Centre for Social Theory

Dummyvariabelen in meervoudige regressie

Een inleiding voor sociale wetenschappers

Ronan Van Rossem
Universiteit Gent

e-doc

CST e-DOCUMENT NR. 1, NOVEMBER 2010

Uitgave van:
Centre for Social Theory
Vakgroep Sociologie
Universiteit Gent
Korte Meer 3-5
9000 Gent

ISSN: 2033-9356

UNIVERSITEIT GENT
Department of Sociology
Korte Meer 3
9000 Gent
Belgium

Phone: +32(0)9 264.67.96
Fax: +32(0)9 264.69.75
Email: socio@ugent.be

Dummyvariabelen in meervoudige regressie

Een inleiding voor sociale wetenschappers

Ronan Van Rossem

Universiteit Gent

CST e-DOCUMENT NR. 1, NOVEMBER 2010

INHOUD

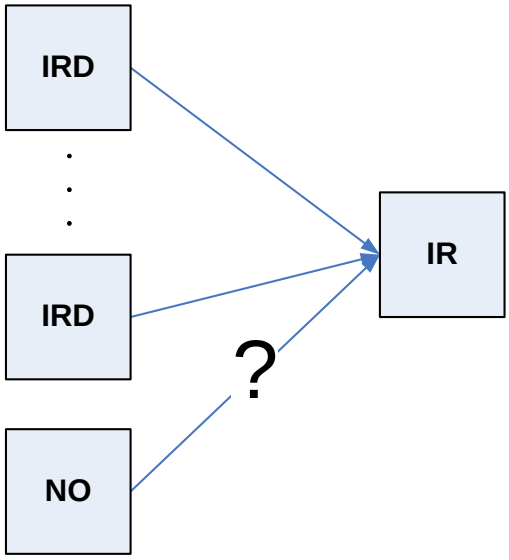
Figuren

Tabellen

Categorische variabelen in regressiemodellen

Regressie technieken, zoals multipele of multivariate lineaire regressie, binomiale of multinomiale logistische regressie, zowel als andere manifestaties van het algemeen lineair model veronderstellen dat de onafhankelijke variabelen allemaal geordend en een constante eenheid hebben. Dit laatste betekent dat het verschil tussen waarden *i* en *i* + 1 even groot is als dit tussen *j* en *j* + 1. Bv. het verschil tussen een score 3 en een score 4 op een variabele moet eenzelfde betekenis hebben als het verschil tussen een score 21 en een score 22. In praktijk betekent dit dat de enige variabelen die als onafhankelijke variabelen in deze regressiemodellen mogen gebruikt worden interval, ratio of dichotome variabelen zijn.

Vele van de variabelen in de sociale wetenschappen zijn echter categorisch van aard, bv. godsdienst, etniciteit, onderwijsniveau, beroep, enz. Zolang er maar twee categorieën zijn kunnen deze als dichotome variabelen in de regressiemodellen opgenomen worden. Zijn er echter meerdere categorieën dan kunnen die variabelen niet zomaar gebruikt worden in regressiemodellen.

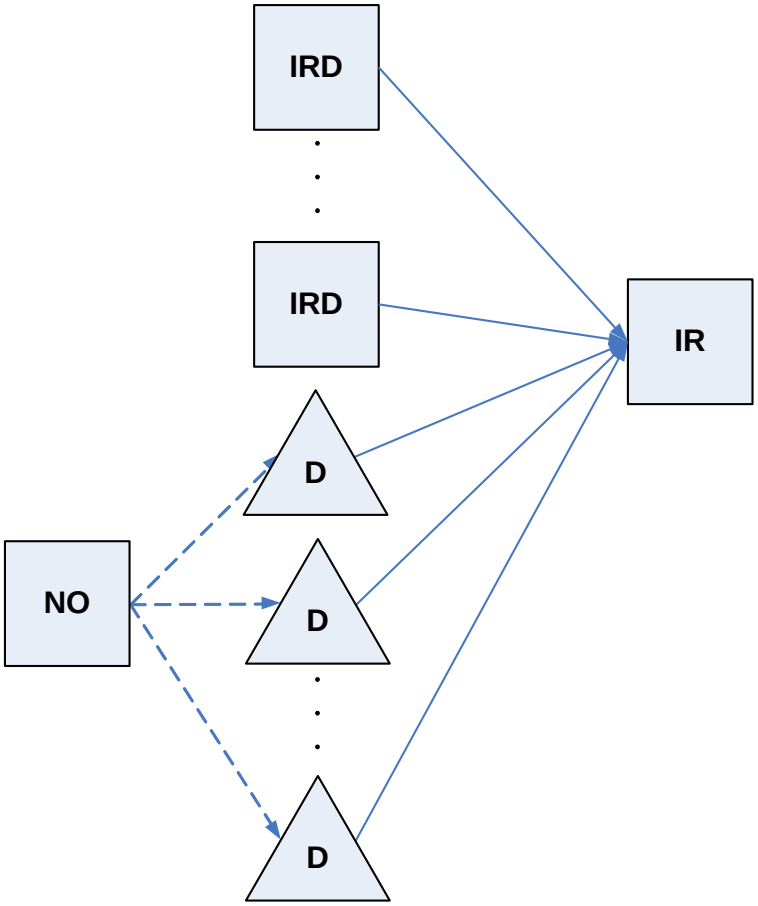


Figuur 1: Het probleem: categorische variabelen in multipele regressie

Het model in Figuur 1 geeft het probleem aan: hoe kan men een categorische onafhankelijke variabele gebruiken in een multipele regressiemodel. De eenvoudigste oplossing om dergelijke variabelen toch te kunnen opnemen is ze om te vormen tot dummyvariabelen.

Principes van dummyvariabelen

Het principe achter dummyvariabelen is zeer eenvoudig: indien men een categorische variabele heeft met *k* categorieën, vervang die dan door *k* – 1 dummyvariabelen die elk één van de categorieën vertegenwoordigen en waarbij de resterende categorie die niet door een dummyvariabele vertegenwoordigd wordt de referentiecategorie genoemd wordt.



Figuur 2: Multipele regressie met dummyvariabelen

Dit wordt uitgebeeld in Figuur 2Figuur , de oorspronkelijke categorische variabele wordt opgebroken in een reeks dummyvariabelen, en deze dummyvariabelen nemen een vorm aan die wel kan gebruikt worden in multipele regressieanalyse, bv. die van een dichotome variabele. Pas op, een dummyvariabele is niet hetzelfde als een dichtome variabelen. Een dummyvariabele neemt inderdaad vaak een dichotome vorm aan (bv. bij eenvoudige dummycodering) maar dit is niet altijd het geval (bv. bij effect dummycodering). De term dummyvariabele betekent feitelijk een eenvoudige plaatsvervangende variabelen, net zoals een modepop (een dummy) een echt levend model vervangt.

De reden waarom men maximaal steeds maar 1 dummy minder mag hebben dan er categorieën in de categorische variabele zijn, is dat indien men evenveel dummyvariabelen zou aanmaken als er categorieën zijn men een probleem van lineaire afhankelijkheid zou veroorzaken. Bij multipele regressie, bv. zou dat leiden tot een singuliere covariantiematrix en zou het model niet kunnen geschat worden. Het komt er op neer dat indien men *k* categorieën heeft, men maar *k* – 1 vrijheidsgraden heeft. Neem bv. een variabele religie met 4 categorieën: christen (C), moslim (M), geen (G) en andere (A). Indien men weet hoe een respondent scoort op 3 van deze categorieën, dan weet men ook hoe hij of zij scoort op de vierde. Bijvoorbeeld, indien een respondent niet tot een andere religie behoort, ook niet tot de geen categorie behoort en ook geen moslim is, dan moet hij of zij wel een christen zijn:

Indien (M = nee) en (G = nee) en (A = nee) dan (C = ja)

Anderzijds weten we bv ook dat indien (C = ja) de andere drie “nee”moeten zijn. Men moet maximaal voorkennis hebben van 3 van de 4 categorieën om alle categorieën te kennen.

Eenvoudige dummycodering

De meest gebruikte dummycodering methode is de eenvoudige dummycodering—ook wel soms gewoon “dummycodering” of “gewone dummycodering” genoemd.

Werkwijze

Neem bijvoorbeeld dat men een categorische variabele voor religie heeft met vier waarden:

1 christen
2 moslim
3 geen
4 ander.

Voor deze vier categorieën kan men nu 3 dummyvariabelen aanmaken. Een dummyvariabele is gewoon een dichotome variabele met scores 0 en 1, waarbij een score van 1 betekent dat de observatie behoort tot de categorie verbonden met deze dummyvariabele, en een score van 0 dat men niet tot die categorie behoort. De vierde categorie krijgt geen eigen dummyvariabele en wordt de referentie of controlecategorie genoemd. Het effect van deze categorie op de uitkomst variabele wordt verrekend via de constante.

Voor deze vier categorieën kan men 3 dummyvariabelen aanmaken: D_1, D_2 en D_3. Als men de categorie “christen” als referentiecategorie neemt, de dummyvariabele D_1 voor de categorie “moslim” gebruikt, D_2 voor “geen” en D_3 voor “ander”, dan krijgt men de volgende codering:

	D_1	D_2	D_3
Christen	0	0	0
Moslim	1	0	0
Geen	0	1	0
Ander	0	0	1

De observaties in de referentiecategorie “christen” scoren 0 op alle drie de dummyvariabelen (D_1, D_2 en D_3); terwijl bv. de observaties in de categorie “moslim” een 1 scoren op D_1, de dummyvariabele die deze categorie vertegenwoordigt en 0 op de twee andere dummyvariabelen D_2 en D_3. Observaties in de “geen” categorie scoren 1 op D_2 en 0 op D_1 en D_3, en deze in de “andere” categorie 1 op D_3 en 0 op D_1 en D_2.

Het aantal dummyvariabelen is steeds beperkt tot $k - 1$ variabelen als k het aantal niet-lege categorieën voor de variabele is. De reden hiervoor is dat men maar $k - 1$ vrijheidsgraden heeft, en het gebruik van k dummyvariabelen zou leiden tot een lineaire afhankelijkheid onder de dummyvariabelen. Veronderstel even dat men in het voorbeeld ook een dummyvariabele D_0 voor de categorie “christen” zou aanmaken. In dat geval zou de waarde van D_0 steeds gelijk zijn aan 1 min de som van D_1, D_2 en D_3:

$$D_0 = 1 - D_1 - D_2 - D_3$$

Op dezelfde manier kan men stellen dat:

$$D_1 = 1 - D_0 - D_2 - D_3$$

$$D_2 = 1 - D_0 - D_1 - D_3$$

$$D_3 = 1 - D_0 - D_1 - D_2$$

Zodra men de scores op $k - 1$ dummyvariabelen kent, kent men automatisch ook die op de k^{de}.

De effecten van deze dummyvariabelen d_1, d_2 en d_3—of in het algemeen d_i, $i = 1, \ldots, k - 1$—kunnen geïnterpreteerd worden als de gemiddelde afwijking van de door de dummyvariabele vertegenwoordigde categorie van de referentiecategorie, controlerend voor andere variabelen in het model. Bijvoorbeeld, het effect (d_1) van de dummyvariabele (D_1) voor de categorie "moslim" is het aangepaste verschil tussen de gemiddelden van de categorieën "moslim" en "christen", of m.a.w. dat controlerend voor de andere variabelen in het model, observaties in de categorie "moslim" gemiddeld d_1 meer scoort op de uitkomstvariabele Y dan observaties in de referentiecategorie "christen".

Keuze referentiecategorie

De keuze van de referentiecategorie is in principe vrij. Dit betekent dat statistisch het niks uitmaakt welke categorie men als referentiecategorie kiest. Of men in het voorgaande voorbeeld nu de categorie "christen" of "moslim"of "geen" of "ander" kiest maakt weinig uit. De geschatte coëfficiënten zullen wel wijzigen daar men met een andere categorie vergelijkt maar het globale effect van de variabele zal niet veranderen.

Toch zijn er enkele strategieën die men kan gebruiken bij het kiezen van een referentiecategorie. Zo kan men een referentiecategorie kiezen op substantiële gronden. Vraag je af welke categorie de meest logische keuze is als referentiecategorie. Als men bijvoorbeeld het effect van verschillende interventieprogramma's op een uitkomst wil onderzoeken is de controlegroep die aan geen enkel programma deelgenomen heeft de meest logische keuze als referentiecategorie. Indien men het effect van etniciteit op schoolprestaties wil nagaan is het zinvol om de autochtonen als referentiecategorie te nemen.

Een tweede criterium dat men kan gebruiken om een referentiecategorie te kiezen is de grootte van de categorieën. Het is altijd een goed idee om een grote categorie te selecteren als referentiecategorie. Dit hoeft niet noodzakelijk de grootste categorie zijn maar moet wel vermijden dat men een kleine categorie kiest. Bij een kleine referentiecategorie zal men niet alleen te maken hebben met een grotere standaardfout op het referentiegroepgemiddelde, en dus ook op het geschatte groepsgemiddelde, maar zal men ook sterkere correlaties krijgen tussen de verschillende dummyvariabelen, wat in bepaalde gevallen tot multicollineariteitsproblemen kan leiden.

Wanneer de categorische variabele ordinaal is kiest men vaak de hoogste of de laagste categorie als referentiecategorie. Dit vergemakkelijkt de interpretatie van de resultaten.

Het is ook aan te raden een homogene categorie als referentiecategorie te kiezen. Men heeft nogal snel de neiging om een residuele categorie zoals "ander" of "weet het niet" als referentiecategorie te kiezen. De motivatie hiervoor is dat men de resultaten voor deze categorieën toch niet goed kan interpreteren daar deze categorieën te heterogeen zijn. Wat bedoelt men met "ander" of "weet het niet". Men hoopt dan ook door deze als referentie-categorie te selecteren dit interpretatieprobleem onder de mat te vegen. Jammer genoeg verergert men de situatie op deze manier. Als men een dergelijke residuele categorie behoudt is er maar één coëfficiënt die moeilijk interpreteerbaar is, terwijl wanneer men deze als referentiecategorie gebruikt men $k - 1$ moeilijk interpreteerbare coëfficiënten krijgt. Men weet dan namelijk niet waarmee men de homogene categorieën aan het vergelijken is.

Voorbeeld 1

Een eerste voorbeeld komt uit de Belgische steekproef van de 1990 *World Values Survey* (World Values Study Group, 1994). In dit voorbeeld wordt multipele lineaire regressie gebruikt maar de gebruikte werkwijze kan eenvoudig geëxtrapoleerd worden naar alle regressietechnieken. De afhankelijke variabele is een maat voor etnocentrisme, gebaseerd op items die aangaven dat men joden, moslims, immigranten of gastarbeiders, mensen van een ander ras, hindoes, of mensen met grote gezinnen liever niet als buren zou hebben.

De onafhankelijke variabelen zijn een dichotome geslacht variabele (vrouw = 0, man = 1), en twee continue variabelen: politieke oriëntatie (1 = links, 10 = rechts) en belang van god in het dagelijks leven (helemaal niet = 1, zeer belangrijk = 10). Er is ook één categorische onafhankelijke variabele, namelijk onderwijs, welke uit vier categorieën bestaat: lager onderwijs, lager middelbaar, hoger middelbaar, en hoger onderwijs. De frequentieverdeling voor deze variabele wordt getoond in Tabel 1.

Tabel 1: Frequentieverdeling voor onderwijsniveau (WVS 1990, België)

Onderwijsniveau	F	f
Lager onderwijs	39	2.1%
Lager middelbaar onderwijs	390	21.5%
Hoger middelbaar onderwijs	683	37.6%
Hoger onderwijs	705	38.8%
Totaal	1817	100%

Om deze variabele te kunnen gebruiken in een regressie analyse moet men dummyvariabelen voor de verschillende onderwijsniveaus aanmaken. Onderwijsniveau heeft vier niet-lege categorieën wat maakt dat drie dummyvariabelen kunnen aangemaakt worden. Vooraleer over te gaan tot het aanmaken van de dummyvariabelen dient eerst beslist te worden welke categorie als referentiecategorie zal fungeren. Onderwijsniveau is een ordinale variabele, dus het heeft zin om de hoogste of laagste categorie als referentiecategorie te gebruiken. De laagste categorie “lager onderwijs” bevat echter weinig observaties (maar 2.1%) en is dus niet echt geschikt om als referentiegroep te fungeren. Men kiest er hier dan ook het beste voor om de hoogste categorie “hoger onderwijs” als referentiecategorie te weerhouden.

Zoals in Tabel 2 getoond wordt, worden er drie dummyvariabelen aangemaakt, één voor “lager onderwijs” (D_1), één voor “lager middelbaar onderwijs” (D_2) en één voor “hoger middelbaar onderwijs” (D_3). Respondenten die alleen lager onderwijs genoten hebben krijgen een score van 1 op D_1 en scores van 0 op D_2 en D_3. Respondenten met lager middelbaar onderwijs scoren 1 op D_2 en 0 op D_1 en D_3, en diegenen met hoger middelbaar onderwijs scoren 1 op D_3 en 0 op D_1 en D_2. Respondent die hoger onderwijs genoten scoren 0 op alle drie dummyvariabelen.

Tabel 2: Eenvoudige dummycoderingschema voor onderwijsniveau

Onderwijsniveau	D_1	D_2	D_3
Lager onderwijs	1	0	0
Lager middelbaar onderwijs	0	1	0
Hoger middelbaar onderwijs	0	0	1
Hoger onderwijs	0	0	0

Of de codering van de dummyvariabelen goed verlopen is kan men checken door middel van de frequentieverdelingen voor de dummyvariabelen (zie Tabel 3). Wanneer de dummyvariabelen correct aangemaakt werden moet de frequentie van de observaties met een waarde 1 gelijk zijn aan het aantal observaties in de categorie die door deze dummyvariabele vertegenwoordigd wordt. De onderstaande tabel toont dat hier de dummyvariabelen correct geconstrueerd werden.

Tabel 3: Frequentietabel voor de dummyvariabelen voor onderwijsniveau

	$F(0)$	$F(1)$	N
D_1 Lager onderwijs	1778	39	1817
D_2 Lager middelbaar onderwijs	1427	390	1817
D_3 Hoger middelbaar onderwijs	1134	683	1817

Daar er altijd een risico voor multicollineariteit is tussen dummyvariabelen doet men er ook goed aan om de correlatie tussen de verschillende dummyvariabelen te berekenen.

Tabel 4: Correlaties tussen dummyvariabelen voor onderwijsniveau

	D_1	D_2	D_3
D_1	1.000		
D_2	-0.077	1.000	
D_3	-0.115	-0.406	1.000

Alle coëfficiënten zijn significant bij $\alpha = 1\%$.

Zoals Tabel 4 toont zijn de drie dummyvariabelen significant gecorreleerd met elkaar. Bij eenvoudige dummycodering is het normaal dat de dummyvariabelen negatief met elkaar gecorreleerd zijn. Namelijk, wanneer de score op één dummyvariabele 1 is dan moet die op de andere 0 zijn, wat voor een negatieve correlatie zorgt. Deze negatieve correlatie is vooral sterk tussen de grote categorieën. Zoals ook blijkt uit Tabel 4, waar de correlatie tussen D_2 en D_3 -0.41 is, terwijl de correlaties met D_1 maximaal -0.12 zijn. In dit geval lijkt er geen gevaar voor multicollineariteit te zijn.

Er is geen verschil tussen de behandeling van dummyvariabelen en die van andere dichotome en continue variabelen in regressiemodellen. Tabel 5 geeft de resultaten van een lineaire regressieanalyse met etnocentrisme als afhankelijke variabele en de drie onderwijsniveau dummyvariabelen onder de onafhankelijke variabelen.

Tabel 5: Regressie resultaten voor etnocentrisme op geslacht, politieke oriëntatie, belang van god, en onderwijsniveau

Variabele	*b* (*β*)
Constante	-0.005
Geslacht (man = 1)	0.010
	(0.003)
Politieke oriëntatie	0.123***
	(0.154)
Belang van god	0.014
	(0.027)
Onderwijsniveau (ref: Hoger onderwijs)	
Lager onderwijs	0.704**
	(0.063)
Lager middelbaar onderwijs	0.413***
	(0.104)
Hoger middelbaar onderwijs	0.354***
	(0.105)
R^2	0.043***

significantie: *: 0.050; **: 0.010; ***: 0.001

Deze resultaten tonen aan dat respondenten met lager onderwijs, lager middelbaar of hoger middelbaar onderwijs gemiddeld significant hogere scores op de etnocentrisme schaal dan deze met hoger onderwijs.

Daar er ofwel slechts één van de dummyvariabele een score van 1 heeft, of bij respondenten die tot de referentiecategorie behoren alle dummyvariabelen een score van 0 hebben, kan men afzonderlijke regressievergelijkingen schrijven voor elk van de onderwijsniveaus. Als ETNOC staat voor een respondent's score op de etnocentrisme schaal, GESLACHT voor de score op de geslacht variabele, POLORIENT voor de politieke oriëntatie en BELGOD voor het belang van god, dan kunnen we de regressievergelijkingen voor de verschillende onderwijsniveaus schrijven als:

Lager onderwijs:

$$ETNOC = (-.005 + .704) + .010 \times GESLACHT + .123 \times POLORIENT + .014 \times BELGOD$$
$$= .669 + .010 \times GESLACHT + .123 \times POLORIENT + .014 \times BELGOD$$

Lager middelbaar onderwijs:

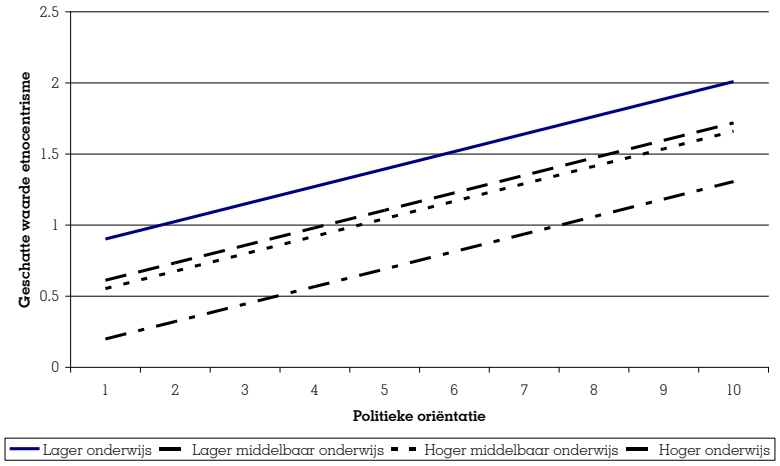
$$ETNOC = (-.005 + .413) + .010 \times GESLACHT + .123 \times POLORIENT + .014 \times BELGOD$$
$$= .408 + .010 \times GESLACHT + .123 \times POLORIENT + .014 \times BELGOD$$

Hoger middelbaar onderwijs:

$$ETNOC = (-.005 + .354) + .010 \times GESLACHT + .123 \times POLORIENT + .014 \times BELGOD$$
$$= .349 + .010 \times GESLACHT + .123 \times POLORIENT + .014 \times BELGOD$$

Hoger onderwijs:

$$ETNOC = -.005 + .010 \times GESLACHT + .123 \times POLORIENT + .014 \times BELGOD$$



Figuur 3: Regressielijnen voor etnocentrisme op politieke oriëntatie bij verschillende onderwijsniveaus en controlerend voor geslacht en belang van god

Zoals hier getoond wordt verschillen de regressievergelijkingen voor de verschillende onderwijsniveaus alleen maar in hun constante. In de vergelijking voor de referentiecategorie (hoger onderwijs) bestaat de constante alleen maar uit de constante zoals die in de regressie geschat werd. Voor de andere onderwijscategorieën, daarentegen bestaat de constante uit twee delen: de constante zoals ze geschat werd in de regressie analyse plus de regressiecoëfficiënt voor de dummyvariabele die de categorie vertegenwoordigde. Bijvoorbeeld, $b_{Constante}$ in de vergelijking was -0.005 en b_{D3} was 0.354, wat maakt dat de constante voor respondenten met hoger middelbaar onderwijs gelijk is aan -0.005 + 0.354 = 0.349. Als men dit grafisch voorstelt zoals in Figuur 3, waar de geschatte waarden voor etnocentrisme bij de verschillende waarden voor politieke oriëntatie en bij de verschillende onderwijsniveaus getoond worden, terwijl men het effect van geslacht en belang van god constant houdt, merkt men dat de dummyvariabelen een set van parallelle regressieruimten genereren. Alleen het snijpunt met de *Y*-as verschilt in deze regressies. Alle andere effecten zijn identiek voor de verschillende onderwijsniveaus, wat maakt dat de regressieruimten voor de verschillende onderwijsniveaus parallel lopen.

Toetsen voor sets van dummyvariabelen

In de vorige sectie werden de dummyvariabelen behandeld als gewone aparte variabelen. Maar dat zijn ze natuurlijk niet. Men moet ze zien als verschillende indicatoren van één enkele onderliggende variabele, de oorspronkelijke categorische variabele die ze vertegenwoordigen. Om, bijvoorbeeld, in het bovenstaande voorbeeld na te gaan of onderwijsniveau een effect heeft op etnocentrisme is het niet voldoende om te toetsen of één of meerdere van de

regressiecoëfficiënten voor de dummyvariabelen significant verschillen van nul. Men moet eerst toetsen of de dummyvariabelen als set al dan niet een significante bijdrage leveren aan de verklaarde variantie of de fit van het model.

In het voorbeeld waren de regressiecoëfficiënten voor de drie onderwijs-dummyvariabelen allen significant. Dit betekent dat respondenten met lager onderwijs, lager middelbaar onderwijs, of hoger middelbaar onderwijs, allen significant hoger op etnocentrisme neigden te scoren dan respondenten met hoger onderwijs. Het betekent echter nog niet automatisch dat men ook kan stellen dat onderwijs een effect heeft op etnocentrisme. Om dit laatste te toetsen moet men nagaan of de set dummyvariabelen voor onderwijs als een geheel iets significant bijdraagt aan de verklaring van etnocentrisme. Dit kan men doen door de verandering in de determinatiecoëfficiënt (R^2) veroorzaakt door de set dummyvariabelen te toetsen. De vraag hierbij is of onderwijs na controle voor de andere variabelen in het model nog een significant deel van de variantie in etnocentrisme verklaart.

Hypothese

De nulhypothese voor deze toets is dat de set dummyvariabelen geen extra deel van variantie in de afhankelijke variabele verklaart, of m.a.w. dat de verandering in de determinatiecoëfficiënt (ΔR^2) gelijk is aan nul. De alternatieve hypothese is dat ΔR^2 wel groter is dan nul. Dit kan men samenvatten als:

$$H_0: \Delta R^2 = 0$$

$$H_1: \Delta R^2 > 0$$

In het voorbeeld betekent dit dat de nulhypothese is dat ΔR^2 wanneer de 3 dummyvariabelen voor onderwijs aan het model toegevoegd worden gelijk is aan nul, terwijl de alternatieve hypothese stelt dat deze ΔR^2 wel significant verschilt van nul.

Steekproevenverdeling

De steekproevenverdeling, is net zoals bij de toets voor R^2 zelf (de ANOVA toets), een *F*-verdeling, waarbij het aantal vrijheidsgraden voor de teller (df_{teller}) gelijk is aan het aantal dummyvariabelen ($k - 1$) en het aantal vrijheidsgraden voor de noemer (df_{noemer}) gelijk is aan het aantal observaties (n) min het aantal variabelen reeds in het model (m), min het aantal dummyvariabelen in de set ($k - 1$) min 1, of $df_{noemer} = n - m - (k - 1) - 1 = n - m - k$. De steekproevenverdeling is dus:

$$F(k - 1, n - m - k)$$

waar k: aantal categorieën in categorische variabele, n: steekproefgrootte, en m: aantal andere variabelen in het model.

In het voorbeeld zijn er vier categorieën in de onderwijsniveau variabele ($k = 4$), zijn er $n = 1817$ observaties in de dataset, en waren er reeds drie variabelen opgenomen in de analyse ($m = 3$). Dit leidt tot de volgende steekproevenverdeling:

$$F(4 - 1 = 3, 1817 - 3 - 4 = 1810)$$

Kritieke waarde

De kritieke waarde voor deze toets wordt bepaald door enerzijds de gekozen Type 1 fout (het α-niveau) en anderzijds de vrijheidsgraden van de steekproevenverdeling. Toetsen voor een

F-toets zijn steeds eenzijdig. Als men in het voorbeeld een α = 5% hanteert, dan is de kritieke waarde $F_\alpha$voor een *F*-verdeling met 3 en 1810 vrijheidsgraden: F_α = 2.61.

Toetsstatistiek

De gebruikte toetsstatistiek is een variant op de gebruikelijke ANOVA toets. In deze versie vergelijken we de schatting van de populatievariantie gebaseerd op de door de dummyvariabelen extra verklaarde variantie met de schatting van de populatievariantie gebaseerd op de residuele variantie. Indien de dummyvariabelen niet meer dan op basis van willekeur verwacht kon worden bijdragen aan de verklaarde variantie, moeten deze twee schattingen gelijk zijn, en hun ratio gelijk zijn aan 1. Indien de dummyvariabelen wel een unieke bijdrage tot de verklaarde variantie hebben dan zal de ratio van de twee schattingen groter dan 1 zijn. De toetsstatistiek kan geschreven worden als:

$$F = \frac{\frac{R_n^2 - R_v^2}{k-1}}{\frac{1-R_n^2}{n-m-k}} = \frac{\frac{\Delta R^2}{k-1}}{\frac{1-R_n^2}{n-m-k}}$$

waar R_n^2 en R_v^2 zijn respectievelijk de determinatiecoëfficiënten voor de regressiemodellen met de set dummyvariabelen en zonder hen. Het verschil tussen beide is de unieke bijdrage van de dummyvariabelenset (ΔR^2). Deze verandering in de R^2 wordt gedeeld daar het aantal dummyvariabelen (*k* – 1). De noemer wordt gevormd door de residuele variantie (of de aliënatiecoëfficiënt) voor het model met de dummyvariabelen inbegrepen, gedeeld door het residuele aantal vrijheidsgraden.

Als men dit toepast op het voorbeeld, waar de determinatiecoëfficiënt van het model zonder de onderwijsniveau dummyvariabelen 2.9% is, en met deze dummyvariabelen 4.3% en onderwijs dus een uniek verklarend effect heeft op etnocentrisme van 1.4%, wordt deze toets:

$$F = \frac{\frac{R_n^2 - R_v^2}{k-1}}{\frac{1-R_n^2}{n-m-k}} = \frac{\frac{.043 - .029}{3}}{\frac{1-.043}{1810}} = 8.83$$

Conclusie

Indien de gevonden toetsstatistiek *F* groter is dan de kritieke waarde F_α dan verwerpt men de nulhypothese en aanvaardt men de alternatieve hypothese. Is *F* kleiner dan of gelijk aan F_α dan aanvaardt men de nulhypothese en besluit men dat de set dummyvariabelen niet bijdragen aan het verklaren van de afhankelijke variabele. Samengevat wordt dit:

$F \leq F_\alpha$: aanvaard H_0

$F > F_\alpha$: verwerp H_0, aanvaard H_1

In het voorbeeld hier is *F* = 8.83 en F_α= 2.61, en is de toetsstatistiek dus beduidend groter dan de kritieke waarde en moet men besluiten de nulhypothese te verwerpen en de alternatieve hypothese te aanvaarden. Het significantieniveau geassocieerd met deze *F* waarde, *p* = 0.0000082319. Men kan hier dus besluiten dat onderwijsniveau wel degelijk een effect heeft op etnocentrisme.

Effect dummycodering

Hoewel de eenvoudige dummycodering het meest voorkomende coderingschema is voor dummyvariabelen, zijn er natuurlijk nog een groot aantal andere mogelijke coderingschema's. Een schema dat ook relatief vaak gebruikt wordt is effect dummycodering. Waar bij de eenvoudige dummycodering men refereerde naar één bepaalde referentiecategorie, wordt bij effect dummycodering het gemiddelde van de groepen als referentie genomen, en zijn de geschatte effecten van de dummyvariabelen afwijkingen van dit gemiddelde groepsgemiddelde.

Werkwijze

De werkwijze om een set effect gecodeerde dummyvariabelen aan te maken is bijna identiek aan deze gevolgd voor de eenvoudige dummycodering. Als men hetzelfde voorbeeld neemt als in sectie 0waar men een vier-categorie godsdienst variabele had die omgezet diende te worden naar dummyvariabelen, dan begint men weer met het kiezen van een referentiecategorie. De regels voor het kiezen van een referentiecategorie zijn identiek aan die bij eenvoudige dummycodering (zie sectie 0). Men kan hier weer dezelfde referentiecategorie kiezen, namelijk "christen". Het aanmaken van de dummyvariabelen is bijna identiek als bij de eenvoudige dummycodering, behalve dan dat de referentiecategorie geen score van 0 krijgt op alle dummyvariabelen, maar een score van -1. Het effect van de referentiecategorie is hier dan ook -1 maal de som van de effecten van alle dummyvariabelen, en geeft aan hoever de observaties in de referentiecategorie gemiddeld afwijken van het gemiddelde van de groepen. Voor deze godsdienstvariabele wordt dit dus:

	D_1	D_2	D_3
Christen	-1	-1	-1
Moslim	1	0	0
Geen	0	1	0
Ander	0	0	1

Drie dummyvariabelen werden aangemaakt, D_1, D_2 en D_3. Elk van deze dummyvariabelen vertegenwoordigd één van de categorieën van de godsdienstvariabele, D_1 de "moslim" categorie, D_2 "geen" en D_3 "ander". Observaties in de "moslim" categorie scoren dan ook een 1 op D_1 en een 0 op D_2 en D_3. Dit was ook zo bij eenvoudige dummycodering. Het verschil zit in de codering voor de referentiecategorie. De observaties in de referentiecategorie "christen" krijgen scores van -1 op alle drie de dummyvariabelen.

Voorbeeld 1

Als men effect dummycodering toepast op de onderwijsvariabele in het voorbeeld op de Belgische steekproef van de *WVS* krijgt men de codering in Tabel 6. Opnieuw werden 3 dummyvariabelen aangemaakt D_1, D_2, en D_3 en ook bij deze werd "hoger onderwijs" als referentiecategorie gekozen. Dummyvariabele D_1 vertegenwoordigt het "lager onderwijs" vs. het gemiddelde van de groepen, D_2 het "lager middelbaar onderwijs", en D_3 het "hoger middelbaar onderwijs". De referentiecategorie "hoger onderwijs" scoorde -1 op alle drie dummyvariabelen.

Tabel 6: Effect dummycoderingschema voor onderwijsniveau

Onderwijsniveau	D_1	D_2	D_3
Lager onderwijs	1	0	0
Lager middelbaar onderwijs	0	1	0
Hoger middelbaar onderwijs	0	0	1
Hoger onderwijs	-1	-1	-1

Weer kan men controleren of de creatie van de dummyvariabelen correct verliep door de frequentieverdelingen van de dummyvariabelen te vergelijken met de frequentieverdeling van

de categorische variabele (zie Tabel 1). Tabel 7 geeft de frequentieverdelingen weer voor de 3 onderwijsniveau dummyvariabelen. Merk op dat voor alle drie de dummyvariabelen het aantal observaties met een -1 score gelijk moet zijn aan het aantal observaties in de referentiecategorie (hier “hoger onderwijs”). Het aantal observaties met een score 1 op de dummyvariabele moet overeenstemmen met het aantal observaties in de categorie geassocieerd met deze dummyvariabele, terwijl het aantal observaties met een score 0 gelijk moet zijn aan het aantal observaties in de categorieën geassocieerd met de andere dummyvariabelen.

Tabel 7: Frequentietabel voor de dummyvariabelen voor onderwijsniveau

	F(-1)	*F*(0)	*F*(1)	*N*
D_1 Lager onderwijs	705	1073	39	1817
D_2 Lager middelbaar onderwijs	705	722	390	1817
D_3 Hoger middelbaar onderwijs	705	429	683	1817

Tabel 8: Correlaties tussen dummyvariabelen voor onderwijsniveau

	D_1	D_2	D_3
D_1	1.000		
D_2	0.817	1.000	
D_3	0.837	0.584	1.000

Alle coëfficiënten zijn significant bij $\alpha = 1\%$.

De correlatie tussen de drie dummyvariabele is hier niet alleen duidelijk sterker dan bij de eenvoudige dummycodering maar ze is ook positief. Dit komt omdat men een grote referentiecategorie koos, waardoor alle observaties in de referentiecategorie gematched zijn op alle drie dummyvariabelen en dusdanig een sterk positieve bijdrage leveren aan de correlatie tussen de dummyvariabelen. Hier dient men dus de mogelijkheid van multicollineariteitsproblemen ernstig te nemen.

In tegenstelling tot de regressiecoëfficiënten van dummyvariabelen met eenvoudige dummycodering, blijken geen van de drie dummyvariabelen met effect dummycodering een significant effect te hebben op etnocentrisme.

Tabel 9: Regressie resultaten voor etnocentrisme op geslacht, politieke oriëntatie, belang van god, en onderwijsniveau

Variabele	*b* (β)
Constante	-0.005
Geslacht (man = 1)	0.010
	(0.003)
Politieke oriëntatie	0.123***
	(0.154)
Belang van god	0.014
	(0.027)
Onderwijsniveau (ref: Hoger onderwijs)	
Lager onderwijs	0.336
	(0.108)
Lager middelbaar onderwijs	0.045
	(0.021)
Hoger middelbaar onderwijs	-0.014
	(-0.007)
R^2	0.043***

significantie: *: 0.050; **: 0.010; ***: 0.001

Niettemin, indien men weer de regressievergelijkingen uitschrijft voor elk van de vier onderwijsniveaus, dan krijgt men de volgende vergelijkingen:

Lager onderwijs of missing:

$$ETNOC = (.363 + .336) + .010 \times GESLACHT + .123 \times POLORIENT + .014 \times BELGOD$$
$$= .669 + .010 \times GESLACHT + .123 \times POLORIENT + .014 \times BELGOD$$

Lager middelbaar onderwijs:

$$ETNOC = (.363 + .045) + .010 \times GESLACHT + .123 \times POLORIENT + .014 \times BELGOD$$
$$= .408 + .010 \times GESLACHT + .123 \times POLORIENT + .014 \times BELGOD$$

Hoger middelbaar onderwijs:

$$ETNOC = (.363 - .014) + .010 \times GESLACHT + .123 \times POLORIENT + .014 \times BELGOD$$
$$= .349 + .010 \times GESLACHT + .123 \times POLORIENT + .014 \times BELGOD$$

Hoger onderwijs:

$$ETNOC = (.363 - .336 - .045 + .014) + .010 \times GESLACHT + .123 \times POLORIENT + .014 \times BELGOD$$
$$= -.004 + .010 \times GESLACHT + .123 \times POLORIENT + .014 \times BELGOD$$

Deze vergelijkingen zijn identiek aan diegene die men bekwam met de eenvoudige dummycodering. Alhoewel men dus verschillende regressiecoëfficiënten verkrijgt voor de twee coderingschema's leiden ze toch tot dezelfde vergelijkingen. Dit komt omdat de verschuivingen in de dummyvariabelen gecompenseerd worden door verschuivingen in de constante. Voor het voorspellen van scores op de afhankelijke variabele doet het er dus niet toe welke dummycodering gebruikt wordt, alle leiden tot dezelfde geschatte waarden.

Vergelijking van gewone en effect codering

In Tabel 10 worden de resultaten van de eenvoudige en effect dummycodering naast elkaar geplaatst. Zoals je merkt maakt de gekozen dummycodering niets uit voor de effecten van de andere variabelen in het model. Alleen de constante wordt beïnvloed door de keuze van het dummycoderingschema. De verklaarde variantie zowel van het ganse model als de unieke bijdrage door de set dummyvariabelen is identiek in beide modellen, en significant in beide gevallen ondanks dat bij de effect dummycodering geen van de dummyvariabelen zelf significant is. Dit brengt mee dat men de significantie van een categorische variabele in een regressiemodel nooit kan bepalen aan de hand van de toetsen voor de regressiecoëfficiënten van de dummyvariabelen, maar dat men steeds een toets moet doen voor de bijdrage aan de verklaarde variantie, de ΔR^2, voor de ganse set dummyvariabelen.

Tabel 10: Vergelijking van resultaten met eenvoudige en effect dummycodering

Etnocentrisme	***Eenvoudige dummy-codering***	***Effect dummy-codering***
Constante	-0.01	0.36*
Geslacht	0.01	0.01
Politieke orientatie	0.12***	0.12***
Belang van god	0.01	0.01
Dummy Lager onderwijs	0.70**	0.34
Dummy Lager middelbaar onderwijs	0.41***	0.05
Dummy Hoger middelbaar onderwijs	0.35***	-0.01
R^2	0.043***	0.043***
ΔR^2 (onderwijs)	0.014***	0.014***

significantie: * p < .050, ** p < .010, *** p < .001

Ook de geschatte waarden zijn, zoals reeds vermeld, identiek voor beide. Het enige verschil tussen de twee modellen is de referentie die gebruikt wordt voor de dummyvariabelen. In de eenvoudige dummycodering vergelijkt men het effect van de andere categorieën met een zelf gekozen referentiecategorie (hier "hoger onderwijs"), terwijl men in de effect dummycodering

ze parametriseert als de afwijkingen van het groepsgemiddelde van het gemiddelde van de groepen. De keuze van het coderingschema en de referentiecategorie kan een grote invloed uitoefenen op de schattingen van de regressiecoëfficiënten voor de dummyvariabelen. Men moet zich dan ook altijd goed bewust zijn van welke vergelijkingen men maakt. De keuze van zowel het coderingschema als de referentiecategorie dient dan ook met de grootste zorg te gebeuren.

Contrast codering

Algemeen

Een andere familie van coderingschema's is wat men contrast coderingen noemt. Dit een zeer flexibele manier van dummycodering maar moet zeer zorgvuldig gebeuren en vergt een goed inzicht in welke vergelijkingen (contrasten) men wil maken. Het is heel gemakkelijk fouten te maken met deze methoden.

Met behulp van contrast codering kan men bepaalde hypothesen over het effect van een categorische variabele op een continue afhankelijke variabele toetsen, en dit door middel van 'geplande vergelijkingen'. De toets voor deze hypothesen wordt dan uitgevoerd door middel van de regressiecoëfficiënt voor de bijhorende dummyvariabele, in plaats van door een algemene *F*-toets. Deze methode kan gebruikt worden zowel om bepaalde trends te toetsen als om post-hoc vergelijkingen te maken.

Om een hypothese te toetsen door contrast codering maakt men gebruik van wat men noemt contrast gewichten. Deze zijn positieve of negatieve getallen die aan de te vergelijken groepen toegekend worden. Het teken duidt aan welke groepen vergeleken worden. De regel is dat de som van alle contrast gewichten over alle categorieën gelijk dient te zijn aan nul. Bijvoorbeeld, neem nu de vier categorieën onderwijsniveau. Veronderstel dat men wil toetsen of de gemiddelde score op de etnocentrisme schaal verschilt tussen respondenten met alleen maar lager onderwijs en zij met hoger onderwijs (H_1), of tussen diegenen met lager middelbaar onderwijs en diegenen met hoger middelbaar onderwijs (H_2), of dat respondenten met lager, lager middelbaar of hoger middelbaar gezamenlijk verschillen van deze met hoger onderwijs (H_3).

	Lager	Lager middel-baar	Hoger middel-baar	Hoger
H_1	+1	0	0	-1
H_2	0	+1	-1	0
H_3	+1	+1	+1	-3

Hypothese H_1 kan men toetsen door een dummyvariabele aan te maken waar respondenten in de categorie "lager onderwijs" een score 1 krijgen en deze in de categorie "hoger onderwijs" een -1 (of omgekeerd). De respondenten in de andere categorieën krijgen een 0. Als het contrast maar twee niet-nul categorieën betreft spreekt men van een eenvoudig contrast, indien er meer dan twee categorieën bij betrokken zijn noemt men het een complex contrast. Het toetsen van hypothese H_2 vraagt ook om een eenvoudig contrast, met een dummyvariabele waarop "lager middelbaar onderwijs" 1 scoort, "hoger middelbaar" -1 (of omgekeerd), en de rest 0. Het toetsen van de derde hypothese H_3 vraagt om een complex contrast. Op deze dummy scoren de respondenten in de "lager onderwijs", "lager middelbaar onderwijs", en "hoger middelbaar onderwijs" allen een 1, en dan moeten de respondenten in de "hoger onderwijs" categorie noodzakelijk -3 scoren, daar de som van de contrast gewichten over alle categorieën steeds nul moet zijn.

Orthogonale en niet-orthogonale contrasten

Met deze methode kan men maximaal *k* – 1 hypothesen toetsen, waar *k* het aantal categorieën in de categorische variabele is. Men kan steeds maximaal maar *k* – 1 dummyvariabelen aanmaken voor een enkele categorische variabele.

Men maakt ook een onderscheid tussen orthogonale en niet-orthogonale contrasten en vergelijkingen. De vergelijkingen zijn orthogonaal indien alle contrasten orthogonaal, d.w.z. onafhankelijk, zijn van elkaar. Of twee contrasten orthogonaal zijn of niet kan nagegaan worden door de som van de kruisproducten van de gewichten van de twee contrasten te berekenen. Indien deze som gelijk is aan nul dan zijn de contrasten orthogonaal, indien niet dan zijn ze niet-orthogonaal. In het voorbeeld is de som van de kruisproducten voor hypothesen H_1 en H_2:

$$(+1)(0) + (0)(+1) + (0)(-1) + (-1)(0) = 0$$

en kan men besluiten dat de twee contrasten orthogonaal zijn. Voor het paar H_1 en H_3 wordt dit:

$$(+1)(+1) + (0)(+1) + (0)(+1) + (-1)(-3) = 4$$

wat betekent dat deze twee contrasten niet-orthogonaal zijn. Voor het resterende paar contrasten (H_2 en H_3) is de som van de kruisproducten:

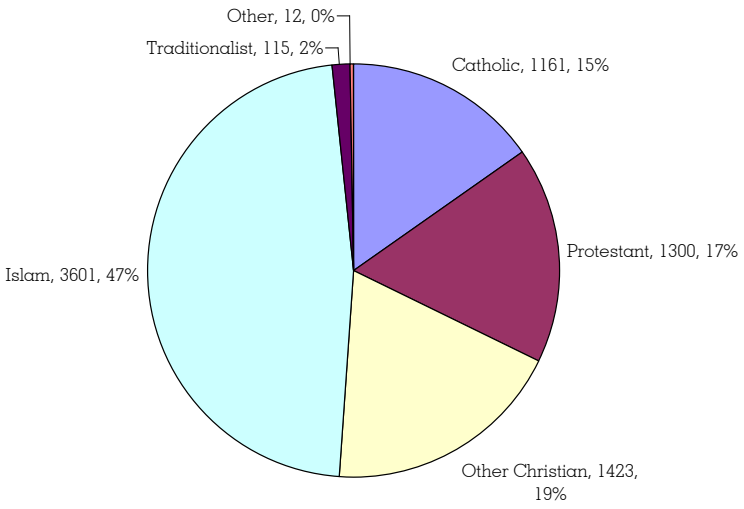
$$(0)(+1) + (+1)(+1) + (-1)(+1) + (0)(-3) = 0$$

en deze twee zijn dan ook orthogonaal. Twee van de drie paren zijn dus orthogonaal maar het derde ook niet. Daardoor wordt ook het ganse stelsel van hypothesen niet-orthogonaal.

Indien de geplande vergelijkingen orthogonaal zijn kan men de significantietoets uitvoeren met de toets voor de regressiecoëfficiënt voor de dummyvariabele corresponderend met de vergelijking. Indien het stelsel echter niet-orthogonaal is onderschat men de Type I fout –daar de resultaten van de vergelijkingen niet onafhankelijk van elkaar zijn—en kan men overwegen om toetsen voor meervoudige vergelijkingen te gebruiken (bv. Bonferroni correcties).

Voorbeeld 2

Een tweede voorbeeld komt uit de *Nigeria Demographic and Health Survey* (DHS+) van 2003 (National Population Commission (NPC) [Nigeria] & ORC Macro, 2004). Alleen de vrouwelijke substeekproef werd hierbij gebruikt. De afhankelijke variabele is hier de kennis van SOA-symptomen (zowel bij mannen als bij vrouwen) die de respondenten hebben. Deze variabele (STDSYMP) werd aangemaakt als het aantal symptomen uit een lijst van 26 die de respondenten kenden. Als onafhankelijke variabelen worden gebruikt de leeftijd van de respondent (V012), de plaats van residentie (urbaan = 1, ruraal = 2) (V102), de burgerlijke staat van de respondent (MARRIED) (gehuwd/samenwonend = 1, andere = 0), en de godsdienst van de respondent (V130). Deze laatste is een categorische variabele met 6 categorieën: katholiek, protestant, andere christenen, moslims, traditionele godsdiensten en andere. De frequentieverdeling voor deze variabele wordt getoond in Figuur 4.



Figuur 4: Frequentieverdeling voor godsdienst in de Nigeria DHS+ 2003

In het totaal zijn er 7612 observaties met geldige gegevens voor deze variabelen en 8 met ontbrekende gegevens. De grootste religieuze groep in deze steekproef zijn de moslims die bijna de helft van de steekproef uitmaken. De christelijke kerken maken samen echter juist iets meer dan de helft van de steekproef uit. Om deze godsdienstvariabele in een regressieanalyse te kunnen gebruiken moet die omgezet worden in een set dummyvariabelen. Als referentiecategorie is “moslim” gekozen daar dit de grootste categorie is met 3601 respondenten uit 7612 of 47.3%.

Dummycodering in SPSS

In tegenstelling tot andere SPSS procedures maak de REGRESSION procedure niet automatisch dummyvariabelen aan[1]. Men moet dus de dummyvariabelen manueel aanmaken. In deze sectie wordt gedemonstreerd hoe zowel eenvoudige dummycodering als effect dummycodering kunnen uitgevoerd worden in SPSS, gebruikmakend van de syntaxis en van de dialoogvensters.

Eenvoudige dummycodering

Godsdienst in de Nigeria 2003 DHS+ is een nominale variabele met 6 categorieën. Dit betekent dat indien men deze variabele in een regressieanalyse wil betrekken men deze moet vervangen door een set van maximaal 5 (aantal categorieën min één) dummyvariabelen. Daar echter de categorie “ander” maar 12 observaties (0.2%) bevat en de categorie “traditionele godsdienst” er ook maar 115 (1.5%) bevat werd besloten deze twee categorieën samen te voegen. Dit betekent dat er niet vijf maar slechts vier dummyvariabelen voor deze categorische variabele dienen aangemaakt te worden. Bij een eenvoudige dummycodering en met de categorie “moslim” als referentiecategorie, wordt het coderingschema:

Tabel 11: Eenvoudig dummycoderingschema voor godsdienstvariabele in Nigeria 2003 DHS+

	REL 1	REL 2	REL 3	REL 4
Catholic	1	0	0	0
Protestant	0	1	0	0
Other Christian	0	0	1	0

[1] Zelfs bij procedures die zelf dummyvariabelen aanmaken moet men nog steeds weten welke dummycoderingschema men hanteert en welke referentiecategorie men kiest. Bij alle procedures kunnen ook—zoals hier gedemonstreerd—zelf aangemaakte dummyvariabelen gebruikt worden.

Islam	0	0	0	0
Traditionalist	0	0	0	1
Other	0	0	0	1

Vier dummyvariabelen: REL1, REL2, REL3 en REL4 worden aangemaakt. REL 1 staat voor "katholiek" vs. "moslim", REL2 voor "protestant" vs. "moslim", REL3 voor "ander christelijk" vs. "moslim", en REL4 voor "traditionele godsdienst" en "ander" vs. moslim. Katholieke respondenten scoren dus 1 op REL1 en 0 op de drie andere dummyvariabelen. Leden van de referentiecategorie "moslims" scoren 0 op alle vier dummyvariabelen.

Het aanmaken van de dummyvariabelen gebeurt in twee stappen. In een eerste stap worden de verschillende dummyvariabelen aangemaakt en voor alle waarden een score van 0 gegeven. Wanneer men de SPSS syntaxis gebruikt kan men dit doen door middel van de COMPUTE instructie. De COMPUTE instructie laat de gebruiker toe nieuwe variabelen aan te maken en de waarde van reeds bestaande variabelen te wijzigen. De syntaxis van de COMPUTE instructie is als volgt:

COMPUTE instructie

COMPUTE *variabele = uitdrukking*.

Het sleutelwoord COMPUTE wordt gevolgd door de naam van de *variabele* die men gaat aanmaken of waarvan men de waarde wil wijzigen, dan het "is gelijk aan" teken, en dan de *uitdrukking* die de waarde van de nieuwe variabele bepaalt. Een dergelijke uitdrukking kan zowel constanten, functies, operatoren als andere variabelen bevatten. Wanneer een bestaande variabele gewijzigd wordt worden alle vroegere waarden voor deze variabele overschreven en zijn ze verloren.

```
compute rel1=0.
compute rel2=0.
compute rel3=0.
compute rel4=0.
```

In dit geval worden vier nieuwe variabelen aangemaakt (REL1, REL2, REL3 en REL4) en elk van deze variabelen krijgt voor alle observaties een waarde 0. In een tweede stap worden dan de waarden voor de betreffende categorieën op de overeenstemmende dummyvariabelen op 1 gezet. Dit kan men doen door middel van een conditionele transformatie, de IF instructie. Bij de IF instructie wordt een bepaalde transformatie alleen maar uitgevoerd indien de gestelde conditie vervuld is. De syntaxis voor de IF instructie is:

IF instructie

IF *conditie variabele = uitdrukking*.

waar *conditie* de voorwaarde is die vervuld moet zijn vooraleer de *variabele* de waarde van *uitdrukking* krijgt. Wanneer *conditie* niet vervuld is behoudt de variabele zijn vroegere waarde. De IF instructie kan ook gebruikt worden om nieuwe variabelen aan te maken. In dit geval krijgen de observaties waarvoor de *conditie* niet vervuld was een systeem-ontbrekende waarde. Bij de constructie van de dummyvariabelen moet afhankelijk van de waarde op de godsdienstvariabele V130 één van de dummyvariabelen een waarde 1 krijgen. Bijvoorbeeld indien een respondent katholiek is en dus de waarde 1 scoort op V130 dient hij of zij ook op de corresponderende dummyvariabele REL1 een waarde 1 te krijgen. Is de respondent echter lid van een "andere christelijke" kerk (V130 = 3) dan dient hij of zij een 1 te scoren op REL3. De code hiervoor vereist is:

```
if v130=1 rel1=1.
if v130=2 rel2=1.
if v130=3 rel3=1.
if v130=5 rel4=1.
if v130=6 rel4=1.
```

Voor elk van de categorieën in de godsdienstvariabele V130 krijgt om beurt de corresponderende dummyvariabele een score 1. Een nadeel van deze werkwijze is dat ook observaties met ontbrekende data op de categorische variabele een score 0 krijgen op alle dummyvariabelen en dus niet kunnen onderscheiden worden van de referentiecategorie. Verschillende oplossingen voor dit probleem zijn mogelijk. Hier wordt gekozen om respondenten met ontbrekende gegevens voor godsdienst uit te sluiten van de analyse. Dit betekent dat deze observaties ook ontbrekende waarden moeten krijgen op alle dummyvariabelen. In de dataset werden de ontbrekende waarden op de godsdienst variabele aangeduid met een score 9, die in SPSS als ontbrekende waarde voor de godsdienstvariabele V130 gedefinieerd was. Men kan dit doen door een extra conditionele transformatie per dummyvariabele waarin observaties met ontbrekende gegevens voor V130 een aparte score krijgen (hier 9) op de dummyvariabele. Deze score zal later dan als ontbrekend voor de dummyvariabelen gedefinieerd worden. Om ontbrekende waarden te specificeren in een conditie kan men gebruik maken van de MISSING(*variabele*) functie. De MISSING VALUES instructie kan vervolgens gebruikt worden om de waarde die men aan de ontbrekende observaties gaf op de dummyvariabelen als ontbrekend te definiëren. De syntaxis van de MISSING VALUES instructie is:

MISSING VALUES instructie

MISSING VALUES *variabelenlijst* (*waarden*).

waarbij *variabelenlijst* een lijst met de namen van de variabelen waarvoor een ontbrekende waarde moet gedefinieerd worden is, en *waarden* het lijstje met waarden die SPSS als ontbrekend zal behandelen. Voor het voorbeeld geeft dit:

```
if missing(V130) rel1=9.
if missing(V130) rel2=9.
if missing(V130) rel3=9.
if missing(V130) rel4=9.
missing values rel1 rel2 rel3 rel4 (9).
```

REL1, REL2, REL3 en REL4 krijgen alle een score van 9 indien de score op V130 ontbreekt, en vervolgens wordt de score 9 voor de vier dummyvariabelen als ontbrekend gedefinieerd.

Wanneer men dummyvariabelen aanmaakt is ook wel handig om ook labels voor deze variabelen aan te maken zodat men ook later nog weet wat de betekenis van deze variabelen was. Variabelen kan men een variabele label geven met de VARIABLE LABELS instructie. De syntaxis van deze instructie is:

VARIABLE LABELS instructie

VARIABLE LABELS *variabele "label"*.

Met één VARIABLE LABELS instructie kunnen meerdere variabelen van een label voorzien worden. Bijvoorbeeld de labels voor de vier dummyvariabelen worden met één enkele instructie aangemaakt:

```
variable labels
 rel1 "Religion: Catholic"
 rel2 "Religion: Protestant"
 rel3 "Religion: Other Christian"
 rel4 "Religion: Traditionalist/Other".
```

Een kortere manier om hetzelfde resultaat te bekomen is door de RECODE instructie te gebruiken. De RECODE instructie laat de gebruiker toe om de waarden van een variabele te hercoderen en deze opslaan in een andere (nieuwe) variabele. De syntaxis voor de RECODE instructie is:

RECODE instructie

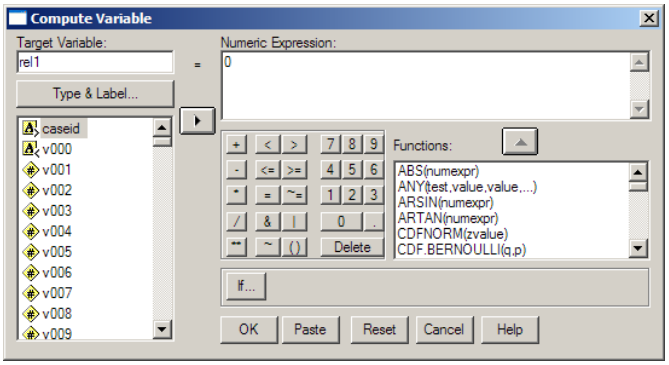
RECODE *oldvar* (*waardenlijst* = *waarde*) (*waardenlijst* = *waarde*)… (*waardenlijst* = *waarde*) INTO *newvar*.

waar *oldvar* de naam is van de variabele die men wenst te hercoderen. Dan volgen de lijstjes met de oude waarden die men wil hercoderen in de nieuwe waarden. Alle observaties met een waarde in *waardenlijst* krijgen de nieuwe *waarde*. Het INTO sleutelwoord duidt aan dat men de gehercodeerde waarden wil opslaan in een nieuwe variabele *newvar*. Indien *newvar* reeds bestaat wordt die overschreven. Om de vier dummyvariabelen aan te maken moet de oorspronkelijke categorische godsdienstvariabele V130 vier maal gehercodeerd worden.

```
recode v130 (1=1)(missing=9)(else = 0) into rel1.
recode v130 (2=1)(missing=9)(else = 0) into rel2.
recode v130 (3=1)(missing=9)(else = 0) into rel3.
recode v130 (5,6=1)(missing=9)(else = 0) into rel4.
missing values rel1 rel2 rel3 rel4 (9).
```

Op de eerste RECODE instructie wordt REL1 aangemaakt en krijgen de respondenten die een waarde 1 scoorden op V130 (“katholiek”) een score 1 op REL1 ‘(1 = 1)’ en alle anderen een 0 ‘(else = 0)’[2]. De tweede RECODE instructie genereert REL2 waarop alle “protestanten” een 1 scoren en alle anderen een 0. De derde instructie doet hetzelfde voor REL3 en de “andere christenen”, terwijl de laatste RECODE instructie REL4 aanmaakt en de categorieën “traditionele godsdienst” en “andere” een waarde 1 geven en de anderen een 0. Bij een RECODE instructie kan men gebruik maken van het MISSING sleutelwoord om ontbrekende waarden aan te duiden. Door ‘(missing=9)’ te specificeren krijgen alle observaties met een ontbrekende waarde op V130 een score van 9 op de dummyvariabelen. Een waarde die vervolgens door een MISSING VALUES instructie als ontbrekende gedefinieerd wordt.

Hetzelfde kan men bereiken door de menu’s en dialoogvensters te gebruiken. De nodige dialoogvensters zijn alle bereikbaar via het [Transform] menu[3]. De COMPUTE instructie wordt bereikt via het [Compute…] item, en dit opent het volgende dialoogvenster:



Figuur 5: Het COMPUTE dialoogvenster in SPSS

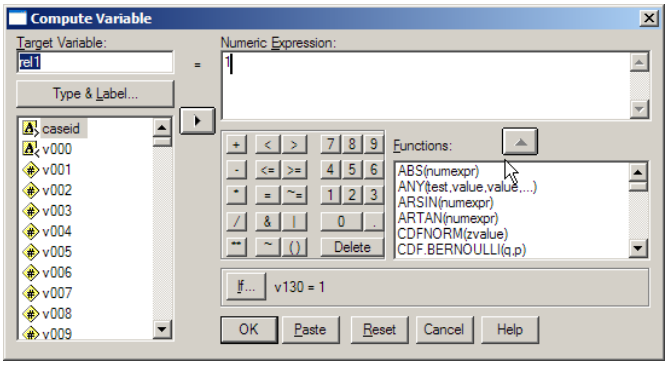
In het tekstvakje “Target Variable:” ① typt men de naam van de aan te maken of te wijzigen variabele. In dit geval is de aan te maken variabele de dummyvariabele REL1 die geassocieerd is met de categorie “katholiek”. In een eerste fase krijgen alle observaties een waarde 0. Het is in het “Numeric Expression:” vak ② dat de uitdrukking getypt wordt. Hier is deze uitdrukking gewoon de constante “0” maar men kan hier complexe formules aanmaken, door ofwel ze rechtstreeks in het “Numeric Expression” vak in te typen of door ze aan te maken met behulp van de variabele lijst ③, het toetsenbord ④, of de functielijst ⑤. Uit de variabelenlijst kan men variabelen selecteren en die in de uitdrukking in ② laten vallen. Dit doet men door op het pijltje tussen de variabelenlijst ③ en het “Numeric Expression” vak te drukken. Het toetsenbord laat toe om nummers en operatoren toe te voegen aan de uitdrukking, en vanuit

[2] Het ELSE sleutelwoord in de RECODE instructie refereert naar alle waarden die nog niet in een eerdere *waardenlijst* genoemd werden.

[3] In oudere versies van SPSS is dit menu alleen toegankelijk van uit het “SPSS Data Editor” venster.

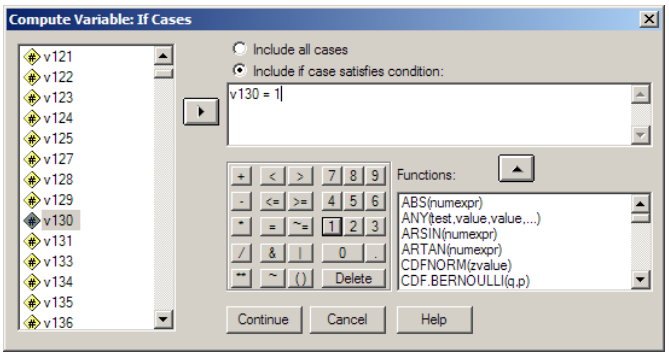
de functielijst kan men door middel van de knop boven de lijst functies toevoegen aan de uitdrukking.

In een tweede stap moeten de dummyvariabelen dan een waarde 1 krijgen voor de observatie in de geassocieerde categorie. De procedure is analoog aan diegene hierboven beschreven, behalve dan dat men er een conditie aan moet verbinden. De conditie voor een conditionele transformatie wordt getoond naast de toets [If…] ⑥. Deze conditie wordt aangemaakt door op deze [If…] knop te klikken en dat opent dan het IF dialoogvenster (zie Figuur 7).



Figuur 6: Het COMPUTE dialoogvenster in SPSS voor conditionele transformaties

Dit IF dialoogvenster gelijkt goed op het COMPUTE dialoogvenster en is er volledig op gericht een conditie aan te maken. Vooraleer men een conditie kan aanmaken dient men de optie om observaties te gebruiken als ze de conditie vervullen (“Include if case satisfies condition:”) ⑦ te selecteren. Wanneer men de optie alle observaties te gebruiken selecteert wordt de transformatie uitgevoerd op alle observaties en is het conditievak ⑧ niet toegankelijk. Weer kan men de conditie manueel invullen of aanmaken met behulp van de variabelenlijst ⑨, toetsenbord ⑩ en functielijst ⑪. In het voorbeeld wordt hier de conditie aangemaakt V130 = 1, dus dat REL1 een waarde 1 krijgt wanneer de observatie een waarde 1 heeft op variabele V130, d.w.z. als de respondent katholiek is.



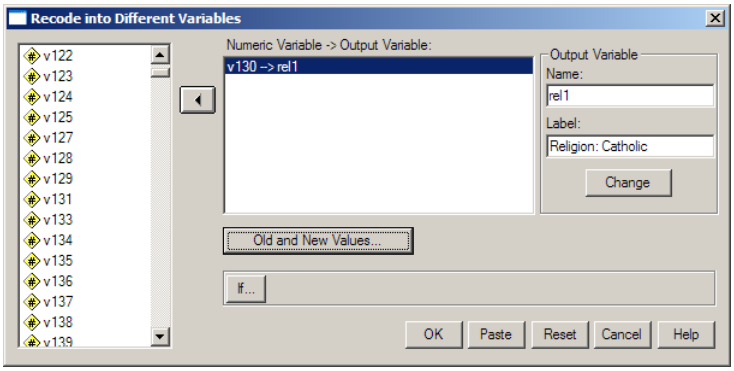
Figuur 7: Het IF dialoogvenster in SPSS

Een label voor de nieuwe variabele kan gedefinieerd worden door op de knop [Type & Label] te klikken in het COMPUTE dialoogvenster. Dit opent het volgende venster, wat je toelaat een nieuw label te specificeren of automatisch een label te laten aanmaken, zowel als te definiëren of de nieuwe variabele een karakter (“String”) of numerieke (“Numeric”) variabele is.

Figuur 8: Toevoegen van variabele labels

Hier werd alleen het voorbeeld voor de eerste dummyvariabele REL1 uitgewerkt. Het spreekt vanzelf dat dit herhaald dient te worden voor de andere dummyvariabelen.

Ook hercoderingen kunnen via dialoogvensters gebeuren. Kies in het [Transform] menu de optie [Recode ►] en vervolgens [Into Different Variables…] wat het dialoogvenster in Figuur 9 opent. Bij deze optie worden de gehercodeerde waarden opgeslagen in een nieuwe variabele. Kiest men de optie [Into Same Variabels…] dan overschrijven de nieuwe waarden de bestaande waarden van de variabele die gehercodeerd wordt.

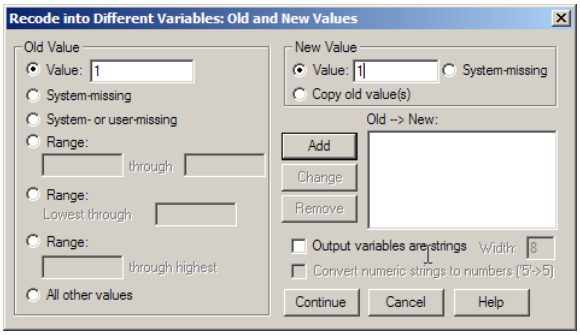


Figuur 9: Het RECODE (nieuwe variabele) dialoogvenster

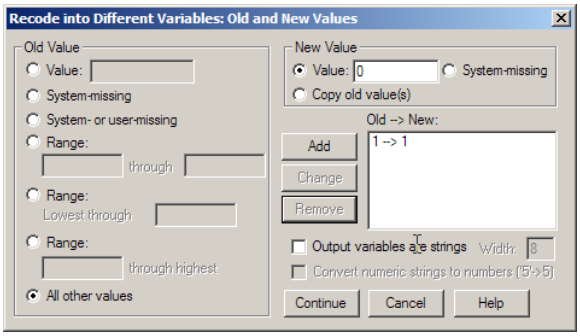
Dit eerste dialoogvenster laat toe de naam van de nieuwe variabele te specificeren. Vanuit de variabelenlijst ⑫ kan men de te hercoderen variabele selecteren en naar het middelste vak verplaatsen ⑯. Hier werd variabele V130 gekozen en nadat men op het pijltje naast de variabelenlijst geklikt heeft verschijnt er in het middelste vak: `V130 --> ?`. Het vraagteken duidt erop dat nog geen nieuwe variabele naam gedefinieerd werd voor de gehercodeerde waarden van V130. De naam en het label voor de nieuwe variabele kan gespecificeerd worden in het “Output Variable” deel van het dialoogvenster ⑬. Door op de [Change] knop te klikken worden de naam en label bevestigd en komt in het middelste vak te staan: `V130 --> REL1`.

Door op de knop [Old and New Values …] ⑭ te drukken opent men een dialoogvenster waarin men de hercodering specificeert (zie Figuur 10). Door middel van de [If …] knop ⑮ kan men deze hercodering conditioneel maken. De hercodering gebeurt door in het “Old en New Values” dialoogvenster de waarden die men wenst te hercoderen (linkerzijde van het venster) en de nieuwe waarden (rechterzijde) te specificeren.

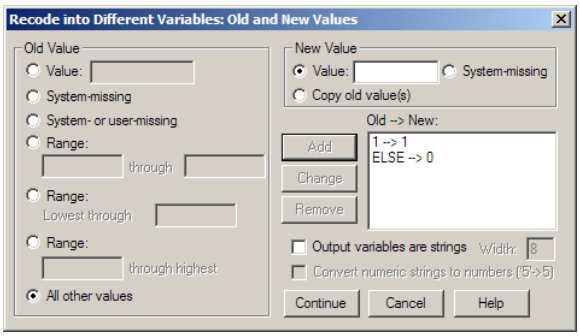
a)



b)



c)



Figuur 10: Het RECODE (oude en nieuwe waarden) dialoogvenster

Zoals blijkt uit Figuur 10a zijn er verschillende mogelijkheden om oude waarden te specificeren. In het voorbeeld hercodeert men V130 (godsdienst) in een dummyvariabele REL1 op een dusdanige manier dat observaties die 1 (katholiek) scoren op V130 een 1 krijgen op REL1, en alle anderen een 0. Als men slechts één enkele waarde wil hercoderen in een nieuwe waarde dan selecteert men de “Value” optie ⑰ in de “Old Value” sectie van het dialoogvenster. Indien de waarde op V130 gelijk is aan 1 wilt men dat de waarde voor REL1 ook gelijk is aan 1. In het tekstvak naast “Value:” in de “Old Value” sectie voert men dan de oude waarde 1 in, terwijl men in de “New Value” sectie ook een 1 invoert in het tekstvak ⑱. Zelfs indien deze waarde niet verandert moet ze expliciet vermeld worden. Waarden die niet gehercodeerd worden, worden automatisch gehercodeerd naar systeem-ontbrekend. Door op de [Add] knop te klikken wordt deze hercodering opgeslagen en verschijnt er in het vak “Old --> New” ㉒ de hercodering, hier samengevat als `1 --> 1`. Andere mogelijkheden zijn dat men de systeem-ontbrekende waarden ⑲ gehercodeerd, dat zowel de gebruiker-gedefinieerde als systeem-ontbrekende waarden ⑳ gehercodeerd worden, of dat men een ganse continue reeks van waarden wil hercoderen ㉑. Wat hier echter dient te gebeuren is dat alle andere waarden gehercodeerd worden naar een nieuwe waarde 0. Dit kan men doen door in de “Old Values” sectie de optie alle andere waarden “All other values” ㉓ te selecten, in de sectie “New Values”

de waarde 0 te specificeren. Klikt men dan op [Add] dan verschijnt in het “Old --> New” vak ㉔: ELSE --> 0.

Resultaten

Vooraleer men dummyvariabelen gebruikt in een analyse is het aangeraden dat men nagaat of men de dummyvariabelen wel correct aangemaakt heeft. De eenvoudigste wijze om dit te doen is door de frequentieverdelingen van de dummyvariabelen te vergelijken met die van oorspronkelijke categorische variabele (zie Figuur 4). Dit kan door middel van de FREQUENCIES procedure in SPSS:

```
FREQUENCIES VARIABLES = REL1 REL2 REL3 REL4.
```

Tabel 12: SPSS frequentietabellen voor dummyvariabelen (eenvoudige dummycodering)

rel1 Religion: Catholic

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	.00	6451	84.7	84.7	84.7
	1.00	1161	15.2	15.3	100.0
	Total	7612	99.9	100.0	
Missing	9.00	8	.1		
Total		7620	100.0		

rel2 Religion: Protestant

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	.00	6312	82.8	82.9	82.9
	1.00	1300	17.1	17.1	100.0
	Total	7612	99.9	100.0	
Missing	9.00	8	.1		
Total		7620	100.0		

rel3 Religion: Other Christian

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	.00	6189	81.2	81.3	81.3
	1.00	1423	18.7	18.7	100.0
	Total	7612	99.9	100.0	
Missing	9.00	8	.1		
Total		7620	100.0		

rel4 Religion: Traditionalist/Other

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	.00	7485	98.2	98.3	98.3
	1.00	127	1.7	1.7	100.0
	Total	7612	99.9	100.0	
Missing	9.00	8	.1		
Total		7620	100.0		

Als de dummyvariabelen correct aangemaakt zijn, moet voor elke dummyvariabele de frequentie van de 1-scores gelijk zijn aan de frequentie van de met de dummyvariabele geassocieerde categorie, en de frequentie van de 0-scores gelijk zijn aan het totaal aantal observaties min het aantal observaties in de geassocieerde categorie.

De frequentietabellen in Tabel 12 tonen aan dat de dummyvariabelen hier correct gemaakt zijn. Bijvoorbeeld, voor REL1, de dummyvariabele geassocieerd met de categorie “katholiek” scoren 1161 respondenten een 1 en 6459 een 0. Het aantal respondenten met score 1 is gelijk aan het aantal respondenten in de “katholiek” categorie in Figuur 4, en het aantal met score 0 is gelijk aan het aantal respondenten in de andere categorieën: 1300 + 1423 + 3601 + 115 + 12 = 6451. Hetzelfde geldt voor de andere dummyvariabelen.

Daar de afhankelijke variabele, aantal gekende SOA symptomen, op het rationiveau gemeten is kiest men hier voor multipele regressie. De volgende instructies werden gegeven om deze analyse uit te voeren:

```
REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS R ANOVA CHANGE
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT stdsymp
  /METHOD=ENTER v012 v102 married
  /METHOD=ENTER rel1 rel2 rel3 rel4  .
```

Zoals u kunt merken werden de dummyvariabelen als een afzonderlijk blok ingevoerd, en dit om te toetsen of religie als een geheeld een unieke bijdrage leverde aan de verklaring van de kennis van SOA symptomen. De verandering in de determinatiecoëfficiënt (R^2) kan gebruikt worden of de set dummyvariabelen (of dus de categorische variabele) als een geheel een unieke bijdrage hebben tot het model. Tabel 13 geeft de SPSS output hiervoor weer. Het totale model verklaart 3.9% van de variantie in kennis over SOA symptomen ❶. De dummyvariabelen set heeft een unieke bijdrage aan deze variantie van 1.4% ❷, als bij toevoeging van deze variabelen de proportie verklaarde variantie toeneemt van 2.5% tot 3.9%.

Tabel 13: SPSS REGRESSION “Model Summary” tabel voor Voorbeeld 2 (eenvoudige dummycodering).

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics
1	.157(a)	.025	.024	2.33166	
2	.198(b)	❶.039	.038	2.31495	

R Square Change	F Change	df1	df2	Sig. F Change
.025	64.139	3	7594	.000
❷.014	❸ 28.509	❹ 4	❹ 7590	❺.000

a Predictors: (Constant), married, v102 Type of place of residence, v012 Current age - respondent
b Predictors: (Constant), married, v102 Type of place of residence, v012 Current age - respondent, rel4 Religion: Traditionalist/Other, rel2 Religion: Protestant, rel1 Religion: Catholic, rel3 Religion: Other Christian

Noot: De layout van de tabel is aangepast zodat die binnen deze bladspiegel paste.

De verandering van de R^2 tengevolge van het toevoegen van de godsdienst dummyvariabelen kan getoetst worden met een *F*-toets. De toetsstatistiek is hier 28.5 ❸, en de steekproevenverdeling heeft 4 en 7590 vrijheidsgraden ❹. De hiermee gepaarde overschrijdingskans p = 0.000 ❺. In dit geval verwerpt men dus de nulhypothese dat godsdienst geen effect heeft op de kennis van SOA symptomen en aanvaardt men de alternatieve hypothese dat godsdienst wel een effect heeft op deze variabele.

Tabel 14 geeft de resultaten voor de regressie met de godsdienst dummyvariabelen weer. Drie van de vier godsdienst dummyvariabelen hebben een significant effect op kennis van SOA symptomen ❻. Alleen de categorieën "traditionele godsdienst" en "andere" verschillen gemiddeld qua kennis van SOA symptomen niet significant van de "moslim" categorie ❼. "Protestanten" kennen gemiddeld en controlerend voor de andere variabelen 0.7 symptomen meer dan "moslims" ❽. De "andere christenen" kennen gemiddeld 0.5 symptomen meer dan de "moslims" ❾ en de "katholieken" 0.3 ❿.

Tabel 14: SPSS REGRESSION output: regressiecoëfficiënten voor Voorbeeld 2 (eenvoudige dummycodering, gedeeltelijke output)

Coefficients(a)

	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
⋮	⋮	⋮	⋮	⋮	⋮
(Constant)	.322	.124		2.601	.009
v012 Current age - respondent	.035	.003	.142	11.114	.000
v102 Type of place of residence	-.234	.055	-.049	-4.263	.000
married	.147	.068	.029	2.166	.030
rel1 Religion: Catholic	❿ .253	.081	.039	3.134	❻ .002
rel2 Religion: Protestant	❽ .712	.077	.114	9.232	❻ .000
rel3 Religion: Other Christian	❾ .534	.075	.088	7.116	❻ .000
rel4 Religion: Traditionalist/Other	-.348	.211	-.019	-1.647	❼ .100

a Dependent Variable: stdsymp

De regressiecoëfficiënten voor de dummyvariabelen kunnen geïnterpreteerd worden als conditionele verschuivingen in de constante. Een dummyvariabele kan namelijk maar twee bijdragen leveren aan de geschatte uitkomstvariabele: namelijk 0 indien de score op de dummyvariabele 0 is, en de regressiecoëfficiënt indien de score op de dummyvariabele 1 is. De algemene regressievergelijking voor Voorbeeld 2 is:

$$\begin{aligned} STDSYMP &= b_0 + b_1 \times V012 + b_2 \times V102 + b_3 \times MARRIED \\ &\quad + d_1 \times REL1 + d_2 \times REL2 + d_3 \times REL3 + d_4 \times REL4 \\ &= 0.322 + 0.035 \times V012 - 0.234 \times V102 \\ &\quad + 0.147 \times MARRIED + 0.253 \times REL1 + 0.712 \times REL2 \\ &\quad + 0.534 \times REL3 - 0.348 \times REL4 \end{aligned}$$

De variabelen REL1, REL2, REL3 en REL4 kunnen slechts twee waarden aannemen (0 & 1). Bijvoorbeeld, voor moslims (de referentiecategorie) is de score of alle vier dummyvariabelen 0 wat volgende regressievergelijking geeft voor de “moslim” categorie:

$$\begin{aligned} STDSYMP &= b_0 + b_1 \times V012 + b_2 \times V102 + b_3 \times MARRIED \\ &\quad + d_1 \times REL1 + d_2 \times REL2 + d_3 \times REL3 + d_4 \times REL4 \\ &= 0.322 + 0.035 \times V012 - 0.234 \times V102 \\ &\quad + 0.147 \times MARRIED + 0.253 \times 0 + 0.712 \times 0 + 0.534 \times 0 - 0.348 \times 0 \\ &= 0.322 + 0.035 \times V012 - 0.234 \times V102 + 0.147 \times MARRIED \end{aligned}$$

De constante voor de referentiecategorie is gelijk aan de constante bekomen in de regressievergelijking. Voor de andere categorieën heeft steeds slechts één dummyvariabele een score 1. Dit geeft volgende regressievergelijkingen voor deze categorieën:

- Katholiek

$$\begin{aligned} STDSYMP &= b_0 + b_1 \times V012 + b_2 \times V102 + b_3 \times MARRIED \\ &\quad + d_1 \times REL1 + d_2 \times REL2 + d_3 \times REL3 + d_4 \times REL4 \\ &= 0.322 + 0.035 \times V012 - 0.234 \times V102 \\ &\quad + 0.147 \times MARRIED + 0.253 \times 1 + 0.712 \times 0 + 0.534 \times 0 - 0.348 \times 0 \\ &= (0.322 + 0.253) + 0.035 \times V012 - 0.234 \times V102 + 0.147 \times MARRIED \\ &= 0.575 + 0.035 \times V012 - 0.234 \times V102 + 0.147 \times MARRIED \end{aligned}$$

- Protestant

$$\begin{aligned} STDSYMP &= b_0 + b_1 \times V012 + b_2 \times V102 + b_3 \times MARRIED \\ &\quad + d_1 \times REL1 + d_2 \times REL2 + d_3 \times REL3 + d_4 \times REL4 \\ &= 0.322 + 0.035 \times V012 - 0.234 \times V102 + 0.147 \times MARRIED \\ &\quad + 0.253 \times 0 + 0.712 \times 1 + 0.534 \times 0 - 0.348 \times 0 \\ &= (0.322 + 0.712) + 0.035 \times V012 - 0.234 \times V102 + 0.147 \times MARRIED \\ &= 1.034 + 0.035 \times V012 - 0.234 \times V102 + 0.147 \times MARRIED \end{aligned}$$

- Andere christenen

$$\begin{aligned} STDSYMP &= b_0 + b_1 \times V012 + b_2 \times V102 + b_3 \times MARRIED \\ &\quad + d_1 \times REL1 + d_2 \times REL2 + d_3 \times REL3 + d_4 \times REL4 \\ &= 0.322 + 0.035 \times V012 - 0.234 \times V102 + 0.147 \times MARRIED \\ &\quad + 0.253 \times 0 + 0.712 \times 0 + 0.534 \times 1 - 0.348 \times 0 \\ &= (0.322 + 0.534) + 0.035 \times V012 - 0.234 \times V102 + 0.147 \times MARRIED \\ &= 0.856 + 0.035 \times V012 - 0.234 \times V102 + 0.147 \times MARRIED \end{aligned}$$

- Tradionele godsdienst/andere

$$
\begin{aligned}
STDSYMP &= b_0 + b_1 \times V012 + b_2 \times V102 + b_3 \times MARRIED \\
&\quad + d_1 \times REL1 + d_2 \times REL2 + d_3 \times REL3 + d_4 \times REL4 \\
&= 0.322 + 0.035 \times V012 - 0.234 \times V102 + 0.147 \times MARRIED \\
&\quad + 0.253 \times 0 + 0.712 \times 0 + 0.534 \times 0 - 0.348 \times 1 \\
&= (0.322 - 0.348) + 0.035 \times V012 - 0.234 \times V102 + 0.147 \times MARRIED \\
&= -0.026 + 0.035 \times V012 - 0.234 \times V102 + 0.147 \times MARRIED
\end{aligned}
$$

De constante voor deze vergelijkingen is steeds gelijk aan de constante geschat in de regressievergelijking plus de regressiecoëfficiënt voor de dummyvariabele geassocieerd met de categorie in kwestie. Voor de categorie "katholiek" wordt dit de algemene constante (b_0) plus de regressiecoëfficiënt voor REL1 (d_1). Het gevolg hiervan is dat men voor de verschillende categorieën van de categorische variabele parallelle regressieruimten verkrijgt die alleen maar verschillen in de waarde van hun constante.

Effect dummycodering

Effect dummycodering werkt op een analoge manier. Daar er weinig verschillen zijn tussen de twee coderingschema's geven we hier alleen het voorbeeld gebruik makend van de SPSS syntaxis. Net zoals bij de eenvoudige dummycodering krijgen alle niet-referentie-categorieën een score 1 op de met hen geassocieerde dummyvariabele en een score 0 op alle andere dummyvariabelen. Het verschil met de eenvoudige dummycodering ligt in de behandeling van de referentiecategorie. Waar die in eenvoudige dummycodering een 0 scoorde op alle dummyvariabelen, krijgt de referentiecategorie in de effect dummycodering een score -1 op alle dummyvariabelen. Het gevolg hiervan is dat het referentiepunt voor de geschatte coëfficiënten het gemiddelde van de verschillende groepen wordt en dat het effect van de referentiecategorie gelijk is aan minus de som van de effecten van alle dummyvariabelen.

Tabel 15 toont het effect dummycoderingschema voor Voorbeeld 2. Respondenten in de categorieën "katholiek", "protestant" en "andere christenen" scoren 1 op REL1, REL2 en REL3, respectievelijk en 0 op de andere dummyvariabelen. Daar ook hier de "traditionele religie" en "andere" categorieën samengenomen worden scoren zij beide 1 op REL4 en 0 op de drie andere dummyvariabelen. De referentiecategorie "moslim" krijgt een score -1 op alle 4 de dummyvariabelen.

Tabel 15: Effect dummycoderingschema voor godsdienstvariabele in Nigeria 2003 DHS+

	REL 1	REL 2	REL 3	REL 4
Catholic	1	0	0	0
Protestant	0	1	0	0
Other Christian	0	0	1	0
Islam	-1	-1	-1	-1
Traditionalist	0	0	0	1
Other	0	0	0	1

De SPSS instructies voor de effect dummycodering van de godsdienstvariabele verschillen weinig van die van de eenvoudige dummycodering. Het enige verschil is dat er per dummyvariabele een extra conditionele transformatie gespecificeerd wordt waarin de observaties in de referentiecategorie een score -1 krijgen. In dit voorbeeld wordt dit:

```
compute rel1=0.
compute rel2=0.
compute rel3=0.
compute rel4=0.
if v130=1 rel1=1.
if v130=2 rel2=1.
if v130=3 rel3=1.
```

```
if v130=5 rel4=1.
if v130=6 rel4=1.
```

waar de volgende vier lijnen de referentiecategorie (4 = “moslim”) scores op -1 zetten:

```
if v130=4 rel1=-1.
if v130=4 rel2=-1.
if v130=4 rel3=-1.
if v130=4 rel4=-1.
if missing(V130) rel1=9.
if missing(V130) rel2=9.
if missing(V130) rel3=9.
if missing(V130) rel4=9.
missing values rel1 rel2 rel3 rel4 (9).
```

Hetzelfde kan ook verwezenlijkt worden met de RECODE instructie. Hierbij dient men gewoon extra te specificeren dat de referentiecategorie omgezet wordt tot -1:

```
recode v130 (1=1)(4=-1)(missing=9)(else = 0) into rel1.
recode v130 (2=1)(4=-1)(missing=9)(else = 0) into rel2.
recode v130 (3=1)(4=-1)(missing=9)(else = 0) into rel3.
recode v130 (5,6=1)(4=-1)(missing=9)(else = 0) into rel4.
missing values rel1 rel2 rel3 rel4 (9).
```

Weer kan men de constructie van de dummyvariabelen checken door de frequenties van de dummyvariabelen te vergelijken met die van de oorspronkelijke categorische variabele. Net zoals bij de eenvoudige dummycodering moet de frequentie van de 1-score gelijk zijn aan de frequentie van de geassocieerde categorie. Bij effect dummycodering kan de dummyvariabele drie waarden aannemen 1, 0 en -1. De frequentie van de -1-score moet altijd gelijk zijn aan de frequentie van de referentiecategorie, terwijl de 0-score frequentie gelijk dient te zijn aan de som van de frequenties van de resterende categorieën.

Zoals men kan vaststellen in **Error! Not a valid bookmark self-reference.** is de frequentie voor de -1-score steeds 3601 wat overeenstemt met de moslim categorie. De 1-score frequenties stemmen steeds overeen met de frequentie van de geassocieerde categorie.

Tabel 16: SPSS frequentietabellen voor dummyvariabelen (effect dummycodering)

rel1 Religion: Catholic

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	-1.00	3601	47.3	47.3	47.3
	.00	2850	37.4	37.4	84.7
	1.00	1161	15.2	15.3	100.0
	Total	7612	99.9	100.0	
Missing	9.00	8	.1		
Total		7620	100.0		

rel2 Religion: Protestant

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	-1.00	3601	47.3	47.3	47.3
	.00	2711	35.6	35.6	82.9
	1.00	1300	17.1	17.1	100.0
	Total	7612	99.9	100.0	
Missing	9.00	8	.1		
Total		7620	100.0		

rel3 Religion: Other Christian

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	-1.00	3601	47.3	47.3	47.3
	.00	2588	34.0	34.0	81.3
	1.00	1423	18.7	18.7	100.0
	Total	7612	99.9	100.0	
Missing	9.00	8	.1		
Total		7620	100.0		

rel4 Religion: Traditionalist/Other

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	-1.00	3601	47.3	47.3	47.3
	.00	3884	51.0	51.0	98.3
	1.00	127	1.7	1.7	100.0
	Total	7612	99.9	100.0	
Missing	9.00	8	.1		
Total		7620	100.0		

De verklarende kracht van het model, de determinatiecoëfficiënt bij effect dummycodering is exact dezelfde als bij eenvoudige dummycodering. In beide gevallen is de totale verklaarde variantie gelijk aan 3.9% en heeft godsdienst een uniek effect van 1.4%. Ook de toets voor de verandering in R^2 bij toevoeging van de dummyvariabelen is identiek bij de beide coderingschema's.

De regressiecoëfficiënten van de andere variabelen worden niet beïnvloed door het dummycoderingschema. Alleen de constante en de regressiecoëfficiënten voor de dummyvariabelen worden beïnvloed door het coderingschema. Tabel 18 toont de regressieresultaten voor het model met de effect dummycodering. Ook hier zijn maar drie van de vier dummyvariabelen effecten significant. Het zijn echter niet dezelfde variabelen als bij de eenvoudige dummycodering. In dit geval is de regressiecoëfficiënt voor REL1 niet significant, wat betekent dat het gemiddelde van de "katholieke" categorie, controlerend voor de andere variabelen, niet significant verschilt van het gemiddelde van de verschillende categorieën. De andere godsdienstscategorieën verschillen wel significant van dit gemiddelde. Protestanten, bijvoorbeeld, kennen gemiddeld 0.48 symptomen meer dan het gemiddelde, andere christenen 0.30 en respondenten in de categorieën "traditionele godsdiensten" en "andere" kennen gemiddeld en na controle voor de andere variabelen 0.58 symptomen minder dan het gemiddelde.

Tabel 17: SPSS REGRESSION "Model Summary" tabel voor Voorbeeld 2[1] (effect dummycodering).

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.157(a)	.025	.024	2.33166
2	.198(b)	.039	.038	2.31495

Change Statistics				
R Square Change	F Change	df1	df2	Sig. F Change
.025	64.139	3	7594	.000
.014	28.509	4	7590	.000

a Predictors: (Constant), married, v102 Type of place of residence, v012 Current age - respondent
b Predictors: (Constant), married, v102 Type of place of residence, v012 Current age - respondent, rel2 Religion: Protestant, rel3 Religion: Other Christian, rel1 Religion: Catholic, rel4 Religion: Traditionalist/Other

Noot: De layout van de tabel is aangepast zodat die binnen deze bladspiegel paste.

Tabel 18: SPSS REGRESSION output: regressiecoëfficiënten voor Voorbeeld 2 (effect dummycodering, gedeeltelijke output)

Coefficients(a)

	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
(Constant)	.552	.130		4.252	.000
v012 Current age – respondent	.035	.003	.142	11.114	.000
v102 Type of place of residence	-.234	.055	-.049	-4.263	.000
Married	.147	.068	.029	2.166	.030
rel1 Religion: Catholic	.023	.071	.007	.321	.748
rel2 Religion: Protestant	.482	.069	.152	6.977	.000
rel3 Religion: Other Christian	.304	.068	.098	4.488	.000
rel4 Religion: Traditionalist/Other	-.578	.168	-.130	-3.450	.001

a Dependent Variable: stdsymp number of STD symptoms known

De pseudo-regressiecoëfficiënt voor de referentiecategorie is gelijk aan minus de som van de regressiecoëfficiënten van de andere dummyvariabelen. In dit voorbeeld wordt dit:

$$b_{moslim} = -(0.023 + 0.482 + 0.304 – 0.578) = -0.230$$

Respondenten in de referentiecategorie “moslim” kennen gemiddeld dus 0.23 SOA symptomen minder dan het gemiddelde van de verschillende groepen. Men kan echter niet onmiddellijk vaststellen of dit effect ook significant is[4]. Ook in dit geval kan men de effecten van

[4] Het is mogelijk om deze coëfficiënt te toetsen door zelf de standaardfout ervan te berekenen. Meestal is het echter eenvoudiger de analyse opnieuw uit te voeren met een andere categorie als referentiecategorie en het statistisch programma de toets te laten uitvoeren.

de dummyvariabelen voorstellen als een set van parallelle regressieruimten, waarbij alleen de constante verschilt:

- Moslim

$$\begin{aligned} STDSYMP &= (0.552 - 0.230) + 0.035 \times V012 - 0.234 \times V102 + 0.147 \times MARRIED \\ &= 0.322 + 0.035 \times V012 - 0.234 \times V102 + 0.147 \times MARRIED \end{aligned}$$

- Katholiek

$$\begin{aligned} STDSYMP &= (0.552 + 0.023) + 0.035 \times V012 - 0.234 \times V102 + 0.147 \times MARRIED \\ &= 0.575 + 0.035 \times V012 - 0.234 \times V102 + 0.147 \times MARRIED \end{aligned}$$

- Protestant

$$\begin{aligned} STDSYMP &= (0.552 + 0.482) + 0.035 \times V012 - 0.234 \times V102 + 0.147 \times MARRIED \\ &= 1.034 + 0.035 \times V012 - 0.234 \times V102 + 0.147 \times MARRIED \end{aligned}$$

- Andere christenen

$$\begin{aligned} STDSYMP &= (0.552 + 0.304) + 0.035 \times V012 - 0.234 \times V102 + 0.147 \times MARRIED \\ &= 0.856 + 0.035 \times V012 - 0.234 \times V102 + 0.147 \times MARRIED \end{aligned}$$

- Traditionele godsdienst/andere

$$\begin{aligned} STDSYMP &= (0.552 - 0.578) + 0.035 \times V012 - 0.234 \times V102 + 0.147 \times MARRIED \\ &= -0.026 + 0.035 \times V012 - 0.234 \times V102 + 0.147 \times MARRIED \end{aligned}$$

Zoals men kan zien zijn deze vergelijkingen identiek aan die met eenvoudige dummycodering. Dit bevestigt nogmaals dat ongeacht het coderingschema de voorspelde waarden identiek zullen zijn. Beide coderingschema's leiden dan ook tot equivalente modellen. Geen is beter dan het andere. De keuze tussen de twee coderingschema's is dan ook geen statistische maar een substantiële keuze. Welke van de twee coderingschema's past het beste bij de gestelde vraag?

Samenvatting

Dummycodering is een methode om categorische variabelen (nominaal of ordinaal) als onafhankelijke variabelen in een regressieanalyse te kunnen gebruiken. De methode bestaat eruit dat een categorische variabele met *k* categorieën in de regressievergelijking vervangen wordt door *k* – 1 dichotome variabelen: de dummyvariabelen. Elk van deze dummyvariabelen representeert één van de oorspronkelijke categorieën. De k^{de} categorie die niet door een dummyvariabele gerepresenteerd wordt noemt men de referentiecategorie. De keuze van de referentiecategorie is belangrijk voor de substantiële interpretatie van de regressiecoëfficiënten voor de dummyvariabelen. Deze dummyvariabelen worden als gewone variabelen in de regressievergelijking gebracht. Het is echter steeds belangrijk om te toetsen of een set dummyvariabelen als een geheel een significant effect hebben op de afhankelijke variabele.

Bibliografie

National Population Commission (NPC) [Nigeria], & ORC Macro. (2004). *Nigeria demographic and health survey 2003*. Calverton, MD: National Population Commission & ORC Macro.

World values survey, 1981-1984 and 1990-1993 [Data file]. (1994). World Values Study Group (producer). Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor].

CENTER FOR SOCIAL THEORY – DEPARTMENT OF SOCIOLOGY – GHENT UNIVERSITY
KORTE MEER 3-5, 9000 GENT, BELGIUM
SOCIO@UGENT.BE

Reeds verschenen in deze reeks

2. Ronan Van Rossem. *PC. vs. PAF. Een enigszins technische inleiding* (Jan. 2011)
1. Ronan Van Rossem. *Dummyvariabelen in meervoudige regressie. Een inleiding voor sociale wetenschappers* (Nov. 2010)

There is nothing more practical than a good theory

(K. Lewin, 1952)

CENTRE FOR SOCIAL THEORY - DEPARTMENT OF SOCIOLOGY - UNIVERSITEIT GENT - KORTE MEER 3-5, 9000 GENT, BELGIUM