieren om na te gaan of in een concreet geval de data voldoen aan die assumpties. Daarnaast kunnen er met de steekproefdata allerhande problemen zijn die opgespoord en verholpen moeten worden om met vertrouwen de resultaten van een regressieanalyse te kunnen interpreteren; dat is het onderwerp van hoofdstuk 8.

In hoofdstuk 9 laten we zien dat naast kwantitatieve, gemeten variabelen ook kwalitatieve variabelen in de rol van onafhankelijke variabelen kunnen meedoen in een regressiemodel. Hierbij leggen we de verbinding met het 'algemene lineaire model' (*general linear model*, GLM). Het algemene lineaire model ligt ten grondslag aan zowel lineaire regressieanalyse als variantieanalyse. Het is zelfs zo dat variantieanalyse (de analyse van verschillen tussen gemiddelden) uitgewerkt kan worden als een speciaal geval van regressieanalyse. Voor meer informatie over variantieanalyse in relatie tot het algemene lineaire model verwijzen we naar Van den Bercken en Voeten (2002).

In hoofdstuk 10 keren we terug naar de specificatie van het model van lineaire regressie dat een onderzoeker zou willen hanteren om een bepaalde vraagstelling te beantwoorden. De onderzoeker moet kiezen welke variabelen van belang zijn om mee in het model te worden opgenomen. Verder moet de onderzoeker keuzes maken met betrekking tot de manier waarop de onafhankelijke variabelen geacht worden met de afhankelijke variabele samen te hangen. In dat verband zullen we zien dat het lineaire model zó omvattend is dat het ook sommige soorten van niet-lineaire relaties tussen variabelen in zich kan opnemen. Ook komt aan de orde hoe met lineaire regressie eventuele interactie-effecten kunnen worden onderzocht. Daarmee gaan we na hoe een variabele invloed heeft op het effect van een andere variabele.

In hoofdstuk 11 is de algemene statistische problematiek van het onder-

scheidingsvermogen van een statistische toets, met name de F-toets, aan de orde, toegepast op lineaire regressie. Dit is van groot praktisch belang in verband met de keuzes die onderzoekers moeten maken aangaande de grootte van hun steekproef en het aantal variabelen dat ze zich in een regressieanalyse kunnen veroorloven.

De hoofdstukken 12 tot en met 14 geven inleidingen op een aantal uitbreidingen van regressieanalyse naar meer geavanceerde toepassingen. Hoofdstuk 12 introduceert de logistische regressie waarmee het algemene lineaire model nog verder veralgemeend wordt naar toepassingen waarbij de afhankelijke variabelen kwalitatief van aard zijn. Wij beperken ons ter introductie tot het geval van een dichotome afhankelijke variabele. Hoofdstuk 13 geeft uitbreiding aan de bespreking in hoofdstuk 6 door te kijken naar een samenhangend stelsel van regressievergelijkingen (padanalyse en structurele vergelijkingen). Hierbij komt ook de analyse van mediatorvariabelen aan de orde. In hoofdstuk 14, ten slotte, besteden we aandacht aan regressieanalyse van data waarin eenheden op verschillende niveaus zijn te onderscheiden, individuen in hun sociale context. In de hoofdstukken 1 tot en met 12 gebruiken we SPSS voor de berekeningen, in hoofdstuk 13 en 14 presenteren we analyses die met andere software zijn verricht.

Een cursus die enkel een inleiding op regressieanalyse wil geven, zou zich in eerste instantie kunnen beperken tot de statistische en methodologische basisprincipes voor regressieanalyse (hoofdstuk 1 tot en met 4) en de analyse van residuen in hoofdstuk 7. Een meer uitgebreide cursus kan zich richten op hoofdstuk 1 tot en met 11, waarbij desgewenst bepaalde onderdelen (bijvoorbeeld uit hoofdstuk 8 en hoofdstuk 10) kunnen worden weggelaten en andere kunnen worden toegevoegd (uit de hoofdstukken 12 tot en met 14). Voor hoofdstuk 12 is het gewenst om eerst hoofdstuk 9 te hebben bestudeerd en voor hoofdstuk 13 is hoofdstuk 6 een voorwaarde.

We hebben ernaar gestreefd het boek ook geschikt te maken voor zelfstudie, waarbij van essentieel belang is gebruik te maken van het oefenmateriaal bij het boek. Het maken van de oefeningen en het werken met een computerprogramma zijn nodig om de stof te kunnen verwerken.

In het boek zijn vele verwijzingen naar literatuur opgenomen die de lezer kan benutten om zich verder in specifieke onderwerpen te verdiepen. De lezer die behoefte heeft aan een andere, meer uitgebreide inleiding op regressieanalyse verwijzen wij graag naar de boeken van Cohen, Cohen, West en Aiken (2003), Fox (1997) en Pedhazur (1997). Ter aanvulling bij dit boek zijn ook nuttig de vele groene boekjes over regressieanalyse uit de Sagereeks *Quantitative Applications in the Social Sciences*, waarvan er diverse in onze literatuurlijst zijn te vinden.

Bij dit boek en bij *Variantieanalyse* hoort een website waar men aanvullend materiaal kan vinden (www.data-analyse.nl). Er zijn oefeningen en uitwerkingen bij de diverse hoofdstukken. Verder is er een handleiding beschikbaar voor het werken met SPSS, met name voor REGRESSION en GLM, en zijn er aansturingen en databestanden te vinden voor alle analyses gerapporteerd in het boek, zodat men alle analyses kan nadoen. Verder zijn er computeropdrachten om te oefenen met verschillende vormen van regressieanalyse. Ook is een cd-rom beschikbaar bij het boek met dit en ander aanvullend materiaal.

Variabelen en beschrijvende statistiek

1

Veel empirisch onderzoek heeft tot doel om relaties tussen variabelen vast te stellen. In dit boek staat regressieanalyse centraal, als een statistische methode om relaties tussen variabelen te analyseren en hypothesen te toetsen over die relaties. Daarbij moeten minstens twee rollen worden onderscheiden die variabelen in onderzoek kunnen hebben. Gewoonlijk speelt een van de variabelen de rol van *afhankelijke variabele* (dependent variable, response variable, criterion variable), de variabele die uit andere variabelen moet worden voorspeld. Die andere variabelen worden *onafhankelijke variabelen* (independent variable) of predictoren genoemd. In dit boek gaan we er over het algemeen van uit dat de afhankelijke variabele gemeten is als een kwantitatieve, continue variabele; men spreekt dan meestal van een variabele gemeten op intervalniveau. We maken hierop één uitzondering door ook te kijken naar regressieanalyses voor dichotome afhankelijke variabelen, die slechts twee waarden kunnen aannemen (hoofdstuk 12). In principe hoeven de onafhankelijke variabelen niet op een kwantitatieve schaal te zijn gemeten, maar in de meeste gevallen zullen we daar toch van uitgaan. We schenken echter ook aandacht aan kwalitatieve onafhankelijke variabelen die in onderzoek zeer frequent voorkomen.

In de gedrags- en sociaal-wetenschappelijke onderzoeksliteratuur worden vormen van regressieanalyse zeer veel toegepast. We geven enkele voorbeelden van dergelijke onderzoeken die in de vorm van tijdschriftartikelen zijn gepubliceerd. Daarbij leggen we de nadruk op het doel van het onderzoek, de vraagstellingen en hypothesen. Deze voorbeelden dienen om enig idee te geven van het gebruik van regressieanalyse als een methode om relaties tussen variabelen na te gaan.

Het begin van elke data-analyse is zich te realiseren wat men wil bereiken, welke variabelen daarbij een rol spelen, welke rollen die variabelen spelen en van welke aard of meetniveau die variabelen zijn. In paragraaf 1.2 werken we een voorbeeld uit van vraagstelling tot data. Vervolgens geven we

een samenvatting van relevante kernbegrippen uit de beschrijvende statistiek en we eindigen met een schema van verwante analyseprocedures.

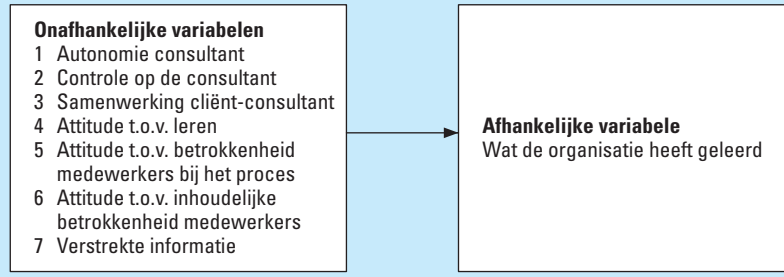
1.1 Voorbeelden van regressieanalyse ontleend aan de literatuur

De beschrijvingen van voorbeelden uit de onderzoeksliteratuur dienen vooral om in vraagstellingen of hypothesen te onderkennen welke variabelen een rol spelen, wat de rol is van deze variabelen (afhankelijke variabele, onafhankelijke variabele, controlevariabele, mediatorvariabele, moderatorvariabele) en wat de aard is van deze variabelen (kwantitatief gemeten variabele of kwalitatieve variabele). Daarnaast dienen deze voorbeelden om een variëteit van toepassingen van regressieanalyse te laten zien. In volgende hoofdstukken keren deze voorbeelden terug om toepassingen van regressieanalyse te illustreren.

1.1.1 Voorbeeld 1: Leren van organisaties

Vraagstelling: Welke factoren bepalen dat een organisatie leert van het inschakelen van externe consultants?
Sommige bedrijven of organisaties schakelen externe consultants in als ze bepaalde doelen van organisatieverandering willen bereiken. Je kunt je afvragen wat een organisatie in feite leert van deze consultants. Kan de organisatie zodanig leren dat in de toekomst soortgelijke doelen van organisatieverandering kunnen worden gerealiseerd, zonder daarbij opnieuw externe consultants nodig te hebben? Het blijkt dat sommige organisaties leren terwijl andere afhankelijk blijven van externe consultants. Halvari, Johansen en Sørhaug (1998) verrichtten een onderzoek naar dit verschijnsel om na te gaan welke factoren bepalen dat een organisatie leert van het consultancyproces. In deze vraagstelling is 'het leren door de organisatie' de *afhankelijke variabele*. De *onafhankelijke variabelen* zijn de mogelijke determinanten van leren die je zou kunnen onderscheiden. Daarbij kun je denken aan de manier waarop het consultancyproces is ingericht en aan kenmerken van de medewerkers van de betrokken organisatie. Figuur 1.1 vat samen welke variabelen door de onderzoekers zijn gekozen en gemeten.

Figuur 1.1 **Schema van variabelen in het onderzoek van Halvari et al. (1998)**



De steekproef in het onderzoek van Halvari et al. bestond uit 109 managers en 22 werknemervertegenwoordigers van één bedrijf. Alle deelnemers hadden met een consultant gewerkt. De data zijn verzameld via een per post toegezonden vragenlijst. De afhankelijke en alle onafhankelijke variabelen

zijn via één en dezelfde vragenlijst gemeten. Op basis van de items van de vragenlijst zijn schalen geconstrueerd, bestaande uit één tot zeven items per schaal. In totaal kwamen zo acht variabelen tot stand: zeven onafhankelijke en één afhankelijke. Alle acht variabelen kunnen als kwantitatief gemeten variabelen worden opgevat. Een belangrijke vraag daarbij is natuurlijk of deze variabelen op een valide en betrouwbare manier zijn gemeten, maar dergelijke vragen vallen buiten het bestek van dit boek.

Een hoofddoel van het onderzoek was na te gaan welke van de zeven onafhankelijke variabelen de variatie in de mate van leren kunnen verklaren en hoe goed de mate van leren uit deze variabelen kan worden verklaard. Met regressieanalyse kun je nagaan welke relaties er bestaan tussen de zeven onafhankelijke variabelen enerzijds en de afhankelijke variabele anderzijds. Gegeven dat het hier om kwantitatief gemeten variabelen gaat, kun je natuurlijk de correlatie uitrekenen van elke onafhankelijke variabele apart met de afhankelijke variabele. Die correlaties zeggen al iets over het antwoord op de vraagstelling. Als je echter naar elke onafhankelijke variabele apart kijkt, negeer je de mogelijkheid dat de onafhankelijke variabelen onderling samenhangen. Regressieanalyse houdt rekening met de onderlinge samenhangen tussen de onafhankelijke variabelen. Een ander verschil tussen correlationele analyse en regressieanalyse is dat correlatie een symmetrisch begrip is en regressie niet. In een regressieanalyse gaat het om relaties van de vorm $X \rightarrow Y$, en deze relatie is niet hetzelfde als $Y \rightarrow X$. Het maakt dus uit welke variabele je de rol van afhankelijke variabele (Y) geeft en welke de rol van onafhankelijke (X). Bij een correlatie tussen twee variabelen maakt dat op zich niet uit.

Wat bepaalt nu of een variabele afhankelijke variabele of onafhankelijke variabele is? Dat hangt geheel af van de vraagstelling of doelstelling van het onderzoek. De onderzoekers zijn in een bepaald verschijnsel geïnteresseerd, in dit geval de verschillende mate waarin organisaties leren van ingeschakelde consultants. Dit verschijnsel wordt de afhankelijke variabele. De factoren die dit leren kunnen bevorderen of belemmeren, zijn dan potentiële onafhankelijke variabelen. Vanuit praktisch oogpunt is het daarbij van belang om speciaal te letten op onafhankelijke variabelen die zodanig beïnvloed kunnen worden dat ze het organisatieleren bevorderen. Wat zijn belangrijke condities voor het optimaal inrichten van het consultancyproces? Als je dergelijke condities kunt vinden die aantoonbaar het organisatieleren bevorderen, dan kun je als onderzoeker praktische aanbevelingen doen.

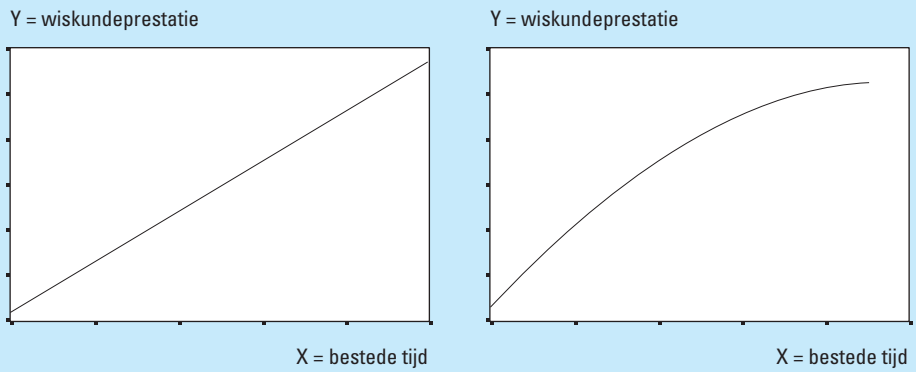
1.1.2 Voorbeeld 2: Informatietechnologie in het wiskundeonderwijs

Vraagstelling: Heeft het integreren van informatietechnologie in het wiskundeonderwijs op de middelbare school een positief effect op de wiskundeprestaties van de leerlingen?
Taylor (1999) onderzocht de effecten van het gebruik van *Learning Expedition*, een softwarepakket voor het wiskundeonderwijs, in één bepaalde Engelse middelbare school op de wiskundeprestaties van de leerlingen in het eerste en tweede leerjaar. Haar aanpak om het programma te evalueren was als volgt. Zij registreerde de tijd die een leerling besteedde aan het werken met Learning Expedition. Vervolgens ging zij na in welke mate de geconstateerde verschillen in bestede tijd samenhingen met de resultaten die leerlingen behaalden op een wiskundetoets aan het eind van het schooljaar. De evaluatievraag is hier een vraag naar de samenhang tussen twee variabelen.

De *afhankelijke variabele* is het resultaat op de wiskundetoets en de *onafhankelijke variabele* is de hoeveelheid tijd besteed aan het werken met het programma. In onderzoek om de effecten van een bepaalde behandeling (treatment) na te gaan, vergelijkt men meestal deelnemers die de behandeling ondergingen (experimentele groep) met deelnemers die de behandeling niet ondergingen (controlegroep). In dat geval zou de onafhankelijke variabele kwalitatief van aard zijn en zou de data-analyse vooral bestaan uit het vergelijken van de gemiddelde wiskundeprestaties, met een t-toets of met variantieanalyse. Het onderzoek van Taylor betrof echter een school waar alle leerlingen gebruik konden maken van het softwarepakket. De onderzoeker verzamelde informatie over de mate waarin de leerlingen dat daadwerkelijk deden. Als de wiskundeprestaties gemiddeld hoger worden naarmate een leerling meer tijd besteed heeft aan het werken met de software, dat wil zeggen, als er een *positief* effect is van bestede tijd op prestaties, dan zou daaruit iets kunnen blijken over de effectiviteit van de software. Nu is ook de onafhankelijke variabele kwantitatief van aard en kan regressieanalyse dienstig zijn om de relatie tussen prestatie en bestede tijd na te gaan.

Een kwantitatieve onafhankelijke variabele noemen we vaak een *predictor*. Als een variabele X samenhangt met een variabele Y, of sterker gezegd, als een variabele X effect heeft op een variabele Y, dan kun je ook zeggen dat je Y kunt voorspellen op basis van X; vandaar de naam 'predictor'. De afhankelijke variabele Y wordt vaak het *criterium* of de *criteriumvariabele* genoemd. In het onderzoek van Taylor was de prestatie op de wiskundetoets het criterium om er de effectiviteit van het softwarepakket aan af te meten. Voor een onafhankelijke variabele gebruiken we meestel het symbool X; een afhankelijke variabele duiden we aan als Y. In een grafiek wordt de onafhankelijke variabele op de X-as uitgezet en de afhankelijke variabele op de Y-as. Figuur 1.2 toont twee mogelijke ideeën over de relatie tussen wiskundeprestatie (Y) en bestede tijd (X).

Figuur 1.2 **Twee mogelijke modellen voor de relatie tussen 'wiskundeprestatie' en bestede tijd: links een lineair stijgend verband, rechts een niet-lineair stijgend verband**



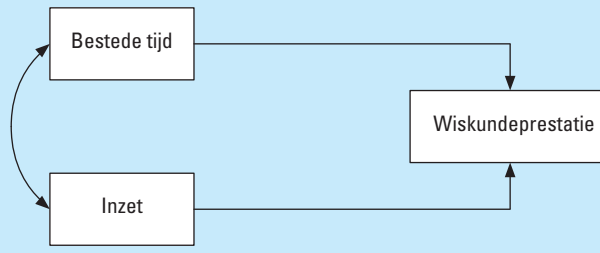
In de linkergrafiek is er sprake van een positieve lineaire relatie tussen de beide variabelen. De wiskundeprestatie stijgt constant naarmate meer tijd is besteed aan het werken met de software. In de rechtergrafiek neemt de wiskundeprestatie ook toe naarmate meer tijd wordt geïnvesteerd, maar die toe-

name wordt steeds minder naarmate de bestede tijd groter is. Er is in de rechtergrafiek sprake van een niet-lineair verband tussen de twee variabelen. Beide gevallen kun je interpreteren als evidentie voor de effectiviteit van het softwarepakket. De onderzoeker moet dus een goed model specificeren voor de relatie tussen de afhankelijke en onafhankelijke variabele.

In dit boek gaan we voornamelijk uit van lineaire verbanden tussen variabelen, maar we laten tevens zien dat je met lineaire regressieanalyse ook een niet-lineair verband als in figuur 1.2 kunt onderzoeken (hoofdstuk 10). Als er sprake is van één afhankelijke variabele en één onafhankelijke variabele, spreken we van *enkelvoudige regressie* (simple regression). Het zal echter zelden het geval zijn dat je in onderzoek met slechts één onafhankelijke variabele kunt volstaan. Dat was ook in het onderzoek van Taylor niet het geval.

Omdat de resultaten op een wiskundetoets met tal van factoren kunnen samenhangen (zoals aanleg voor wiskunde, leerbereidheid, inzet, kwaliteit van het onderwijs en ondersteuning door ouders), kon het onderzoek niet beperkt blijven tot slechts de twee variabelen genoemd in de vraagstelling. Als je die andere variabelen zou negeren, kun je niet het specifieke effect van het programma bepalen. Deze andere mogelijke invloeden op de toetsresultaten kunnen immers samenhangen met de hoeveelheid tijd die een leerling besteedt aan het programma. Dat kan de samenhang tussen de afhankelijke en de onafhankelijke variabele vertekenen. Het effect van de tijd besteed aan het programma zou (deels) een effect kunnen zijn van de inzet die studenten vertonen bij het leren van wiskunde. In de methodologie wordt dit *contaminatie* of, in het Engels, *confounding* genoemd. Je zou de variabele 'inzet' hier een storende variabele mogen noemen; immers de relatie die we met het onderzoek tussen 'bestede tijd' en 'wiskundeprestaties' willen vaststellen, wordt vertekend of verstoord door de variabele 'inzet'. De relaties tussen deze drie variabelen zijn schematisch in beeld gebracht in figuur 1.3. De potentieel *storende* variabele 'inzet' wordt geacht net als de bestede tijd een effect te hebben op de wiskundeprestaties. Tot zover is dat nog niet storend. Het wordt pas storend als 'inzet' en bestede tijd ook nog eens gecorreleerd zijn. Dan kan immers een effect dat we menen te zien van 'bestede tijd' in feite (mede) een effect van 'inzet' zijn. Dat zou de validiteit van onze conclusie over de effectiviteit van het softwarepakket aantasten. Men zegt dan dat het effect van 'bestede tijd' gecontamineerd (confounded) is met het effect van 'inzet'.

Figuur 1.3 **Voorbeeld van een potentieel storende variabele**

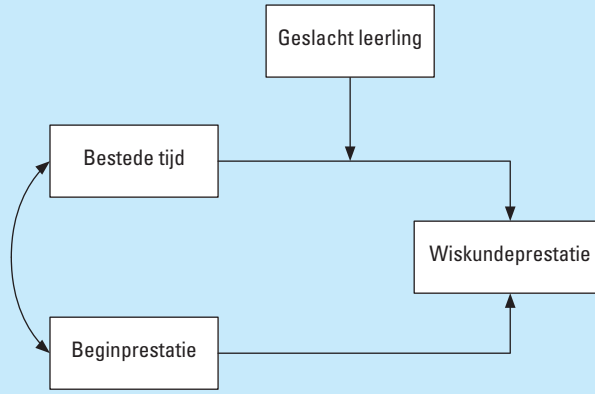


Een potentieel storende variabele is een variabele die gecorreleerd is met de onafhankelijke variabele en die tevens een effect heeft op de afhankelijke variabele.

Een manier om *storende variabelen* uit te schakelen, is ze te meten en ze als onafhankelijke variabele in de regressieanalyse op te nemen. Zo'n variabele noemen we een *controlevariabele* (control variable). Het probleem van de onderzoekster was echter dat zij niet al deze, mogelijk storende, invloeden kon uitschakelen of meten. Haar oplossing was om de wiskundeprestaties bij het begin van het eerste jaar op de middelbare school als controlevariabele in het onderzoek mee te nemen. Je mag aannemen dat de beginprestaties een weerspiegeling vormen van de meeste van de mogelijk storende invloeden (behalve natuurlijk de kwaliteit van het wiskundeonderwijs gedurende die eerste jaren op de middelbare school). Zo waren er in dit onderzoek dus twee onafhankelijke variabelen: de beginprestatie en de tijd besteed aan het programma. We spreken nu van een tweevoudige of, algemener, een *meervoudige* regressieanalyse (multiple regression): er is één afhankelijke variabele en er is meer dan één onafhankelijke variabele. De controlevariabele 'beginprestatie' is dan net als 'bestede tijd' een onafhankelijke variabele in de analyse. Van de variabele 'beginprestatie' kun je ook zeggen dat deze hier de rol heeft van *covariabele* (covariable, covariate). Met de meervoudige regressieanalyse kun je het effect vaststellen dat de 'bestede tijd' heeft op de wiskundeprestaties bovenop het effect van de controlevariabele 'beginprestatie'. Dit is een *statistische* manier van uitschakelen van storende variabelen, te onderscheiden van de controles uitgeoefend in experimentele designs.

In het onderzoek van Taylor werd ook nog rekening gehouden met de mogelijkheid dat het effect van de bestede tijd op de wiskundeprestaties voor meisjes anders zou kunnen zijn dan voor jongens. Dit betekent dat het geslacht van de leerling in het onderzoek is betrokken als een *moderatorvariabele*. De veronderstelde relaties tussen de variabelen in het onderzoek kunnen we ons nu voorstellen als in figuur 1.4. Het effect van de moderatorvariabele is voorgesteld door een pijl van geslacht leerling naar de pijl tussen bestede tijd en wiskundeprestaties aan het eind van het schooljaar.

Figuur 1.4 **Variabelenschema in het onderzoek van Taylor (1999)**



De wiskundeprestatie bij het begin van het schooljaar functioneert als controlevariabele voor het effect van bestede tijd op de wiskundeprestatie bij het eind van het schooljaar. Het geslacht van de leerling fungeert als moderatorvariabele.

Het geslacht van de leerling kan daarnaast een effect hebben op de wiskundeprestaties, dat wil zeggen dat de gemiddelde wiskundeprestaties van jongens en meisjes zouden kunnen verschillen. Verder kan het ook zijn dat jongens en meisjes gemiddeld verschillende tijden hebben besteed aan het werken met het softwarepakket. Deze twee mogelijke relaties zijn in de figuur niet ingetekend. Het ging er de onderzoekster primair om na te gaan of het effect van het werken met de software voor meisjes anders zou zijn dan voor jongens: dat is het moderatoreffect van geslacht.

Voor het onderzoek waren gegevens beschikbaar bij 148 leerlingen uit het eerste en 117 leerlingen uit het tweede leerjaar. Per leerjaar werd een regressieanalyse uitgevoerd. Hieruit kon worden geconcludeerd dat de tijd besteed aan het programma in beide leerjaren een positief effect had op de wiskundeprestaties voor leerlingen met een gelijk beginniveau. Dit effect bleek sterker te zijn voor meisjes dan voor jongens. Meisjes profiteerden dus meer van het werken met de software dan jongens. De conclusie van de onderzoekster was dat het voor leerlingen loont veel tijd te investeren in het programma Learning Expedition.

1.1.3 Voorbeeld 3: Betrokkenheid van ouders bij kind in een instelling

Baker, Blacher en Pfeiffer (1996) onderzochten de volgende drie vragen:

1 In welke mate voelen ouders en andere gezinsleden zich betrokken bij een kind dat vanwege een verstandelijke handicap of een psychiatrische stoornis in een instelling is geplaatst?
2 Neemt de gezinsbetrokkenheid af in de tijd?
3 Hoe is de gezinsbetrokkenheid gerelateerd aan voorzieningen geboden door de instelling, aan kenmerken van het gezin en aan kenmerken van het uit huis geplaatste kind, met name de diagnostische status?

Voor het beantwoorden van de eerste vraag maakten de onderzoekers onderscheid tussen betrokkenheid bij het gezinslid (kindbetrokkenheid) en betrokkenheid bij de door de instelling geboden faciliteiten (programmabetrokkenheid). Van beide vormen van betrokkenheid gingen ze na hoe vaak die voorkwamen. De tweede vraag betreft de correlatie tussen de frequentie van contacten met de instelling en de tijd verstreken sinds de uithuisplaatsing. De derde vraag werd opgesplitst in 3a: de vraag naar de verschillen in betrokkenheid per diagnostische status, en 3b: de vraag hoe kindbetrokkenheid en programmabetrokkenheid zijn gerelateerd aan faciliteiten, gezinskenmerken en kindkenmerken.

Zowel vraag 2 als de vragen 3a en 3b zijn vragen naar de samenhang tussen afhankelijke variabelen enerzijds en onafhankelijke variabelen anderzijds. In al deze gevallen zijn er *twee afhankelijke variabelen*: kindbetrokkenheid en programmabetrokkenheid. In vraag 2 is de *onafhankelijke variabele* de tijd verstreken sinds de uithuisplaatsing. Deze vraag is beantwoord door correlaties (met de t-toets voor de correlatiecoëfficiënt) te berekenen. Ook in vraag 3a is er slechts één *onafhankelijke variabele* (diagnostische status). Deze variabele is echter *kwalitatief* van aard; er werden drie globale diagnostische categorieën onderscheiden (psychiatrische stoornis, verstandelijke handicap, combinatie van psychiatrische stoornis en verstandelijke handicap). Vanwege het kwalitatieve karakter van de onafhankelijke variabele is vraag 3a via (enkelvoudige) variantieanalyses beantwoord. Vraag 3b omvat meer-

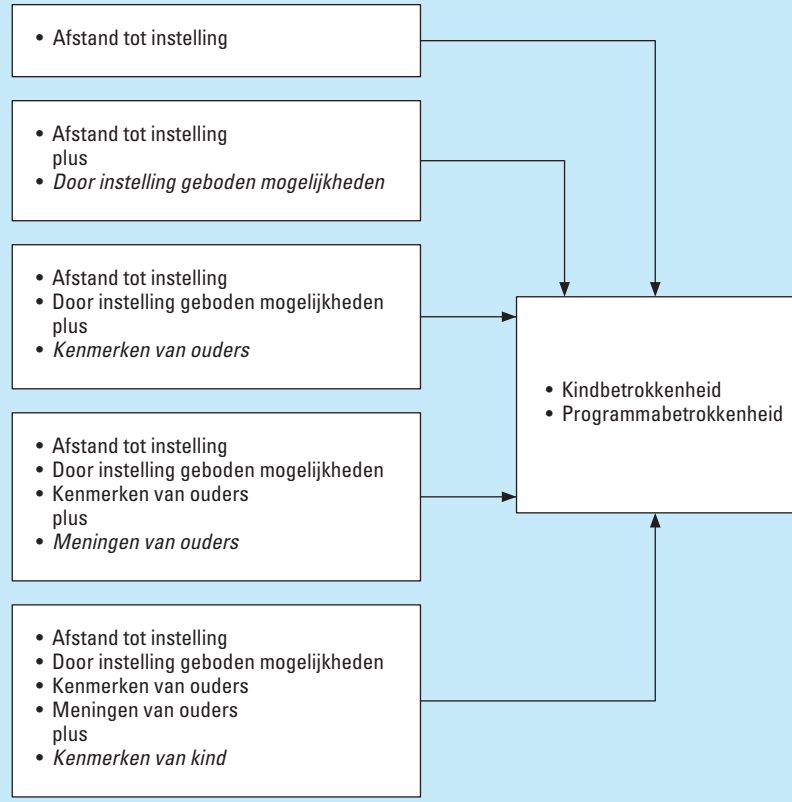
dere *kwantitatieve* onafhankelijke variabelen en leent zich daarom voor meervoudige regressieanalyse.

Deelnemers aan het onderzoek waren gezinsleden (meestal ouders) van 163 personen die in behandeling waren in drie grote instellingen in de Verenigde Staten en die een vragenlijst hadden beantwoord. De onderzoekers voerden regressieanalyses uit ter beantwoording van vraag 3b voor kindbetrokkenheid en programmabetrokkenheid apart, waarbij zij op inhoudelijke gronden hun elf onafhankelijke variabelen verdeelden in vijf subsets:

1 afstand tot de instelling (één variabele);
2 gepercipieerde mogelijkheden tot betrokkenheid geboden door de instelling (één variabele);
3 kenmerken van de ouders (vier variabelen);
4 meningen over de uithuisplaatsing (drie variabelen);
5 kenmerken van het kind (leeftijd en IQ).

In totaal werden dus elf onafhankelijke variabelen in de regressieanalyses betrokken. De analyses werden uitgevoerd door de vijf subsets onafhankelijke variabelen successievelijk toe te voegen in de hier aangegeven volgorde (zie figuur 1.5).

Figuur 1.5 **Schematisch overzicht van vijf regressieanalyses met toevoeging van predictoren in elke stap**



Per subset variabelen stelden de onderzoekers vast of deze subset een statistisch significante *additionele* bijdrage had aan het voorspellen van de betrokkenheid en hoe groot die bijdrage was. Van de variantie van kindbetrokkenheid kon in totaal 52% worden verklaard. De eerste twee subsets, die betrekking hadden op de instelling waar het kind verbleef, verklaarden 34%. Hoe groter de afstand tot de instelling, hoe minder de kindbetrokkenheid. De ouderkenmerken (het derde blok van variabelen) leverden geen statistisch significante extra bijdrage. Dat was wel het geval voor de meningen van de ouders, met name voor het antwoord op de vraag of men dacht dat het verblijf in de instelling permanent zou zijn (een variabele toegevoegd in de vierde analyse). Hoe langer het verwachte verblijf, hoe minder de kindbetrokkenheid. Ook de kindkenmerken hadden een statistisch significante bijdrage, met name IQ. Ouders waren meer betrokken naarmate de verstandelijke handicap minder was. Van de variantie in programmabetrokkenheid kon 43% worden verklaard. Daarbij werd de belangrijkste bijdrage geleverd door de gepercipieerde mogelijkheden geboden door de instelling (28%). Hoe meer de ouders mogelijkheden tot participatie zagen, hoe groter de programmabetrokkenheid was. Andere statistisch significante predictoren waren de afstand en de huwelijkse staat van de respondent (grotere programmabetrokkenheid bij intacte gezinnen). De onderzoekers gaven implicaties van deze bevindingen voor het vergroten van de ouderbetrokkenheid.

In dit onderzoek werden de onafhankelijke variabelen dus in een door de onderzoekers van tevoren bepaalde volgorde geplaatst, beginnend bij kenmerken van de instelling (het eerste en tweede blok van variabelen), dan kenmerken van de ouders (het derde en vierde blok) en als laatste kenmerken van het kind zelf (zie figuur 1.5). Net als in het eerste voorbeeld (subparagraaf 1.1.1) gaat het bij deze studie om de vraag welke van de predictoren bijdragen aan het verklaren van verschillen in scores op de afhankelijke variabele. In het eerste voorbeeld is gekeken naar de eigen of *unieke* bijdrage die elke predictor kan leveren in het voorspellen van de score op de afhankelijke variabele, uniek ten opzichte van alle andere predictoren. In het huidige voorbeeld is bekeken welke bijdrage een predictor kan leveren bovenop de bijdragen van de predictoren uit een vorige stap. De resultaten van de analyses zullen mede afhangen van de ordening die de onderzoekers in de predictoren aanbrachten. De onderzoekers kozen ervoor om kenmerken van de instelling als eerste aan bod te laten komen. Variabelen die later in de volgorde zijn geplaatst, kunnen hun effect pas tonen nadat de effecten van de eerder geplaatste variabelen zijn verdisconteerd. Deze werkwijze vereist van de onderzoekers dat ze een heldere motivering geven voor de gekozen volgorde van onafhankelijke variabelen; hiervoor verwijzen we naar het artikel zelf. Op regressieanalyses van deze vorm wordt nader ingegaan in hoofdstuk 5.

1.1.4 **Voorbeeld 4: Kindermishandeling en emotionele steun door moeder**

Wanneer iemand slachtoffer is geweest van kindermishandeling kan dat ook in het volwassen leven (negatieve) psychologische gevolgen hebben. In het onderzoek van Weismann, Wind en Silvern (1994) is nagegaan welke gezinsvariabelen mediëren tussen kindermishandeling en de aard en sterkte van de langetermijneffecten. Wanneer een slachtoffertje warmte en ondersteuning van (een van) de ouders heeft ervaren en is opgegroeid in een gezin

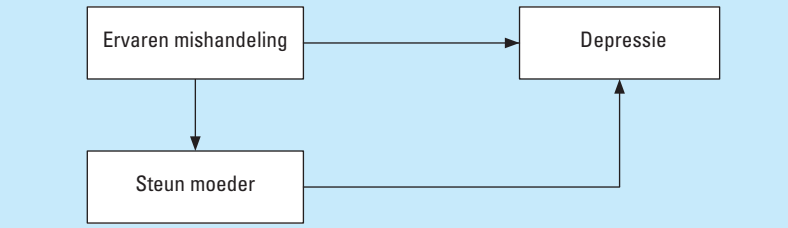
met weinig stress, kan dat het optreden van negatieve effecten op volwassen leeftijd afzwakken. Het onderzoek concentreerde zich op vrouwen die in hun jeugd binnen het gezin fysiek of seksueel waren mishandeld. Verschijnselen die op volwassen leeftijd optreden als gevolg van kindermishandeling omvatten onder meer depressie en traumasymptomen. Dit waren twee *afhankelijke variabelen* in het onderzoek. De verwachtingen van de onderzoekers waren onder meer:

a dat een laag niveau van ouderlijke emotionele ondersteuning en mishandeling binnen het gezin aan elkaar gerelateerd zijn en
b dat elk van deze gerelateerd is aan depressie op volwassen leeftijd.

De *onafhankelijke variabelen* waren: mishandeld zijn en ondersteuning door de moeder. Alle variabelen werden opgevat als gemeten op intervalniveau, behalve de variabele mishandeld zijn. Deze variabele was nominaal van aard, voor ons doel nu beperkt tot wel of niet mishandeld binnen het gezin.

De veronderstelde relaties tussen enkele van de onderzochte variabelen zijn weergegeven in figuur 1.6.

Figuur 1.6 **'Steun moeder' als interveniërende variabele tussen 'ervaren mishandeling' en latere depressieverschijnselen**



Volgens figuur 1.6 wordt de mate van depressie zowel beïnvloed door het wel of niet mishandeld zijn als door de steun ervaren van de moeder. De gedachte was dat 'steun moeder' hier een *mediatorvariabele* of *interveniërende variabele* kan zijn. Dat betekent dat de samenhang tussen het wel of niet mishandeld zijn en depressieverschijnselen op volwassen leeftijd geheel of gedeeltelijk zou verdwijnen als de variabele 'steun moeder' constant wordt gehouden. De relaties in de figuur houden in dat de variabele 'steun moeder' twee rollen heeft. Deze variabele is een onafhankelijke variabele voor het verklaren van variatie in depressieverschijnselen. Tevens heeft 'steun moeder' ook de rol van afhankelijke variabele: 'steun moeder' als afhankelijk van 'ervaren kindermishandeling'. Deze dubbelrol van afhankelijke en onafhankelijke variabele is typerend voor een interveniërende variabele.

De steekproef bestond uit 259 vrouwelijke stafleden van een bepaalde Amerikaanse universiteit. Deze vrouwen in de leeftijd van 19 tot 70 jaar beantwoordden een vragenlijst waarin alle genoemde variabelen plus enkele demografische kenmerken aan bod kwamen. Het onderzoek is dus retrospectief van aard en alle variabelen stammen uit één en dezelfde vragenlijst. De relatie tussen 'mishandeld zijn' en de gezinsvariabele 'steun moeder' werd vastgesteld door met een t-toets de gemiddelden van de twee groepen

op ‘steun moeder’ te vergelijken. Er bleek inderdaad een statistisch significante samenhang tussen deze twee variabelen te bestaan. Vervolgens werden regressieanalyses uitgevoerd voor elk van de afhankelijke variabelen apart, met twee onafhankelijke variabelen: wel of niet mishandeld zijn en mate van ondersteuning ervaren van de moeder.

1.1.5 Voorbeeld 5: Schoolsucces en cultureel kapitaal van de ouders

Vraagstelling: Hangt schoolsucces af van het cultureel kapitaal van de ouders?
Het is al vaak vastgesteld dat schoolsucces samenhangt met het sociale milieu van leerlingen. Leerlingen uit hogere sociale milieus hebben een succesvoller schoolloopbaan dan leerlingen uit lagere sociale milieus. In verklaringen voor dit verband wordt naast het economisch kapitaal ook gewezen op het cultureel kapitaal van de ouders. Milieuspecifieke verschillen in de beschikking over culturele hulpbronnen kunnen bijdragen tot verklaring van verschillen in schoolloopbanen. In een studie van De Graaf, De Graaf en Kraaykamp (2000) werd onderscheid gemaakt tussen twee varianten van de ‘culturele kapitaal’-hypothese. De eerste variant betreft het verschil tussen een hoge en lage culturele habitus. Een hoge culturele habitus houdt een hoge mate van deelname aan de hogere cultuurvormen in, zoals het bezoeken van klassieke-muziekconcerten en theaters of musea. De hypothese is dat ouderlijk cultureel kapitaal in deze zin kinderen de symbolische macht verschaft om zich de culturele codes van de hogere onderwijsvormen eigen te maken, waardoor ze meer succes op school hebben dan kinderen uit de lagere sociale milieus die deze culturele codes niet van hun ouders hebben meegekregen. De tweede variant van cultureel kapitaal legt de nadruk op cognitieve vaardigheden (met name lees- en taalvaardigheid) die door ouders aan hun kinderen worden doorgegeven. Vertrouwdheid met een leescultuur zal kinderen helpen het goed te doen op school, want leesmateriaal en lezen zijn essentiële elementen van de leeromgeving. De onderzoekers pasten de volgende twee varianten van de culturele-kapitaalhypothese toe voor het Nederlandse onderwijssysteem:

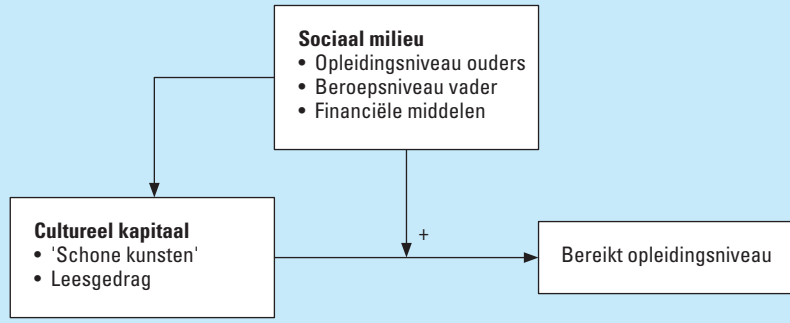
1a Deelname van de ouders aan ‘highbrow’ culturele activiteiten heeft een positief effect op de schoolloopbaan van de kinderen.
1b Het leesgedrag van de ouders heeft een sterker positief effect op de schoolloopbaan van de kinderen dan de deelname van de ouders aan ‘highbrow’ culturele activiteiten.

Deze hypothesen gaan over één *afhankelijke variabele*: schoolloopbaan, in het onderzoek geoperationaliseerd als het hoogst bereikte niveau van onderwijs uitgedrukt in jaren oplopend van vijf (basisschool niet afgemaakt) tot eenentwintig (gepromoveerd aan een universiteit). Er zijn twee (onderling gecorreleerde) *onafhankelijke variabelen*: ‘schone kunsten’-participatie en leesgedrag, elk geoperationaliseerd door vijf items uit een vragenlijst.

De onderzoekers brachten nog een verfijning aan in de theorie met betrekking tot effecten van cultureel kapitaal: de culturele-reproductietheorie versus de culturele-mobiliteitstheorie. Dit leidde tot de volgende hypothesen vanuit de culturele reproductietheorie (zie figuur 1.7):

2a Het effect van het ouderlijk cultureel kapitaal op de schoolloopbaan is sterker voor kinderen uit hogere sociale milieus dan voor kinderen uit lagere sociale milieus.
3a Het ouderlijk cultureel kapitaal medieert de effecten van sociaal milieu op de schoolloopbaan.

Figuur 1.7 **Voorbeeld met zowel een mediator- als een moderatorvariabele**



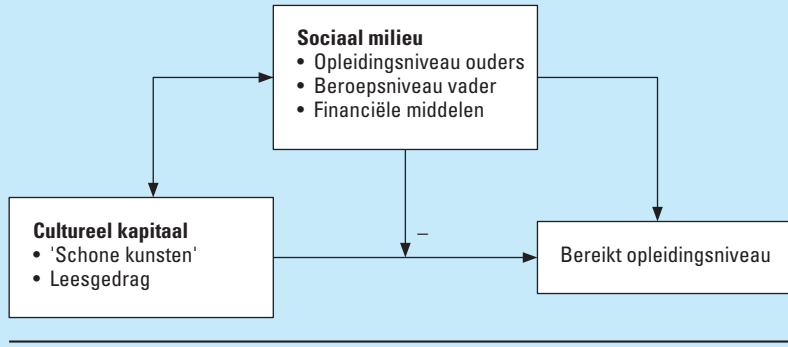
Cultureel kapitaal en sociaal milieu interacteren in hun effect op de schoolloopbaan, en tevens werkt het cultureel kapitaal van de ouders als een mediatorvariabele tussen sociaal milieu van de ouders en schoolloopbaan van het kind (culturele reproductietheorie).

In deze hypothesen speelt, naast de drie al genoemde variabelen, ook sociaal milieu van de ouders een rol als *onafhankelijke variabele*. Drie aspecten van sociaal milieu werden in het onderzoek betrokken: het opleidingsniveau van de ouders, het beroepsniveau van de vader en de financiële middelen van het gezin. In hypothese 2a krijgt sociaal milieu de rol van *moderatorvariabele*. Deze hypothese gaat over een *interactie-effect* van cultureel kapitaal en sociaal milieu op schoolloopbaan, in figuur 1.7 aangegeven als een pijl van 'sociaal milieu' naar het effect van 'cultureel kapitaal' van de ouders op de schoolloopbaan van het kind. Tegelijkertijd geeft figuur 1.7 overeenkomstig hypothese 3a aan, dat 'cultureel kapitaal' werkt als een interveniërende variabele of mediatorvariabele. Dat wil zeggen dat de samenhang tussen sociaal milieu en schoolloopbaan zou kunnen worden verklaard door het 'culturele kapitaal' van de ouders.

Vanuit de culturele-mobiliteitstheorie formuleerden de auteurs de volgende hypothesen (zie figuur 1.8):

2b Het effect van het ouderlijk cultureel kapitaal op de schoolloopbaan is sterker voor kinderen uit lagere sociale milieus dan voor kinderen uit hogere sociale milieus.
3b De effecten van het ouderlijk cultureel kapitaal komen additioneel op de effecten van sociaal milieu op de schoolloopbaan.

Figuur 1.8 **Twee gecorreleerde onafhankelijke variabelen, waarvan één moderatorvariabele**



Cultureel kapitaal en sociaal milieu interacteren in hun effect op de schoolloopbaan, en tevens levert het cultureel kapitaal van de ouders naast het sociale milieu van de ouders additioneel een bijdrage aan de schoolloopbaan van het kind (culturele-mobiliteitstheorie).

De hypothese 2a en 2b spreken elkaar tegen; dat is in de figuren aangegeven door bij het moderatoreffect van sociaal milieu in figuur 1.7 een plusteken te plaatsen en in figuur 1.8 een minteken. Slechts één van de hypothesen kan waar zijn, maar ze kunnen ook beide onwaar zijn. Hypothese 3a en 3b spreken elkaar eveneens tegen. In hypothese 3b zijn cultureel kapitaal en sociaal milieu twee, mogelijk gecorreleerde, *onafhankelijke variabelen* die samen mede de schoolloopbaan bepalen. De correlatie tussen sociaal milieu en cultureel kapitaal is in figuur 1.8 aangegeven door een pijl in twee richtingen. Deze correlatie wordt in figuur 1.8 als een gegeven geaccepteerd zonder er een verklaring voor te zoeken.

Naast de genoemde variabelen waren ook nog drie *controlevariabelen* in het onderzoek betrokken, namelijk geslacht, geboortecohort en al dan niet opgegroeid zijn in een eenoudergezin. De data-analyses waren gebaseerd op een steekproef van 1479 personen van 25 jaar of ouder die in 1992–1993 deelnamen aan de Nederlandse Gezinssurvey. De hypothesen werden getoetst met behulp van een aantal meervoudige regressieanalyses, waarbij steeds het bereikte opleidingsniveau de afhankelijke variabele was. De onderzoekers concludeerden dat voor het Nederlandse schoolsysteem niet de ouderlijke participatie aan de schone kunsten, maar wel het leesgedrag van de ouders een effect had op de schoolloopbaan van de kinderen. Verder concludeerden zij dat de data geen steun gaven aan de culturele-reproductietheorie maar wel aan de culturele-mobiliteitstheorie.

1.1.6 Voorbeeld 6: Seksuele intimidatie op middelbare scholen

Lee, Croninger, Linn en Chen (1996) onderzochten de volgende drie vragen: Welke scholieren worden seksueel lastiggevallen? Hoe hangt de ernst van de seksuele intimidatie samen met kenmerken van het individu en van de schoolcontext? Welke ongewenste gevolgen (leerproblemen, psychologische problemen, vermijdingsgedrag) heeft seksuele intimidatie? Het onderzoek betrof een secundaire analyse van data verzameld in een grootschalig survey onder leerlingen van de eerste vier leerjaren van openbare scholen voor voortgezet onderwijs in de Verenigde Staten.

Bij de eerste vraagstelling is de *afhankelijke variabele* het al dan niet seksueel lastiggevallen zijn. De onderzoekers wilden nagaan welke kenmerken van de scholieren zelf (geslacht, leeftijd, etniciteit, sociaal milieu en studieprestaties) en welke kenmerken van de schoolcontext samenhangen met seksueel lastiggevallen worden. Er zijn dus *twee sets onafhankelijke variabelen* te onderscheiden: persoonskenmerken en schoolkenmerken. Als schoolkenmerken hanteerden de onderzoekers vier variabelen die aangeven in welke mate er een cultuur van seksuele intimidatie op de school bestaat:

1 de perceptie die leerlingen hebben van seksuele intimidatie op hun school;
2 de ervaringen van vrienden met ongewenste seksuele intimiteiten;
3 of de leerlingen seksueel lastiggevallen werden door een staflid van de school;
4 of leerlingen ooit betrokken zijn geweest in het zelf seksueel lastigvallen van anderen op school.

De interesse van de onderzoekers ging vooral naar deze laatste set variabelen om te kunnen aantonen dat ongewenste seksuele intimiteiten en agressie niet als puur individuele verschijnselen kunnen worden verklaard, maar dat ze cultureel zijn bepaald.

De tweede en de derde vraagstelling betreffen enkel de studenten die zeiden seksueel te zijn lastiggevallen. In de tweede vraagstelling is de ernst van seksueel lastiggevallen zijn de *afhankelijke variabele*. Deze ernst werd gescoord naar hoe vaak het voorkwam en naar hoe erg het was ervaren door de leerling. De *onafhankelijke variabelen* zijn dezelfde als bij de eerste vraagstelling.

In de derde vraagstelling zijn de ervaren negatieve consequenties de *afhankelijke variabelen*. Hierbij waren de onderzoekers er vooral in geïnteresseerd om na te gaan of deze negatieve consequenties verschillen voor jongens en meisjes, als wordt gecontroleerd op de ernst van seksueel lastiggevallen zijn. Er zijn dus bij deze vraagstelling twee *onafhankelijke variabelen* te onderscheiden. Merk op dat de afhankelijke variabele van de tweede vraagstelling nu bij de derde vraagstelling als onafhankelijke variabele naar voren komt.

Bij de eerste vraagstelling is de afhankelijke variabele *dichotoom* (binair) van aard; er zijn slechts twee categorieën onderscheiden. De consequentie hiervan is dat deze vraagstelling er zich niet voor leent om met lineaire regressieanalyse te worden beantwoord. In alle voorgaande voorbeelden was de afhankelijke variabele steeds een kwantitatieve variabele die we als een variabele gemeten op intervalniveau zouden kunnen beschouwen. Het ‘wel of niet zijn lastiggevallen’ voldoet daar niet aan. Daarom pasten de onderzoekers bij deze vraagstelling *logistische* regressieanalyse toe (zie hoofdstuk 12). De tweede vraagstelling leent zich wel voor beantwoording via meervoudige lineaire regressie, omdat voor de ‘ernst van lastiggevallen zijn’ scores werden bepaald. Deze afhankelijke variabele kan als een kwantitatieve variabele worden beschouwd. Ten behoeve van de derde vraagstelling beschikten de onderzoekers over scores die de mate van ervaren problemen aangaven, maar ze besloten deze scores te dichotomiseren, omdat ze extreem scheef waren verdeeld; een sterke concentratie van problemen kwam namelijk veel minder voor dan de ervaring hebben van geen of slechts een paar problemen. Daardoor ontstonden afhankelijke variabelen met slechts twee categorieën, zodat ook bij deze vraagstelling logistische regressie de aangewezen analyseprocedure was.

Welke rollen kunnen variabelen spelen in data-analyse?

Afhankelijke variabele
De afhankelijke variabele representeert het verschijnsel dat men wil beschrijven of verklaren, of de variabele die men uit andere variabelen wil voorspellen. De afhankelijke variabele wordt ook wel criteriumvariabele genoemd.

Onafhankelijke variabele
Een variabele die moet dienen om een verschijnsel te beschrijven of te verklaren, of om een andere variabele te voorspellen. De onafhankelijke variabele wordt ook wel predictor genoemd, een naam die stamt uit het gebruik van regressieanalyse voor predictiedoeleinden maar die ook buiten die context wel wordt gehanteerd.

Mediatorvariabele
Een variabele Z die in een causale keten tussen twee andere variabelen X en Y gedacht wordt: $X \rightarrow Z \rightarrow Y$. De mediatorvariabele medieert het effect van X op Y en geeft daarmee een verklaring voor het bestaan van een samenhang tussen X en Y. Als de mediatorvariabele constant gehouden wordt, dan verdwijnt de samenhang tussen X en Y geheel of gedeeltelijk. Een mediatorvariabele, ook wel interveniërende of mediërende variabele genoemd, speelt dus twee rollen: is tegelijk afhankelijke en onafhankelijke variabele.

Moderatorvariabele
Een variabele die invloed heeft op de relaties tussen twee of meer andere variabelen. Men zegt dan ook wel dat er sprake is van interactie-effecten. Als de moderatorvariabele verschil in groepen van onderzoekseenheden beschrijft (kwalitatieve variabele), dan houdt deze interactie in dat de samenhang tussen afhankelijke en onafhankelijke variabelen verschillend is in de verschillende groepen. Een moderatorvariabele is een onafhankelijke variabele.

Controlevariabele
Een variabele Z die als onafhankelijke variabele in een analyse meegenomen wordt om de samenhangen tussen de andere variabelen te kunnen bestuderen met constanthouding van de waarde van Z. Het maakt daarbij niet uit op welke waarde Z constant gehouden wordt; als dat wel zou uitmaken, dan fungeert Z als een moderatorvariabele. Soms worden controlevariabelen ook covariabelen genoemd. Een controlevariabele is een onafhankelijke variabele die in de vraagstelling vaak niet expliciet benoemd wordt.

Verwaarloosde variabele (storende variabele)
Een variabele die ten onrechte niet in het regressiemodel is opgenomen. Dit is een variabele die de afhankelijke variabele beïnvloedt en die samenhangt met een of meer onafhankelijke variabelen.

1.2 Van vraagstelling naar data

Voordat een onderzoeker toe is aan de data-analyse is er al heel wat gebeurd: er is een vraagstelling ontwikkeld, een onderzoeksopzet uitgewerkt, de data zijn verzameld, gecodeerd en in een computerbestand ondergebracht. De eerste stap van de data-analyse is dan om de data te screenen, samen te vatten en in beeld te brengen met behulp van beschrijvende statistiek. In deze sectie lopen we dit voorwerk na aan de hand van een voorbeeld. Daarmee verschaffen we tevens een reële context aan de artificiële dataset waarmee wij in de hoofdstukken 2 tot en met 5 de techniek van meervoudige regressieanalyse zullen illustreren. Deze context is ontleend aan daadwerkelijk verricht onderzoek (Mommers, 1987).

Bij het onderwijs in leren lezen moeten leerkrachten rekening houden met verschillen in de beginsituaties van hun leerlingen. Daarbij is van groot belang om zo vroeg mogelijk de leerlingen te identificeren die risico lopen op leesmoeilijkheden. Er bestaan toetsen om na te gaan in hoeverre kinderen voldoen aan auditieve en visuele voorwaarden voor het leren lezen, zoals het herkennen van beginklanken van woorden en het visueel kunnen onderscheiden van combinaties van letters. Voor dergelijke toetsen is het van belang te onderzoeken wat de voorspellende waarde ervan is voor de lees-

prestaties na een jaar leesonderwijs. Met name zullen we willen weten *welke* toetsen een bijdrage leveren aan het voorspellen van de leesprestaties. We werken zo'n onderzoek in een eenvoudige vorm uit en illustreren daarbij vooral de exploratie van de data die gewenst is voor je met regressieanalyse aan de slag gaat.

1.2.1 Vraagstelling: afhankelijke variabele en onafhankelijke variabelen

Eerst specificeren we de *vraagstelling* van het onderzoek. We willen nagaan hoe groot de voorspellende waarde is van leesvoorwaardentoetsen afgenomen vóór het begin van het eigenlijke leesonderwijs. Daartoe gaan we de relatie na tussen de scores op dergelijke toetsen en de prestaties op een toets voor leesvaardigheid, afgenomen na het eerste jaar leesonderwijs op de basisschool. De *afhankelijke variabele* die we in dit geval proberen te voorspellen, is de leesprestatie. De variabelen waarmee we die voorspelling proberen te doen, zijn de *onafhankelijke variabelen*, in dit geval de toetsen voor leesvoorwaarden. We beperken ons in eerste instantie tot een toets voor de auditieve leesvoorwaarden (fonemische analyse). De vraagstelling is dan hoe goed we de verschillen in prestaties op de leesvaardigheidstoets kunnen voorspellen uit de verschillen in scores op de toets voor auditieve leesvoorwaarden die één jaar eerder bij dezelfde leerlingen werd afgenomen.

In een onderzoek naar de relatie tussen leesvaardigheid en auditieve leesvoorwaarden kan de eenvoudigste vorm van regressieanalyse worden toegepast, namelijk enkelvoudige regressieanalyse, omdat het slechts om de relatie tussen twee (kwantitatieve) variabelen gaat, een afhankelijke en een onafhankelijke variabele. Als we naast auditieve ook visuele leesvoorwaarden onderscheiden, dan ontstaat een onderzoeksvraag naar de relatie tussen aan de ene kant de leesvaardigheid als afhankelijke variabele en aan de andere kant de auditieve en de visuele leesvoorwaarden als onafhankelijke variabelen. In deze *vraagstelling* gaat het om drie variabelen. Om precies te zijn, het gaat om de vraag of we de prestaties op een toets voor leesvaardigheid (Y), afgenomen aan het einde van een jaar leesonderwijs, kunnen voorspellen op basis van toetsen voor auditieve en visuele leesvoorwaarden (X_1 en X_2), afgenomen aan het begin van het onderwijs. Een dergelijke vraagstelling met een kwantitatieve afhankelijke variabele en twee of meer onafhankelijke variabelen leent zich voor *meervoudige regressieanalyse*.

De onderzoeksvraag van het voorbeeld zou nog verder kunnen worden uitgebreid; niet alleen met meer predictoren van leesvaardigheid, maar ook met meerdere aspecten van leesvaardigheid, bijvoorbeeld technisch lezen en begrijpend lezen. Wanneer er twee of meer kwantitatieve afhankelijke variabelen zijn, spreekt men wel van *multivariate* regressieanalyse. Sommige auteurs spreken van 'multivariate regression' of in het algemeen van multivariate analyse zodra de data-analyse over meer dan twee variabelen tegelijk gaat. Anderen reserveren de term 'multivariate regression' voor regressieanalyses waarin twee of meer *afhankelijke* variabelen tegelijk worden geanalyseerd. Zij kijken dus alleen naar het aantal afhankelijke variabelen; bij meer dan één onafhankelijke variabele wordt dan de term multipele of meervoudige regressie gebruikt. In de gedragswetenschappen is dit laatste gebruikelijk (Grimm & Yarnold, 1995, 2000; Tabachnick & Fidell, 2001). Wij gebruiken de term univariate analyse in geval van één enkele afhankelijke variabele, ongeacht het aantal onafhankelijke variabelen. Van multivariate analyse is sprake als twee of meer *afhankelijke* variabelen tegelijk worden geanalyseerd. Bij één onafhankelijke variabele gebruiken we de term enkelvoudige regressieanalyse; bij twee of meer onafhankelijke variabelen spreken we van meervoudige of multipele regressieanalyse (zie ook *Variantieanalyse*, Van den Bercken & Voeten, 2002, voor dezelfde terminologie).

De vraag of we de latere leesprestatie kunnen voorspellen uit een leesvoorwaardentoets, is een vraag naar de samenhang tussen de leesprestaties enerzijds en de leesvoorwaarden anderzijds. Over het algemeen stellen we zo'n vraag, omdat we een samenhang verwachten. Meer in het bijzonder verwachten we een positieve samenhang: bij hogere scores op de leesvoorwaardentoetsen zullen eerder hogere dan lagere leesprestaties horen. We hebben dus een meer of minder uitgewerkte *hypothese* over de samenhang tussen de variabelen. Wat we willen onderzoeken is, hoe deze samenhang er precies uitziet en hoe sterk deze samenhang is. Kunnen we een formule vinden waarmee we de (toekomstige) score van een leerling op de leestoets kunnen bepalen, als we de scores op de leesvoorwaardentoets kennen?

Wat is multipele regressie?

Multipele regressie is een statistische methode om de relaties te bestuderen tussen een kwantitatieve afhankelijke variabele enerzijds en een of meer onafhankelijke variabelen anderzijds. Met de term 'kwantitatieve variabele' bedoelen we een variabele die gemeten is op het niveau van minstens een intervalschaal. De onafhankelijke variabelen in een multipele regressie zijn meestal ook kwantitatief van aard, maar in een multipele regressie kunnen ook kwalitatieve onafhankelijke variabelen (nominaal of ordinaal meetniveau) voorkomen.

1.2.2 Methode van onderzoek

Gegeven de vraagstelling werken we de manier uit waarop we de vraagstelling gaan onderzoeken: de *methode* van onderzoek. Daarbij moeten we antwoord geven op de volgende vragen: wat is de *opzet* van het onderzoek (het design): wie, wat en wanneer gaan we observeren (deelnemers, variabelen en meetinstrumenten, tijdstippen of condities)? Wat is de concrete *procedure* voor het verzamelen van de gegevens? En op welke manier gaan we de data analyseren, welke *statistische technieken* gaan we gebruiken? Niet alleen de vraagstelling maar ook de manier waarop het onderzoek is opgezet, bepalen welke statistische analyse geschikt is.

De *deelnemers* die we in het onderzoek betrekken, moeten natuurlijk representatief zijn voor de *relevante populatie*: de groep waarop we de uitkomsten van ons onderzoek van toepassing achten. In het voorbeeld kunnen dat kinderen zijn in het derde jaar van de basisschool, kinderen die beginnen met systematisch leesonderwijs. Hoeveel kinderen betrekken we daadwerkelijk in ons onderzoek en hoe vinden we die? Het antwoord op deze vraag hangt af van twee soorten overwegingen: theoretische en praktische. Theoretisch (statistisch) geldt: hoe meer aselect (random) gekozen kinderen we observeren, hoe beter onze onderzoeksresultaten de hele populatie zullen weergeven. Praktisch gesproken zullen tijd en middelen mede bepalen hoeveel kinderen we werkelijk kunnen observeren. In een onderzoek als het onderhavige ligt het voor de hand om uit te gaan van een steekproef van scholen voor basisonderwijs; de steekproef bestaat dan in principe uit alle leerlingen die in het schooljaar van de eerste meting in groep 3 zitten van de deelnemende scholen. Hoeveel scholen en welke scholen we kunnen kiezen, hangt af van de beschikbare financiële middelen en van de bereidheid van scholen om aan het onderzoek deel te nemen. Aan het huidige onderzoek namen 24 scholen uit Nijmegen en omstreken deel.

Welke *variabelen* we willen meten hebben we al aangegeven in de vraagstelling. De vraag is nu: hoe gaan we die meten? De manier waarop relevante variabelen kunnen worden gemeten, is een vak op zich (testtheorie, psychometrie, meet- en schaaltheorie). Laten we aannemen dat we beschikken over geëigende *meetinstrumenten*; 'geëigend' betekent hier: betrouwbaar en valide in de zin van de testtheorie. Voor de auditieve en visuele leesvoorwaarden maken we gebruik van de fonemische analysetest en de letterclustertest (Boland & Mommers, 1991). Voor het meten van de leesvaardigheid gebruiken we een toets voor begrijpend lezen (Cito Lees en Begrijp 1A). We veronderstellen dat onze tests voor leesvaardigheid en voor auditieve en visuele leesvoorwaarden valide scores opleveren, waarbij een hogere score een hogere vaardigheid betekent. Bij dergelijke variabelen is regressieanalyse een geschikte techniek om de vraag naar aard en mate van samenhang tussen de variabelen te beantwoorden.

In het voorbeeld volgt het antwoord op de vraag op welke *tijdstippen* we de metingen verrichten gemakkelijk uit de vraagstelling zelf. De leesvoorwaardentoets wordt afgenomen vóór het begin van het leesonderwijs, in het begin van groep 3 van het basisonderwijs, en de leesprestatie wordt een jaar later gemeten, in het begin van groep 4. In de praktijk houdt een dergelijke opzet, waarbij variabelen met een grote tussentijd worden gemeten, het risico in van uitval van deelnemers. Alleen al daarom moet de steekproefomvang van meet af aan zo groot mogelijk worden genomen.

Met het voorgaande hebben we globaal beschreven hoe we ons onderzoek kunnen opzetten. In een onderzoeksverslag zouden we meer concrete details moeten geven over de samenstelling van de steekproef, over de testbatterij en over de afnameprocedure. In principe moet het verslag de replicatie van het onderzoek door anderen mogelijk maken.

1.2.3 **Data**

De data bestaan uit drie scores voor elk van de kinderen die deelnamen aan het onderzoek. De eerste zorg is altijd of de data correct zijn ingevoerd in het databestand. Bij een klein databestand kan dit blijken door alle ingevoerde scores te controleren. Bij een groot databestand is dat niet praktisch, maar dan kunnen invoerfouten blijken uit het berekenen van de samenvattende descriptieve statistieken en uit grafische exploratie van de data. Fouten die men op die manier op het spoor kan komen, betreffen in ieder geval de scores die buiten het bereik van mogelijke scores op de variabelen vallen.

Een grote zorg zijn altijd de ontbrekende scores. Dit is in het bijzonder een zorg bij longitudinale data, zoals in het geval van ons voorbeeld. Het gaat immers om twee tijdstippen van meting die een jaar uit elkaar liggen. In het bijzonder is in dit geval het feit van belang dat het om data gaat uit twee verschillende schooljaren. Er zijn verschillende oorzaken waardoor scores kunnen ontbreken. In het huidige voorbeeld zijn met name de volgende van belang:

- Leerlingen die bij de eerste meting aanwezig waren, zijn niet meer aanwezig bij de tweede meting omdat ze in groep 3 zijn blijven zitten of zijn verwezen naar het speciaal onderwijs.
- Leerlingen die bij de eerste meting aanwezig waren, zijn niet meer aanwezig bij de tweede meting omdat ze naar een andere school zijn verhuisd.

- Leerlingen die in groep 4 zijn blijven zitten, zijn bij de tweede meting aanwezig maar niet bij de eerste.
- Leerlingen die na het tijdstip van de eerste meting op een van de onderzoeksscholen zijn ingestroomd, hebben een score op de tweede meting maar niet op de eerste.
- Leerlingen die om toevallige redenen afwezig waren op een van de dagen dat een test werd afgenomen, hebben een ontbrekende score op die test.

Het is van groot belang dat een onderzoeker nagaat welke patronen van ontbrekende scores in de data voorkomen. De SPSS-procedure *Missing Value Analysis* (MVA) biedt mogelijkheden om de patronen van ontbrekende scores te onderzoeken en om desgewenst ontbrekende scores door geschatte waarden te vervangen. Tabel 1.1 toont twee tabellen uit de output van MVA voor de data van het voorbeeld.

Tabel 1.1 **Patronen van ontbrekende scores vastgesteld met SPSS MVA**

Tabulated Patterns

	Missing Patterns [a]			
Number of Cases	*Visueel*	*Auditief*	*Lezen*	*Complete if ...* [b]
571				571
47			X	618
6	X			577
30	X	X		614
7		X		578

Patterns with less than 0.5% cases (3 or fewer) are not displayed.
a Variables are sorted on missing patterns.
b Number of complete cases if variables missing in that pattern (marked with X) are not used.

Separate Variance t Tests[a]

Begrijpend Lezen		*Visueel*	*Auditief*	*Lezen*
	t	2.8	1.9	.
	df	51.6	56.2	.
	P(2-tail)	.008	.059	.
	# Present	578	577	614
	# Missing	48	48	0
	Mean(Present)	30.79	16.31	19.29
	Mean(Missing)	27.81	14.44	.

For each quantitative variable, pairs of groups are formed by indicator variables (present, missing).
a Indicator variables with less than 5% missing are not displayed.

Het bovenste deel van tabel 1.1 laat zien dat volledige data op de drie variabelen beschikbaar zijn voor 571 kinderen in de steekproef. Er zijn 47 kinderen (7% van de totale steekproef) voor wie scores op beide leesvoorwaardentoetsen beschikbaar zijn, maar niet op de criteriumvariabele. Deze kinderen waren dus om een of andere reden afwezig bij het tweede meetmoment. Het onderste deel van tabel 1.1 toont de resultaten van een t-toets

waarbij de kinderen die aanwezig waren op het tweede meetmoment worden vergeleken met de kinderen van wie bij de tweede meting een score ontbreekt. Uit deze resultaten blijkt dat de kinderen met een ontbrekende score op de meting in groep 4 gemiddeld lager scoorden op de leesvoorwaardentoetsen dan de kinderen die bij beide meetmomenten een score hadden; voor de visuele leesvoorwaarden was dit verschil statistisch significant. We zien dat de 578 kinderen die een score hadden op de toets voor begrijpend lezen gemiddeld 30.79 scoorden op de toets voor visuele leesvoorwaarden, terwijl de 48 kinderen met een ontbrekende score op begrijpend lezen een lager gemiddelde behaalden op visuele leesvoorwaarden, namelijk 27.81. Het laatste is in de tabel aangeduid als Mean(Missing), het gemiddelde van de groep met een ontbrekende score (missing value) op begrijpend lezen. Die laatste groep bestond kennelijk uit 48 kinderen: de 47 kinderen met alleen een 'missing' op begrijpend lezen in groep 4 plus één kind dat daarnaast ook een 'missing' had op visuele leesvoorwaarden in groep 3. De resultaten duiden er dus op dat sprake was van een systematische uitval; bij de leerlingen met ontbrekende scores op de criteriumvariabele waren dat juist leerlingen met lage scores op de leesvoorwaardentoetsen. Dit verschijnsel kan betekenen dat de analyse die de voorspelbaarheid van de leesprestaties op basis van de leesvoorwaardentoetsen wil vaststellen, een enigszins vertekend beeld oplevert. Sommige 'risicokinderen' die we willen opsporen, waren bij de tweede meting al niet meer aanwezig. We hadden dit probleem misschien (deels) kunnen voorkomen door te kiezen voor een criteriumvariabele die op een eerder tijdstip is gemeten.

Het tweede meest frequente patroon van ontbrekende scores betreft de 30 kinderen zonder score op de leesvoorwaardentoetsen maar met wel een score op de toets voor begrijpend lezen. Dit betreft dus kinderen die afwezig waren bij de metingen in het begin van groep 3, waarschijnlijk omdat ze later in de betrokken scholen zijn ingestroomd of omdat ze in groep 4 zijn blijven zitten. Deze leerlingen toonden geen statistisch significante verschillen op de toets voor begrijpend lezen, vergeleken met de leerlingen die wel een score hadden op de leesvoorwaardentoetsen (niet getoond in de tabel). Deze ontbrekende scores kunnen mogelijk als toeval worden opgevat. De overige patronen van ontbrekende scores komen weinig voor en zijn dus geen bron van zorg.

Deze analyse van de ontbrekende scores laat zien dat de onderzoeker niet zomaar aan het verschijnsel voorbij kan gaan. In hoofdstuk 8 komen we terug op de vraag wat hieraan eventueel kan worden gedaan. We negeren nu de ontbrekende scores en gaan door met de beschrijvende statistieken van de beschikbare data.

1.2.4 Beschrijvende statistieken

In een onderzoeksverslag presenteert men zelden of nooit de ruwe data; men volstaat met de beschrijvende statistieken, en wel die welke het de lezer mogelijk maken de gerapporteerde analyses na te doen. Soms is dat niet goed doenlijk, omdat er bijvoorbeeld te veel variabelen in een onderzoek betrokken zijn. Om anderen de gelegenheid te geven de data zelf ook te analyseren – eventueel vanuit een ander gezichtspunt – met alternatieve of aanvullende analyses, is het gebruikelijk dat men de ruwe data beschikbaar houdt gedurende ten minste vijf jaar nadat men er een rapport over heeft gepubliceerd (*American Psychological Association*, 2001, p. 137).

De relevante descriptieve statistieken voor de data bestaan uit de kenmerken van univariate verdelingen, namelijk de gemiddelden en varianties of standaardafwijkingen van alle variabelen, en uit de kenmerken van de bivariate verdelingen, namelijk de covarianties of correlaties voor elk paar van variabelen. Deze gegevens voor de 571 kinderen met volledige data staan in tabel 1.2. Deze tabel bevat alles wat nodig is om een meervoudige regressieanalyse te kunnen uitvoeren, althans de hoofdzaken. Ook computerprogramma's kunnen vaak een regressieanalyse uitvoeren op basis van de samenvattende statistieken van tabel 1.2. Dat is vooral praktisch als men gepubliceerde data wil heranalyseren. Men kan dan volstaan met het invoeren van een correlatiematrix, de gemiddelden en standaarddeviaties en het aantal waarnemingen.

Tabel 1.2 **Beschrijvende statistieken voor de data (cases met een score op alle variabelen) van het longitudinaal onderzoek naar leesvaardigheid**

	Auditieve leesvoorwaarden X_1	Visuele leesvoorwaarden X_2	Begrijpend lezen Y
Aantal	625	626	614
Effectief aantal[a]	571	571	571
Minimum	1	0	0
Maximum	29	38	27
Gemiddelde	16.33	30.91	19.37
Standaardafwijking	6.86	5.39	6.08
Correlaties			
Auditieve leesvoorwaarden	1.00	.38	.39
Visuele leesvoorwaarden	.38	1.00	.46
Begrijpend lezen	.39	.46	1.00

a Aantal kinderen met een score op alledrie variabelen (listwise deletion).

Door het aantal te vergelijken met het 'effectief aantal' in de tabel wordt duidelijk hoeveel ontbrekende scores er waren bij elke variabele. Uit de tabel blijkt dat met name bij de toets voor visuele leesvoorwaarden en bij de leesvaardigheidstoets het gemiddelde vrij dicht nadert tot de maximumscore op de toets; het gemiddelde ligt weinig meer dan één standaardafwijking van het maximum vandaan. Deze twee toetsen vertonen dus een frequentieverdeling die scheef is naar links. Hoe deze verdelingen er precies uitzien, kunnen we beter met een grafische voorstelling van de data nagaan.

De drie variabelen blijken onderling positief te zijn gecorreleerd. Een veelvoorkomende fout bij rapportages van onderzoek is om niet alle correlaties te rapporteren, maar alleen de correlaties die statistisch significant zijn op 5%-niveau. Als niet alle correlaties worden vermeld, dan hinder je daarmee een lezer die de analyses (op een andere manier) zou willen herhalen. Het is ook fout vanuit het oogpunt van het toetsen van statistische hypothesen (Humphreys, 2002). Als je een nulhypothese niet kunt verwerpen, dan betekent dit nog niet dat je moet accepteren dat de correlatie in de populatie gelijk is aan 0. Verder zouden best twee correlaties statistisch significant van elkaar kunnen verschillen, ook als een ervan of zelfs beide niet significant

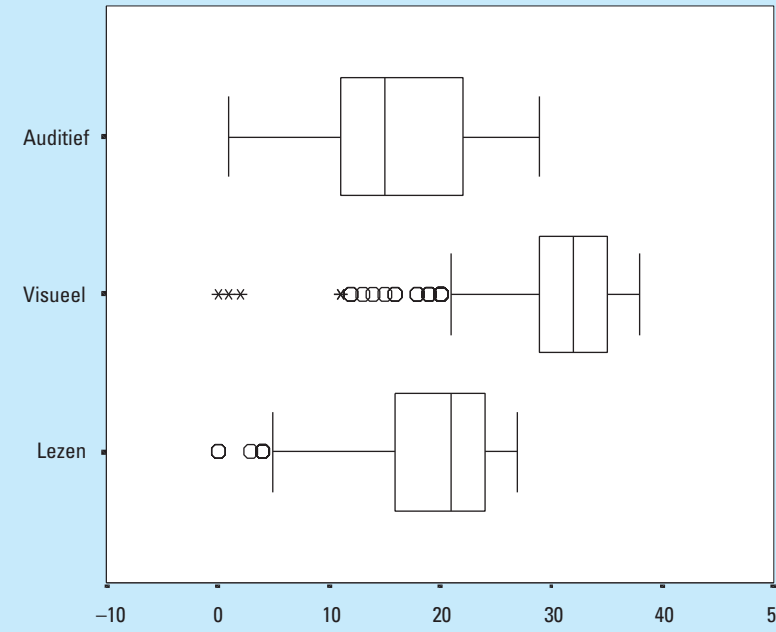
van 0 verschillen. In tabel 1.2 verschillen alle correlaties significant van 0 (niet getoond in de tabel).

Deze correlaties betreffen de mate van lineaire samenhang tussen variabelen. Of de samenhang tussen variabelen lineair is dan wel een ander patroon vertoont, kunnen we echter niet direct uit de correlaties aflezen. Ook hiervoor is het nodig om naar een grafische voorstelling van de data te kijken.

1.2.5 Grafische weergave van de data

Figuur 1.9 laat de univariate verdelingen van de variabelen zien in de vorm van boxplots (Hamilton, 1992; Jacoby, 1997; Van Peet et al., 1995). De verdelingen van de leestoets en vooral van de toets voor visuele leesvoorwaarden zijn enigszins scheef naar links. Bij beide toetsen is er sprake van uitbijters (outliers) aan de onderkant van de verdeling, vooral bij de toets voor visuele leesvoorwaarden. Scores die verder dan anderhalve keer de lengte van de box (de interkwartielbreedte) onder het eerste kwartiel van de verdeling liggen, zijn apart aangegeven. Bij de visuele leesvoorwaarden is er nog verschil tussen uitbijters en extreme uitbijters. De laatste liggen meer dan driemaal de interkwartielbreedte (in het Engels ook 'midspread' genoemd) buiten de box en zijn aangeduid met een asterisk. De locaties van de boxen op de scoreschaal zijn niet goed vergelijkbaar voor de drie variabelen, omdat het scorebereik van de drie variabelen enigszins verschilt (zie tabel 1.2).

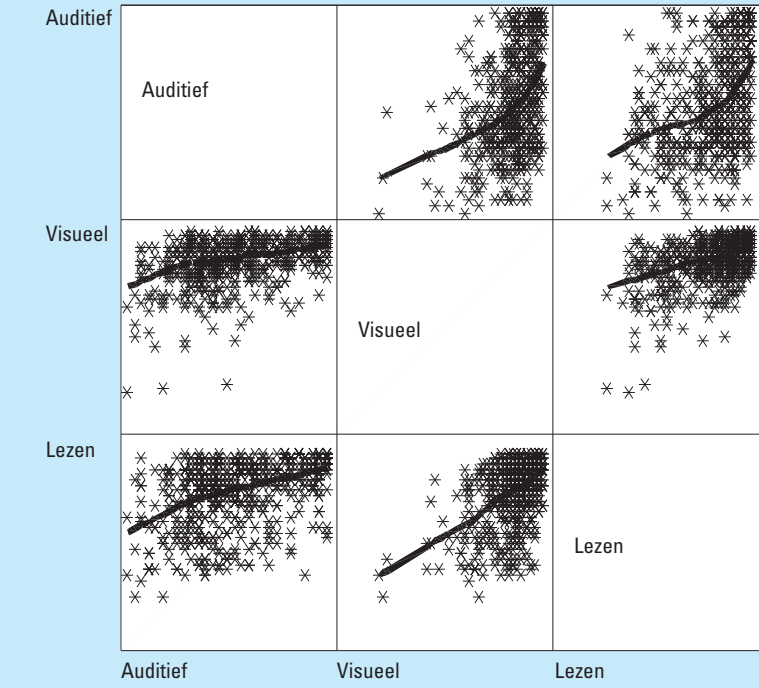
Figuur 1.9 **Boxplots van de drie variabelen uit tabel 1.2**



De boxplot is ontwikkeld als een van de technieken van de zogenoemde 'exploratieve data-analyse' (Tukey, 1977; zie voor een inleiding Velleman & Hoaglin, 1981, en Behrens, 1997). Dit is, zoals de naam zegt, een explorerende benadering die niet vertrekt vanuit een model of assumpties over de data. In deze benadering wordt een sterk beroep gedaan op grafische methoden om een visuele indruk te krijgen van de belangrijkste aspecten van de data. Exploratieve data-analyse kan belangrijke diensten bewijzen om een goed beeld van de data te krijgen voordat je een regressieanalyse gaat uitvoeren.

Zoals gezegd heb je aan correlaties nog niet genoeg om te zien hoe de afhankelijke variabele begrijpend lezen samenhangt met de auditieve en visuele leesvoorwaarden. Om te zien hoe de samenhang tussen twee variabelen is, kun je het beste de spreidingsdiagrammen bekijken voor de relaties tussen de diverse variabelen. Figuur 1.10 toont het spreidingsdiagram (scatterplot) voor elk paar variabelen, geordend in een matrix, de *scatterplotmatrix*.

Figuur 1.10 **Matrix van spreidingsdiagrammen voor drie paren van variabelen**



De naam op de diagonaal benoemt rijgewijs de Y-as en kolomgewijs de X-as.

Een matrix van spreidingsdiagrammen geeft de relaties weer tussen alle paren van variabelen. In feite betreft het voorbeeld echter data in een driedimensionale ruimte. Elke plot in de matrix geeft de relatie tussen twee variabelen weer met voorbijgaan aan de relaties met de andere variabele(n). De scatterplots zijn dus direct gerelateerd aan enkelvoudige regressie, maar zijn ook nuttig om te bekijken ter voorbereiding op een meervoudige regressie. Voor meer informatie over de scatterplotmatrix en over het visualiseren van data op meer dan twee variabelen tegelijk (multivariate data), zie Jacoby (1998).

Bij drie variabelen zoals in het voorbeeld zijn er drie paren van variabelen die in de matrix van scatterplots in figuur 1.10 dubbel zijn weergegeven, net als in de correlatiematrix van tabel 1.2. Neem het paar lezen en auditieve voorwaarden. In de ene weergave (eerste rij, derde vakje) staat lezen op de horizontale as uitgezet tegen auditieve voorwaarden op de verticale as; in de andere weergave (derde rij, eerste vakje) is het omgekeerd. Deze laatste weergave past het best, want het is gebruikelijk om de afhankelijke variabele op de verticale Y-as te zetten en de onafhankelijke variabele op de horizontale X-as. In elke plot is een curve ingetekend die de trend in de data aangeeft; gebruikt is een zogenoemde 'lowess'- (of 'loess'-)curve; zie voor meer informatie: Fox (2000). Deze curve geeft aan waar ongeveer het gemiddelde van Y ligt bij de verschillende waarden van X. 'Lowess' is een acroniem voor 'locally weighted scatterplot smoother'. We maken er ook gebruik van in hoofdstuk 7.

De twee plots met de auditieve en visuele leesvoorwaarden op de X-as en begrijpend lezen op de Y-as (derde rij, eerste en tweede vakje) laten zien dat de relaties van de afhankelijke variabele met elk van de twee onafhankelijke variabelen nagenoeg lineair zijn; de ingetekende curve is praktisch een rechte lijn. Er is overigens een lichte kromming zichtbaar in de plot voor de relatie tussen auditieve leesvoorwaarden en begrijpend lezen. Bij de visuele leesvoorwaarden spelen de enkele lage scores op deze toets een belangrijke rol in het bepalen van de lineaire trend.

De plots waaieren tamelijk breed uit; bij hoge scores op de leesvoorwaarden komen hoge maar ook nog veel lage scores op begrijpend lezen voor. Bij de hele lage scores op visuele leesvoorwaarden zie je enkel lage scores op begrijpend lezen; bij de lage scores op auditieve leesvoorwaarden zie je echter ook hoge scores op begrijpend lezen. Het brede uitwaaieren van de plots komt in tabel 1.2 tot uiting in de slechts middelmatige correlaties tussen de variabelen. De al in figuur 1.9 geconstateerde uitbijters op begrijpend lezen en vooral op visuele leesvoorwaarden zijn ook in de plots te zien, maar er zijn geen uitbijters zichtbaar die sterk van de trend in de data afwijken. De plots tonen duidelijk de scheve verdelingen van de variabelen.

De matrix van spreidingsdiagrammen brengt het asymmetrische karakter van regressie goed in beeld. De relaties tussen de leesvoorwaarden en begrijpend lezen zijn ook in de derde kolom afgebeeld boven de diagonaal, maar nu zijn de rollen van afhankelijke en onafhankelijke variabele omgekeerd. Met name in de plot voor auditieve leesvoorwaarden verticaal uitgezet en begrijpend lezen horizontaal uitgezet, zie je een andere en niet-lineaire trend, dan in de plot met de assen omgekeerd. Toch wordt de samenhang in beide plots uitgedrukt in dezelfde correlatiecoëfficiënt, namelijk $r = .39$. Voor de onderlinge samenhang tussen auditieve en visuele leesvoorwaarden kun je geen keuze uit beide plots maken, want we beschouwen deze twee variabelen enkel als onderling gecorreleerde onafhankelijke variabelen. In deze plots is duidelijk de hoge concentratie van scores zichtbaar bij hoge scores op de toets voor visuele leesvoorwaarden. Kennelijk was deze toets voor veel kinderen gemakkelijker dan de toets voor auditieve leesvoorwaarden. Deze scheefheid van de verdelingen zal de correlaties tussen de variabelen enigszins hebben verlaagd.

Samenvattend betekent dit dat de exploratieve analyse van de data in het voorbeeld vooral twee problemen naar voren brengt. Als gevolg van het longitudinale karakter is er sprake van uitval van deelnemers, waarbij waarschijnlijk juist zwakke lezers oververtegenwoordigd zijn. Verder is er sprake van scheve verdelingen voor met name de leesvaardigheidstoets en de toets voor visuele leesvoorwaarden, die wellicht voor de meeste leerlingen vrij gemakkelijk zijn en daardoor onvoldoende goed discrimineren. Deze twee problemen zullen leiden tot een verlaging van het voorspellend vermogen dat we kunnen vinden voor de twee leesvoorwaardentoetsen.

1.3 Samenvatting beschrijvende statistiek

De belangrijkste beschrijvende statistische grootheden voor een meervoudige regressieanalyse zijn de kenmerken van de univariate verdelingen van elke variabele en de kenmerken van de paarsgewijze of bivariate verdelingen. Aan de hand van de descriptieve statistieken voor Y en voor X_1 in tabel 1.2 in de vorige subparagraaf lopen we de definities van de voornaamste univariate en bivariate verdelingskenmerken na, hetgeen tevens de gelegenheid biedt om vertrouwd te raken met de symbolen en de notatie die we in deze tekst gebruiken. Zie tabel 1.3.

Tabel 1.3 **Overzicht van de belangrijkste univariate en bivariate verdelingskenmerken (geïllustreerd aan de hand van Y en X_1 uit tabel 1.2)**

Naam	Symbool	Definitie	Voorbeeld
Steekproefgrootte	n	aantal waarnemingen	571
Som	ΣY	som van de ruwe scores	11063
Gemiddelde	gem(Y), $\bar{Y}$	$(\Sigma Y)/n$	19.37
Kwadratensom	SS(Y)	$\Sigma(Y-\bar{Y})^2$	21091.80
Variantie	var(Y), s_y^2	$SS(Y)/(n-1)$	37.00
Standaarddeviatie	sd(Y), s_y	$\sqrt{s_y^2}$	6.08
Kruisproductensom	SCP(Y, X)	$\Sigma(Y-\bar{Y})(X-\bar{X})$	9282.92
Covariantie	cov(Y, X)	$\Sigma(Y-\bar{Y})(X-\bar{X})/(n-1)$	16.29
Standaardscore	z_y	$(Y-\bar{Y})/s_y$	
Correlatie	r_{yx}	$cov(Y,X)/s_y s_x$	.39
	r_{yx}	$\Sigma z_y z_x/(n-1)$	.39

Een centrale grootheid is de kwadratensom (SS = Sum of Squares) van een variabele. Hiermee bedoelen we doorgaans, zoals aangegeven in tabel 1.3, de som van gekwadrateerde afwijkingsscores. De variantie van een variabele Y volgt uit de kwadratensom van de afwijkingsscores SS(Y) en n; omgekeerd kun je de kwadratensom SS(Y) uitrekenen uit var(Y) en n; bijvoorbeeld: var(Y) = 37.00 = 21091.80/(571 – 1) en SS(Y) = 21091.80 = (571 – 1) × 37.00 Welke grootheden men gebruikt hangt af van de context van de analyse. Bij het beschrijven van de spreiding van scores heeft de standaarddeviatie meestal de voorkeur, omdat deze maat in de oorspronkelijke schaaleenheid uitgedrukt is. Maar bij het analyseren van variabiliteit zijn kwadratensommen praktischer. Het is dus goed om met alle vermelde grootheden vertrouwd te zijn en hun onderlinge relaties te kennen.

De meest bewerkelijke grootheden zijn die waarbij alle scores moeten worden gesommeerd, al dan niet na aftrek van het gemiddelde. Er zijn drie van zulke grootheden: de som van de scores, de kwadratensom en de kruisproductensom. Heb je die eenmaal berekend, dan is de rest betrekkelijk eenvoudig. De formules van de kwadratensom en de kruisproductensom in tabel 1.3 zijn niet de handigste voor het uitrekenen van deze grootheden. Er bestaan handiger rekenformules voor, maar die hebben we sinds het computertijdperk niet meer nodig. Wij geven in deze tekst bij voorkeur alleen definitieformules. Die laten heel duidelijk zien hoe de gedefinieerde grootheid conceptueel opgebouwd is en wat de onderlinge relaties zijn van de bestanddelen; ze omschrijven als het ware een begrip. Dat ze niet praktisch zijn, is niet zo erg, omdat we in beginsel alle berekeningen aan een computerprogramma over laten.

In tabel 1.3 is SS(Y) gedefinieerd als de zogenoemde *gecorrigeerde kwadratensom*, de kwadratensom van afwijkingsscores ten opzichte van het steekproefgemiddelde. Deze staat tegenover de ongecorrigeerde of ruwe kwadratensom, die simpelweg bestaat uit de som van de gekwadrateerde scores: $\Sigma Y^2 = \Sigma \overline{Y}^2$. De gecorrigeerde kwadratensom is gelijk aan $\Sigma(Y - \overline{Y})^2 = \Sigma Y^2 - n\overline{Y}^2$. In output van computerprogramma's wordt soms alleen de gecorrigeerde kwadratensom afgedrukt; het programma REGRESSION van SPSS bijvoorbeeld drukt de gecorrigeerde kwadratensom af onder de naam 'total sum of squares'. Soms rapporteert een programma, zoals GLM van SPSS, beide grootheden: 'total sum of squares' (= de ruwe kwadratensom) en 'corrected total sum of squares' (= de gecorrigeerde kwadratensom). Als we in het vervolg over de kwadratensom spreken, bedoelen we deze gecorrigeerde kwadratensom.

De (gecorrigeerde) kwadratensom SS(Y) is een maat voor de totale hoeveelheid variatie rondom het gemiddelde. Ze kan worden gedeeld door het aantal onafhankelijke scores dat die variatie bepaalt (het aantal vrijheidsgraden, degrees of freedom, df), hetgeen resulteert in de variantie; de variantie is dus het gemiddelde van de gekwadrateerde afwijkingen, een gemiddeld kwadraat (mean square).

Bij de voorbeelddata is de centrale vraag hoe Y varieert in relatie tot de onafhankelijke variabelen, de X-variabelen. Belangrijke informatie voor het beantwoorden van die vraag leveren de zogenoemde bivariate verdelingskenmerken, de covariantie en de correlatie. We kunnen het samen variëren van scores op twee kwantitatieve variabelen, Y en X, zoals dat in een spreidingsdiagram als in figuur 1.10 in beeld is gebracht, samenvatten in één enkel getal. In tabel 1.3 is dat in drie vormen gedaan: kruisproductensom, covariantie en correlatie. In de praktijk is de correlatie het belangrijkst, omdat deze de mate van samenhang uitdrukt als een getal tussen –1 en +1; de andere twee maten zijn niet begrensd en daardoor moeilijker te interpreteren. Voor de berekening van de correlatie beginnen we met de som van de kruisproducten, aangeduid met SCP (sum of cross products). Net zoals de kwadratensom het hoofdbestanddeel is van de variantie, zo is de kruisproductensom het voornaamste element voor de covariantie. Met een kruisproduct bedoelen we een product van de afwijkingsscore op de ene variabele Y met de afwijkingsscore op de andere variabele, X_1, bij dezelfde persoon. De covariantie is gelijk aan het gemiddelde van de kruisproducten. Delen we de covariantie van Y en X door de standaarddeviaties van Y en X, dan krijgen we de correlatie (voluit: de Pearson productmomentcorrelatiecoëfficiënt).

De Pearson correlatie wordt gewoonlijk met r genoteerd en is genoemd naar de Engelse wetenschapper Karl Pearson (1857–1936). Pearson was een leerling van Francis Galton (1822–1911) die als de grondlegger van correlationeel onderzoek en regressieanalyse kan worden beschouwd.

De correlatie is in feite een speciaal geval van de covariantie: het is de covariantie van gestandaardiseerde variabelen. De formules voor de covariantie en correlatie verschillen alleen in de elementen van de kruisproducten: bij de covariantie gaat het om ruwe scores, bij de correlatie om gestandaardiseerde scores. We vinden de correlatie tussen Y en X, dat wil zeggen de covariantie van de gestandaardiseerde scores, door simpelweg de definitie van de covariantie toe te passen op z_y en z_x, de standaardscores van Y en X, als volgt:

$$r_{yx} = \text{cov}(z_y, z_x) = \frac{\Sigma z_y z_x}{(n-1)}$$

Voor een covariantie hebben we de kruisproductensom nodig, in dit geval van de variabelen z_y en z_x: $\Sigma(z_y - \bar{z}_y)(z_x - \bar{z}_x)$. Maar het gemiddelde van gestandaardiseerde variabelen is gelijk aan 0, dus de kruisproductensom wordt gelijk aan $\Sigma z_y z_x$, waarmee we bij de gegeven formule terechtkomen.

Wanneer werk je met correlaties en wanneer werk je met covarianties?

Correlaties hebben het voordeel van standaardisering. Interpreteerbaarheid wordt bevorderd doordat de waarde van een correlatiecoëfficiënt voor positieve samenhang ligt tussen 0 (geen lineaire samenhang) en 1 (perfecte positieve lineaire samenhang). Evenzo wordt negatieve samenhang uitgedrukt door negatieve getallen tussen 0 en –1. Daarom worden over het algemeen correlaties gerapporteerd en geen covarianties.
Wanneer men echter de samenhang tussen variabelen wil vergelijken voor twee of meer groepen, dan ligt het niet voor de hand om dat te doen door de correlaties in de verschillende groepen met elkaar te vergelijken. Dat komt omdat een correlatiecoëfficiënt mede bepaald wordt door de spreidingen van de twee variabelen. Als die spreidingen verschillen in de groepen, dan vergelijk je met het vergelijken van correlaties niet alleen de sterkte van de samenhangen tussen variabelen, maar tegelijk ook de verschillen in spreiding tussen de groepen. Door nu covarianties te vergelijken in plaats van correlaties, maak je een zuivere vergelijking van alleen de mate van lineaire samenhang in de verschillende groepen.
Een covariantie is onafhankelijk van de n van de steekproef, het is het steekproefgemiddelde van de kruisproducten. Kruisproductensommen zijn niet direct vergelijkbaar tussen groepen omdat de hoogte ervan mede bepaald wordt door de groepsgrootte.

1.4 Keuze van een analyseprocedure

Bij de keuze van een analyseprocedure zijn de volgende elementen essentieel:

- het doel of de vraagstelling van het onderzoek;
- identificatie van de afhankelijke en onafhankelijke variabelen in de vraagstelling;
- de aard van de afhankelijke en onafhankelijke variabelen: gaat het om variabelen gemeten op intervalniveau of gaat het om variabelen die kwalitatief van aard zijn. Met *kwalitatieve* variabelen bedoelen we variabelen die de deelnemers aan het onderzoek naar een aantal categorieën onderscheiden, daarom ook wel *categorische* variabelen genoemd (nominaal of

ordinaal meetniveau) (Analyseprocedures die specifiek zijn voor ordinale data laten we in dit boek buiten beschouwing.);
- de rollen die de verschillende onafhankelijke variabelen hebben te spelen.

Bij de voorbeelden in paragraaf 1.1 en 1.2 was het doel steeds om een samenhang vast te stellen tussen een afhankelijke variabele aan de ene kant en een of meer onafhankelijke variabelen aan de andere kant. De voorbeelden verschillen in de meer of minder uitgewerkte hypothesen die de onderzoekers hadden en in de aard van de gebruikte variabelen. Tabel 1.4 geeft een overzicht van de analyseprocedures die aan bod komen in het onderhavige boek over regressieanalyse en in het hieraan gerelateerde boek *Variantieanalyse*. Regressie- en variantieanalyse zijn varianten van het algemene lineaire model en hangen dus zeer nauw samen.

Tabel 1.4 **Overzicht van analyseprocedures**

Afhankelijke variabele	**Onafhankelijke variabele**		
	Aard	*Aantal*	*Analyse*
Vragen over samenhang tussen variabelen			
Kwantitatief	Kwantitatief	1	Enkelvoudige regressieanalyse Hoofdstuk 2–4
Kwantitatief	Kwantitatief	>1	Meervoudige regressieanalyse Hoofdstuk 2–8, 10 en 11
Kwantitatief	Kwantitatief en kwalitatief	>1	Meervoudige regressieanalyse met dummyvariabelen Hoofdstuk 9
Dichotoom	Kwantitatief	1 of meer	Logistische regressieanalyse Hoofdstuk 12
Meerdere kwantitatief	Kwantitatief	1 of meer	Padanalyse Hoofdstuk 13
Vragen over verschillen tussen gemiddelden			
Kwantitatief	Kwalitatief	1	Enkelvoudige variantieanalyse Van den Bercken & Voeten (2002), hoofdstuk 3, 4, 12 en 13
Kwantitatief	Kwalitatief	>1	Meervoudige variantieanalyse Van den Bercken & Voeten (2002), hoofdstuk 5, 6 en 13
Kwantitatief	Kwalitatief en kwantitatief	>1	Covariantieanalyse Van den Bercken & Voeten (2002), hoofdstuk 7
Meerdere kwantitatief	Kwalitatief	0 of meer	Multivariate variantieanalyse Van den Bercken & Voeten (2002), hoofdstuk 8–11

In tabel 1.4 is het onderscheid gemaakt in termen van vragen naar samenhang tussen variabelen versus vragen naar verschillen tussen gemiddelden. Dat onderscheid is een beetje kunstmatig, want je kunt vragen naar verschillen tussen gemiddelden meestal gemakkelijk vertalen als vragen naar samenhang tussen variabelen. Maar het is wel een praktisch hanteerbaar onderscheid. Vragen naar samenhang tussen variabelen doen zich gewoonlijk voor in observationeel onderzoek, waarbij alle variabelen, afhankelijke en

onafhankelijke, door de onderzoeker zijn gemeten. Vragen naar verschillen tussen gemiddelden doen zich vooral voor in vergelijkend onderzoek waarbij de onderzoeker zelf groepsindelingen heeft gemaakt in (quasi-) experimenteel onderzoek. Variantieanalyse is echter ook bruikbaar in observationeel onderzoek, als de onafhankelijke variabelen groepsindelingen betreffen, categorieën waarin de deelnemers aan het onderzoek zijn te onderscheiden. Regressieanalyse kan ook in vergelijkend en experimenteel onderzoek goede diensten bewijzen.

Er is ook een procedure die regressie- en variantieanalyse combineert, namelijk covariantieanalyse. Hierbij is er sprake van een mix van kwalitatieve en kwantitatieve onafhankelijke variabelen. De analyseprocedures die worden besproken in hoofdstuk 9 van dit boek zijn dan ook nauw verwant met de covariantieanalyse zoals besproken in hoofdstuk 7 van Van den Bercken en Voeten (2002).

Tabel 1.4 geeft slechts een grof onderscheid in analyseprocedures en geeft zeker geen uitputtend overzicht van alle beschikbare analyseprocedures. Voor meer omvattende indelingen, zie bijvoorbeeld Tabachnick en Fidell (2001, p. 17–29) en Sheskin (2000, p. 1–49).

Tabel 1.4 maakt alleen onderscheid naar afhankelijke en onafhankelijke variabelen. De andere rollen van variabelen die we hebben onderscheiden, komen in de tabel niet tot hun recht. In feite kunnen deze rollen zich bij alle genoemde analyseprocedures voordoen.

Wat voor soort data heb je nodig om multipele regressie te kunnen toepassen?

Je hebt eerst een steekproef nodig van onderzoekseenheden (cases). In de sociale wetenschappen zijn de onderzoekseenheden meestal personen, maar het kunnen ook groepen zijn (bijvoorbeeld gezinnen), instituten (bijvoorbeeld scholen) of organisaties. Per onderzoekseenheid zijn metingen nodig van de afhankelijke en de onafhankelijke variabelen die je in de analyse wilt gebruiken. Het aantal onderzoekseenheden moet groter zijn dan het aantal variabelen en liefst beduidend groter, bijvoorbeeld minstens 15 keer zoveel onderzoekseenheden als variabelen (zie hoofdstuk 11). Grote steekproeven zijn altijd beter dan kleine steekproeven. Steekproeven met n kleiner dan bijvoorbeeld 200 noemen we kleine steekproeven. Multipele regressie kan goed werken bij kleine steekproeven, maar er is dan enige terughoudendheid nodig bij het kiezen van onafhankelijke variabelen. Ook zijn er eerder technische problemen te verwachten bij kleine dan bij grote steekproeven. Idealiter wordt de steekproef verkregen via aselecte (random) trekking uit een goed gedefinieerde populatie. In de onderzoekspraktijk bestaat de steekproef echter vaak uit de personen (of andere eenheden) die voorhanden waren (bijvoorbeeld schoolklassen). Dat maakt het gevaar groter dat niet voldaan is aan de assumpties van de regressieanalyse (zie hoofdstuk 7 en 8).

In een regressieanalyse kunnen variabelen worden gebruikt die gemeten zijn op intervalniveau. In ieder geval gaan we ervan uit dat de afhankelijke variabele van intervalniveau is. Voor kwalitatieve afhankelijke variabelen (nominaal, ordinaal) bestaan overigens vergelijkbare statistische technieken (bijvoorbeeld logistische regressie voor dichotome afhankelijke variabelen, zie hoofdstuk 12). Gewoonlijk zijn de onafhankelijke variabelen bij multipele regressie ook van intervalniveau, maar variabelen van nominaal niveau kunnen ook gebruikt worden. Nominale variabelen moeten dan echter wel omgezet worden in zogenoemde dummyvariabelen, variabelen die enkel de waarde 0 of 1 kunnen aannemen (zie hoofdstuk 9).