SVR - Technisch rapport 2008 / 1

De interpretatie van interactie-effecten in regressiemodellen

Jan Pickery

Studiedienst van de Vlaamse Regering

Vlaamse overheid

De interpretatie van interactie-effecten in regressiemodellen

Jan Pickery

Vlaamse overheid

Samenstelling
Diensten voor het Algemeen
Regeringsbeleid
Studiedienst van de Vlaamse Regering

Jan Pickery

Verantwoordelijke uitgever
Josée Lemaître
Administrateur-generaal
Boudewijnlaan 30 bus 23
1000 Brussel

Lay-out cover
Diensten voor het Algemeen
Regeringsbeleid
Communicatie
Patricia Van Dichel

Druk
Agentschap voor Facilitair Management

Depotnummer
D/2008/3241/253

http://www4.vlaanderen.be/dar/svr

I N H O U D S T A F E L

1. Inleiding

De Studiedienst van de Vlaamse Regering (SVR) wil bijdragen aan een kwaliteitsverhoging van de statistiekproductie binnen de Vlaamse overheid. Onze brochure over de principes van kwaliteitszorg in het statistische productieproces met heel wat aanbevelingen in verband met het verzamelen, verwerken en documenteren van statistische gegevens, kadert daarin (APS, 2003). Verder willen wij ook concretere handleidingen aanbieden over het juiste gebruik van statistische technieken. Zo was er reeds een *Technisch rapport* over de analyse van een evolutie in een ongelijke participatie (Pickery, 2006) en volgt er binnenkort een rapport over het gebruik van contextuele regressiemodellen bij het vergelijken van landen (Callens, 2008).

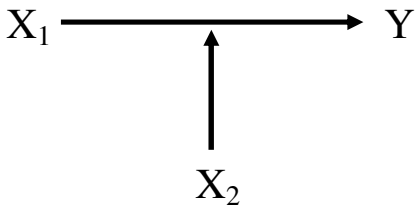
In dit rapport gaan we dieper in op de inhoudelijke interpretatie van interactie-effecten in regressiemodellen. Deze tekst richt zich voornamelijk op de interpretatie. De meer technische kant van regressieanalyse wordt voldoende behandeld in verschillende handboeken, zie bijvoorbeeld McClendon (2002) en Welkenhuysen-Gybels & Loosveldt (2002) voor meervoudige regressie en Pampel (2000) en Hosmer & Lemeshow (2000) voor logistische regressie. Die teksten behandelen meestal ook interactie-effecten, maar toch blijken vele onderzoekers nog problemen te ondervinden bij de interpretatie ervan. De sterke inhoudelijke focus van deze tekst kan voor die onderzoekers een hulp betekenen.

In de volgende paragraaf bekijken we statistische interactie vanuit een theoretische invalshoek. Die paragraaf verduidelijkt ook waarom een statistische analyse rekening moet houden met het bestaan van interacties. Paragraaf 3 illustreert enkele interacties met eenvoudige cijfervoorbeeldjes, uitgewerkt in verschillende tabellen. In paragraaf 4 nemen we interactie-effecten op in een meervoudige lineaire regressie. In het eerste deel van die paragraaf (4.1) bespreken we de interpretatie van een regressiemodel zonder interactie-effecten. Het daaropvolgende deel (4.2) bevat de eigenlijke focus van deze nota. Die paragraaf 4.2 beschrijft uitvoerig de betekenis van verschillende soorten interactie-effecten in een meervoudige lineaire interactie. Paragrafen 5 en 6 bespreken interactie-effecten in respectievelijk binaire en multinomiale logistische regressie en paragraaf 7 ten slotte toont enkele bijkomende toepassingsmogelijkheden.

2. Interactie als statistisch concept

In onderzoek dat gebruik maakt van de statistische methoden wordt de term “interactie” meestal voorbehouden voor situaties waarbij de impact van een onafhankelijke variabele op een afhankelijke variabele beïnvloed wordt door een derde variabele. In formele termen: het effect van X_1 op Y varieert naargelang de waarden van X_2. Grafisch ziet dat er dan zo uit:

Grafiek 1 Voorstelling van een interactie-effect



Jaccard (2001) en Jaccard en Turrisi (2003) reiken een conceptueel kader aan om zo’n situatie te beschrijven. Volgens dat kader wordt Y de afhankelijke variabele genoemd, X_1 de onafhankelijke *focusvariabele* en X_2 de *moderatorvariabele*. X_2 verandert immers de wijze waarop X_1 Y bepaalt.

Dit conceptueel kader veronderstelt dat de onderzoeker vooraf bepaalt van welke X-variabele in eerste instantie het effect op Y wordt onderzocht en welke X-variabele die relatie wijzigt. Soms ligt die keuze voor de hand. Jaccard en Turrisi (2003, 3) geven het voorbeeld van een experimenteel opzet, waarbij patiënten al dan niet een bepaalde behandeling krijgen. De impact van die behandeling op de gezondheidstoestand vormt het eigenlijke voorwerp van het onderzoek en het al dan niet krijgen van die behandeling is bijgevolg de focusvariabele. Als het effect van de behandeling anders blijkt te zijn voor vrouwen dan voor mannen is geslacht een moderatorvariabele.

In andere voorbeelden is de keuze minder eenduidig. Wat voor de ene onderzoeker een focusvariabele is, kan voor de andere onderzoeker een moderatorvariabele zijn en vice versa. Stel dat iemand

geïnteresseerd is in de impact van het opleidingsniveau en het geslacht op het loon van werknemers. Bij het bestaan van een interactie kan de interpretatie luiden dat het effect van het geslacht van een werknemer op zijn/haar loon varieert naargelang het opleidingsniveau van die werknemer. Maar de verklaring dat het scholingsniveau bij vrouwen een andere impact heeft op het loon dan bij mannen is waarschijnlijk evenwaardig. Het is ook niet fout om beide perspectieven te belichten. Maar zelfs als dat gebeurt, kan het conceptueel kader van Jaccard en van grafiek 1 zijn nut bewijzen. Het wordt dan gewoon twee keer toegepast.

Dit conceptueel kader geeft ook aan waarom het bestuderen van interactie-effecten nuttig en/of nodig kan zijn. Een model zonder interactie-effecten maakt impliciet de veronderstelling dat het effect van de onafhankelijke focusvariabele op de afhankelijke variabele constant is. Die veronderstelling is vaak niet realistisch. Als er redenen zijn – theoretisch of op basis van vorig onderzoek – om aan te nemen dat de aard van dat effect samenhangt met een derde variabele, is een model met interactie-effecten aangewezen. Alternatieve benaderingen in deze situatie, zoals bvb. het schatten van verschillende modellen voor één of meerdere subpopulaties, vormen een minder goede oplossing. Stel dat X_2 een dichotome variabele is (bvb. geslacht) dan zou als alternatief een aparte regressievergelijking voor mannen en voor vrouwen geschat kunnen worden. Als X_2 meerdere categorieën telt, dan zouden voor alle onderscheiden groepen van die variabele regressievergelijkingen geschat kunnen worden. Als X_2 metrisch zou zijn, zou de variabele vanuit dat oogpunt gecategoriseerd kunnen worden. Zulke afzonderlijke regressies zijn misschien makkelijker te interpreteren en laten toe om op eenvoudige wijze specifieke groepen nader te onderzoeken. Maar een werkwijze met gescheiden regressievergelijkingen geeft geen indicatie over de significantie van de gevonden verschillen tussen de groepen. Bovendien verkleint het statistische onderscheidingsvermogen niet alleen voor de effecten van de interagerende variabelen, maar voor alle in de regressie opgenomen effecten. Voornamelijk om die reden vinden Jaccard et al. (1991, 48-49) het opnemen van interactie-effecten in het regressiemodel de beste keuze.

3. Eenvoudige voorbeelden van interactie

Enkele eenvoudige beschrijvende tabellen kunnen statistische interacties illustreren. Hieronder volgen twee voorbeelden die gebaseerd zijn op de SCV-survey van 2002 en de SILC-enquête van 2005. Meer informatie over deze surveys is te vinden in Carton et al. (2003) en in een uitgebreid kwaliteitsrapport van de Algemene Directie Statistiek en Economische Informatie (2006).

Het **eerste voorbeeld** kijkt naar het aantal uren dat de respondenten vrij hebben op een dag in de week- of een werkdag. De vraag die in de SCV-survey 2002 gesteld werd aan de 18- tot 85-jarige respondenten, luidde:

Wanneer u de voorbije maand bekijkt, hoeveel uren had u dan gewoonlijk vrij tijdens een weekdag of een werkdag?
Tijd die gebruikt wordt voor betaald werk, verzorging van de kinderen of andere huisgenoten, het vervullen van taken waartoe men zich verplicht voelt en essentiële behoeften zoals eten, persoonlijke verzorging en slapen, tellen niet mee als vrije tijd.

Uit tabel 1 blijkt dat het gemiddelde aantal gerapporteerde uren vrije tijd iets meer dan 3,7 bedraagt (bijna 3 uur en drie kwartier). Maar er zijn natuurlijk grote verschillen. Sommige respondenten zeggen helemaal geen vrije tijd te hebben op een weekdag terwijl andere respondenten zeggen over maar liefst 16 vrije uren te beschikken[1].

Tabel 1 Aantal uren vrije tijd op een dag in de week

	N	Minimum	Maximum	Gemiddelde	Standaard-afwijking
vrij_weekdag	1460	0	16	3,71	2,82

Bron: SCV-survey 2002

[1] Uit de oorspronkelijke dataset zijn enkele extreme antwoorden (20 uren of meer) verwijderd.

Deze beschikbare hoeveelheid vrije tijd verschilt volgens een aantal kenmerken. Twee van die kenmerken zijn geslacht en het al dan niet hebben van betaald werk. Tabel 2 toont inderdaad het bestaan van die verschillen.

Tabel 2 Gemiddeld aantal uren vrije tijd op een dag in de week volgens geslacht en volgens het al dan niet hebben van betaald werk

	Gemiddelde	Standaardafwijking	N
Geslacht			
vrouw	3,36	2,60	751
man	4,08	3,00	709
Betaald werk			
nee	5,08	3,24	652
ja	2,60	1,79	808

Bron: SCV-survey 2002

Mannen hebben gemiddeld meer vrije tijd dan vrouwen (meer dan 4 uur tegenover iets minder dan 3 uur en half). Mensen met betaald werk hebben begrijpelijkerwijze gemiddeld minder uren vrij op een weekdag dan mensen zonder betaald werk (2 uur en half tegenover meer dan 5 uur gemiddeld). Het is duidelijk dat zowel geslacht als het hebben van betaald werk een effect hebben op de vrije tijd. Maar het effect van het hebben van betaald werk is daarom niet noodzakelijk hetzelfde voor mannen en voor vrouwen en het verschil tussen mannen en vrouwen is niet noodzakelijk hetzelfde voor de mensen zonder betaald werk en voor de mensen met betaald werk. Dat kan blijken uit een meer gedetailleerde tabel, zoals tabel 3 die het onderscheid volgens het al dan niet hebben van betaald werk apart maakt voor mannen en voor vrouwen.

Tabel 3 Aantal uren vrije tijd op een dag in de week volgens de gecombineerde verdeling van geslacht en het al dan niet hebben van betaald werk

Geslacht	Betaald werk	Gemiddelde	Standaard-afwijking	N
vrouw	nee	4,32	2,88	396
	ja	2,28	1,68	355
man	nee	6,25	3,42	256
	ja	2,85	1,82	453

Bron: SCV-survey 2002

Uit tabel 3 kunnen enkele vaststellingen afgeleid worden. Zo blijken de mannen zonder betaald werk over de meeste vrije tijd te beschikken. Vanuit het oogpunt van de interactie zijn er echter twee bevindingen relevant. Ten eerste is het verschil naargelang het al dan niet hebben van betaald werk duidelijk groter bij mannen dan bij vrouwen. Bij mannen bedraagt dit verschil bijna 3 uur en half (6,25 – 2,85 = 3,40); bij vrouwen net 2 uur (4,32 – 2,28 = 2,04). De tweede relevante bevinding is hiervan een logische afgeleide. Het verschil tussen mannen en vrouwen is groter bij mensen zonder betaald werk dan bij mensen met betaald werk. Bij de mensen zonder betaald werk is er een verschil van bijna 2 uur (6,25 – 4,32 = 1,93), bij de werkenden bedraagt dit verschil volgens geslacht een dik half uur (2,85 – 2,28 = 0,57). Er is dus duidelijk sprake van interactie. Zowel het hebben van betaald werk als het geslacht hebben een effect op het volume vrije tijd, maar het effect van betaald werk verschilt volgens geslacht en het effect van geslacht verschilt volgens het al dan niet hebben van betaald werk.

Een **tweede voorbeeld** kijkt naar het al dan niet hebben van betaald werk als afhankelijke variabele. De data komen deze keer uit de SILC-enquête van 2005 en uit het oorspronkelijke bestand werden enkel de 20- tot 64-jarigen geselecteerd. Van die groep blijkt ongeveer 68% betaald werk te hebben (zie tabel 4).

Tabel 4 Al dan niet hebben van betaald werk

	Frequentie	Percentage
Nee	1215	31,6%
Ja	2624	68,4%
N	3839	100,0%

Bron: SILC2005

Ook hier zijn er makkelijk twee kenmerken te vinden die samenhangen met de kans om betaald werk te hebben, zo bvb. scholingsniveau en geslacht. We bekijken eerst de verschillen volgens opleidingsniveau. Voor de eenvoud werd het hoogste diploma van de respondenten opgedeeld in twee categorieën. In de gehanteerde tweedeling hebben laaggeschoolden ten hoogste een diploma lager secundair onderwijs. Hogergeschoolden hebben minstens een diploma hoger secundair onderwijs. Tabel 5 toont grote verschillen. Van de hoger opgeleide respondenten, heeft meer dan 74% betaald werk. Bij de respondenten met ten hoogste een diploma lager secundair onderwijs is dit minder dan 50%. De verschillen volgens geslacht zijn een beetje kleiner, maar toch ook nog opvallend. Van de 20- tot 64-jarige mannen heeft zo'n 78% betaald werk, bij de vrouwen is dit 60%.

Tabel 5 Al dan niet hebben van betaald werk volgens opleidingsniveau en geslacht

	% met betaald werk	N waarop % berekend is
Opleidingsniveau		
laag	49,1%	899
hoger	74,3%	2940
Geslacht		
vrouw	59,8%	1958
man	77,8%	1881

Bron: SILC2005

Om het al dan niet bestaan van een interactie na te gaan, kijken we naar een tabel die het onderscheid volgens opleidingsniveau binnen de twee geslachten maakt. Uit tabel 6 blijkt dat bij de hogergeschoolden het verschil tussen mannen en vrouwen duidelijk beperkter is dan bij de lageropgeleiden. Van de hogeropgeleide vrouwen is ongeveer 68% aan het werk, bij de hogeropgeleide mannen ligt dit 12 procentpunten hoger. Bij de laaggeschoolden bedraagt het verschil tussen mannen en vrouwen daarentegen ruim 31 procentpunten (66,1% – 34,4%). Deze interactie wordt weerspiegeld in de verschillen volgens opleidingsniveau binnen de geslachten. Bij de vrouwen loopt het verschil tussen hoog- en laaggeschoolden op tot bijna 34 procentpunten (68,1% – 34,4%), terwijl datzelfde verschil bij de mannen zo'n 14 procentpunten bedraagt (80,4% – 66,1%).

Tabel 6 Al dan niet hebben van betaald werk volgens de gecombineerde verdeling van opleidingsniveau en geslacht

Geslacht	Opleidingsniveau	% met betaald werk	N waarop % berekend is
vrouw	laag	34,4%	483
	hoger	68,1%	1475
man	laag	66,1%	416
	hoger	80,4%	1465

Bron: SILC2005

Ook in dit tweede voorbeeld is er dus sprake van interactie. Opleidingsniveau en geslacht bepalen beide de kans op het hebben van betaald werk. Maar het verschil tussen mannen en vrouwen is groter voor lageropgeleiden dan voor hogeropgeleiden en het verschil volgens opleidingsniveau is niet hetzelfde bij mannen als bij vrouwen.

Bij deze twee voorbeelden hebben we telkens een dubbele interpretatie weergegeven, waarbij focus- en moderatorvariabele onderling gewisseld werden. Het onderscheid tussen beide is inderdaad conceptueel en van nut bij de inhoudelijke duiding. Maar statistisch zijn de variabelen en de interpretaties – in dit geval – inwisselbaar.

4. Interactie-effecten in een meervoudige lineaire regressie

Deze uitwerking van interactie-effecten in tabellen met verschillende niveaus is interessant en illustratief, maar stuit op beperkingen wanneer het effect van meerdere onafhankelijke variabelen op één afhankelijke variabele onderzocht wordt. Zeker als één of meerdere van die onafhankelijke variabelen metrisch zijn, wordt een tabelweergave moeilijk tot onmogelijk.

Een regressiemodel laat wel toe om het effect van verschillende onafhankelijke variabelen op één afhankelijke variabele gelijktijdig te onderzoeken en kan eveneens interactie-effecten opnemen. Vóór we echter regressiemodellen met interactie-effecten bespreken, kijken we eerst nog eens naar de algemene interpretatie van een regressievergelijking.

4.1 Interpretatie van een meervoudig lineair regressiemodel zonder interactie-effecten

4.1.1 Algemeen model

Een meervoudige lineaire regressievergelijking heeft volgende vorm:

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2 + ... + b_k X_k \qquad (1)$$

Hierin is $\hat{Y}$ de voorspelde waarde van de afhankelijke variabele, b_0 het intercept, X_1 tot X_k zijn de onafhankelijke variabelen en b_1 tot b_k zijn de richtingscoëfficiënten of regressiecoëfficiënten. De interpretatie van de verschillende coëfficiënten volgt uit deze vergelijking. Zo is b_0 de voorspelde waarde voor de afhankelijke variabele als alle onafhankelijke variabelen gelijk zijn aan 0 en b_1 tot b_k geven het effect aan van respectievelijk X_1 tot X_k op de afhankelijke variabele. Maar omdat dit een meervoudige regressievergelijking is (met meerdere onafhankelijke variabelen) is dat het effect van X op Y gecontroleerd voor de effecten van alle andere onafhankelijke variabelen in het model. Letterlijk is b_1 de voorspelde wijziging in Y bij één eenheid wijziging in X_1 en onder controle van de effecten van X_2 tot X_k op Y.

Een voorbeeldje kan deze theoretische uitleg wat verduidelijken. We grijpen hiervoor terug naar de SCV-survey van 2002 en de daarin gemeten vrije tijd op een dag in de week. Als we die variabele in een regressie afhankelijk maken van leeftijd en aantal kinderen ten laste, krijgen we volgende vergelijking:

$$\widehat{\text{vrij\_weekdag}} = 1{,}36 + 0{,}06 \text{ leeftijd} - 0{,}59 \text{ aantal kinderen ten laste} \qquad (2)$$

Uit deze regressievergelijking leren we dus dat het voorspelde volume vrije tijd op een dag in de week voor nuljarigen zonder kinderen ten laste gelijk is aan 1,36 uren. Voor elk jaar dat onze respondent ouder is, voorspellen we 0,06 uren meer vrije tijd. Voor elk kind ten laste gaat er meer dan een half uur vrije tijd af (-0,59). Dit model veronderstelt dat deze verbanden lineair zijn en de voorspellingen gelden onder controle van het effect van respectievelijk aantal kinderen ten laste en leeftijd. Het zijn bijgevolg "netto-effecten".

Bemerk dat de SCV-survey slechts een momentopname is. Eigenlijk kunnen we daarmee moeilijk of niet onderzoeken wat de impact is van ouder worden of van het krijgen van een (bijkomend) kind op het volume vrije tijd. We zien wel dat op het moment van de meting (2002) jongeren en mensen met kinderen ten laste minder vrije tijd hebben dan hun tegenpolen. Een interpretatie in termen van "bij een stijging van de leeftijd met tien jaar stijgt het aantal voorspelde uren vrije tijd met 0,6 uren" is aanlokkelijk, maar strikt genomen hebben we zo'n stijging niet geobserveerd en kunnen we er dus ook moeilijk een uitspraak over doen.

4.1.2 Werken met deviatiescores voor metrische onafhankelijke variabelen

In vergelijking (2) verwijst de voorspelde waarde van het intercept naar nuljarigen. Maar de SCV-survey heeft alleen maar 18- tot 85-jarigen bevraagd. Deze voorspelling van het intercept wordt dus niet ondersteund door data en is eigenlijk ook zinloos. Dat is vaak het geval bij metrische variabelen waarbij nulwaarden onmogelijk zijn of niet geobserveerd werden. Dit probleem kan verholpen worden door te werken met deviatiescores. Een deviatiescore trekt van de oorspronkelijke variabele het gemiddelde – al dan niet afgerond – af. Tabel 7 illustreert de deviatiescore van leeftijd.

Tabel 7 Leeftijd en deviatiescore van leeftijd

	N	Minimum	Maximum	Gemiddelde	Standaard-afwijking
leeftijd	1460	18	85	47,84	17,47
dev_leeftijd	1460	-30	37	-0,16	17,47

Bron: SCV-survey 2002

Van de oorspronkelijke leeftijd hebben we 48 afgetrokken. Als gemiddelde hebben we dus een afgeronde waarde genomen, die ook geobserveerd is in de data. Het exacte gemiddelde zou ook kunnen, maar levert soms een iets minder handige interpretatie op. Als we in onze regressie nu dev_leeftijd zouden opnemen i.p.v. leeftijd, ziet de resulterende vergelijking er zo uit:

$$\widehat{\text{vrij\_weekdag}} = 4{,}09 + 0{,}06\ \text{dev\_leeftijd} - 0{,}59\ \text{aantal kinderen ten laste} \qquad (3)$$

De geschatte effecten van leeftijd en het aantal kinderen ten laste zijn exact dezelfde als in vergelijking (2). Het intercept is echter wel duidelijk veranderd. Dat is nu de voorspelde waarde van het aantal uren vrije tijd als het aantal kinderen ten laste en dev_leeftijd gelijk zijn aan 0, ofwel voor 48-jarige respondenten zonder kinderen ten laste.

4.1.3 Categorische onafhankelijke variabelen

De standaardinterpretatie van de regressiecoëfficiënten gaat eigenlijk uit van metrische onafhankelijke variabelen: "de voorspelde wijziging in Y bij één eenheid wijziging van X". Categorische onafhankelijke variabelen kunnen echter ook opgenomen worden als onafhankelijke variabelen in een meervoudige regressie. Mits hercodering is de interpretatie dan eveneens zinvol en eenvoudig. Verschillende hercoderingen zijn mogelijk (zie McClendon, 2002, 198-229). In combinatie met interactie-effecten is dummycodering echter vaak het eenvoudigst. In deze tekst beperken we ons daar ook toe. Dummycodering vormt een categorische variabele om tot een aantal dummies, variabelen die enkel de waarden 0 of 1 kunnen aannemen. Er is één dummy minder dan het aantal categorieën van de oorspronkelijke variabele. Een dichotome variabele wordt dus omgevormd tot één dummy, een variabele met drie categorieën tot twee dummies, ... Zo kan de variabele geslacht bijvoorbeeld omgevormd worden tot de dummy *vrouw* die waarde 0 aanneemt voor mannen en waarde 1 voor vrouwen. *Betaald werk* is eveneens een dummy die waarde 0 aanneemt bij mensen zonder betaald werk en waarde 1 bij respondenten met betaald werk. Opname van vrouw en betaald werk in regressievergelijking (3) resulteert in het volgende resultaat:

$$\begin{aligned}\widehat{\text{vrij\_weekdag}} = {} & 5{,}48 + 0{,}04\ \text{dev\_leeftijd} - 0{,}38\ \text{aantal kinderen ten laste} \\ & - 1{,}03\ \text{vrouw} - 1{,}79\ \text{betaald werk}\end{aligned} \qquad (4)$$

De regressiecoëfficiënt bij vrouw is gelijk aan -1,03. Een wijziging van één eenheid bij deze dummy komt overeen met de stap van man (= 0) naar vrouw (= 1). We kunnen die regressieparameter dus ook als volgt interpreteren: bij controle voor de effecten van leeftijd, aantal kinderen ten laste en het al dan niet hebben van betaald werk voorspellen we ongeveer één uur minder vrije tijd voor vrouwen dan voor mannen. Analoog hieraan voorspellen we voor mensen met betaald werk één uur en drie kwartier minder vrije tijd dan voor mensen zonder betaald werk.

Bemerk dat ook de andere regressieparameters wijzigingen hebben ondergaan. Dat is een logisch gevolg van de opname van die extra variabelen. De voorspelling van het intercept duidt nu bijvoorbeeld op 48-jarige ***mannen zonder kinderen ten laste*** en ***zonder betaald werk***. Het is de voorspelling die geldt als alle onafhankelijke variabelen gelijk zijn aan 0, dus ook de dummies vrouw en betaald werk. De effecten van leeftijd en aantal kinderen ten laste zijn iets kleiner geworden in vergelijking met model (3), wat verklaard kan worden door de samenhang van die variabelen met het al dan niet hebben van betaald werk.

4.1.4 Significantietesten

Meervoudige regressie kan gebruikt worden als louter beschrijvende techniek, maar de resultaten zijn vaak pas echt interessant als ze veralgemeend kunnen worden naar een populatie. Voor regressieanalyse gebeurt deze statistische inferentie doorgaans met behulp van twee significantietesten: een test op basis van (het verschil in) de verklaarde variantie en een test op basis van de standaardfout van een parameter. We bespreken hier kort de werkwijze van de laatste test, ook wel de Wald-test genoemd. Van de eerste, de F-test, wordt enkel het principe beschreven. Meer gedetailleerde informatie kan gevonden worden in McClendon (2002, 133 – 197) en Welkenhuysen-Gybels en Loosveldt (2002, 112-118 en 173-183).

De **standaardfout** van een parameter is de standaarddeviatie van de steekproevenverdeling van die parameter. Deze steekproevenverdeling toont welke waarden de betreffende parameter aanneemt in alle mogelijke steekproeven van een bepaalde omvang uit de populatie.

Om dit te verduidelijken vertrekken we van een regressieparameter voor een volledige populatie, bijvoorbeeld b_1 die het effect van X_1 op Y in de populatie weergeeft[2]. Als we uit die populatie een steekproef van een bepaalde omvang (n) hebben getrokken en we berekenen nadien dezelfde b_1 voor deze steekproef, zullen beide parameters waarschijnlijk niet 100% identiek zijn, maar (hopelijk) wel vergelijkbaar. Als we nu uit die populatie alle mogelijke steekproeven met een omvang gelijk aan n trekken en telkens die b_1 berekenen, dan is de verzameling van alle mogelijke waarden die die b_1 aanneemt in de verschillende steekproeven de steekproevenverdeling. De spreiding van deze verdeling kan geduid worden met een standaarddeviatie en deze standaarddeviatie wordt dus de standaardfout genoemd, in dit geval de standaardfout van b_1.

Natuurlijk is dit theorie, omdat de populatiewaarde (vrijwel) nooit gekend is en omdat het onhaalbaar is om alle mogelijke steekproeven van een bepaalde omvang te trekken. Maar omdat een aantal eigenschappen van de steekproevenverdeling wiskundig afgeleid kan worden en omdat de standaardfout geschat kan worden op basis van een steekproef, biedt deze steekproeventheorie wel de bouwstenen om veralgemeningen naar een populatie mogelijk te maken of om met andere woorden significantietesten uit te voeren. Deze theorie maakt het immers mogelijk om bijvoorbeeld te bepalen welk aandeel van alle mogelijke parameterschattingen groter is dan een bepaalde waarde, gegeven de waarde van die parameter in de populatie. Deze populatiewaarde kan dan een onderdeel zijn van de hypothese.

De meeste significantietesten willen bepalen of een effect (bvb. b_1) al dan niet aanwezig is in de populatie. De (te verwerpen) hypothese stelt in dat geval dat de regressiecoëfficiënt in de populatie gelijk is aan 0. Op basis van de steekproeventheorie kunnen we dan stellen dat, als b_1 in de populatie inderdaad gelijk is aan 0, we in alle mogelijke steekproeven van die bepaalde omvang in 95% van de gevallen een b_1-schatting zullen bekomen die zich bevindt in het interval [-1,96*standaardfout; +1,96*standaardfout][3]. In de praktijk deelt men daarom doorgaans de parameter door zijn geschatte standaardfout. Als het bekomen resultaat groter is dan 2 of kleiner dan -2, is het besluit dat het effect

[2] De standaardnotatie voorziet voor populatieparameters Griekse tekens i.p.v. Latijnse (en dus β_1). Voor eenvoud behouden we hier b_1.

[3] Dit geldt voor steekproeven die voldoende groot zijn (bvb. n > 200). Bij kleinere steekproeven is het interval wat groter. Omdat de meeste software p-waarden meegeeft, is het niet nodig om zelf het interval of de kans te berekenen. Voor de eenvoud ronden we deze waarde in de rest van de tekst af tot 2. Uiteindelijk is die 1,96 en de daarbijhorende 95% immers ook een conventie.

"significant is". De volledige interpretatie luidt in dat geval: "Als het betreffende effect in de populatie inderdaad gelijk is aan 0, hebben we in een steekproef van deze omvang minder dan 5% kans om louter als gevolg van toeval een effect te bekomen dat zo groot is als, of groter dan het effect dat nu geschat werd. Deze kans vinden we te klein en dus verwerpen we de hypothese dat het effect in de populatie inderdaad gelijk is aan 0."

Natuurlijk zijn andere vooropgestelde significantieniveaus mogelijk dan de vaak gebruikte 0,05. Dikwijls wordt ook de kans of de zgn. p-waarde meegegeven.

We verduidelijken deze significantietest voor het voorbeeld van hierboven. Tabel 8 geeft dezelfde resultaten als vergelijking (4), maar in tabelvorm en met inbegrip van standaardfouten en p-waarden.

Tabel 8 Resultaten van de regressie met aantal uren vrij in de week als afhankelijke variabele en leeftijd (deviatiescore), aantal kinderen ten laste, geslacht en het al dan niet hebben van betaald werk als onafhankelijke variabelen

	b	Standaardfout	p-waarde
intercept	5,48	0,13	0,000
dev_leeftijd	0,04	0,00	0,000
aantal kinderen ten laste	-0,38	0,06	0,000
vrouw	-1,03	0,13	0,000
betaald werk	-1,79	0,15	0,000

Bron: SCV-survey 2002

Uit tabel 8 blijkt dat de standaardfouten telkens vele malen kleiner zijn dan de geschatte parameters. De p-waarden zijn bijgevolg zeer klein. Op basis van deze tabel komen we tot de volgende letterlijke duiding van de significantietest van bvb. *vrouw*: "Stel dat er in de populatie (alle Nederlandstalige Belgen wonend in het Vlaamse Gewest of het Brusselse Hoofdstedelijke Gewest) geen verschil is tussen vrouwen en mannen in het volume beschikbare vrije tijd (en dit onder controle van de andere variabelen in het model), dan hebben we minder dan 1 kans op 100 om in een steekproef van 1460 personen door toeval een effect te vinden dat in absolute waarde gelijk is aan of groter dan 1,03. Deze kans is zeer klein. We verwerpen dus de hypothese dat er in de populatie geen verschil is tussen mannen en vrouwen en besluiten m.a.w. dat er wel degelijk een verschil is volgens geslacht"[4]. Ook alle andere effecten zijn duidelijk "significant".

Volgens dezelfde principes is een test mogelijk op basis van de (bijkomende) **verklaarde variantie**. Zo'n test vergelijkt steeds 2 modellen, waarbij het tweede model dezelfde effecten bevat als het eerste en daarbovenop nog één of meerdere effecten extra. De bijkomende onafhankelijke variabelen zorgen steeds voor een stukje bijkomende verklaring van de afhankelijke variabele. De verklaarde variantie of de R^2 zal stijgen. De test gaat dan na of deze toename in de verklaarde variantie significant is. Letterlijk: stel dat dit bijkomende effect in de populatie gelijk is aan 0, wat is dan de kans om louter als gevolg van toeval een dergelijke toename in de verklaarde variantie te bekomen. Als die kans te klein is, besluiten we dat we de hypothese moeten verwerpen en dat het effect in de populatie niet gelijk is aan 0. Hoe dit technisch in zijn werk gaat, wordt beschreven in de hoger vermelde publicaties.

Een voordeel van deze test op basis van de verklaarde variantie is dat hij gebruikt kan worden voor verschillende effecten tegelijkertijd. Dit is o.a. interessant bij gehercodeerde categorische onafhankelijke variabelen. Stel dat iemand burgerlijke staat als onafhankelijke variabele met vijf categorieën (ongehuwd, samenwonend, gehuwd, verweduwd, gescheiden) wil opnemen in een regressie. Na hercodering in dummies neemt het model vier variabelen op. Met deze significantietest kan dan nagegaan worden of burgerlijke staat "als geheel" een significante impact heeft op de afhankelijke variabele.

In de volgende paragraaf zullen we ons regressiemodel voort uitbouwen met interactie-effecten. We vertrekken daarvoor steeds opnieuw van model (4) of tabel 8. Puur inhoudelijk zou een verdere modellering van het effect van leeftijd nochtans correcter zijn. Zulke bijkomende modellering met een kwadratische term zou aantonen dat het leeftijdseffect immers niet lineair maar curvilineair is. Het zijn niet de jongste leeftijdsgroepen die over de minste vrije tijd beschikken, maar de middengroepen.

[4] Bemerk dat de geschatte p-waarde nog vele malen kleiner is dan 0,001. De tabel geeft slechts 3 cijfers na de comma, de exacte p-waarde is gelijk aan 0,0000000000000011.

Ouderen én jongeren hebben meer vrije tijd dan de middencategorie (30- tot 45-jarigen). Een modellering van niet-lineaire effecten valt echter buiten het bereik van dit rapport. Meer informatie daarover kan bijvoorbeeld gevonden worden bij McClendon (2002, 230-270). Ook de mogelijke alternatieve aanpak waarbij de leeftijdsvariabele gecategoriseerd wordt, laten we hier buiten beschouwing.

4.2 Een meervoudige regressie met interactie-effecten

4.2.1 Werken met producttermen

Het schatten van een interactie-effect in een regressiemodel gebeurt door de opname van een productterm van de inter-agerende variabelen in het model. Als er twee onafhankelijke variabelen zijn, ziet het model er zo uit:

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2 + b_3 X_1 X_2 \qquad (5)$$

Hoewel verschillende softwaretoepassingen menugestuurd mogelijkheden aanbieden om interactie-effecten op te nemen in regressiemodellen, is het vaak niet onverstandig om zelf de productterm te berekenen en op te nemen in het model. Zo behoud je de volledige controle en zie je ook duidelijk wat je doet. Tabellen 9 tot 11 tonen de berekening van zo'n productterm voor de twee dummyvariabelen: betaald werk en geslacht.

Tabel 9 Al dan niet hebben van betaald werk

	Frequentie	Percentage
0 (= nee)	652	44,7%
1 (= ja)	808	55,3%
N	1460	100,0%

Bron: SCV-survey 2002

Tabel 10 Geslacht (dummy vrouw)

	Frequentie	Percentage
0 (= man)	709	48,6%
1 (= vrouw)	751	51,4%
N	1460	100,0%

Bron: SCV-survey 2002

Tabel 11 Product van de variabelen vrouw en betaald werk

	Frequentie	Percentage
0 (= man, of vrouw zonder betaald werk)	1105	75,7%
1 (= vrouw met betaald werk)	355	24,3%
N	1460	100,0%

Bron: SCV-survey 2002

De interpretatie van tabellen 9 en 10 is evident. Zoals uit tabel 11 blijkt, is het product van twee 0/1-variabelen natuurlijk ook een 0/1-variabele. In die laatste tabel duidt de waarde 0 aan dat de respondent ofwel een man is, ofwel een vrouw zonder betaald werk. Daarnaast zijn er 335 vrouwen met betaald werk in het bestand, zij kregen code 1.

Deze eenvoudige productberekening werkt op exact dezelfde manier voor twee metrische variabelen en ook voor een metrische en een categorische variabele. Voor de volledigheid tonen tabel 12 en 13 een voorbeeld van deze laatste situatie. Het betreft de productterm van geslacht en aantal kinderen ten laste.

Tabel 12 Aantal kinderen ten laste

	Frequentie	Percentage
0	976	66,8%
1	187	12,8%
2	202	13,8%
3	67	4,6%
4	15	1,0%
5	8	0,5%
6	2	0,1%
7	1	0,1%
8	2	0,1%
N	1460	100,0%

Bron: SCV-survey 2002

Tabel 13 Product van de variabelen vrouw en aantal kinderen ten laste

	Frequentie	Percentage
0	1191	81,6%
1	108	7,4%
2	105	7,2%
3	41	2,8%
4	9	0,6%
5	3	0,2%
7	1	0,1%
8	2	0,1%
N	1460	100,0%

Bron: SCV-survey 2002

De verdeling van de variabele vrouw werd al in tabel 10 weergegeven. Tabel 12 toont eenvoudigweg de verdeling van het aantal kinderen ten laste. Ongeveer 2/3 van de respondenten in het bestand heeft geen kinderen ten laste. De betekenis van de waarden bij de productterm in tabel 13 vloeit hier logisch uit voort. De groep met waarde 0 is samengesteld uit twee subgroepen: (1) mannen, ongeacht het aantal kinderen ten laste en (2) vrouwen zonder kinderen ten laste. De daaropvolgende waarden 1 t.e.m. 8 gelden alleen voor vrouwen. Zo zijn er in het bestand 108 vrouwen met 1 kind ten laste. Het is door de opname van deze producttermen dat we een interactie-effect van de twee respectievelijke variabelen kunnen schatten.

4.2.2 Theoretische interpretatie van een interactieterm

De interpretatie van een interactie-effect kan afgeleid worden uit vergelijking (5). Voor de eenvoud hernemen we die hier even.

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2 + b_3 X_1 X_2 \qquad (5)$$

In deze vergelijking worden b_1 en b_2 wel eens de hoofdeffecten ("main effects") genoemd en b_3 het interactie-effect. Omdat de benaming hoofdeffect ingeburgerd is, wordt ze ook in deze tekst gebruikt. Ze is echter enigszins misleidend omdat het over een conditioneel of specifiek effect gaat (zie hieronder)!

De interpretatie van het intercept blijft in deze vergelijking dezelfde: de voorspelde waarde voor de afhankelijke variabele als alle onafhankelijke variabelen gelijk zijn aan 0. Voor de andere effecten wijzigt er wel één en ander. Uit vergelijking (5) volgt dat de voorspelde wijziging in Y bij één eenheid wijziging in X_1 gelijk is aan:

$$b_1 + b_3 X_2 \qquad (6)$$

De voorspelde wijziging in Y bij één eenheid wijziging in X_2 is gelijk aan:

$$b_2 + b_3 X_1 \qquad (7)$$

Uit (6) en (7) blijkt dat b_1 nog altijd het effect is van X_1 op Y, maar enkel in bepaalde situaties, namelijk als X_2 gelijk is aan 0. Het is dus inderdaad een conditioneel effect. Het effect van X_1 op Y bij andere waarden van X_2 kan wel makkelijk berekend worden met behulp van vergelijking (6). Equivalent hieraan geldt dat b_2 het effect is van X_2 op Y als X_1 gelijk is aan 0. Het is duidelijk dat het effect van X_1 op Y afhankelijk is van X_2 en omgekeerd. Dit is in overeenstemming met de definitie van een interactie-effect die we gaven in paragraaf 2.

Je kan b_3 zelf ook interpreteren. Zo'n interpretatie is wel behoorlijk ingewikkeld. Letterlijk is het het verschil in het effect van X_1 op Y bij één eenheidswijziging van X_2, waarbij dat effect zelf de voorspelde wijziging in Y is bij één eenheidswijziging in X_1. Ook hier geldt natuurlijk de symmetrie; b_3 is eveneens

het verschil in het effect van X_2 op Y bij één eenheidswijziging van X_1. Enkele voorbeelden moeten deze interpretatie verder verduidelijken.

4.2.3 Interpretatie van een interactie tussen twee categorische variabelen

We vertrekken opnieuw van model (4) in tabel 8. We schatten in dat model bijkomend een interactie tussen geslacht en betaald werk. Het resultaat van die regressie wordt weergegeven in tabel 14.

Tabel 14 Resultaten van de regressie met aantal uren vrij in de week als afhankelijke variabele en leeftijd (deviatiescore), aantal kinderen ten laste, geslacht, betaald werk en de interactie van die twee laatste als onafhankelijke variabelen

	b	Standaardfout	p-waarde
intercept	5,91	0,15	0,000
dev_leeftijd	0,04	0,00	0,000
aantal kinderen ten laste	-0,37	0,06	0,000
vrouw	-1,74	0,19	0,000
betaald werk	-2,48	0,20	0,000
vrouw*betaald werk	1,26	0,25	0,000

Bron: SCV-survey 2002

Hoewel de waarde van het intercept veranderd is, geldt dat niet voor de interpretatie ervan. Het is nog steeds het voorspelde aantal uren vrij op een dag in de week voor 48-jarige mannen zonder kinderen ten laste en zonder betaald werk. Ook de effecten van leeftijd en aantal kinderen ten laste interpreteren we op dezelfde manier als voorheen.

Deze volkomen overeenstemming geldt echter niet voor de effecten van geslacht en betaald werk. In vergelijking met (4) is het effect van vrouw in tabel 14 veel groter. Dat komt omdat dit "hoofdeffect" in deze laatste analyse geldt voor één bepaalde groep, namelijk de mensen zonder betaald werk (= waarde 0 op de andere dummy in het interactie-effect). Dus ***bij de mensen zonder betaald werk*** voorspellen we 1 uur en 3 kwartier minder vrije tijd voor vrouwen dan voor mannen (en dit onder controle van de effecten van leeftijd en aantal kinderen ten laste). Analoog hieraan geldt het hoofdeffect van betaald werk enkel voor ***mannen***. Bij mannen met betaald werk voorspellen we 2 uur en half minder vrije tijd dan bij mannen zonder betaald werk.

Het effect van geslacht bij mensen met betaald werk en het effect van betaald werk bij vrouwen kan berekend worden met (6) en (7). Voor vrouwen met betaald werk voorspellen we een half uur vrij minder dan voor mannen met betaald werk (–1,74 + 1,26 = –0,48). Voor vrouwen met betaald werk voorspellen we 1 uur en een kwartier minder vrij dan voor vrouwen zonder betaald werk (–2,48 + 1,26 = –1,22).

Als we het interactie-effect zelf willen verklaren, kunnen we zeggen dat het effect van geslacht bij mensen met betaald werk 1,26 hoger uitvalt dan bij mensen zonder betaald werk, of ook dat het effect van betaald werk bij vrouwen 1,26 hoger ligt dan bij mannen. Omdat de respectievelijke hoofdeffecten negatief zijn, betekent "hoger" in beide gevallen wel dat de effecten kleiner worden.

Deze eerder ingewikkelde interpretatie laat zich gelukkig toch ook makkelijk vertalen naar enkele bevattelijke conclusies. Zo kan gesteld worden dat het verschil tussen mannen en vrouwen in het volume beschikbare vrije tijd groter is bij mensen zonder betaald werk dan bij mensen met betaald werk. Analoog hieraan betekent het hebben van betaald werk voor mannen een grotere aanslag op hun vrije tijd dan voor vrouwen.

Ook voor interactie-effecten is een significantietest op basis van de standaardfout mogelijk. Zoals de tabel toont is de p-waarde hier zo klein dat we inderdaad van een significant interactie-effect mogen spreken. Een volledige interpretatie van deze significantie, die betaald werk als focusvariabele beschouwt en geslacht als moderator variabele, zou als volgt luiden: als we veronderstellen dat het effect van betaald werk op het beschikbare volume vrije tijd in de populatie gelijk is voor mannen en vrouwen, dan hebben we in een steekproef van 1460 eenheden een heel kleine kans om een interactie-effect te bekomen dat groter dan of gelijk is aan 1,26. We verwerpen bijgevolg de hypothese dat het al dan niet hebben van betaald werk bij mannen en vrouwen hetzelfde effect heeft op het aantal uren vrij op een dag in de week.

Bij het bestuderen van interactie-effecten wordt er ook vaak gebruik gemaakt van de significantietest op basis van de bijkomende verklaarde variantie. Zo'n test zou dan bijvoorbeeld tegelijkertijd de twee hoofdeffecten en het interactie-effect in overweging kunnen nemen en aldus nagaan of geslacht én betaald werk een significante bijdrage hebben aan de verklaarde variantie. Een test op basis van de verklaarde variantie, die enkel de bijdrage van het interactie-effect nagaat, is equivalent aan de test op basis van de standaardfout.

Om dit voorbeeld te besluiten tonen we nog dat het mogelijk is om met andere dummies maar zonder interactie-effect een model te bekomen dat volledig equivalent is aan het model in tabel 14. Met het oog daarop categoriseren we de vier onderscheiden groepen (mannen/vrouwen met/zonder betaald werk) in drie dummies zoals in tabel 15. Uit de tabel blijkt dat we mannen zonder betaald werk als referentiecategorie kozen. Zij krijgen telkens waarde 0 op de dummies voor de drie andere categorieën.

Tabel 15 Dummycodering van 4 verschillende groepen

OORSPRONKELIJKE VARIABELEN		NIEUWE DUMMIES		
geslacht	**betaald werk**	**man met betaald werk**	**vrouw zonder betaald werk**	**vrouw met betaald werk**
man	nee	0	0	0
	ja	1	0	0
vrouw	nee	0	1	0
	ja	0	0	1

Deze drie dummies nemen we op in de regressievergelijking, naast de variabelen leeftijd en aantal kinderen ten laste. Tabel 16 toont de resultaten van die regressieanalyse.

Tabel 16 Resultaten van de regressie met aantal uren vrij in de week als afhankelijke variabele en leeftijd (deviatiescore), aantal kinderen ten laste en drie dummies voor de gecombineerde verdeling van geslacht en het al dan niet hebben van betaald werk als onafhankelijke variabelen

	b	**Standaardfout**	**p-waarde**
intercept	5,91	0,15	0,000
dev_leeftijd	0,04	0,00	0,000
aantal kinderen ten laste	-0,37	0,06	0,000
man met betaald werk	-2,48	0,20	0,000
vrouw zonder betaald werk	-1,74	0,19	0,000
vrouw met betaald werk	-2,96	0,22	0,000

Bron: SCV-survey 2002

Eerst en vooral valt op dat zowel het intercept als de effecten van leeftijd en aantal kinderen ten laste identiek zijn aan deze in tabel 14. De andere drie effecten vergelijken de desbetreffende categorie met de referentiecategorie (mannen zonder betaald werk). Mannen met betaald werk hebben 2 uur en half minder vrije tijd dan mannen zonder betaald werk; vrouwen zonder betaald werk hebben 1 uur en drie kwartier minder en vrouwen met betaald werk bijna 3 uur minder. Hoewel het andere dummies zijn, vonden we deze eerste twee effecten hierboven ook terug. In tabel 14 waren de dummies algemener (geslacht en betaald werk voor de hele populatie), maar door de opname van het interactie-effect, werden de gemeten hoofdeffecten specifiek. In tabel 16 hebben we ineens de specifieke dummies opgenomen in het model. Zo kunnen we ook rechtstreeks de laatste categorie (vrouwen met betaald werk) vergelijken met de referentiecategorie.

Ook in tabel 16 zijn alle effecten significant. Hoewel het model equivalent is aan het model in tabel 14 is de betekenis van deze test wel niet exact dezelfde. De p-waarde van het laatste effect geeft bijvoorbeeld aan dat het voorspelde verschil tussen mannen zonder betaald werk en vrouwen met betaald werk significant is. In tabel 14 is het verschil tussen deze twee categorieën niet rechtstreeks af te lezen en bijgevolg geeft tabel 14 ook geen test voor dat verschil.

De modellen van tabel 14 en tabel 16 zijn 100% equivalent, bvb. ook voor de hier niet behandelde verklaarde variantie (bvb. R^2). In dit geval is een keuze tussen beide eerder een kwestie van smaak. Het voordeel van de werkwijze met de andere dummies is dat er meer verschillen tussen de onderscheiden groepen direct af te lezen zijn uit de resultaten. Langs de andere kant is het bij een verdere uitbouw van het model vaak handiger en logischer om te vertrekken van het eerste, meer algemene model. In dat model kan je immers probleemloos bijkomende interacties opnemen bvb. tussen geslacht en aantal kinderen ten laste. In het laatste model is dat theoretisch ook mogelijk, maar het is minder logisch en de interpretatie zal nog ingewikkelder worden. Zo'n alternatieve modellering met andere dummies is ook alleen mogelijk als de onafhankelijke variabelen betrokken in de interactie categorisch zijn.

4.2.4 Interpretatie van een interactie tussen een categorische en een metrische variabele

In een volgend voorbeeld onderzoeken we de interactie tussen een categorische variabele (geslacht) en een metrische variabele (aantal kinderen ten laste). Om deze illustratie relatief eenvoudig te houden, vertrekken we opnieuw van model (4) (tabel 8). In dat model nemen we nu dus de productterm van vrouw en aantal kinderen ten laste op. Tabel 17 toont de resulterende regressievergelijking en maakt duidelijk dat de interactieterm een (rand)significant effect heeft. De p-waarde is net kleiner dan 0,05.

Tabel 17 Resultaten van de regressie met aantal uren vrij in de week als afhankelijke variabele en leeftijd (deviatiescore), betaald werk, geslacht, aantal kinderen ten laste en de interactie van die laatste twee als onafhankelijke variabelen

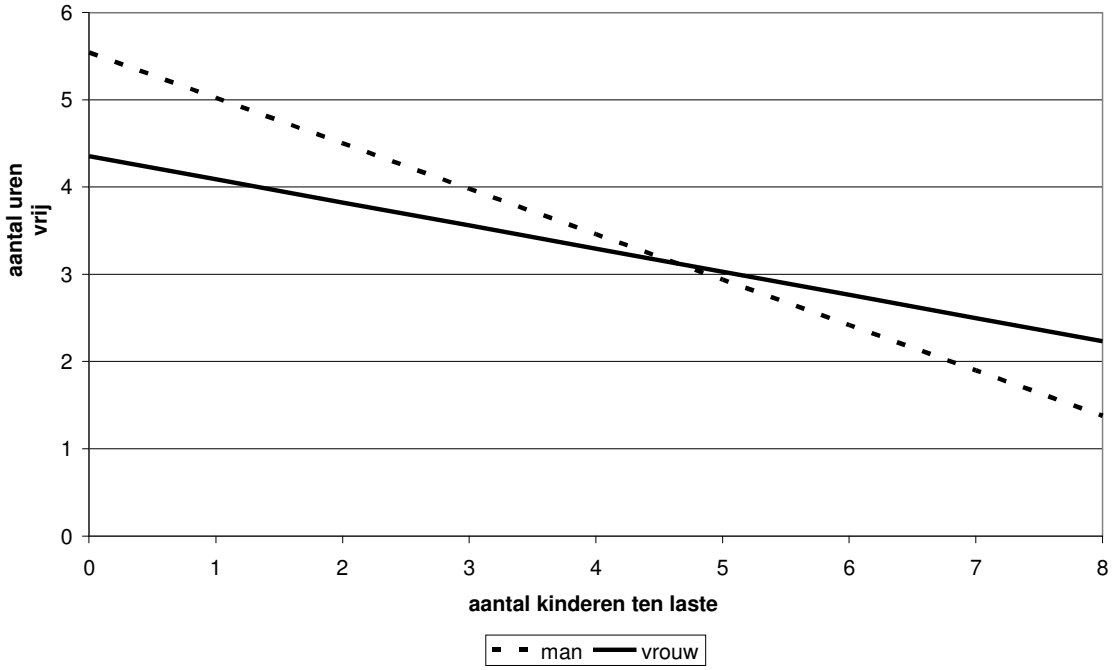
	b	Standaardfout	p-waarde
intercept	5,54	0,13	0,000
dev_leeftijd	0,04	0,00	0,000
aantal kinderen ten laste	-0,52	0,09	0,000
vrouw	-1,19	0,15	0,000
betaald werk	-1,77	0,15	0,000
vrouw*aantal kinderen ten laste	0,26	0,12	0,028

Bron: SCV-survey 2002

De interpretatie van het intercept en van de effecten van leeftijd en betaald werk blijft identiek aan deze in tabel 8. Het effect van vrouw geldt nu voor respondenten zonder kinderen ten laste. Bij die respondenten voorspellen we voor vrouwen ongeveer 1 uur en 10 minuten minder vrij dan voor mannen (–1,19). Het effect van aantal kinderen ten laste geldt nu alleen voor mannen. Bij de mannen voorspellen we dat er voor elk kind ten laste iets meer dan een half uur vrije tijd minder is.

Het effect van het aantal kinderen ten laste bij vrouwen is gelijk aan –0,26 (= –0,52 + 0,26). Het aantal kinderen zorgt bij vrouwen dus minder voor een inperking van de vrije tijd dan bij mannen. Het verschil tussen mannen en vrouwen wordt bijgevolg ook kleiner naarmate er meer kinderen zijn. Bij 1 kind voorspellen we nog een verschil van 0,95 uur (–1,19 + 0,26), bij 2 kinderen wordt dat 0,69 uur... Het resultaat is dat bij een groot aantal kinderen het voorspelde verschil tussen mannen en vrouwen zelfs omgekeerd is. In dat geval zouden mannen minder vrije tijd hebben dan vrouwen. Dat wordt ook aangetoond in grafiek 2 die het interactie-effect grafisch weergeeft.

Grafiek 2 Voorspeld aantal uren vrij op een dag in de week volgens aantal kinderen ten laste en volgens geslacht bij 48-jarigen zonder betaald werk



Bron: SCV-survey 2002

Grafiek 2 toont het voorspelde aantal uren vrij volgens geslacht en aantal kinderen. Zulke grafieken zijn wel specifiek, zij veronderstellen ook een waarde op de andere onafhankelijke variabelen in het model. In dit geval geldt die grafiek voor 48-jarigen zonder betaald werk. Voor andere groepen zullen de voorspellingslijnen hoger of lager liggen, maar het getoonde interactie-effect zal wel hetzelfde zijn. Voor mannen is de voorspellingsrechte dus gebaseerd op de vergelijking 5,54 – 0,52 *aantal kinderen ten laste* en voor vrouwen op de vergelijking 4,35 – 0,26 *aantal kinderen ten laste*.

In deze grafiek lopen de voorspellingsrechten voor mannen en vrouwen niet parallel. Dat is eigen aan grafieken die een interactie-effect weergeven. Het effect van de X-variabele op de betreffende as is niet constant. Zonder interactie-effect zouden beide lijnen evenwijdig lopen. Zo maakt de grafiek inderdaad duidelijk dat het volume vrije tijd bij mannen sneller daalt naarmate zij meer kinderen ten laste hebben dan bij vrouwen. Het verschil tussen mannen en vrouwen is dan ook het grootst bij de mensen zonder kinderen. Bij (zeer) veel kinderen ten laste voorspelt het model dat het geslachtsverschil omgekeerd wordt.

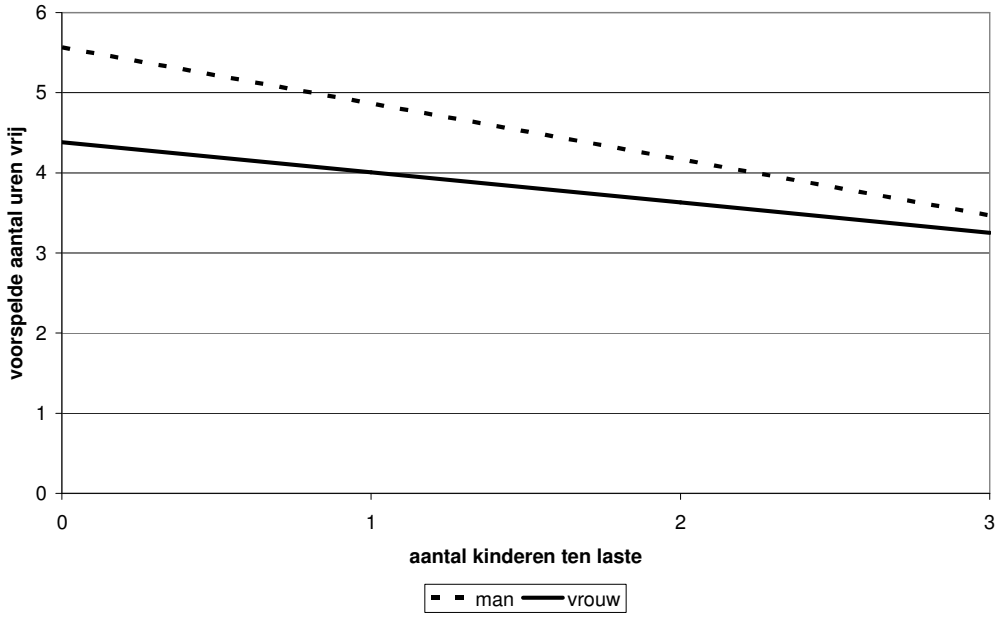
Bij deze laatste voorspelling dient echter een belangrijke kanttekening geplaatst te worden. Voorspellingscurves zijn vaak minder betrouwbaar naar de uiteinden toe. In dit geval komt daar bij dat er in het volledige bestand slechts 28 personen zijn met 4 of meer kinderen ten laste. Uiteindelijk zijn er dus weinig data om deze voorspelling te ondersteunen. Bij de interpretatie is het daarom beter om te focussen op het linkerdeel van de grafiek. Dat dat deel behoorlijk betrouwbaar is, kan afgeleid worden uit een model dat alleen de respondenten in ogenschouw neemt met drie kinderen ten laste of minder. De regressievergelijking van dat model voor een beperktere dataset is vergelijkbaar, net als het geschatte interactie-effect (zie tabel 18 en grafiek 3).

Tabel 18 Resultaten van de regressie met aantal uren vrij in de week als afhankelijke variabele en leeftijd (deviatiescore), betaald werk, geslacht, aantal kinderen ten laste en de interactie van die laatste twee als onafhankelijke variabelen voor alle respondenten met maximum 3 kinderen ten laste

	b	Standaardfout	p-waarde
intercept	5,57	0,13	0,000
dev_leeftijd	0,04	0,00	0,000
aantal kinderen ten laste	-0,70	0,10	0,000
vrouw	-1,19	0,15	0,000
betaald werk	-1,70	0,15	0,000
vrouw*aantal kinderen ten laste	0,32	0,14	0,022

Bron: SCV-survey 2002

Grafiek 3 Voorspeld aantal uren vrij op een dag in de week volgens aantal kinderen ten laste en volgens geslacht bij 48-jarigen zonder betaald werk – analyse gebaseerd op uitsluitend de respondenten met 3 kinderen ten laste of minder



Bron: SCV-survey 2002

Een mogelijke inhoudelijke interpretatie van het interactie-effect dat we in deze paragraaf vinden, is dat in grotere gezinnen het volume vrije tijd evenrediger verdeeld is dan in kleinere gezinnen. Maar we moeten natuurlijk opletten met zo'n interpretatie omdat we geen huishoudens hebben bevraagd maar individuen.

4.2.5 Interpretatie van een interactie tussen twee metrische variabelen

In een laatste voorbeeld bespreken we een interactie tussen twee metrische variabelen. We vertrekken opnieuw van model (4) maar nemen nu in dat model bijkomend de interactie tussen leeftijd en aantal kinderen ten laste op. Tabel 19 toont de resulterende regressievergelijking.

Tabel 19 Resultaten van de regressie met aantal uren vrij in de week als afhankelijke variabele en betaald werk, geslacht, leeftijd (deviatiescore), aantal kinderen ten laste en de interactie van die laatste twee als onafhankelijke variabelen

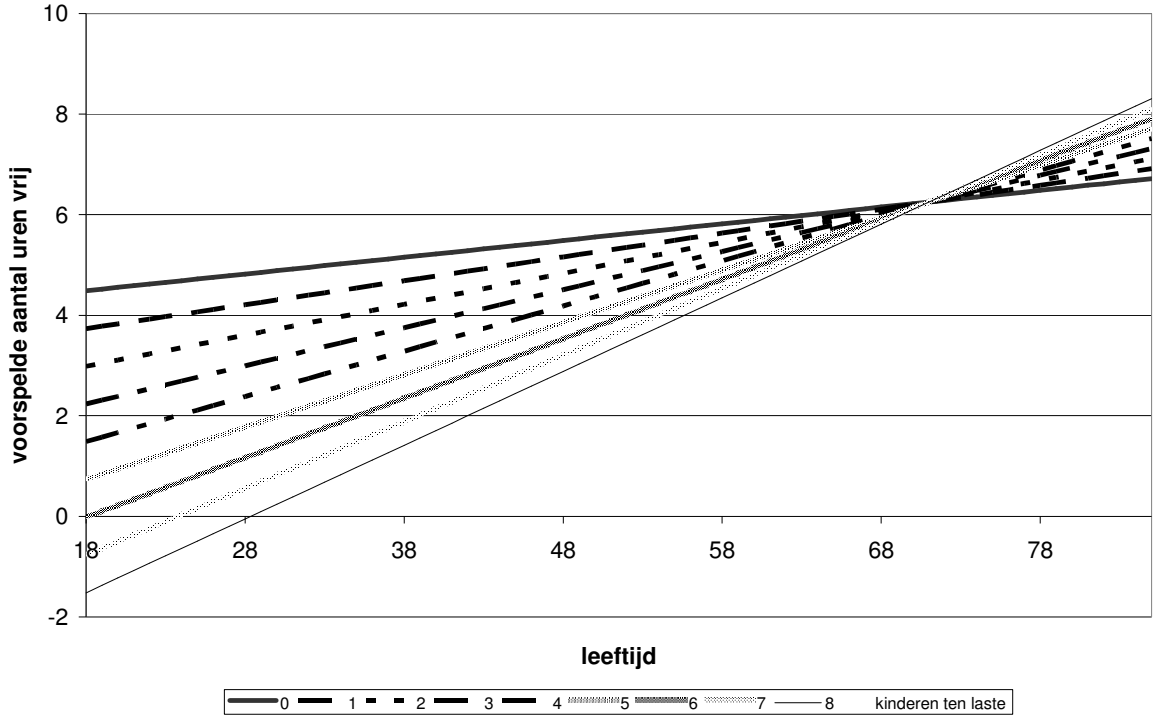
	b	Standaardfout	p-waarde
intercept	5,48	0,13	0,000
dev_leeftijd	0,03	0,00	0,000
aantal kinderen ten laste	-0,33	0,06	0,000
vrouw	-1,01	0,13	0,000
betaald werk	-1,79	0,15	0,000
dev_leeftijd*aantal kinderen ten laste	0,01	0,01	0,008

Bron: SCV-survey 2002

De interpretatie van het intercept en van de effecten van vrouw en betaald werk blijft ongewijzigd. Het hoofdeffect voor leeftijd geldt nu voor respondenten zonder kinderen ten laste. Het hoofdeffect van aantal kinderen ten laste geldt bij gemiddelde leeftijd (48 jaar, als dev_leeftijd gelijk is aan 0). Bijkomend is er een positieve interactie van beide variabelen. Omdat het hoofdeffect van leeftijd positief is, betekent dit dat het leeftijdseffect sterker is bij meer kinderen ten laste. Het hoofdeffect van aantal kinderen ten laste is echter negatief. Hier is de conclusie dus dat het effect van aantal kinderen ten laste kleiner is bij oudere respondenten dan bij jongere respondenten.

Dit interactie-effect wordt ook grafisch weergegeven in grafiek 4. Ook hier is het feit dat de lijnen niet parallel lopen een illustratie van het interactie-effect. De lijn die in de grafiek links bovenaan ligt, geldt voor de respondenten zonder kinderen ten laste. Deze voorspellingslijn is vlakker dan de onderliggende lijnen, die gelden voor respondenten met 1, 2, 3,... kinderen ten laste. De hellingsgraad toont de sterkte van het effect dat dus het kleinst is voor de respondenten zonder kinderen ten laste. De afstand tussen de lijnen geeft een indicatie voor de verschillen volgens aantal kinderen ten laste. Helemaal links is die afstand duidelijk het grootst. Het effect van het aantal kinderen ten laste is dus het grootst bij de jongste respondenten.

Grafiek 4 Voorspeld aantal uren vrij op een dag in de week volgens leeftijd (X-as) en volgens aantal kinderen ten laste (verschillende rechten – zie legende onderaan) bij mannen zonder betaald werk



Bron: SCV-survey 2002

Ook bij deze grafiek hoort een relativerende kanttekening thuis. Er zijn natuurlijk geen 18-jarige respondenten met veel kinderen ten laste, en er zijn ook maar vier 70-plussers in het bestand met kinderen ten laste. Een aantal voorspellingen is dus op weinig of geen data gebaseerd. We voorspellen ook voor een bepaalde groep een negatief aantal uren vrij, wat onmogelijk is. Bovendien was het interactie-effect klein en houdt het niet stand als we onze analyse uitvoeren op een beperktere groep (kleinere leeftijdsspanne, minder dan 4 kinderen ten laste). Dat interactie-effect blijkt eerder het gevolg te zijn van enkele invloedrijke respondenten met veel kinderen ten laste. In die zin moeten tabel 19 en de grafische weergave van het interactie-effect in grafiek 4 vooral als illustratief gezien worden en moet er niet teveel waarde gehecht worden aan de inhoudelijke conclusies ervan. Deze bedenking vloeit ook voort uit de beperkte modellering van het leeftijdseffect. Een niet-lineair effect zou ook hier tot een modelverbetering leiden en natuurlijk ook een impact hebben op het interactie-effect.

Dit voorbeeld gaf een bijkomend argument om te werken met deviatiescores. Niet alleen het intercept wordt daardoor betekenisvoller, ook het hoofdeffect van aantal kinderen ten laste werd zo berekend voor een zinvolle groep en niet voor nuljarigen. Er is overigens ook nog een meer technisch argument om met deviatiescores te werken als het model interactie-effecten bevat. Doordat deviatiescores zowel positieve als negatieve waarden (kunnen) aannemen, is een productterm van deviatiescores in regel minder gecorreleerd met de afzonderlijke variabelen dan een productterm van de oorspronkelijke scores. Een beperktere correlatie en bijgevolg een lagere multicollineariteit zorgen voor kleinere standaardfouten en bijgevolg statistische testen met een groter onderscheidend vermogen (zie ook Jaccard, et al. 1991, 30-31; 74).

5. Interactie-effecten in een binaire logistische regressie

Een meervoudige lineaire regressie kan wel categorische onafhankelijke variabelen opnemen in het model, maar de afhankelijke variabele is noodzakelijk metrisch. Als de afhankelijke variabele categorisch is, is logistische regressie een aangewezen techniek. In deze sectie 5 beperken we ons tot afhankelijke variabelen met 2 categorieën, waarvoor binaire logistische regressie geschikt is. In sectie 6 bekijken we multinomiale logistische regressie. In een eerste stap staan we stil bij de interpretatie van een binaire logistische regressie, zonder interactie-effecten.

5.1 Interpretatie van de parameters in een binaire logistische regressie zonder interactie-effecten

Als de dichotome afhankelijke variabele dummy gecodeerd wordt, kan de verwachte waarde worden uitgedrukt als een probabiliteit, de kans dat die afhankelijke variabele gelijk is aan 1. Deze kans wordt ook wel met de letter *p* aangeduid. De kans dat de afhankelijke variabele gelijk is aan 0 is dan natuurlijk gelijk aan 1 – *p*.

$$\begin{aligned} \Pr\,(Y=1) &= p \\ \Pr\,(Y=0) &= 1-p \end{aligned} \qquad (8)$$

Omwille van een aantal statistische redenen (vnl. de assumpties m.b.t. de lineariteit van de voorspellingsrechte en de verdeling van de residuelen, zie Agresti (1996, 72-87)) is het echter niet deze *p* die in een regressiemodel als afhankelijke variabele wordt geplaatst, maar wel de logit-transformatie ervan. Die logit-transformatie gebeurt in twee stappen. In een eerste stap wordt de probabiliteit omgezet in een odds, waarbij de geschatte probabiliteit ($\hat{p}$) gedeeld wordt door het complement ervan ($1-\hat{p}$). In een tweede stap wordt van deze odds de natuurlijke logaritme genomen. De logistische regressie ziet er dan als volgt uit:

$$\ln\left(\frac{\hat{p}}{1-\hat{p}}\right) = \text{logit}\left(\hat{p}\right) = b_0 + b_1 X_1 + b_2 X_2 + \ldots + b_k X_k \qquad (9)$$

Vergelijking (9) levert regressiecoëfficiënten op die vergelijkbaar zijn met de coëfficiënten uit een meervoudige regressie (zoals in vergelijking (1)). Alleen is een analoge interpretatie weinig informatief omdat het effecten op de logit betreft. De "voorspelde wijziging van de logit bij één eenheid wijziging in *X*" zegt eigenlijk niet zoveel. Daarom worden vaker de effecten op de odds, de zogenaamde oddsratio's, geïnterpreteerd. Deze effecten op de odds bekomt men door de effecten op de logit in te

geven in de exponentiële functie. Dat is de inverse bewerking van het natuurlijke logaritme en verheft het getal *e* tot de macht *b*: e^b. Dit volgt uit de herformulering van vergelijking (9), waarbij zowel het linker als het rechterdeel van de vergelijking "geëxponentieerd" werden:

$$\begin{aligned}\frac{\hat{p}}{1-\hat{p}} &= e^{b_0+b_1X_1+b_2X_2+\ldots+b_kX_k} \\ &= e^{b_0} * e^{b_1X_1} * e^{b_2X_2} * \ldots * e^{b_kX_k}\end{aligned} \tag{10}$$

e^{b_0} kan begrepen worden als de voorspelde odds of kansverhouding $\left(\frac{\hat{p}}{1-\hat{p}}\right)$ voor leden in de populatie met waarde 0 in de vergelijking. e^{b_1} is dan het voorspelde *multiplicatieve* effect op die (baseline) odds bij één eenheid wijziging in X_1 en onder controle van de effecten van X_2 tot X_k.

Ook voor logistische regressie zijn verschillende significantietesten mogelijk. Een test op basis van de standaardfout verloopt op exact dezelfde manier als bij een lineaire regressie. Een test die twee modellen vergelijkt, waarbij het tweede enkele extra effecten bevat, kijkt echter niet naar de verklaarde variantie maar naar de likelihood. Dit hangt samen met de schattingsmethode voor een logistische regressie (Maximum Likelihood Estimation)[5]. Wij zullen alleen testen op basis van de standaardfout bespreken. Details over de andere test zijn te vinden in Hosmer en Lemeshow (2000, 36-40; 145-147).

Bij deze theoretische uitleg is een voorbeeld onontbeerlijk. Voor dit voorbeeld kijken we naar het al dan niet hebben van betaald werk, zoals gemeten in de SILC-enquête van 2005 (20- tot 64-jarigen). We nemen dit op als afhankelijke variabele in een logistische regressie met leeftijd (deviatiescore rond het gemiddelde van 42) en opleidingsniveau (dummy, 0 = laag en 1 = hoger) als onafhankelijke variabelen. Tabel 20 toont de resultaten van die logistische regressie.

Tabel 20 Resultaten van de logistische regressie met betaald werk als afhankelijke variabele en leeftijd en opleidingsniveau als onafhankelijke variabelen

	b	Standaardfout	p-waarde	e^b
intercept	0,27	0,07	0,000	1,31
dev_leeftijd	-0,05	0,00	0,000	0,96
hoger opgeleid	0,80	0,08	0,000	2,22

Bron: SILC2005

Tabel 20 is de weergave van volgende vergelijkingen:

$$\ln\left(\frac{\hat{p}}{1-\hat{p}}\right) = 0{,}27 - 0{,}05 dev\_leeftijd + 0{,}80 hogeropgeleid \tag{11}$$

$$\begin{aligned}\frac{\hat{p}}{1-\hat{p}} &= e^{0{,}27} * e^{-0{,}05 dev\_leeftijd} * e^{0{,}80 hogeropgeleid} \\ &= 1{,}31 * 0{,}96^{dev\_leeftijd} * 2{,}22^{hogeropgeleid}\end{aligned} \tag{12}$$

De interpretatie van de verschillende parameters volgt hieruit. Zo is 0,27 de voorspelde logit "betaald werk" voor 42-jarige lageropgeleiden (waarde 0 op beide onafhankelijke variabelen). Voor elk jaar dat onze respondent ouder is, voorspellen we dat de logit daalt met 0,05. Verder is volgens dit model de logit van hogeropgeleiden 0,80 eenheden hoger dan de logit van lageropgeleiden. De effecten van

[5] Er is discussie over of en hoe de verklaarde variantie bij een logistische regressie berekend kan worden. Vandaar dat een test op basis van de verklaarde variantie ongebruikelijk is. Bemerk dat een lineaire regressie ook geschat kan worden met Maximum Likelihood Estimation, zodat een test op basis van de likelihood daar ook mogelijk is.

zowel leeftijd als opleidingsniveau zijn netto-effecten. Dat wil zeggen dat het effect van de andere onafhankelijke variabele onder controle gehouden wordt.

Makkelijker interpreteerbaar zijn de odds en de oddsratio's (effecten op de odds). De voorspelde odds voor de 42-jarige lageropgeleiden is gelijk aan 1,31. Die odds is de kans op betaald werk in verhouding tot de kans op het niet hebben van betaald werk. Een mogelijke interpretatie is dat het model voorspelt dat voor iedere 42-jarige lagere opgeleide die geen betaald werk heeft er 1,31 lageropgeleiden zijn die wel betaald werk hebben. Ook bij verklaring van de effecten van leeftijd en opleidingsniveau kan makkelijker teruggegrepen worden naar de effecten op de odds. Het effect van één jaar leeftijdsverschil is gelijk aan 0,96. Deze oddsratio betekent dat we voor elk jaar dat onze respondent ouder is, voorspellen dat de odds ***vermenigvuldigd*** worden met een factor 0,96. De kansverhouding betaald werk/geen betaald werk is dus kleiner voor de oudere leeftijdsgroepen. Een verschil van 17 jaren resulteert volgens dit model in een halvering van de odds. We moeten de odds dan immers vermenigvuldigen met 0,50 [= $(0,96)^{17}$]. Tot slot voorspellen we dat de kansverhouding "betaald werk/geen betaald werk" voor hogeropgeleiden gelijk is aan liefst 2,22 keer dezelfde kansverhouding bij lageropgeleiden. Als we deze factor combineren met het intercept, zouden we dus kunnen stellen dat voor iedere 42-jarige hogeropgeleide zonder betaald werk er 2,91 (= 1,31 * 2,22) hogeropgeleide 42-jarigen met betaald werk zijn.

Ook in deze multiplicatieve interpretatie gaat het natuurlijk om netto-effecten en zoals tabel 20 toont zijn alle gevonden effecten significant. De p-waarden zijn zeer klein. Als er in de populatie geen verschil zou zijn volgens respectievelijk leeftijd en opleidingsniveau dan zou de kans zeer klein zijn om in een steekproef van 3839 personen effecten te bekomen die zo groot zijn als of groter dan de effecten die wij gevonden hebben. Bijgevolg concluderen we dat er in de populatie wel leeftijds- en opleidingsverschillen zijn.

5.2 Een binaire logistische regressie met interactie-effecten

De opname van interactie-effecten in een logistische regressie gebeurt op exact dezelfde manier als in een meervoudige regressie, namelijk door de opname van producttermen. In tabel 21 rapporteren we de resultaten van een logistische regressie met één metrische onafhankelijke variabele en verder twee dummies en twee significante interactietermen. De afhankelijke variabele is opnieuw het al dan niet hebben van betaald werk.

Tabel 21 Resultaten van de logistische regressie met betaald werk als afhankelijke variabele en leeftijd, opleidingsniveau, geslacht en twee interactietermen als onafhankelijke variabelen

	b	Standaardfout	p-waarde	e^b
intercept	1,37	0,14	0,000	3,94
dev_leeftijd	-0,09	0,01	0,000	0,91
hoger opgeleid	0,06	0,15	0,680	1,07
vrouw	-1,47	0,16	0,000	0,23
vrouw * hoger opgeleid	0,76	0,18	0,000	2,15
dev_leeft * hoger opgeleid	0,05	0,01	0,000	1,06

Bron: SILC2005

Omdat een interpretatie in termen van odds en oddsratio's het interessantst is, beperken we ons daartoe. Het intercept geeft ons de odds voor mannelijke laagopgeleide respondenten van 42 jaar. Die odds zijn gelijk aan 3,94. Dit wil zeggen dat er voor elke laagopgeleide man van 42 jaar die niet werkt, 3,94 laagopgeleide mannen van die leeftijd zijn die wel werken. Bemerk dat dit intercept veel groter is dan in het model van tabel 20. Dat komt door de opname van geslacht in het model. Het intercept is in dit model dus specifieker, het geldt nu voor mannen.

Het hoofdeffect van leeftijd is groter geworden. Dit kan verklaard worden door het interactie-effect leeftijd*opleidingsniveau. Het hoofdeffect van leeftijd geldt daardoor alleen voor de lageropgeleiden. Voor elk jaar dat de lageropgeleide respondent ouder is, voorspellen we een vermenigvuldiging van de odds met 0,91. Dat impliceert dat voor lageropgeleiden een verschil van 7 jaar al resulteert in een halvering van de odds [$(0,91)^7$ = 0,51]. Het geschatte netto-effect van opleidingsniveau is in tegenstelling tot in tabel 20 zeer klein. Dat lijkt eigenaardig, maar is eveneens een resultaat van de interactie-effecten in het model. In het model zitten interactie-effecten van opleidingsniveau met geslacht én met leeftijd. Het geschatte effect (oddsratio = 1,07) geldt bijgevolg voor 42-jarige mannen.

Bij die groep is er volgens het model amper een verschil tussen lager- en hogeropgeleiden en is dat verschil ook niet significant! Het hoofdeffect van geslacht (vrouw) is wel nog zeer groot. De oddsratio is gelijk aan 0,23. Dat wil zeggen dat de odds "betaald werk/geen betaald werk" bij vrouwen gelijk is aan 0,23 keer dezelfde odds bij mannen, of ook dat die odds bij mannen dus ongeveer 4 keer groter is dan bij vrouwen. Let wel, dit laatste geldt enkel voor de lageropgeleiden, want er is inderdaad ook nog een interactie tussen geslacht en opleidingsniveau.

Om de effecten te berekenen voor de andere groepen moeten de oddsratio's vermenigvuldigd worden. Zo is het effect van leeftijd voor hogeropgeleiden gelijk aan 0,96 (= 0,91 * 1,06). Bij hogeropgeleiden voorspellen we dus een ***vermenigvuldiging*** van de odds met 'slechts' 0,96 voor elk jaar dat de respondent ouder is. Het negatieve effect van leeftijd op de kans op betaald werk versus geen betaald werk is bijgevolg kleiner voor hogeropgeleiden dan voor lageropgeleiden. Analoog hieraan is het zo dat het effect van opleidingsniveau toeneemt met de leeftijd. Voor 52-jarige mannen (10 jaar ouder dan gemiddeld) is het effect van opleidingsniveau gelijk aan 1,92 (= 1,07 * 1,79 waarbij 1,79 gelijk is aan 1,06 tot de 10de macht). De kansverhouding of odds "betaald werk/geen betaald werk" is bij de 52-jarige hogeropgeleiden dus gelijk aan bijna 2 keer dezelfde kansverhouding bij de 52-jarige lageropgeleiden. Het effect van het opleidingsniveau is ook veel groter voor de vrouwen dan voor de mannen. Bij de vrouwen van 42 jaar is de oddsratio voor opleidingsniveau gelijk aan 2,30 (= 1,07 * 2,15). Bij 42-jarige hogeropgeleide vrouwen is de odds "betaald werk/geen betaald werk" dus 2,30 keer diezelfde odds van 42-jarige lageropgeleide vrouwen. Bij jongere vrouwen zal de oddsratio (en dus het effect van opleidingsniveau) kleiner zijn, bij oudere vrouwen (nog) groter.

Achter deze complexe berekeningen en interpretaties schuilen dus enkele interessante en bevattelijke conclusies, die zich als volgt laten samenvatten:
- het negatieve effect van leeftijd op de kans op tewerkstelling is groter bij de lageropgeleiden dan bij de hogeropgeleiden;
- het verschil in werkzaamheid tussen mannen en vrouwen is groter bij de lageropgeleiden dan bij de hogeropgeleiden;
- het verschil tussen lageropgeleiden en hogeropgeleiden is bij ouderen en vrouwen groter dan bij de jongeren en mannen (als spiegelbeeld van de twee voorgaande).

Het zijn de interactie-effecten in het regressiemodel die toelaten om zulke conclusies te trekken.

6. Interactie-effecten in een multinomiale logistische regressie

Als de nominale afhankelijke variabele meer dan twee categorieën telt, is een multinomiale logistische regressie aangewezen. De opbouw van het model en de interpretatie daarvan zijn heel gelijklopend als bij een binaire logistische regressie, maar vergen nu zeer uitdrukkelijk de bepaling van een referentiecategorie vooraf. Dit wordt aangetoond in de volgende paragraaf, waarin we in eerste instantie terug een model zonder interactie-effecten bespreken.

6.1 Interpretatie van een multinomiale logistische regressie zonder interactie-effecten

Stel dat de afhankelijke variabele drie categorieën telt, voor de eenvoud a, b en c. In onze analyse zijn we dan geïnteresseerd in de kans dat die afhankelijke variabele één van die onderscheiden waarden aanneemt. Die kans wordt ook aangeduid met de letter *p*.

$$\Pr\ (Y = a)\ = p^{(a)}$$

$$\Pr\ (Y = b)\ = p^{(b)} \qquad (13)$$

$$\Pr\ (Y = c)\ = p^{(c)}$$

Net zoals bij een binaire logistische regressie zijn het niet deze verschillende *p*'s die in een regressiemodel als afhankelijke variabele wordt geplaatst, maar wel een logit-transformatie ervan. Ook hier gebeurt deze transformatie in twee stappen. In een eerste stap wordt de probabiliteit omgezet in een kansverhouding of odds, waarbij de probabiliteit om in een categorie terecht te komen gedeeld wordt door de probabiliteit om in de zelf gekozen referentiecategorie terecht te komen. In een tweede stap wordt van deze odds de natuurlijke logaritme genomen. Deze transformatie en de daaruitvolgende vergelijking wordt twee keer uitgevoerd; twee keer omdat dat gelijk is aan het aantal

categorieën van de afhankelijke variabele minus 1. Stel dat we c als referentie nemen, dan schat de multinomiale logistische regressie gelijktijdig volgende vergelijkingen:

$$\ln\left(\frac{\hat{p}^{(a)}}{\hat{p}^{(c)}}\right) = b_0^{(a)} + b_1^{(a)} X_1 + b_2^{(a)} X_2 + ... + b_k^{(a)} X_k$$

$$\ln\left(\frac{\hat{p}^{(b)}}{\hat{p}^{(c)}}\right) = b_0^{(b)} + b_1^{(b)} X_1 + b_2^{(b)} X_2 + ... + b_k^{(b)} X_k \quad (14)$$

We krijgen dus afzonderlijke regressiecoëfficiënten voor de verschillende categorieën (bvb. $b_1^{(a)}$ en $b_1^{(b)}$). In theorie is het niet noodzakelijk om voor elke categorie dezelfde onafhankelijke variabelen op te nemen in het model, maar meestal wordt dat wel gedaan. Anders kan het model zeer complex worden.

Deze vergelijkingen leveren opnieuw regressiecoëfficiënten op die vergelijkbaar zijn met de coëfficiënten uit een meervoudige regressie. Maar omdat effecten op die logits minder informatief zijn, worden ze – net zoals bij een binaire logistische regressie – geëxponentieerd om makkelijker begrijpbare effecten op de odds of oddsratio's te bekomen: $e^{b_1^{(a)}}$ is dan het voorspelde *multiplicatieve* effect op de kansverhouding of odds $p^{(a)} \Big/ p^{(c)}$ bij één eenheid wijziging in X_1 en onder controle van de effecten van X_2 tot X_k.

Een voorbeeld kan deze theoretische uitleg hopelijk wat verduidelijken. Voor dit voorbeeld kijken we naar het al dan niet deelnemen aan de SCV-survey van 2006. Meer informatie over deze data is te vinden in Carton et al. (2007). In 2006 werd in het kader van de SCV-survey geprobeerd om bij 2323 mensen een interview af te nemen. Dat lukte 1540 keren; 421 mensen weigerden en 144 mensen konden niet gecontacteerd worden. Voor dit voorbeeld kijken we alleen naar deze drie groepen, in totaal 2105 personen. De restgroep waarbij bvb. geen interview kon afgenomen worden omwille van een taalbarrière of omdat de potentiële respondent ziek of verhuisd was, laten we buiten beschouwing.

Tabel 22 Resultaat van de contactpogingen

	Frequentie	Percentage
interview	1540	73,2%
weigering	421	20,0%
geen contact	144	6,8%
N	2105	100,0%

Bron: SCV-survey 2006

We weten niet zoveel van deze mensen. Van de respondenten is er natuurlijk veel informatie beschikbaar uit het interview, maar van de anderen hebben we die informatie niet. Wat we bijvoorbeeld wel weten, is het geslacht van de potentiële respondenten, de leeftijd en een inschatting van de woonomgeving door de interviewer. Deze drie variabelen zullen we opnemen in onze analyse.

Zoals uit tabel 23 blijkt, zaten er in dit deel van de steekproef iets minder vrouwen dan mannen.

Tabel 23 Geslacht

	Frequentie	Percentage
man	1057	50,2%
vrouw	1048	49,8%
N	2105	100,0%

Bron: SCV-survey 2006

De verdeling volgens woonomgeving wordt weergegeven in tabel 24. De 7 categorieën van deze variabele worden herleid tot 3 dummies door enerzijds de drie categorieën "verstedelijkt" samen te nemen in de referentiegroep en door anderzijds de categorie "andere" buiten beschouwing te laten. Zo verliezen we wel nog 22 respondenten bij deze analyse, maar deze vereenvoudiging zal het voorbeeld verduidelijken.

Tabel 24 Woonomgeving

	Frequentie	Percentage
landelijk bosrijk	277	13,2
niet al te grote dorpskom	355	16,9
woongebied – eengezinswoning	713	33,9
verstedelijkt – eengezinswoning	450	21,4
verstedelijkt – appartement	252	12,0
verstedelijkt – ander	36	1,7
ander	22	1,0
Total	2105	100,0

Bron: SCV-survey 2006

Tabel 25 Dummy landelijk

	Frequentie	Percentage
0	1806	86,7
1	277	13,3
N	2083	100,0

Bron: SCV-survey 2006

Tabel 26 Dummy - dorpskom

	Frequentie	Percentage
0	1728	83,0
1	355	17,0
N	2083	100,0

Bron: SCV-survey 2006

Tabel 27 Dummy woongebied

	Frequentie	Percentage
0	1370	65,8
1	713	34,2
N	2083	100,0

Bron: SCV-survey 2006

Tot slot hebben we dus nog leeftijd als onafhankelijke variabele. Ook bij deze analyse nemen we hiervan de deviatiescore.

Tabel 28 Leeftijd en deviatiescore van leeftijd

	N	Minimum	Maximum	Gemiddelde	Standaard-afwijking
leeftijd	2105	18,00	85,00	48,14	17,54
dev_leeftijd	2105	-30,00	37,00	0,14	17,54

Bron: SCV-survey 2006

Deze vijf variabelen, leeftijd, geslacht en de drie dummies voor woonomgeving nemen we op in een multinomiale logistische regressie met resultaat van de contactpogingen als afhankelijke variabele. Voor deze laatste variabele moeten we ook een referentiecategorie kiezen. In dit geval valt die keuze logischerwijze op "een afgenomen interview", het ideale en tevens meest frequent voorkomende resultaat van de contactpogingen.

Tabel 29 Resultaten van de multinomiale logistische regressie met resultaat van de contactpoging(en) als afhankelijke variabele en geslacht, leeftijd en de drie dummies voor woonomgeving als onafhankelijke variabelen

Afhankelijke variabele	Onafhankelijke variabele	b	Standaard-fout	p-waarde	e^b
weigering	intercept	-1,40	0,11	0,000	0,25
	vrouw	0,46	0,11	0,000	1,58
	dev_leeftijd	0,02	0,00	0,000	1,02
	landelijk	-0,42	0,19	0,026	0,66
	dorpskom	-0,31	0,17	0,066	0,74
	woongebied	-0,21	0,13	0,107	0,81
geen contact	intercept	-1,79	0,15	0,000	0,17
	vrouw	-0,13	0,18	0,448	0,87
	dev_leeftijd	-0,01	0,01	0,172	0,99
	landelijk	-0,97	0,31	0,002	0,38
	dorpskom	-0,82	0,27	0,002	0,44
	woongebied	-0,99	0,22	0,000	0,37

Bron: SCV-survey 2006

Tabel 29 toont de resultaten van die multinomiale logistische regressie. Deze tabel maakt duidelijk dat er inderdaad verschillende regressiecoëfficiënten geschat worden voor de onderscheiden categorieën van de afhankelijke variabele (behalve de referentiecategorie).

Voor de interpretatie beperken we ons tot de laatste kolom, de geëxponentieerde coëfficiënten. De geëxponentieerde intercepten zijn odds. Zo is 0,25 de voorspelde odds of kansverhouding weigering/interview voor mensen die waarde 0 hebben op alle onafhankelijke variabelen. Voor elke man van 48 jaar, woonachtig in een verstedelijkt gebied, bij wie een interview mogelijk is, zijn er 0,25 48-jarige mannen met een gelijkaardige woonomgeving die een interview weigeren. Uit de andere geëxponentieerde coëfficiënten, de oddsratio's, blijkt dat deze odds kleiner is in andere woonomgevingen (het verschil is echter alleen significant voor landelijke gebieden), maar stijgt met de leeftijd en merkelijk groter is bij vrouwen dan bij mannen. Letterlijk: onder controle van geslacht en leeftijd is de kansverhouding weigering/interview in landelijke gebieden gelijk aan 2/3 van dezelfde kansverhouding in verstedelijkte gebieden. Onder controle van de effecten van leeftijd en woonomgeving voorspellen we dat die kansverhouding bij vrouwen gelijk aan 1,6 keer dezelfde kansverhouding bij mannen. En onder controle van geslacht en woonomgeving is deze odds hoger voor oudere mensen. We voorspellen dat die kansverhouding moet vermenigvuldigd worden met een factor 1,02 voor elk jaar dat de potentiële respondent ouder is. Dat betekent bvb. dat de kansverhouding voor een 70-jarige ongeveer dubbel zo groot is als voor een 30-jarige [$(1,02)^{40} = 2$]. De conclusie van het eerste deel van de tabel is dus dat vrouwen en ouderen vaker weigeren dan mannen en jongeren en dat het aandeel weigeringen nergens hoger ligt dan in de stad.

Het tweede deel van de tabel kan op een vergelijkbare wijze geïnterpreteerd worden. Daar zullen de conclusies zijn dat er geen (significante) verschillen zijn volgens geslacht en leeftijd in het al dan niet kunnen contacteren van potentiële respondenten. Woonomgeving speelt echter wel een rol, met drie significante negatieve effecten. Mensen die in de stad wonen zijn moeilijker bereikbaar en kunnen vaker niet gecontacteerd worden.

6.2 Interpretatie van een multinomiale logistische regressie met interactie-effecten

In de in tabel 29 gerapporteerde analyse nemen we in een volgende stap een interactie-effect op tussen geslacht en leeftijd. Net zoals bij de andere statistische modellen, doen we dat door een productterm te berekenen en deze op te nemen in het model. Tabel 30 toont dat deze productterm ook een metrische variabele is. Voor alle mannen zal hij gelijk zijn aan 0, voor alle vrouwen is die variabele gelijk aan de deviatiescore (en dus identiek aan dev_leeftijd).

Tabel 30 Leeftijd en deviatiescore van leeftijd

	N	Minimum	Maximum	Gemiddelde	Standaard-afwijking
vrouw * dev_leeftijd	2105	-30,00	37,00	0,05	12,44

Bron: SCV-survey 2006

Opname van deze variabele in de multinomiale regressie geeft de resultaten van tabel 31. De intercepten en de effecten van de woonomgevingvariabelen worden op exact dezelfde manier geïnterpreteerd als die effecten in tabel 29. Maar de opname van de interactieterm wijzigt wel de effecten van vrouw en dev_leeftijd. Die hoofdeffecten gelden nu respectievelijk voor 48-jarigen en voor mannen. Bij de 48-jarige vrouwen is de kansverhouding weigering/interview gelijk aan 1,5 keer dezelfde kansverhouding bij 48-jarige mannen. Verder voorspellen we bij de mannen dat die kansverhouding moet vermenigvuldigd worden met een factor 1,01 voor elk jaar dat de potentiële respondent ouder is. Dit alles onder controle van de effecten van de woonomgeving.

Het leeftijdseffect is bij de vrouwen beduidend groter. Bij de vrouwen voorspellen we dat dezelfde kansverhouding moet vermenigvuldigd worden met een factor 1,03 voor elk jaar dat de potentiële respondente ouder is (1,01 * 1,02 = 1,03). Een logische afgeleide hiervan is dat het effect van geslacht groter is naarmate de gecontacteerde persoon ouder is. Het verschil tussen vrouwen en mannen in de kans om een interview te weigeren is bij ouderen groter dan bij jongeren. Voor elk jaar dat de potentiële respondente ouder is vergroot het (multiplicatieve) effect met een factor 1,02. Hieruit volgt bvb. dat de oddsratio, die aangeeft hoeveel groter de kansverhouding weigering/interview is voor vrouwen t.o.v. mannen, bij 30-jarigen gelijk is aan 1,06 [$=1,51*(1,02)^{-18}$], terwijl die bij 70-jarigen gelijk is aan 2,34 [$=1,51*(1,02)^{22}$].

Tabel 31 Resultaten van de multinomiale logistische regressie met resultaat van de contactpoging(en) als afhankelijke variabele en de drie dummies voor woonomgeving, geslacht, leeftijd en de interactie tussen de twee laatste als onafhankelijke variabelen

Afhankelijke variabele	Onafhankelijke variabele	b	Standaard-fout	p-waarde	e^b
weigering	intercept	-1,39	0,11	0,000	0,25
	vrouw	0,41	0,12	0,000	1,51
	dev_leeftijd	0,01	0,01	0,098	1,01
	landelijk	-0,42	0,19	0,025	0,66
	dorpskom	-0,29	0,17	0,089	0,75
	woongebied	-0,21	0,13	0,121	0,81
	vrouw*dev_leeftijd	0,02	0,01	0,016	1,02
geen contact	intercept	-1,82	0,15	0,000	0,16
	vrouw	-0,09	0,18	0,608	0,91
	dev_leeftijd	-0,02	0,01	0,026	0,98
	landelijk	-0,98	0,31	0,002	0,38
	dorpskom	-0,80	0,27	0,003	0,45
	woongebied	-0,99	0,22	0,000	0,37
	vrouw*dev_leeftijd	0,02	0,01	0,068	1,02

Bron: SCV-survey 2006

De inhoudelijke conclusies van dit model met interactie-effect zijn dus drievoudig. Vrouwen weigeren vaker dan mannen. Zowel bij mannen als bij vrouwen stijgt de kans om een interview te weigeren naarmate ze ouder zijn. Maar die stijging is bij vrouwen veel groter dan bij mannen. Bijgevolg is het verschil in de kans op weigering tussen mannen en vrouwen bij ouderen ook groter dan bij jongeren.

Voor het onderste gedeelte van tabel 31 gaan we op dezelfde manier tewerk. Doordat hoofdeffecten en interactie-effect van teken verschillen, is de interpretatie een beetje gecompliceerder, maar de werkwijze is wel helemaal analoog. De belangrijkste conclusie is hier dat het effect van leeftijd voor mannen negatief is, terwijl er voor vrouwen (vrijwel) geen leeftijdseffect is. Voor elk jaar dat een man ouder is, voorspellen we dat de kansverhouding geen contact/interview vermenigvuldigd wordt met 0,98. Bij oudere mannen is de kans dat er geen contact kan gelegd worden dus wat kleiner dan bij

jongere mannen. Bij vrouwen verschilt deze multiplicatieve factor amper van 1 (0,98 * 1,02 = 1,00). Het effect van leeftijd op de kans om al dan niet gecontacteerd te kunnen worden is bijgevolg klein en waarschijnlijk niet significant. Dit model geeft hiervoor geen significantietest. De eenvoudigste manier om zo'n test te bekomen, hercodeert geslacht op een andere manier. Als vrouwen waarde 0 krijgen bij de dummy en mannen waarde 1, geldt het hoofdeffect van dev_leeftijd en de bijhorende significantietest immers voor vrouwen.

7. Bijkomende toepassingsmogelijkheden

In de voorgaande paragrafen toonden we een aantal voorbeelden van regressiemodellen waarbij de interactie-effecten een inhoudelijke verrijking van het model betekenden. Er zijn echter nog andere toepassingsmogelijkheden van interactie-effecten, die misschien minder evident lijken. In deze paragraaf beschrijven wij er twee. In de eerste toepassing gaan we met behulp van een interactieterm na of een effect veranderd is in de tijd. In de tweede toepassing vormt de interactieterm een oplossing voor het probleem dat een bepaalde variabele niet van toepassing is voor een deel van de populatie.

7.1 Samenvoegen van databestanden en kijken of het effect verschilt tussen beide bestanden

Deze toepassingsmogelijkheid kwam al uitvoerig aan bod in een vorig SVR-Technisch rapport (Pickery, 2006). Dat rapport toonde een manier om na te gaan of de ongelijkheid in participatie toeneemt of daalt. Meer specifiek werd de internetpenetratie bij mensen met betaald werk en mensen zonder betaald werk onderzocht voor de jaren 2001, 2003 en 2005. Het is bekend dat het internetgebruik stijgt, het is eveneens geweten dat werkenden vaker internet gebruiken dan niet-werkenden. Maar hoe kan je op een statistisch verantwoorde manier nagaan of die ongelijkheid tussen mensen met betaald werk en mensen zonder betaald werk toeneemt dan wel afneemt? Een model met interactie-effecten is hiervoor één van de mogelijkheden. In eerste instantie worden hiervoor de databestanden van de verschillende jaargangen samengevoegd ("pooled cross sections"). Hierbij is het wel nodig dat de afhankelijke en onafhankelijke variabele(n) dezelfde zijn. In een regressiemodel wordt dan naast de oorspronkelijke onafhankelijke variabele(n) ook het jaartal van de verschillende surveys opgenomen (al dan niet gecategoriseerd en gehercodeerd in dummies) evenals de interactietermen die de wisselwerking tussen die jaartallen en de onafhankelijke (focus)variabelen tot uitdrukking brengen. Het voorbeeld in tabel 32, dat overgenomen is uit het *SVR-Technisch rapport*, kan dit illustreren.

Tabel 32 Resultaten van de logistische regressie van de kans op internetgebruik voor de verschillende surveys (2001, 2003 en 2005) samen met interactie-effecten

	b	Standaardfout	p-waarde	e^b
intercept	-1,45	0,10	0,000	0,23
betaald werk	1,35	0,12	0,000	3,85
jaar_2003	0,56	0,13	0,000	1,75
jaar_2005	0,89	0,13	0,000	2,44
betaald werk * jaar_2003	0,05	0,17	0,759	1,05
betaald werk * jaar_2005	0,39	0,17	0,019	1,48

Bron: SCV-survey

De afhankelijke variabele van de analyse in tabel 32 is het al dan niet gebruiken van internet. Omdat dat een dichotome variabele is, gebruiken we een logistische regressie. We hebben eigenlijk maar één inhoudelijke onafhankelijke variabele, namelijk het al dan niet hebben van betaald werk. Dat is een dummy die waarde 0 heeft bij de mensen zonder betaald werk en waarde 1 bij de mensen met betaald werk. De verschillende surveyjaren hebben we ook opgenomen als dummies. 2001 is de referentiecategorie en er zijn dus dummies voor 2003 en 2005. De producttermen betaald werk * jaar_2003 en betaald werk* jaar_2005 zijn dan natuurlijk ook dummies. Door de opname van die interactietermen in het model zijn de hoofdeffecten in tabel 33 ook conditioneel of specifiek. Het effect van betaald werk geldt enkel voor 2001 (een oddsratio van 3,85) en de effecten van jaar_2003 en jaar_2005 gelden enkel voor de mensen zonder betaald werk.

Vanuit de vraagstelling van een evolutie in de ongelijkheid zijn de interactie-effecten relevanter. Zij tonen het verschil in het effect van betaald werk tussen de verschillende surveyjaren, en meer specifiek

hoeveel groter die oddsratio is in respectievelijk 2003 en 2005. Zo blijkt dat het effect tussen 2001 en 2003 een klein beetje groter geworden is. In 2003 is de oddsratio gelijk aan 4,04 (= 3,85 * 1,05). Het model toont echter dat deze toename niet significant is. Maar het effect van betaald werk is wel sterk en significant toegenomen in 2005, met een oddsratio van 5,70 (= 3,85 * 1,48). De ongelijkheid is dus toegenomen: een klein beetje van 2001 tot 2003, maar duidelijk van 2001 tot 2005 (met een factor vrijwel gelijk aan 1,5).

In dit voorbeeld voegden we bestanden van verschillende jaargangen samen en is de afhankelijke variabele dichotoom, waardoor we ons wendden tot logistische regressie. Maar de werkwijze is natuurlijk algemener toepasbaar. De afhankelijke variabele kan evengoed metrisch zijn (en het model bijgevolg een meervoudige lineaire regressie) en de methode kan ook gebruikt worden om te zien of effecten die gemeten zijn in verschillende populaties of verschillende regio's gelijk zijn.

7.2 Opname van variabelen in het model die bij een deel van de populatie niet van toepassing zijn

In het laatste voorbeeld van dit *SVR-Technisch rapport* gebruiken we een interactie-effect om een variabele die voor een deel van de populatie niet van toepassing is toch op te nemen in het model. We grijpen daarvoor terug naar het eerste voorbeeld dat het volume vrije tijd probeerde te verklaren. Stel dat we in een volgende stap in dat model de partnersituatie van de respondent willen opnemen, bvb. het al dan niet samenwonen met een partner en de tewerkstellingsstatus van die partner (van de respondent). Die eerste variabele vormt natuurlijk geen probleem. Dat is gewoon een dummy die waarde 1 krijgt voor mensen met partner en waarde 0 voor de overigen. Maar de tweede variabele is wel problematischer. Voor respondenten zonder partner is de vraag naar de tewerkstellingsstatus van de partner natuurlijk niet van toepassing. Toch is het met behulp van een interactieterm mogelijk om die tewerkstellingsstatus op te nemen in het model. Tabellen 33 tot 36 tonen hoe dit in zijn werk gaat.

Tabel 33 Al dan niet samenwonen met een partner

	Frequentie	Percentage
0 (= nee)	426	29,2%
1 (= ja)	1034	70,8%
N	1460	100,0%

Bron: SCV-survey 2002

Tabel 34 Tewerkstelling van partner (partner_betaald werk)

	Frequentie	Percentage N (totaal)	Percentage N (partner)
0 (= nee)	429	29,4%	41,5%
1 (= ja)	605	41,4%	58,5%
N (met partner)	1034	70,8%	100,0%
Niet van Toepassing	426	29,2%	
N (totaal)	1460	100,0%	

Bron: SCV-survey 2002

Bijna 71% van de respondenten woont samen met een partner. Van die partners heeft een kleine 60% betaald werk. Maar tabel 34 toont ook de 29% van de respondenten waarvoor deze vraag niet van toepassing was. Die "niet van toepassing"-categorie verdwijnt echter bij de productterm die gebruikt wordt om het interactie-effect te meten. Mensen zonder partner hebben immers waarde 0 op de variabele "partner" en een vermenigvuldiging met 0 is natuurlijk gelijk aan 0. Zo krijgen bij de productterm alleen de mensen met een partner die werkt waarde 1 en alle anderen waarde 0. Of ze een partner hebben die niet werkt, of geen partner, doet er niet toe.

Tabel 35 Productterm van het al dan niet hebben van een partner en het al dan niet hebben van betaald werk door de partner

	Frequentie	Percentage
0 (= geen partner of partner zonder betaald werk)	855	58,6%
1 (= partner met betaald werk)	605	41,4%
N	1460	100,0%

Bron: SCV-survey 2002

In ons regressiemodel – voor de eenvoud vertrekken we terug van het eenvoudige model (4) / tabel 8 – kunnen we nu het gewone hoofdeffect van partner opnemen en de interactie van partner * tewerkstellingsstatus partner. Het hoofdeffect van de variabele "tewerkstelling partner" nemen we echter ***niet*** op. Dat zou problematisch zijn o.w.v. de groep "niet van toepassing", maar ook niet zinvol. Tabel 36 toont de resulterende regressievergelijking.

De eerste vijf parameters in tabel 36 worden op exact dezelfde manier geïnterpreteerd als bij model (4). We staan er dus niet verder bij stil. Het effect van partner is een hoofdeffect, de interpretatie is specifiek. Bij mensen *met een partner die **niet** werkt*, voorspellen we ongeveer drie kwartier minder vrije tijd dan bij mensen zonder partner (0,77 uur). Het interactie-effect heeft hetzelfde teken en telt hierbij op. Dus bij mensen *met een partner die werkt*, voorspellen we een dik uur vrije tijd minder dan bij mensen zonder partner (–0,77 – 0,29 = –1,06). Bemerk wel dat het interactie-effect niet significant is. De p-waarde is gelijk aan 0,12 en de conclusie is dus voorbarig.

Tabel 36 Resultaten van de regressie met aantal uren vrij in de week als afhankelijke variabele en leeftijd (deviatiescore), aantal kinderen ten laste, geslacht, betaald werk, partner en de interactie van partner en partner_betaald werk als onafhankelijke variabelen

	b	Standaardfout	p-waarde
intercept	5,90	0,16	0,000
dev_leeftijd	0,04	0,04	0,000
aantal kinderen ten laste	-0,26	0,06	0,000
vrouw	-1,00	0,13	0,000
betaald werk	-1,52	0,16	0,000
partner	-0,77	0,18	0,000
partner * partner_betaald werk	-0,29	0,19	0,122

Bron: SCV-survey 2002

Uit deze interpretatie blijkt dat we hier ook voor een oplossing met andere dummies hadden kunnen kiezen. We hebben drie groepen (mensen zonder partner, mensen met een partner die niet werkt en mensen met een partner die werkt) en opname van twee dummies voor deze verschillende groepen zou leiden tot een 100% equivalent model. Maar als de variabele die bij de partner gemeten werd metrisch was, zou dit niet mogelijk zijn. Stel dat bvb. de leeftijd van de partner een impact zou hebben op het volume vrije tijd, dan is alleen de werkwijze met een interactie-effect mogelijk. Het model met interactie-effect is dus algemener. Bovendien kan dat effect zelf ook geïnterpreteerd worden. Bij een significant interactie-effect zou de inhoudelijke interpretatie ervan luiden dat bij de mensen met een partner, het hebben van betaald werk door die partner een bijkomende vermindering van het volume vrije tijd inhoudt.

Besluit

Dit SVR-Technisch rapport had de bedoeling om de interpretatie van interactietermen in regressiemodellen te verduidelijken. De tekst beperkte zich daarbij tot "eenvoudige" tweewegsinteracties. Meer gecompliceerde modellen zoals een niet-lineaire interactie of een driewegsinteractie werden niet behandeld. In het eerste geval heeft X_2 een impact op het effect van X_1 op Y, maar is die impact niet-lineair. In het tweede geval is er een derde variabele (X_3) die een impact heeft op de interactie van X_1 en X_2 bij hun effect op Y. Zulke meer gecompliceerde modellen vallen buiten het bereik van dit rapport, maar worden bijvoorbeeld wel behandeld in Jaccard et al. (1991, 40-42; 50-59).

Een andere beperking is dat er niet werd stilgestaan bij de technische finesses van de verschillende regressiemodellen. Zo kwamen de schattingsmethoden van de modellen bijvoorbeeld niet aan bod. Voor die meer technische kant kan de lezer zich echter makkelijker richten tot bestaande handleidingen in (logistische) regressieanalyse.

Vanuit de vaststelling dat de interpretatie van interactie-effecten vaak voor problemen blijft zorgen, koos dit rapport ervoor om, met enkele eenvoudige voorbeelden en een vrijwel exclusieve focus op het inhoudelijke, die interpretatie scherp te stellen. Hopelijk draagt deze tekst op die manier bij tot een correcter gebruik van interactie-effecten in regressiemodellen en tot een treffende duiding van de resultaten van die modellen.

Bibliografie

Agresti, A. (1996) *An Introduction to Categorical Data Analysis*. New York: Wiley.

Algemene Directie Statistiek en Economische Informatie (2006) *Quality Report Belgian SILC 2005.* Brussel: Algemene Directie Statistiek en Economische Informatie, ook online raadpleegbaar via http://statbel.fgov.be/silc/.

APS (2003) *Kwaliteitszorg Statistisch Productieproces. Aanbevelingen*. Brussel: Ministerie van de Vlaamse Gemeenschap.

Callens, M. (2008) *Contextuele regressiemethoden voor internationaal vergelijkend onderzoek. SVR-Technisch rapport 2008/X*. Brussel: Studiedienst van de Vlaamse Regering (in druk).

Carton, A., L. Hegemann & H. Van Geel (2003) *Basisdocumentatie: Sociaal Culturele Verschuivingen in Vlaanderen 2002*. Brussel: Ministerie van de Vlaamse Gemeenschap, Administratie Planning en Statistiek.

Carton, A., T. Vander Molen & J. Pickery (2007) *Sociaal Culturele Verschuivingen in Vlaanderen 2006. Basisdocumentatie. SVR-Technisch rapport 2007/2*. Brussel: Vlaamse overheid, Studiedienst van de Vlaamse Regering.

Jaccard, J. (2001) *Interaction Effects in Logistic Regression*. Thousand Oaks/London: Sage.

Jaccard, J., R. Turrisi (2003) *Interaction effects in multiple regression. Second Edition.* Thousand Oaks/London: Sage.

Jaccard, J., R. Turrisi & C.K. Wan (1991) *Interaction effects in multiple regression.* Newbury Park/London: Sage.

Hosmer, D. & S. Lemeshow (2000) *Applied Logistic Regression, 2nd Edition.* New York: Wiley.

McClendon, M.J. (2002) *Multiple Regression and Causal Analysis.* Prospect Heights, IL: Waveland Press.

Pampel, F. (2000) *Logistic Regression. A Primer. Sage University Papers. Series: Quantitative Applications in the Social Sciences, 07-132*. Thousand Oakes: Sage.

Pickery, J. (2006) *Een statistische analyse van een toenemende of dalende ongelijkheid in participatie. Van kruistabellen naar oddsratio's en van oddsratio's naar een logistische regressie (en terug). SVR-Technisch rapport 2006/3.* Brussel: Studiedienst van de Vlaamse Regering.

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