Multicausaliteit en multicollineariteit bij meervoudige regressie

Jan Van Bavel

> "[C]ollinearity is not so much a problem as a state of nature – like the law of gravity – and railing against collinearity is rather like complaining about not being able to fly by flapping your arms" (anonieme reviewer van *Journal of the American Statistical Association*).[1]

1. **Inleiding**

Een socioloog die de verdeling van bepaalde kenmerken in een populatie wil verklaren (bijvoorbeeld de inkomensverdeling of het aantal kinderen), moet over het algemeen rekening houden met twee fenomenen. Ten eerste met het feit dat sociale fenomenen zelden of nooit maar één oorzaak hebben, en ten tweede, met het gegeven dat verschillende oorzakelijke factoren meestal onderling correleren. In de wereld van de sociale wetenschappen is het immers veeleer uitzonderlijk dat causale factoren onderling onafhankelijk zijn.

Deze twee vaststellingen vormen samen twee doorslaggevende argumenten om bij de statistische verklaring van een afhankelijke variabele de voorkeur te geven aan een meervoudig regressiemodel boven één of meerdere enkelvoudige regressie-analyses.[2] De schatting van de effecten van de verklarende variabelen gebeurt immers efficiënter en preciezer wanneer ze voor alle predictoren tegelijkertijd wordt uitgevoerd en bovendien is het enkel met meervoudige regressietechnieken mogelijk om tot geldige schattingen te komen van het unieke en directe effect van predictoren wanneer die predictoren onderling samenhangen (Welkenhuysen-Gybels & Loosveldt 2002, 99-119; Wooldridge 2003, 68-89). Dankzij meervoudige regressietechnieken kunnen we dus de empirische implicaties van causale uitspraken toetsen in situaties waar enkelvoudige regressie tot misleidende conclusies zou leiden. Hoe sterker de samenhang tussen verklarende variabelen, hoe groter de vertekening van de schattingen verkregen via enkelvoudige regressie-analyse en dus hoe sterker de motivatie om aan meervoudige regressie te doen.

Op deze manier motiveren handboeken multivariate analyse de studenten meestal, en terecht, tot de studie van meervoudige regressietechnieken. Tot

daar geen probleem. Maar dan komt het: na een min of meer omstandige behandeling van het onderliggende model, de Gauss-Markov-assumpties en de toe te passen schattingsmethoden komen de problemen aan bod die bij toepassing kunnen opduiken. Op dat moment valt voor het eerst "the M word" (Morrow-Howell 1994): multicollineariteit. Eén van de problemen bij meervoudige regressie, aldus het gros van de handboeken, is de situatie waarin twee of meer verklarende variabelen onderling sterk samenhangen. Het is die situatie die men omschrijft met de term multicollineariteit.

Daarmee is de cirkel rond en de verwarring compleet. Eerst zegt men dat samenhang tussen de verklarende variabelen een belangrijke reden is om aan meervoudige regressie te doen. Vervolgens deelt men mee dat betrouwbare regressie-analyse bemoeilijkt wordt door hoge correlaties tussen de verklarende variabelen. Een voorbeeld uit het eindeloze aanbod van inleidingen tot regressie-analyse:

> "A final problem associated with data used in a regression is multicollinearity. It arises whenever two or more independent variables used in a regression are not independent but are correlated. Unfortunately, in the social sciences this problem arises often, since many socioeconomic variables such as education, social status, political preference, income, and wealth are likely to be interrelated. [...] When two or more independent variables are correlated, the statistical estimation techniques discussed earlier are incapable of sorting out the independent effects of each on the dependent variable" (Schroeder, Sjoquist & Stephan 1986, 71-72).

Net zoals de meeste inleidingen behandelt dit boekje multicollineariteit dus niet als aanleiding tot maar wel als probleem bij meervoudige regressie. Als gevolg daarvan wordt multicollineariteit vaak ten onrechte als schuldige aangewezen voor onbetrouwbare regressie-analyses met onverwacht niet-significante effecten. Sommige aanbevelingen over wat in zulke situaties te doen staat (bijvoorbeeld een sterk correlerende variabele laten vallen of, wat op hetzelfde neerkomt, "multicollineariteit in de mate van het mogelijke vermijden", of een stapsgewijze selectie van variabelen, zie Lewis-Beck 1980, 61-62; Hutcheson & Sofroniou 1999, 84-85; Cohen e.a. 2003, 425-430), zijn als remedie dan ook meestal erger dan de kwaal.

Deze bijdrage gaat uit van het standpunt dat multicollineariteit, naast multicausaliteit, een belangrijk positief argument is om aan meervoudige regressie-analyse te doen. Toegegeven: multicollineariteit heeft inderdaad ongewenste gevolgen bij de schatting van regressieparameters, net zoals een kleine steekproef of een hoge conditionele variantie van de afhankelijke variabele betrouwbare schatting bemoeilijkt. Hoe sterker de multicollineariteit, hoe meer, of hoe meer doordacht verzamelde, data er zullen nodig zijn om de directe causale invloed van de correlerende variabelen empirisch aan te tonen. Maar niettemin: hoe sterker de multicollineariteit, hoe meer reden er is om een

meervoudig regressiemodel op te stellen, tenzij een experimenteel onderzoeksdesign mogelijk is, wat in sociologische toepassingen zelden het geval is.

De volgende paragraaf behandelt meer in detail wat met multicollineariteit bedoeld wordt en wat de gevolgen zijn voor de schatting van regressieparameters. Belangrijker dan de sterkte van de multicollineariteit te kennen, is het om te weten wat de bron ervan is. We onderscheiden drie verschillende oorzaken en gaan in paragraaf 3 dieper in op die vorm van multicollineariteit die vanuit het theoretische standpunt van de multicausaliteit het meest relevant is. Het gaat dan om lineaire afhankelijkheid "as a state of nature": de verklarende variabelen correleren omdat de onderscheiden eigenschappen in de sociale werkelijkheid nu eenmaal samenhangen. Meer bepaald zet paragraaf 3 een aantal misverstanden op een rij die maken dat multicollineariteit soms ten onrechte gevreesd en beschuldigd wordt.

2. Wat is multicollineariteit en wat zijn de gevolgen?

Er is sprake van collineariteit tussen twee verklarende variabelen in een regressiemodel als er een lineaire relatie bestaat tussen die twee variabelen. Wanneer meer dan twee verklarende variabelen lineair samenhangen, spreken we van multicollineariteit (Tacq 2004). Hieronder wordt voor het gemak altijd de term multicollineariteit gebruikt maar alle stellingen gelden even goed voor collineariteit tussen twee predictoren en vice versa. Voor de eenvoud zullen illustraties zich trouwens meestal beperken tot een situatie met twee predictoren. In de praktijk worden de twee termen trouwens volstrekt inwisselbaar gebruikt.

2.1 Exacte versus hoge multicollineariteit

Algebraïsch betekent multicollineariteit dat één of meerdere predictoren zonder fout of bijna zonder fout te herschrijven is als een lineaire combinatie van één of meerdere andere predictoren. Neem bijvoorbeeld de volgende drie variabelen: jaar van huwelijk (h_i), jaar van huwelijksontbinding (o_i) en huwelijksduur bij ontbinding (d_i). Veronderstel dat we deze variabelen hebben opgemeten voor drie eenheden. De data kunnen dan in een datamatrix **X** van dimensie 3 x 3 (een rij voor elke eenheid en een kolom voor elke variabele) worden weergegeven:

$$\mathbf{X} = \begin{bmatrix} 1970 & 1990 & 20 \\ 1975 & 1985 & 10 \\ 1990 & 1998 & 8 \end{bmatrix}$$

waarin het huwelijksjaar in de eerste kolom staat, het ontbindingsjaar in de tweede en de duur bij ontbinding in de derde kolom. Het is duidelijk dat, wanneer we om het even welke twee van deze variabelen (kolommen) kennen, we de derde exact kunnen afleiden. Wanneer bijvoorbeeld de variabelen h_i en o_i in de eerste twee kolommen gekend zijn, dan weten we dat de derde kolom $d_i = o_i - h_i$. In het algemeen is een variabele x_{ik} een lineaire combinatie van een reeks andere variabelen x_{i1}, ..., x_{ij} als er een rij van j getallen (c_1, ..., c_j) bestaat zodat $x_{ik} = c_1 x_{i1} + ... + c_j x_{ij}$; zodat we de ene variabele dus als een lineaire functie van de andere variabelen kunnen schrijven (Carroll& Green 1997).

Wanneer een verklarende variabele in een regressiemodel een perfecte lineaire combinatie is van één of meerdere andere verklarende variabelen, dan ontstaat exacte multicollineariteit. Bij exacte multicollineariteit kunnen de regressieparameters niet geschat worden omdat er geen mathematisch unieke oplossing bestaat voor het criterium dat men bij schatting minimaliseert (bijvoorbeeld de som van de gekwadrateerde residuelen bij klassieke lineaire regressie of de *deviance* bij *maximum likelihood estimation*). Afwezigheid van perfecte multicollineariteit is daarom een basisvoorwaarde voor zowel klassieke lineaire regressie (Berry 1993) als voor andere regressietechnieken (voor een behandeling in de context van veralgemeende lineaire modellen, zie McCullagh & Nelder 1983; Dobson 2002). Hierover bestaat geen discussie. Het vervolg beperkt zich daarom tot situaties waar de multicollineariteit niet exact is maar wel "hoog".

Volgens sommigen is hoge multicollineariteit nog verraderlijker dan exacte multicollineariteit omdat de ziekte dan kan woekeren zonder dat de patiënt het voelt: "Indeed, extreme multicollinearity is often a more difficult problem because it frequently goes undetected" (Allen 1997, 177). In vele handboeken kan men lezen dat "hoge" multicollineariteit tot "het probleem van niet significante regressieparameters" leidt (bijvoorbeeld Allen 1997, 176). Niettemin zijn gemakkelijk voorbeelden te geven van situaties met zeer hoge multicollineariteit maar toch zeer betrouwbare en significante schattingen enerzijds, en situaties met lage multicollineariteit en hoge onbetrouwbaarheid anderzijds.

Sommige, meer diepgravende handboeken betreuren daarom het gebruik van de term multicollineariteit voor niet-exacte lineaire afhankelijkheid. "Confusingly, the words collinearity and multicollinearity are also often used in regression when there is a 'near dependency' in the columns of **X**", aldus Draper en Smith (1998, 369), eraan toevoegend: "What 'near' means is also a problem." (p. 369). Zij behandelen daarom "hoge multicollineariteit" onder een algemenere noemer, namelijk die van de "ill-conditioned data", in het Nederlands "data met zwakke conditie" of "slecht geconditioneerde data" genoemd (ISI 2005). Dit is een term uit de lineaire algebra die de situatie beschrijft waar er weliswaar een unieke oplossing bestaat voor een algebraïsch probleem maar waarin die oplossing instabiel is. Instabiliteit impliceert dat een zeer klei-

ne verandering in de data een zeer grote impact kan hebben op de verkregen oplossing voor het algebraïsche probleem (Öztürk & Akdinez 2000). Hoge multicollineariteit impliceert soms, maar lang niet altijd, dat de data slecht geconditioneerd zijn voor meervoudige regressie (Draper & Smith 1998, 369-386). "Multicollinearity can be one source of weak data, but the strength of the data cannot be measured solely by the orthogonality of the data" (Smith & Campbell 1980, 75). En Kendall en Stuart (1973) besteden in hun meer dan 700 bladzijden dikke standaardwerk over "Inference and Relationship" welgeteld één zin aan multicollineariteit: "If **X'X** is 'nearly' singular, the elements of its inverse will be large, and estimation therefore imprecise" (p. 84).

2.2 **Precisie en instabiliteit**

Daarmee zijn we bij de kern van de zaak aanbeland. Ten eerste, voor alle duidelijkheid: hoge maar imperfecte multicollineariteit is *niet* in strijd met de assumpties van (veralgemeende) lineaire regressie. Multicollineariteit zorgt dan ook niet voor een vertekening (ongeldigheid, *bias*) van de schatting van de regressieparameters. Het effect van multicollineariteit beperkt zich tot de mate van stabiliteit van, en dus zekerheid over, die schattingen.

Het is onbetwistbaar zo dat, alle andere factoren gelijk blijvend, de schattingen van regressieparameters instabieler zijn wanneer de verklarende variabelen correleren dan wanneer ze niet correleren. Hoe sterker de onderlinge correlatie, hoe instabieler de schattingen *ceteris paribus* zijn. Die instabiliteit valt af te lezen aan de hogere standaardfouten van de geschatte regressieparameters. En een grote standaardfout impliceert dat de kans groot is dat de werkelijke populatiewaarden van de regressieparameters β vér van de geschatte waarden **b** liggen (Silvey 1969; Tacq 1992, 122-124; Fox 1991; Draper & Smith 1998, 369-376; Welkenhuysen-Gybels & Loosveldt 2002, 299-301). Kort gezegd: er is een grote kans op foute schattingen.

Er kan worden aangetoond dat de varianties van de geschatte richtingscoëfficiënten b_j in een meervoudig lineair regressiemodel gelijk zijn aan (Goldberger 1991; Fox & Monette 1992):

$$\sigma^2_{b_j} = \frac{\sigma^2}{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2} \cdot \frac{1}{1 - R_j^2} \qquad (1)$$

waarin σ^2 staat voor de conditionele populatievariantie van de afhankelijke variabele, x_{ij} voor de waarde van de *j*-de variabele voor eenheid *i* (dus de *i*-de rij van de *j*-de kolom in datamatrix **X**) en $\bar{x}_j$ voor het steekproefgemiddelde van variabele *j*. De invloed van multicollineariteit op de variantie van de effectparameters komt tot uiting in de rechterterm van vergelijking (1). R_j^2 is immers de meervoudige determinatiecoëfficiënt van de lineaire regressie waarin variabele x_{ij} als afhankelijke variabele fungeert en alle overige predictoren als onaf-

hankelijke. Dit wordt de auxiliaire regressie genoemd. Met andere woorden: R_j^2 is de proportie van de variantie van x_{ij} die met behulp van een lineaire regressie verklaarbaar is op basis van de andere predictoren. Het aandeel van de variantie van x_{ij} dat niet als een lineaire combinatie van de overige onafhankelijke variabelen verklaard kan worden, namelijk $1 - R_j^2$, wordt de tolerantie genoemd ("tolerance" in het Engels) en is een maat voor de afwezigheid van multicollineariteit (Tacq 1992, 123; Cohen e.a. 2003, 423-424). De verhouding $1/(1 - R_j^2)$ wordt in de literatuur bedacht met de titel variantie-inflatiefactor ("Variance-Inflation Factor", verder afgekort tot VIF). Zij geeft aan in welke mate de variantie van b_j verhoogt als gevolg van correlaties tussen x_{ij} en de overige verklarende variabelen (Fox & Monette 1992; Cohen e.a. 2003, 423). R_j^2 ligt sowieso tussen 0 en 1 en de maximale waarde van 1 komt overeen met een situatie van exacte multicollineariteit. Bij exacte multicollineariteit is de VIF oneindig groot omdat door nul gedeeld wordt, wat er opnieuw op wijst dat schatting in die situatie zinloos is. Als R_j^2 niet gelijk is aan 1 maar er niet ver onder ligt, dan zal de VIF weliswaar niet oneindig maar wel heel groot zijn. Als alle andere termen en factoren in vergelijking (1) constant gehouden worden, dan impliceert een hoge multicollineariteit (een hoge R_j^2) een hoge VIF en dus een hoge standaardfout σ_{b_j}, met als gevolg dat het betrouwbaarheidsinterval rond b_j breed is. De precisie van de schatting van b_j is dus overeenkomstig klein, wat wil zeggen dat de kans groot is dat b_j ver naast de werkelijke populatiewaarde σ_j ligt (Goldberger 1991, 245-248; Tacq 1992, 122-124; Fox 1991; Draper & Smith 1998, 369-376; Welkenhuysen-Gybels & Loosveldt 2002, 299-301). Vandaar die grote kans op foute schattingen bij hoge multicollineariteit, alle andere termen en factoren in vergelijking (1) constant houdend.

2.3 **Het belang van het absolute aantal**

Er is echter geen reden om die andere termen en factoren constant te houden. Uit vergelijking (1) blijkt enerzijds dat hoge standaardfouten van regressieparameters ook andere oorzaken dan multicollineariteit kunnen hebben en anderzijds dat sterke multicollineariteit en een hoge VIF niet noodzakelijk hoge standaardfouten impliceren. Kort gezegd wordt de standaardfout van een richtingscoëfficiënt b_j door drie factoren beïnvloed: de conditionele populatievariantie van de afhankelijke variabele, de totale omvang van de in de steekproef geobserveerde variatie in de verklarende variabele x_{ij} en de mate waarin de geobserveerde variatie in x_{ij} met een lineaire combinatie van de andere predictoren voorspeld kan worden. Enkel die laatste factor verwijst naar multicollineariteit. De conditionele populatievariantie σ^2 in de teller van (1) is haast altijd onbekend en strikt genomen niet door het onderzoeksdesign beïnvloedbaar. In de praktijk zal bij de schatting van die populatievariantie gebruikgemaakt worden van de *Mean Squared Errors* (MSE) (zie Welkenhuysen-Gybels & Loosveldt

2002, 113-115) en die wordt wél door het onderzoeksdesign beïnvloed (onder meer door de operationele definitie van de variabelen in het model en het aggregatieniveau van de observatie-eenheden). We kunnen op deze factor hier niet verder ingaan, hoe belangrijk dit ook is.

Van cruciaal belang voor de verdere argumentatie is de factor $\sum(x_{ij} - \bar{x}_j)^2$ in de noemer van (1). In navolging van Tacq (1992) noemen we dit de hoeveelheid geobserveerde variatie in de betreffende verklarende variabele. Het verschil met de geobserveerde variantie is dat niet gedeeld wordt door de steekproefomvang n (of door n-1 als men op basis hiervan de variantie in de populatie zou willen schatten). De hoeveelheid variatie in x_{ij} staat in de noemer van (1), dus hoe meer variatie in de verklarende variabele, hoe kleiner *ceteris paribus* de standaardfout van de regressieparameter van diezelfde variabele. Als $1/(1 - R_j^2)$ aanspraak mag maken op de titel variantie-inflatiefactor, dan heeft $1/\sum(x_{ij} - \bar{x}_j)^2$ dus even veel (of even weinig) recht op de titel van variantie-deflatiefactor (verder VDF genoemd). Daaruit kan alvast één conclusie getrokken worden: hoge multicollineariteit leidt niet tot grote standaardfouten (en dus onnauwkeurige schattingen) als een sterke VDF de invloed van een hoge VIF countert. De vraag is of zo'n compenserend mechanisme in de praktijk mogelijk is.

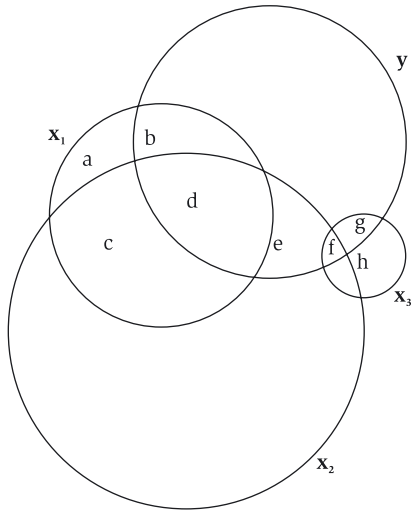
Een eerste element om tot een antwoord te komen is de vaststelling dat de VIF en de VDF onafhankelijk van elkaar kunnen variëren: bij een gegeven VIF kan de VDF groot of klein zijn en vice versa. Om de relatie tussen VIF en VDF te verduidelijken, is het handig om eerst de determinatiecoëfficiënt R_j^2 te ontrafelen. Hij kan als volgt geformuleerd worden:

$$R_j^2 = 1 - \frac{\mathrm{SSE_j}}{\mathrm{SST_j}} = 1 - \frac{\sum_{i=1}^{n}(x_{ij} - \hat{x}_{ij})^2}{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2} \qquad (2)$$

waarin $\hat{x}_{ij}$ staat voor de waarde die voorspeld wordt door de auxiliaire regressie met x_{ij} als afhankelijke en de overige predictoren als verklarende variabelen. $\mathrm{SST_j}$ staat voor de totale kwadratensom (de "Total Sum of Squares") van x_{ij}, zijnde de reeds vermelde maat voor de totale hoeveelheid variatie in die variabele, die we ook in de VDF terugvinden. $\mathrm{SSE_j}$ staat voor de hoeveelheid variatie in x_{ij} die niet door de overige onafhankelijke variabelen verklaard kan worden. (Voor meer uitleg bij deze conventionele kwadratensommen, zie Tacq 1992, 115-116 of Welkenhuysen-Gybels & Loosveldt 2002, 119-125). Door (2) in te vullen in de formule van de VIF, namelijk $1/(1 - R_j^2)$, en te vereenvoudigen, kan die maat voor multicollineariteit geherformuleerd worden als volgt:

$$\mathrm{VIF_j} = \frac{1}{1 - \left(1 - \frac{\mathrm{SSE_j}}{\mathrm{SST_j}}\right)} = \frac{\mathrm{SST_j}}{\mathrm{SSE_j}} \qquad (3)$$

Met andere woorden: de VIF geeft aan hoeveel keer de *totale* variatie in x_{ij} groter is dan de variatie die *uniek* is voor x_{ij}; uniek in de betekenis dat ze niet door de andere predictoren voorspeld kan worden. Figuur 1 poogt dit te verduidelijken. In de figuur stelt elke cirkel een variabele voor (de namen van de variabelen zijn weergegeven in vectornotatie zodat we het subscript *i* voor de eenheden kunnen laten vallen). De oppervlakte van elke cirkel is in verhouding met de omvang van de bijhorende totale kwadratensom SST_j: hoe meer variatie, hoe groter de cirkel. Variabele $\mathbf{x_1}$ covarieert duidelijk met variabele $\mathbf{x_2}$, wat te zien is aan de overlappende delen van de variatiecirkels, namelijk de delen c en d. De gemeenschappelijkheid van een cirkel van een variabele met één of meerdere andere onafhankelijke variabelen, dus de doorsnede van de betrokken cirkels, noemen we de communaliteit van die variabele. De niet-overlappende delen noemen we de uniciteit. (Voor meer uitleg bij deze grafische voorstellingswijze en haar beperkingen, zie Edward 2001). Er is geen overlapping tussen $\mathbf{x_1}$ en $\mathbf{x_3}$, dus de communaliteit van $\mathbf{x_3}$ beperkt zich tot de overlapping met $\mathbf{x_2}$ (we laten **y** voorlopig buiten beschouwing).



Figuur 1. Variatiecirkels voor drie verklarende en één afhankelijke variabele.

De VIF_1 voor de eerste predictor kan op basis van de oppervlakten van delen van deze cirkels berekend worden als de verhouding van a + b + c + d = SST_1 tot a + b = SSE_1. Aangezien de variatie van de eerste onafhankelijke variabele duidelijk voor meer dan de helft overlapt met die van de tweede onafhankelijke variabele (en dus $R^2_1 > 0.50$), zal VIF_1 zeker groter zijn dan 2.

De variantie-deflatiefactor VDF_j kan als volgt in kwadratensommen worden weergegeven:

$$VDF_j = \frac{1}{SSE_j + SSR_j} = \frac{1}{SST_j} \quad (4)$$

waarin de kwadratensommen SST_j en SSE_j dezelfde betekenis hebben als in (2) en (3) en SSR_j staat voor de hoeveelheid variatie in $\mathbf{x_j}$ die op basis van een auxiliaire regressie op de andere predictoren voorspeld kan worden (de "Sum of Squares due to Regression"). In de noemer van de VDF vinden we dus dezelfde kwadratensom terug als in de teller van de VIF.

We kunnen nu formule (1) voor de variantie van regressieparameter b_j herschrijven als het product van VDF en VIF met de populatievariantie van **y** en vereenvoudigen:

$$\sigma^2_{b_j} = \sigma^2 \cdot VDF_j \cdot VIF_j = \sigma^2 \frac{1}{SST_j} \cdot \frac{SST_j}{SSE_j} = \frac{\sigma^2}{SSE_j} \quad (5)$$

Dit is een basisformule voor de variantie van b_j bij klassieke lineaire OLS-regressie (Goldberger 1991, 150-167; vgl. Cohen e.a. 2003, 86-87). Deze formulering is in de argumentatie van dit artikel van centraal belang omdat het een aantal dingen duidelijk maakt waar we later dieper op in gaan. De kern van de zaak is dit: de sterkte van de multicollineariteit, als *proportionele* maat voor het aandeel variantie dat die variabele gemeenschappelijk heeft met collega-predictoren (R^2_j), is niet van fundamenteel belang voor de precisie van de schattingen van een effectparameter (en dus voor de kans om niet al te ver van de werkelijke populatiewaarde te zitten). Gegeven de conditionele populatievariantie van de afhankelijke variabele is de *absolute* hoeveelheid unieke variatie die voor x_{ij} beschikbaar is van doorslaggevend belang: SSE_j is geen relatieve maar een absolute maat voor de hoeveelheid unieke variatie, met een minimum van nul aan de onderkant maar onbegrensd aan de bovenkant. De schatting van de variantie van de effectparameters wordt enkel beïnvloed door de absolute omvang van de uniciteit van de betreffende variabele. Welk aandeel die uniciteit uitmaakt van de totale variatie van die variabele (namelijk 1-R^2_j), doet niet terzake.

In praktische termen vertaald: bij multicollineariteit is de precisie van de schattingen van het directe effect van een predictor x_{ij} afhankelijk van het absolute aantal (niet: het aandeel) observatie-eenheden dat buiten het multicollineaire patroon van x_{ij} valt. Niet-exacte multicollineariteit mag zo hoog zijn als men wil, zolang een voldoende absoluut aantal eenheden beschikbaar is dat buiten het collineaire patroon valt, zorgt het niet voor slecht geconditioneerde data. Een voorbeeld ter verduidelijking: je wil de unieke effecten van twee sterk correlerende verklarende variabelen X en Z schatten op een afhankelijke variabele Y.[3] Stel dat de samenhang tussen de twee predictoren in de populatie is zoals weergegeven in de eerste kolommen van tabel 1.Een toevalssteekproef van 1000 eenheden uit deze populatie levert naar verwachting in totaal slechts 10 + 10 = 20 eenheden op die buiten het collineaire patroon vallen. De schatting van de unieke effecten van X en Z op Y zal dus op slechts weinig unieke

variatie van X en Z gebaseerd zijn en daarom weinig betrouwbaar. In een 10 keer grotere steekproef is het aantal eenheden buiten het collineaire patroon 10 keer zo groot, namelijk 200 eenheden. De multicollineariteit is exact even groot maar de schatting van de unieke effecten van X en Z zal veel betrouwbaarder zijn.

Als er een dergelijke, sterke samenhang tussen predictoren verwacht wordt en men wil het unieke effect van beiden schatten, dan zou een efficiënter want gerichter steekproefdesign er in bestaan om de uitzonderlijke gevallen (X = 0 en Z = 1 of X = 1 en Z = 0) te gaan oververtegenwoordigen. De steekproef in de twee laatste kolommen van tabel 1 levert zo evenveel "nuttige" gevallen buiten het collineaire patroon op als de veel duurdere steekproef van 10000 eenheden.

Tabel 1. Hypothetisch voorbeeld van steekproeven uit een populatie met twee sterk samenhangende variabelen.

	Populatie kansverdeling		Steekproeven					
			N = 1000		N = 10000		N = 1180	
	X = 0	X = 1	X = 0	X = 1	X = 0	X = 1	X = 0	X = 1
Z = 0	0.49	0.01	490	10	4900	100	490	100
Z = 1	0.01	0.49	10	490	100	4900	100	490

Niet multicollineariteit of de verhouding $VIF_j = SST_j/SSE_j$ moet voor een multivariate onderzoeker de grootste zorg zijn maar wel de verhouding σ^2/SSE_j – gewoonweg de foutenmarge van de regressieparameters, dus. "Researchers should not be concerned with whether or not 'there really is collinearity.' They may well be concerned with whether the variances of the coefficient estimates are too large – for whatever reason – to provide useful estimates of the regression coefficients" (Goldberger 1991, 251). Als de afhankelijke variabele maar weinig varieert, volstaat een beperkte SSE_j om tot een betrouwbare schatting van het effect van $\mathbf{x_j}$ te komen – er is geen kanon nodig om een mug te vangen. Hoe hoger de variantie van **y**, hoe hoger SSE_j moet zijn. Een hoge SSE_j: dat is wat nodig is om tot betrouwbare schattingen te komen, ongeacht of de onderlinge samenhang tussen de predictoren nu heel sterk, tamelijk sterk, zwak, of afwezig is (in het laatste, orthogonale geval geldt $SSE_j = SST_j$). (Het spreekt voor zich dat het gaat om de hoogte van SSE_j, uitgedrukt op dezelfde schaal van de variabele x_{ij} als die waarop de regressiecoëfficiënt b_j zelf wordt uitgedrukt. Gewoon SSE_j verhogen door de schaal van x_{ij} te veranderen (bijvoorbeeld $x'_{ij} = 100x_{ij}$) heeft uiteraard geen enkel effect op de precisie van de schattingen, want de verhouding van de puntschatting van de regressiecoëfficiënt tot zijn standaardfout blijft in dat geval hetzelfde).

Terug verwijzend naar figuur 1: hoewel de multicollineariteit en dus de VIF_1 van $\mathbf{x_1}$ duidelijk hoger is dan die van $\mathbf{x_3}$, zal de schatting van het directe effect van $\mathbf{x_1}$ op **y** toch betrouwbaarder zijn dan de schatting van het directe effect van $\mathbf{x_3}$ op **y**. De reden is dat de oppervlakte van de uniciteit van $\mathbf{x_1}$ (namelijk a + b) groter is dan de oppervlakte van de uniciteit van $\mathbf{x_3}$ (g + h). In dit voorbeeld zou de regressieparameter van de variabele met de hoogste VIF van de twee ($\mathbf{x_1}$ en $\mathbf{x_3}$) dus toch de laagste standaardfout hebben. (De mate van overlapping met **y** is van belang voor de sterkte van het effect. Voor b_j zelf, dus).

Sommige handboeken beweren dat een grotere steekproef niet helpt als de sterkte van de multicollineariteit in de grotere steekproef even groot is als in de beperkte. Zo helpt een grotere steekproef volgens Koutsoyiannis en Carter (1973) alleen "if multicollinearity is due to errors of measurement, as well as when intercorrelation happens to exist only in our original sample but not in the population" (p. 250). Of: "More data is no help in multicollinearity if it is simply 'more of the same'. What matters is the structure of the **X'X** matrix, and this will only be improved by adding data which are less collinear than before" (Johnston 1984, 250). Fox (1991) schrijft dat de ideale oplossing voor multicollineariteit zou zijn "to collect new data in such a manner that the problem is avoided – for example, by experimental manipulation of the *x*s" (p. 14). Uit het voorgaande blijkt integendeel dat bij een gelijkblijvende multicollineariteit een grotere steekproef een grotere SSE_j impliceert en dus meer precieze schattingen (Goldberger 1991, 252-253 bevat een interessante oefening die deze stelling bewijst). Het klopt uiteraard wel dat de mate van multicollineariteit niet daalt als de lineaire afhankelijkheid in de additionele data even groot is als in de oorspronkelijke steekproef. Het punt waar we in deze bijdrage op zullen blijven hameren is echter dat de sterkte van de multicollineariteit op zich niet van belang is maar wel het *absolute aantal* observaties dat buiten het multicollineaire patroon valt. Elke extra observatie buiten het multicollineaire patroon zorgt sowieso voor een verhoging van de SSE_j en dus van de precisie van de schattingen, ongeacht de sterkte van de multicollineariteit.

2.4 **Bronnen van multicollineariteit**

Lineaire afhankelijkheid aan de rechterkant van een regressievergelijking kan verschillende oorzaken hebben. We noemen er hier drie. (Voor een andere indeling, toegespitst op de analyse van autocorrelerende tijdreeksen, zie Koutsoyiannis & Carter 1973, 234). De eerste is vanuit het theoretische standpunt van de multicausaliteit het meest relevant. Daarom beperkt de rest van deze bijdrage zich tot die eerste vorm: lineaire afhankelijkheid die het gevolg is van het feit dat duidelijk verschillende eigenschappen van analyse-eenheden in de sociale werkelijkheid nu eenmaal samenhangen. Voorbeelden van deze vorm van multicollineariteit zijn de samenhang tussen leeftijd, huwelijksduur en

kindertal of tussen opleidingsniveau en inkomen. Wanneer we in de sociale wetenschappen een meervoudig regressiemodel met dit soort van predictoren opstellen, dan is dat vaak om theoretisch-causale uitspraken over het unieke netto-effect van dergelijke eigenschappen op de afhankelijke variabele empirisch te toetsen. Daarom behandelen we multicollineariteit van deze soort in de volgende paragraaf onder de noemer "samenhang tussen causale factoren".

Een tweede oorzaak van multicollineariteit is het opnemen van latente constructen als verklarende variabelen waarbij verschillende constructen deels op dezelfde manifeste variabelen gebaseerd zijn. Een voorbeeld is een regressie met zowel de beroepspositie als de sociaal-economische status als verklarende variabelen, waarbij beide gebaseerd zijn op hetzelfde gerapporteerde beroep (bijvoorbeeld De Weerdt & De Witte 2001). Dit is een gevaarlijke praktijk. Ook het omgekeerde komt voor, namelijk de situatie waar men verschillende manifeste variabelen in een regressie opneemt die in feite verschillende indicatoren zijn voor eenzelfde achterliggende, latente factor en als gevolg daarvan sterk correleren. Als een dergelijke analyse onverwachte of geen significante resultaten oplevert, dan is de juiste conclusie niet dat multicollineariteit in dit geval helaas problemen stelt voor meervoudige regressie. De gepaste conclusie is dat er nog theoretisch en conceptueel werk aan de winkel is. Technisch is in dit soort van situaties trouwens schatting van een structureel model meestal aangewezen. Daar gaan we hier verder niet op in.

Een derde vorm van multicollineariteit is het gevolg van modelformuleringen waarin producttermen zijn opgenomen waarvan de samenstellende factoren ook zelfstandig tot de regressievergelijking behoren. Dat is bijvoorbeeld het geval wanneer men interactie-effecten wil toetsen (bijvoorbeeld $E(Y) = \beta_1 A + \beta_2 B + \beta_3 AB$) of niet-lineaire effecten met behulp van een kwadratische term (bijvoorbeeld $E(Y) = \beta_1 A + \beta_2 A^2$). In beide gevallen kan de multicollineariteit vaak sterk teruggedrongen worden door de onafhankelijke variabelen te centreren: multicollineariteit die louter het gevolg is van de schaal waarop de variabelen uitgedrukt zijn (in de literatuur *niet-essentiële* multicollineariteit genoemd), verdwijnt daardoor immers (Marquardt 1980). (Als A en B een perfect symmetrische frequentieverdeling hebben, zal de correlatie tussen A en A^2, of tussen A en B enerzijds en het product AB anderzijds, na centreren helemaal verdwijnen (nul worden)). Centreren heeft echter geen enkele invloed op *essentiële* multicollineariteit, die het gevolg is van scheefheid in de verdeling van A en/of B. We gaan hier niet verder op in (zie Marquardt 1980). Bijzonder nuttige lectuur voor wie interactie-effecten wil opnemen en vreest voor multicollineariteit, is Friedrich (1982). Zie verder ook Jaccard & Turrisi (2003) of Cohen e.a. (2003, 255-267). Draper & Smith (1998, 369-386) behandelen "ill-conditioning" en multicollineariteit op een manier die bijzonder verhelderend is over de kwestie van het centreren bij het gebruik van veeltermen. Voor specifieke uitleg over multicollineariteit bij polynome regressie, zie Cohen e.a. (2003, 193-221).

3. Misverstanden over multicollineariteit

In sociaalwetenschappelijk onderzoek komt het vaak voor dat twee variabelen die mogelijk een causaal effect hebben op het onderzochte fenomeen, onderling sterk correleren. Ongeacht of de samenhang tussen die kandidaat-verklarende variabelen van causale aard is of niet, sluit die collineariteit uit dat het unieke, directe effect van elk van de twee variabelen met bivariate analyse geschat kan worden: enkelvoudige regressie kan in zo'n geval voor zowel een overschatting als een onderschatting zorgen van het effect van de verklarende variabelen op *Y*, of voor een schatting in de verkeerde richting. Daarom, omwille van die multicollineariteit, doen we geen enkelvoudige maar een meervoudige regressie.

Helaas is generaties van sociale wetenschappers ingepeperd dat meervoudige regressie enkel goed werkt als de verklarende variabelen waarvan men het unieke netto-effect wil schatten niet te sterk correleren. Men zou er wanhopig van worden:

> "In mijn onderzoek naar de oorzaken van echtscheiding zijn een groot aantal factoren opgenomen, waaronder huwelijksduur, huwelijksleeftijd, leeftijdsverschil, kindertal, [...]. Sommige factoren, zoals bijvoorbeeld huwelijksduur en kindertal, zijn evenwel zo sterk gecorreleerd dat een multiple regressieanalyse niet meer op acceptabele wijze kan worden toegepast. In de literatuur wordt gesteld dat een correlatie tussen oorzakelijke factoren van 0.60 of meer in een multicausaal model ontoelaatbaar is, omdat de geschatte effecten dan zeer onbetrouwbaar worden [...]. Een voor de hand liggende oplossing voor dit probleem is het weglaten van één van deze sterk samenhangende factoren. Maar, u zult begrijpen dat dit in theoretisch opzicht weinig bevredigend is. Want zowel huwelijksduur [...] als kindertal [...] bieden een bijzondere en vrij aparte verklaring voor het echtscheidingsverschijnsel, ook al zijn zij statistisch sterk gecorreleerd." (citaat van de Jager in Tacq 2001, 101).

Uit de discussie in paragraaf 2 volgt dat het onzin is dat een meervoudige regressie hier a priori niet meer op acceptabele wijze zou kunnen worden toegepast. Correlaties tussen oorzakelijke factoren van ruimschoots meer dan 0.60 zijn perfect verzoenbaar met heel betrouwbare schattingen van partiële regressieparameters. De angst voor multicollineariteit berust op een aantal misverstanden. De volgende alinea's zetten er een aantal op een rij.

"Multicollineariteit leidt tot te hoge standaardfouten"

Of er nu sprake is van hoge correlaties of niet, zoals altijd bij een inferentiële statistische analyse is de betrouwbaarheid maar gewaarborgd in de mate dat voldoende informatie voorhanden is om de beoogde hypothesen te kunnen toetsen met een vooropgestelde kans op een type I fout α. Kortweg: de data moeten geschikt zijn voor de beoogde doeleinden. In technische termen: de data mogen niet "ill-conditioned" zijn (zie supra). Men kan geen betrouwbare statistische analyses van de loonverschillen tussen Brusselaars naar sociale

klasse uitvoeren als men maar een steekproef heeft van 1000 toevallig gekozen Belgen. In zo'n steekproef zal de hoeveelheid nuttige informatie (in casu het aantal Brusselaars) immers te klein zijn. Om dezelfde reden kan men evenmin het onafhankelijke effect van, bijvoorbeeld, het aantal kinderen dat een koppel in leven heeft enerzijds en het aantal geboorten dat bij hen heeft plaatsgevonden anderzijds op hun verdere vruchtbaarheid betrouwbaar schatten als het absolute aantal observaties waarvoor die twee verklarende variabelen niet aan elkaar gelijk zijn, te klein is. Nogmaals: niet het relatieve maar het absolute aantal is hier belangrijk. Er zal dus een grotere of een meer gerichte steekproef nodig zijn naarmate verklarende factoren sterker samenhangen. Als de zuigelingen- en kindersterfte zeer laag is, dan zal het aantal geboorten en het aantal kinderen in leven voor de meeste koppels gelijk zijn en dus zal de correlatie tussen die twee verklarende variabelen zeer hoog zijn. Om de onafhankelijke effecten van die twee predictoren op betrouwbare wijze te schatten, zal dus ofwel een heel grote steekproef nodig zijn, ofwel een meer gerichte, selectieve steekproef, waarbij specifiek wordt gezocht naar koppels die sommige van hun geboren kinderen verloren (en waarbij het aantal geboorten dus niet gelijk is aan het aantal kinderen in leven).

Als multicollineariteit tot onbetrouwbaarheid leidt, dan gaat het dus in wezen om een geval van regressie met een steekproef die te klein is voor de gestelde doeleinden. Om die reden behandelt Chipman (1964) het probleem van multicollineariteit in een artikel met als welgekozen titel: "On least squares with insufficient observations". Het praktische probleem dat kan ontstaan wanneer men het onafhankelijke effect van sterk lineair afhankelijke variabelen wil schatten, is perfect vergelijkbaar met het praktische probleem dat opduikt wanneer men de verdeling van partijpolitieke stemmen in België wil schatten op basis van een steekproef van 100 eenheden: te weinig onafhankelijke observaties. Gegeven de gestelde doeleinden, is het werkelijke probleem dus een onvolkomen onderzoeksdesign. Het argument dat onderzoekers nu eenmaal uit moeten gaan van een gegeven steekproefomvang, is vals. Waarom wordt dit argument alleen gebruikt als er samenhang tussen verklarende factoren in het spel is? Waarom zou een verstandig steekproefdesign wél een issue zijn wanneer we univariate populatieparameters willen schatten en niet bij meervoudige regressie?

In een ondertussen klassiek geworden, hilarisch stuk stelt Goldberger (1991, 245-253) met veel gevoel voor ironie voor om het begrip multicollineariteit te vervangen door het neologisme "micronumerosity": het micronumerositeitsprobleem treedt op wanneer het aantal onafhankelijke eenheden n in de steekproef te laag is. Wanneer dit het geval is, dan zijn schattingen van populatieparameters onbetrouwbaar en loopt men dus het risico om de nulhypothese dat een parameter nul is ten onrechte niet te verwerpen. Het extreme geval van exacte multicollineariteit is perfect analoog aan het geval van exacte micronumerositeit:

> "The extreme case, 'exact micronumerosity', arises when n = 0, in which case the sample estimate of μ is not unique. [...] The extreme case is easy enough to recognize. 'Near micronumerosity' is more subtle, and yet very serious. It arises when the rank condition n > 0 is barely satisfied. [...] The consequences of micronumerosity are serious. Precision of estimation is reduced. [...] The estimate of μ will be very sensitive to sample data, and the addition of a few more observations can sometimes produce drastic shifts in the sample mean. [...] Tests for the presence of micronumerosity require the judicious use of various fingers. Some researchers prefer a single finger, others use their toes, still others let their thumbs rule. A generally reliable guide may be obtained by counting the number of observations." (Goldberger 1991, 249)

Het is een ontnuchterende oefening om een willekeurig handboek meervoudige regressie ter hand te nemen en overal het begrip multicollineariteit te vervangen door het woord micronumerositeit. Bijvoorbeeld in: "high degrees of multicollinearity can result in regression coefficients being incorrectly estimated and even having the wrong signs" (Hair e.a. 1998, 189) of in de volgende passage: "Multicollinearity is a problem because it undermines the statistical significance of an independent variable. Other things being equal, the larger the standard error of a regression coefficient, the less likely it is that this coefficient will be statistically significant" (Allen 1997, 176). Waarom vinden we in het geciteerde handboek wel een hoofdstuk over "The Problem of Multicollinearity" en niet over het ware achterliggende en veel fundamentelere probleem van een te kleine steekproef?

Het is een opvallende en fascinerende gewoonte in handboeken meervoudige regressie om te stellen dat multicollineariteit tot een "inflatie van de standaardfouten" van de regressieparameters leidt ("inflated standard errors", cf. de eerder besproken "variance inflation factor"). Dit is opvallend en fascinerend omdat het woord inflatie niet van stal gehaald wordt wanneer men het heeft over de gevolgen van een te kleine steekproef. In dat laatste geval vindt men de hoge standaardfouten blijkbaar terecht. De waarheid is dat de verhoging van de standaardfouten bij stijgende multicollineariteit even terecht is als de verhoging van de standaardfouten bij dalende steekproefomvang. De gewoonte om het alleen in de context van multicollineariteit te hebben over "inflated standard errors" suggereert ten onrechte dat multicollineariteit de geldigheid van de geschatte standaardfouten zou ondergraven. Van Dale vertaalt het Engelse "to inflate" als "opblazen" of "kunstmatig opdrijven", waarbij de connotatie is dat de verhoging artificieel of onterecht zou zijn. Op die manier doet men uitschijnen dat hoge multicollineariteit een overtreding van een regressie-assumptie zou betekenen. Men erkent dan dat "het multicollineariteitsprobleem" weliswaar geen bedreiging vormt voor de geldigheid van de schattingen van de richtingscoëfficiënten maar helaas wel leidt tot een inflatie van de geschatte standaardfouten (zie bijvoorbeeld Schroeder e.a. 1986, 72). Dat wekt de indruk dat de schatting van die standaardfouten als gevolg van multicollineariteit ten onrechte zo groot is. Nochtans brengt multicollineariteit noch

de geldigheid van de richtingscoëfficiënten in het gedrang, noch de geldigheid van de standaardfouten (Berry 1993).

Er worden inderdaad minder snel significante verbanden tussen de verklarende en afhankelijke variabelen gevonden wanneer de verklarende variabelen sterk samenhangen. Maar als dat als een "probleem" wordt geduid, dan kunnen we net zo goed alle regels van de inferentiële statistische theorie, waarop probabilistische hypothesetoetsen gebaseerd zijn, als probleem gaan beschouwen. De standaardfouten zijn bij multicollineariteit niet ten onrechte, maar geheel terecht hoger dan bij gebrek aan multicollineariteit. Om *precies* dezelfde reden is de standaardfout voor een geschat gemiddelde of een geschatte proportie groter naarmate het absolute aantal steekproefeenheden kleiner is.

De voorstelling van multicollineariteit als een fundamenteel probleem of een "hardnekkige ziekte" die tot opgeblazen standaardfouten leidt, vindt mogelijk zijn oorsprong in het feit dat vele handboeken de variantie-analyse van experimentele, orthogonale data als referentie nemen. Als de randomisatie in een experiment zijn werk naar verwachting heeft gedaan, dan correleert de experimentele manipulatie inderdaad niet (significant) met andere, mogelijk causale kenmerken en ligt de VIF dicht bij de neutrale waarde 1. In dat geval is inderdaad een zeer efficiënte schatting van het unieke, directe effect van de experimentele manipulatie mogelijk en volstaat een kleine steekproef vaak voor betrouwbare hypothesetoetsing (omdat de matrix **X'X** dan een diagonale matrix is waarin geen informatie verloren gaat om de regressieparameters te schatten, zie Kendall & Stuart 1973, 371). Dat is mooi maar er is geen enkele reden waarom sociologen de orthogonale situatie, waarin verklarende variabelen niet lineair samenhangen, als referentie zouden moeten nemen. Orthogonale regressie-analyse is een speciaal geval met gunstige, maar uitzonderlijke, gevolgen. Orthogonaliteit maakt geen deel uit van de algemene theorie van regressie (Kendall & Stuart 1973, Chapter 28; Berry 1993). En of we dat nu graag hebben of niet, het maakt ook zelden deel uit van de sociale realiteit.

"Zelfs het teken van het effect wordt instabiel bij multicollineariteit"

Een ander argument dat vaak tegen multicollineariteit gebruikt wordt, is de vaststelling dat het teken van een geschatte richtingscoëfficiënt vaak verandert wanneer één van twee collineaire predictoren uit de regressievergelijking verwijderd wordt. Erger nog: het nagaan van de stabiliteit van schattingen bij regressieformuleringen mét en zonder collineaire variabelen wordt soms verkocht als een toets om te kijken of de "robuustheid" van de schattingen niet door multicollineariteit bedreigd wordt. Ook dit is een misverstand. Multicollineariteit vormt juist een argument *pro* meervoudige regressie omdat enkel door het gelijktijdig opnemen van collineaire variabelen verdoken effecten en schijneffecten aan het licht gebracht kunnen worden. Net zoals bij de schatting van univariate parameters kan dit uiteraard enkel op betrouwbare manier

als de steekproef groot genoeg is voor de beoogde doeleinden. Als het effect van variabele A na controle voor variabele B significant negatief is, terwijl het zonder die controle voor B significant positief is, dan wijst dat niet op een onbetrouwbare regressie als gevolg van multicollineariteit, maar dan heeft men dankzij de meervoudige regressie een positief schijneffect van A ontmaskerd. Multicollineariteit, als "a state of nature", leidt niet tot problemen voor meervoudige regressie maar leidt juist bij enkelvoudige regressie tot schijneffecten. Het slechtste wat men in geval van multicollineariteit kan doen, hoewel het vaak wordt aanbevolen, is om correlerende variabelen weg te laten.

"Bij multicollineariteit kunnen de F- en de T-test tegenstrijdig zijn"

Als derde teken aan de wand wordt vaak aangehaald dat multicollineariteit kan leiden tot een contradictie tussen een globale test van de significantie van een regressiemodel enerzijds en het toetsen van de significantie van specifieke regressieparameters anderzijds. Deze kritiek wijst opnieuw op twijfels aan de geldigheid van de normaal toepasselijke, inferentieel-statistische regels in geval van "ernstige" multicollineariteit. Opnieuw gaat het om een misverstand.

De F-test voor de significantie van het regressiemodel is bedoeld om na te gaan of de onafhankelijke variabelen een significant deel van de globale variantie in de afhankelijke variabele verklaren. Daarbij wordt de variantie in rekening gebracht die gezamenlijk door alle onafhankelijke variabelen samen verklaard wordt. De T-toetsen voor de partiële regressieparameters hebben daarentegen enkel betrekking op de unieke bijdragen van de onafhankelijke variabelen aan de verklaring van de variantie in de afhankelijke variabele (Welkenhuysen-Gybels & Loosveldt 2002).

Multicollineariteit leidt soms tot de schijnbaar contradictorische situatie waarin de globale F-test aangeeft dat het regressiemodel een significant deel van de variantie in de afhankelijke variabele verklaart terwijl geen enkele van de individuele onafhankelijke variabelen een significant effect heeft. De schijnbare contradictie berust op het niet consequent doordenken van het verschil tussen enerzijds de totale hoeveelheid variantie in y die door de onafhankelijke variabelen samen verklaard wordt en anderzijds de unieke bijdrage van elke onafhankelijke variabele afzonderlijk. De variantie van y wordt bij meervoudige regressie zowel door de communaliteit van de verklarende variabelen verklaard als door de unieke, niet covariërende delen van die variabelen. Er is helemaal geen tegenstrijdigheid wanneer verklarende variabelen gemeenschappelijk wél iets significants verklaren van y maar hun unieke delen (de variantie die niet gemeenschappelijk is met andere variabelen) niet. De partiële regressieparameters worden enkel berekend op basis van de uniciteit van de predictoren, dus op basis van het deel van de predictor dat met geen enkele andere verklarende variabele overlapt (zie figuur 1) – en dat is ook de bedoeling, want we zoeken het unieke, directe effect van de predictoren. Bij de berekening van de

proportie verklaarde variantie worden daarentegen niet alleen observaties gebruikt die buiten het multicollineaire patroon vallen (de niet-overlappende delen van de verklarende variabelen) maar ook de observaties die wél in het multicollineaire patroon passen (de wél overlappende delen van de verklarende variabelen).

Verschillen tussen de F- en de T-test wijzen dus op twee verschillende, inhoudelijk belangrijke zaken in plaats van op een multicollineariteitsprobleem. Wat we nooit uit het oog mogen verliezen, is dat een partiële regressieparameter enkel het unieke, *directe* effect van de betreffende variabele reflecteert en niet het *totale* effect van een variabele. Vaak is het totale effect van een variabele een pak groter dan aangegeven door de regressieparameter. Dat volgt uit de aard der zaak van een partiële regressieparameter en is geen "fout" als gevolg van multicollineariteit. Op theoretische gronden kan dan eventueel het gemeenschappelijke deel van de verklarende variabelen aan één, causaal prioritaire variabele worden toegewezen (wat gebeurt in een padanalyse, zie Welkenhuysen-Gybels & Loosveldt 2002, 144-173), maar dat is een theoretische kwestie die we niet aan de regressietechniek kunnen overlaten. Als onzekerheid bestaat over welke variabele causaal voorrang heeft, dan is dat geen "multicollineariteitsprobleem" maar wel een theoretisch probleem. Men mag van meervoudige regressie niet verwachten dat het problemen zou oplossen waarvoor het niet ontworpen is.

"Multicollineariteit leidt soms tot resultaten die klinkklare nonsens zijn"

Een laatste onterecht verwijt, dat zonder meer impliceert dat multicollineariteit tot grove ongeldigheid zou leiden bij meervoudige regressie, is dat zich bij multicollineariteit de gekste dingen kunnen voordoen wanneer verklarende variabelen sterk samenhangen. Onder meer wordt dan verwezen naar een oefening in het tweede deel van *The Advanced Theory of Statistics* van Kendall & Stuart (1973, 359). Daarin is sprake van twee verklarende variabelen die onderling zeer sterk correleren. Eén van deze twee correleert helemaal niet met de afhankelijke variabele terwijl de andere slechts gematigd met *y* correleert. Nochtans is de meervoudige determinatiecoëfficiënt van de regressie met beide predictoren in de regressie gelijk aan 1 ($R_y^2 = 1$) (Tacq 2001, 103; 2004, 668). Dit voorbeeld wordt nogal eens aangehaald om te illustreren hoe multicollineariteit tot absurde situaties leidt. In feite is er niets absurds aan de beschreven situatie. Het gaat gewoon om een voorbeeld van een schijnbaar noneffect: een effect dat bij bivariate analyse aan het zicht onttrokken wordt door een onderdrukker-variabele ("suppressor variable" in het Engels) (Hamilton 1987). Het komt in de sociologie heel vaak voor dat een variabele in een bivariate analyse geen verband lijkt te hebben met de afhankelijke variabele, terwijl dat na controle voor derde variabelen wel het geval blijkt te zijn. De data in tabel 2 voldoen, bijvoorbeeld, aan deze beschrijving.

Tabel 2. Voorbeeld van een datamatrix met een onderdrukker (zie tekst).

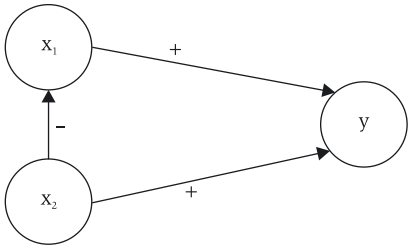
x_1	2.23	2.57	3.87	3.1	3.39	2.83	3.02	2.14	3.04	3.26	3.39	2.35	2.76	3.9	3.16
x_2	9.66	8.94	4.4	6.64	4.91	8.52	8.04	9.05	7.71	5.11	5.05	8.51	6.59	4.9	6.96
y	12.37	12.66	12.00	11.93	11.06	13.03	13.13	11.44	12.86	10.84	11.2	11.56	10.83	12.63	12.46

Veronderstel dat x_1 staat voor de mate van angst voor diefstal en inbraak en veronderstel dat x_2 de scores bevat op een armoedeschaal (hoe hoger de score, hoe armer), telkens gemeten op intervalniveau. Afhankelijke variabele y is de mate waarin men instemt met autoritaire beleidsmaatregelen, opnieuw gemeten met een intervalschaal. De bivariate correlatie tussen x_1 en y in tabel 2 is zo goed als gelijk aan nul, terwijl de correlatie tussen x_2 en y ongeveer 0.43 bedraagt. Het gebrek aan correlatie tussen x_1 en y wil echter niet zeggen dat x_1 geen causaal effect op y heeft. Om de unieke, directe effecten van x_1 en x_2 op y te schatten, doet men een meervoudige lineaire regressie-analyse, gefit met de kleinste-kwadratenmethode. Dit geeft de volgende resultaten.

Ondanks het totale gebrek aan correlatie met y is de geschatte regressiecoëfficiënt voor x_1 gelijk aan 3,09. Ondanks de hoge multicollineariteit (de correlatie tussen x_1 en x_2 bedraagt -0,90) en de beperkte steekproefomvang (n = 15) is deze schatting zeer precies: de geschatte standaardfout bedraagt slechts 0,012 (en er is niets mis met deze schatting: hij is geldig en het is de laagste standaardfout die we op basis van de data kunnen maken). Het netto-effect van x_2 op y bedraagt 1.03 (met geschatte standaardfout kleiner dan 0.004). In tabel 2 is er voor gezorgd dat x_1 en x_2 samen 100% van de variantie van y verklaren (R^2 = 1.00).

Ondanks de lage of afwezige bivariate correlaties tussen de onafhankelijke en de afhankelijke variabelen, verklaren de twee causale factoren samen toch perfect de afhankelijke variabele. In dit voorbeeld is er duidelijk sprake van een onderdrukkingseffect: x_2 is een onderdrukker voor het effect van x_1 op y (Hamilton 1987; Edward 2001). Zolang niet voor x_2 gecontroleerd wordt, blijft het effect van x_1 op y onzichtbaar omdat x_1 samenhangt met x_2 (om wat voor reden dan ook, causaal of niet). Statistisch is het mechanisme identiek aan de situatie bij een schijneffect, alleen leidt bivariate analyse bij aanwezigheid van een niet in rekening gebrachte onderdrukker tot de foute conclusie dat er géén in plaats van wél een effect is. Neem terug het voorbeeld waar de instemming met autoritaire beleidsmaatregelen als afhankelijke variabele fungeert. De twee verklarende variabelen, armoede (x_2) en angst voor diefstal en inbraak (x_1), hangen sterk negatief samen, misschien omdat wie arm is, volgens de gebruikte schaal minder schrik heeft voor inbraak (zie figuur 2). Hoe dan ook, de samenhang is sterk genoeg om bij bivariate analyse te zorgen voor een schijnbaar non-effect: het positieve effect van angst voor diefstal (x_1) op instemming met autoritaire maatregelen (y) kan pas worden onthuld na controle voor een derde variabele (x_2), namelijk armoede. De collineariteit impliceert immers dat

wie angst heeft voor diefstal meestal niet arm is en om die reden minder vaak met autoritaire beleidsmaatregelen instemt. Wie én angst heeft én arm is, zal volgens het model en de data sterk met autoritaire maatregelen instemmen.



Figuur 2. Causaal model met x_2 als onderdrukker.

Dit voorbeeld onderstreept nog maar eens dat multicollineariteit veeleer als een aanleiding tot dan als een probleem voor meervoudige regressie opgevat zou moeten worden. Het illustreert dat meervoudige regressie onontbeerlijk is als de verklarende variabelen onderling correleren. (Voor een meer technische en ook geometrische behandeling van dit fenomeen, zie Hamilton 1987). Dergelijke onderdrukkingseffecten zijn schering en inslag in de sociologie. Enkel met meervoudige regressie is het mogelijk om de reële, unieke en directe effecten van de twee variabelen aan het licht te brengen, juist omdat die twee zo sterk samenhangen. Als multicollineariteit ("a state of nature") voor iets een probleem vormt, dan is het wel voor bivariate correlatieanalyse. Maar zoals altijd geldt dat wie een gedetailleerde hypothese statistisch wil toetsen daarvoor voldoende data moet hebben. Om te vliegen heb je vleugels nodig.

De verwarring tussen correlaties en meervoudige (en dus partiële) regressiecoëfficiënten zorgt voor nog een misverstand over multicollineariteit. Sommige publicaties geven het volgende argument als ultiem bewijs dat multicollineariteit problemen oplevert voor meervoudige regressie: bij hoge multicollineariteit zijn de "gestandaardiseerde regressiecoëfficiënten" soms kleiner dan –1 of groter dan +1. Die zogenaamd gestandaardiseerde regressiecoëfficiënten, ook wel bèta's genoemd, verkrijgt men door zowel de afhankelijke als de onafhankelijke variabelen te standaardiseren alvorens de regressieparameters te schatten (al kunnen ze ook op andere manieren worden berekend). Het is echter verwarrend om dit "gestandaardiseerde regressiecoëfficiënten" te noemen (zie Bring 1994; Welkenhuysen-Gybels & Loosveldt 2002, 89). Hieronder wordt voor het gemak de term 'bèta-coëfficiënt' gebruikt. Na standaardisering valt het intercept weg en in het geval van enkelvoudige, bivariate regressie kan de richtingscoëfficiënt als een correlatiecoëfficiënt geïnterpreteerd worden. Zoals bekend geeft een gekwadrateerde correlatiecoëfficiënt de hoeveelheid variantie van de ene variabele die lineair verklaarbaar is op basis van de variantie in de andere variabele. Een correlatiecoëfficiënt r groter dan +1 of kleiner dan –1 is

absurd omdat het een proportie verklaarde variantie r^2 van meer dan 100% impliceert.

Het punt is dat de richtingscoëfficiënten die men verkrijgt bij meervoudige regressie op gestandaardiseerde data (de bèta's) niet als correlatiecoëfficiënten geïnterpreteerd mogen worden en hun kwadraat al evenmin als proportie door die variabele verklaarde variantie. Bij meervoudige regressie zijn bèta's die kleiner zijn dan –1 of groter dan +1 dan ook geen absurditeit maar een mogelijke realiteit. Het probleem is niet de multicollineariteit. Het probleem is de onterechte suggestie dat met behulp van de bèta's kan worden bepaald welk deel van de meervoudige determinatiecoëfficiënt mag toegeschreven worden aan elke verklarende variabele. Dat wekt de indruk dat de meervoudige determinatiecoëfficiënt niet meer is dan de som van de gekwadrateerde bèta's. Dat is echter zelden of nooit waar. Het klopt alleen in de zelden geziene gevallen waar de correlatie tussen alle verklarende variabelen gelijk is aan nul. Dat is te zien in volgende formule van de meervoudige determinatiecoëfficiënt, geschreven in functie van de bèta's en de correlatiecoëfficiënten tussen verklarende variabelen (Cohen e.a. 2003, 83):

$$R^2 = \sum_{k=1}^{p} \beta_{Zk}^2 + 2\sum_{j=1}^{p-1}\sum_{k=j+1}^{p} (\beta_{Zj}\beta_{Zk}r_{jk}) \qquad (6)$$

waarin β_{zk} staat voor de regressieparameter van variabele z_k (d.i. de gestandaardiseerde waarde van variabele x_k) en waarin er sprake is van p verschillende, verklarende variabelen. Enkel wanneer geen enkel paar (z_j, z_k) van verklarende variabelen correleert en r_{jk} dus gelijk is aan nul voor alle $j \neq k$, vervalt de rechterterm van formule (6). Enkel bij perfect ongecorreleerde verklarende variabelen kan de verklaarde variantie dus op basis van de regressieparameters zomaar aan elk der predictoren worden toegewezen.

De populatiewaarden van de bèta's kunnen gerust groter zijn dan +1 of kleiner dan –1. Dit is trouwens het geval in de populatie waaruit de gegevens in tabel 2 een steekproef vormen. Geldige schattingen van die bèta's moeten dus ook groter dan +1 of kleiner dan –1 kunnen zijn – anders zou dat het ironische bewijs van de ongeldigheid van die schattingen zijn. De meervoudige determinatiecoëfficiënt R^2 wordt, volgens de decompositie van formule (6), enerzijds bepaald door de unieke bijdragen van de onafhankelijke variabelen (de gekwadrateerde β's) en anderzijds door combinaties van en samenhangen tussen variabelen (de $\beta_j \beta_k r_{jk}$-termen). De β's kunnen net als de r's zowel positief als negatief zijn. Het is dus best mogelijk dat een $\beta_j \beta_k r_{jk}$-combinatie een negatief resultaat oplevert. De totale meervoudige determinatiecoëfficiënt, en dus de optelsom van alle termen in formule (6), zal en kan echter nooit groter zijn dan 1, onafhankelijk van de mate van multicollineariteit.

Het voorgaande impliceert trouwens een argument tegen het gebruik van regressie op gestandaardiseerde data om zogenaamd het relatieve belang van

predictoren in te schatten (voor andere argumenten, zie Bring 1994). Het relatieve causale belang van predictoren kan onmogelijk op basis van een meervoudige regressievergelijking nagegaan worden, zelfs niet wanneer de variabelen allemaal gestandaardiseerd zijn, tenzij in het sociologisch zeer uitzonderlijke geval waarin alle predictoren orthogonaal zijn (Sharma 1996, 273). Het totale causale belang van een variabele x_1 voor y bestaat immers niet alleen in het directe effect van x_1 op y maar ook in het eventuele indirecte effect dat x_1 heeft op y via de omweg van andere variabelen. Een meervoudig regressiemodel beperkt zich tot het inschatten van louter de directe effecten. Wie meer wil, moet een structureel model bouwen.

4. **Besluit**

Multicausaliteit en multicollineariteit vormen voor sociologen twee belangrijke motivaties om aan meervoudige regressie te doen. De schatting van de effecten van de verklarende variabelen gebeurt immers efficiënter en preciezer wanneer ze voor alle predictoren tegelijkertijd wordt uitgevoerd en bovendien is het enkel met meervoudige regressietechnieken mogelijk om tot geldige schattingen te komen van het unieke effect van predictoren wanneer die predictoren onderling samenhangen. Hoe sterker de samenhang tussen verklarende variabelen, hoe sterker de motivatie om aan meervoudige regressie te doen. Een fundamentele vereiste is wel dat de onafhankelijke variabelen geen exacte lineaire combinatie van elkaar vormen, want dan is er een onoverkomelijk probleem van gebrek aan informatie (wat algebraïsch blijkt uit het feit dat de datamatrix **X'X** singulier is).

Naarmate de verklarende variabelen in een meervoudige regressie sterker lineair samenhangen, zijn de standaardfouten van de regressieparameters *ceteris paribus* groter en dus de schattingen onbetrouwbaarder. De oorzaak van dit fenomeen is echter niet inherent aan de gebruikte techniek. Het "probleem" van multicollineariteit is een probleem van gebrek aan overeenstemming tussen de verzamelde data en het beoogde model (Draper & Smith 1998, 369-370). Wanneer er door multicollineariteit geen betrouwbare schatting van regressieparameters mogelijk is, dan is dat een symptoom dat er ofwel iets schort aan de data, ofwel aan het theoretische model en/of aan de operationalisering van de concepten. Het is een probleem van ofwel een tekort aan informatie, ofwel een tekort aan theoretische doordachtheid. Het wijst er niet op dat de multicollineariteit te hoog is voor meervoudige regressie-analyse als zodanig.

Wie met sterke samenhang tussen predictoren geconfronteerd wordt enerzijds, en met onbetrouwbare regressieparameters anderzijds, hoeft dus haar of zijn tijd niet te verdoen met het zoeken naar de sterkte van de multicollineariteit en het berekenen van allerhande indicatoren voor de sterkte van het feno-

meen. Men moet in zo'n geval vooral nagaan of de theorie en de operationalisering van de concepten in orde zijn enerzijds, en of de data wel geschikt zijn voor de hypothesen die men wil toetsen anderzijds. Om het unieke effect van samenhangende predictoren te kunnen schatten, is meer informatie nodig dan om het globale effect van alle predictoren samen te kunnen schatten, of om het unieke effect van niet lineair afhankelijke predictoren te bepalen.

Als we mogen aannemen dat we een goede theorie en valide operationalisering van duidelijk onderscheiden maar niettemin samenhangende concepten hebben en die theorie op een geldige manier in een regressiemodel vertaald hebben, dan wijzen hoge standaardfouten als gevolg van multicollineariteit op een probleem in het onderzoeksdesign. Een goed onderzoeksdesign is er een dat leidt tot de verzameling van data die geschikt zijn om de onderzoeksvragen te beantwoorden. Als een onderzoeker dus het directe causale effect van *X* op *Y* wil onderzoeken met behulp van meervoudige regressie (om zo te kunnen controleren voor de storende invloed van andere verklarende factoren), dan houdt een goed onderzoeksdesign dus in dat rekening gehouden wordt met de samenhang tussen *X* en andere predictoren van *Y*. Wie in een multicausale context het unieke effect van diverse oorzakelijke factoren wil nagaan, moet een voldoende hoog absoluut aantal cases hebben waarin de ene oorzakelijke factor niet voorspelbaar is op basis van de andere. Als de oorzakelijke factoren sterk samenhangen en er geen experimentele manipulatie mogelijk is, dan zullen er dus meer, of gerichter verzamelde, data nodig zijn om de theorie te toetsen.

Wanneer een onderzoeker gebruik maakt van een bestaande dataset die door de multicollineariteit ongeschikt blijkt om de beoogde hypothesen te toetsen, dan zal geen enkele kunstgreep daar iets aan veranderen: de data missen nu eenmaal de power die nodig is om de theorie te testen. Zeggen dat men een bepaalde hypothese niet betrouwbaar kan toetsen als gevolg van multicollineariteit is even terecht en even nietszeggend als zeggen dat men de inkomensongelijkheid in België niet betrouwbaar kan meten op basis van een steekproef van 50 eenheden. Is het cynisch om zo'n onderzoeker te adviseren om bijkomende data te verzamelen? Wie ambitieuze analyses voor ogen heeft, moet daar de nodige data voor in huis halen.

NOTEN

1. Geciteerd in Fox & Monette (1992, 183).
2. We verkiezen de term "meervoudige regressie" als Nederlands equivalent voor het Engelse "multiple regression" boven uitdrukkingen als "multipele -" of "multiple regressie" die in sommige (oudere) handboeken gebruikt worden.
3. Met hartelijke dank aan één van de anonieme reviewers van het *Tijdschrift voor Sociologie* voor de suggestie om dit verhelderende voorbeeld toe te voegen.

BIBLIOGRAFIE

Allen, M.P. (1997), *Understanding Regression Analysis*. New York/London: Plenum Press.

Berry, W.D. (1993), *Understanding Regression Assumptions*. Newbury Park/London: Sage. (Quantitative Applications in the Social Sciences, 92)

Bring, J. (1994), How to standardize regression coefficients, *The American Statistician,* 48(3), 209-213.

Carroll, J.D. & P.E. Green (1997), *Mathematical Tools For Applied Multivariate Analysis. Revised edition*. New York/London: Academic Press.

Chipman, J.S. (1964), On least squares with insufficient observations, *Journal of the American Statistical Association,* 59(308), 1078-1111.

Cohen, J., P. Cohen, S.G. West & L.S. Aiken (2003), *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences. Third Edition*. Mahwah, New Jersey: Lawrence Erlbaum Associates.

De Weerdt, Y. & H. De Witte (2001), Economische progressiviteit bij werknemers. Geworteld in hun arbeid?, *Tijdschrift Voor Sociologie,* 22(3), 217-241.

Dobson, A.J. (2002), *An Introduction to Generalized Linear Models. Second Edition*. London/New York: Chapman & Hall/CRC.(Texts in Statistical Science).

Draper, N. R. & H. Smith (1998), *Applied Regression Analysis*. New York: John Wiley & Sons.

Edward, H.S.I. (2001), Visualizing Multiple Regression, *Journal of Statistics Education,* 9(1). (http://www.amstat.org/publications/jse/v9n1/ip.html)

Fox, J. (1991), *Regression diagnostics*. Thousand Oaks: Sage.(Quantitative Applications in the Social Sciences, 79)

Fox, J. & G. Monette (1992), Generalized collinearity diagnostics, *Journal of the American Statistical Association,* 87(417), 178-83.

Friedrich, R.J. (1982), In defense of multiplicative terms in multiple regression equations, *American Journal of Political Science,* 26(4), 797-833.

Goldberger, A.S. (1991), *A Course in Econometrics*. Cambridge (Massachusetts)/London (England): Harvard University Press.

Hair, J.F. jr., R.E. Anderson, R.L. Tatham & W.C. Black (1998), *Multivariate Data Analysis. Fifth Edition. International Edition*. Upper Saddle River, New Jersey: Prentice Hall.

Hamilton, D. (1987), Sometimes $R^2 > r^2_{yx_1} + r^2_{yx_2}$. Correlated variables are not always redundant, *The American Statistician,* 41(2), 129-132.

Hutcheson, G. & N. Sofroniou (1999), *The Multivariate Social Scientist*. London/Thousand Oaks: Sage.

ISI International Statistical Institute (2005), *ISI Multilingual Glossary of Statistical Terms*. Voorburg: ISI (http://isi.cbs.nl/glossary/).

Jaccard, J. & R. Turrisi (2003), *Interaction Effects in Multiple Regression*. Thousand Oaks/London: Sage.(Quantitative Applications in the Social Sciences, 72)

Johnston, J.J. (1984), *Econometric Methods*. New York: McGraw-Hill.

Kendall, M.G. & A. Stuart (1973), *The Advanced Theory of Statistics. Volume 2: Inference and Relationship. Third Edition*. London: Griffin.

Koutsoyiannis, A. & C.F. Carter (1973), *Theory of Econometrics: An Introductory Exposition of Econometric Models*. London: Macmillan.

Lewis-Beck, M.S. (1980), *Applied Regression. An Introduction*. Beverly Hills/London: Sage.

Marquardt, D.W. (1980), Comment: You should standardize the predictor variables in your regression models, *Journal of the American Statistical Association,* 75(369), 87-91.

McCullagh, P. & J.A. Nelder (1983), *Generalized Linear Models*. New York: Chapman and Hall. (Monographs on statistics and applied probability).

Morrow-Howell, N. (1994), The M word: Multicollinearity in multiple regression, *Social Work Research,* 18(4), 247-251.

Öztürk, F. & F. Akdeniz (2000), Ill-conditioning and multicollinearity, *Linear Algebra and Its Applications,* 321(1-3), 295-305.

Schroeder, L.D., D.L. Sjoquist & P.E. Stephan (1986), *Understanding regression analysis. An introductory guide*. Beverly Hills/London: Sage.(Quantitative Applications in the Social Sciences, 57)

Sharma, S. (1996), *Applied Multivariate Techniques*. New York: John Wiley & Sons.

Silvey, S.D. (1969), Multicollinearity and imprecise estimation, *Journal of the Royal Statistical Society. Series B (Methodological),* 31(3), 539-552.

Smith, G. & F. Campbell (1980), A critique of some ridge regression methods, *Journal of the American Statistical Association,* 75(369), 74-81.

Tacq, J. (1992), *Van probleem naar analyse: de keuze van een gepaste multivariate analysetechniek bij een sociaalwetenschappelijke probleemstelling*. Rotterdam: RISBO.

Tacq, J. (2001), *Het methodologisch atelier. Adviezen en beschouwingen voor de sociale wetenschappen*. Leuven: Acco.

Tacq, J. (2004), Multicollinearity, pp. 667-69 in: M.S. Lewis-Beck, A. Bryman & T.F. Liao (eds.), *The Sage Encyclopedia of Social Science Research Methods* Thousand Oaks: Sage.

Van Ruysseveldt, J. (2003), Werkgelegenheidsbevorderende maatregelen in CAO's: een onderzoek naar de macroresponsiviteit van het sectorale CAO-overleg in België, *Tijdschrift Voor Sociologie,* 24(4), 331-63.

Welkenhuysen-Gybels, J. & G. Loosveldt (2002), *Regressieanalyse: een introductie in de multivariabelenanalyse*. Leuven: Acco.

Wooldridge, J.M. (2003), *Introductory Econometrics. A Modern Approach*. Mason (Ohio): Thomson/South-Western.