SVR - Methoden en technieken 2012 / 1

Het onderzoeken van veranderingen in de tijd met cross-sectionele surveys

Jan Pickery

Studiedienst van de Vlaamse Regering

Vlaamse overheid

Het onderzoeken van veranderingen in de tijd met cross-sectionele surveys

Jan Pickery

Vlaamse overheid

Samenstelling
Diensten voor het Algemeen Regeringsbeleid
Studiedienst van de Vlaamse Regering (SVR)

Jan Pickery

Leescomité
Marc Callens, Ann Carton, SVR

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/2012/3241/078
http://www.vlaanderen.be/svr

Inhoudstafel

1. Inleiding

In deze tekst proberen we duidelijk te maken hoe je op basis van verschillende jaargangen van een survey iets kan zeggen over een (significante) evolutie in de tijd. De tekst en toepassingen hebben betrekking op cross-sectionele surveys, dat wil zeggen surveys bij steeds andere respondenten. Panelsurveys, waarbij eenzelfde groep respondenten gevolgd wordt in de tijd, laten we buiten beschouwing. Duidelijke voorbeelden van cross-sectionele surveys in opdracht of onder supervisie van de Studiedienst van de Vlaamse Regering zijn de SCV-survey en de survey voor de Stadsmonitor. De SCV-survey 'Sociaal-culturele verschuivingen in Vlaanderen' peilt sinds 1996 jaarlijks naar de waarden, opvattingen en overtuigingen van Vlamingen, wat het mogelijk maakt om voor een aantal beleidsrelevante thema's te kijken naar verschuivingen en veranderingen. De survey van de stadsmonitor kende in 2011 zijn vierde editie. Die survey peilt bij inwoners van de 13 Vlaamse centrumsteden naar houdingen en gedrag met betrekking tot de woonomgeving, participatie, stedelijk beleid…

Deze nota brengt een aantal gekende statistische weetjes of problemen en oplossingen samen. De vele, eenvoudige, voorbeelden maken de nota ook zeer praktijkgericht.

Alle voorbeelden gaan uit van categorische (afhankelijke) variabelen. De technieken zijn dan ook daarop toegepast. Maar er wordt op voldoende plaatsen duidelijk gemaakt, welke alternatieven van toepassing zijn als de afhankelijke variabelen metrisch zijn.

2. Twee surveys, (ruw) verschil tussen percentages – ongewogen analyse

We beginnen met de meest eenvoudige situatie. We hebben met twee verschillende surveys een gedrag, houding of attitude gemeten en gaan na of er een significant verschil is tussen beide surveys.

Uit de SCV-surveys van 2008 en 2009 selecteerden we het item van de tevredenheid met de woning. De vier categorieën van deze variabele hergroeperen we voor de analyse naar twee. Om de berekeningen te vereenvoudigen, tonen we in eerste instantie alleen ongewogen analyses.

Tabel 1 Tevredenheid met de woning in 2008

	Aantal	Percentage
heel ontevreden	10	0,7
ontevreden	46	3,1
tevreden	743	50,4
heel tevreden	676	45,8
Totaal	1.475	100,0

Tabel 2 Tevredenheid met de woning in 2008, gedichotomiseerd en met betrouwbaarheidsintervallen

		Aantal	Percentage	95%-betr.interv.
niet heel tevreden		799	54,2	51,6 - 56,7
heel tevreden		676	45,8	43,3 - 48,4
	Totaal	1.475	100,0	

Uit tabel 2 maken we op dat bijna 46% heel tevreden is met de eigen woning en iets meer dan 54% dus niet heel tevreden. In tabel 2 hebben we voor de onderscheiden percentages ook betrouwbaarheidsintervallen berekend. We gebruikten daarvoor de meest eenvoudige formule, die vertrekt van een benadering gebaseerd op de Normaalverdeling[1]:

$$\hat{p} \pm 1{,}96 \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Waarbij

$\hat{p}$ = steekproefschatter voor de populatieproportie π
= p, de waargenomen steekproefproportie
1,96 = z-waarde bij een 95%-betrouwbaarheidsniveau (α = 0,05)
n = steekproefomvang

Er zijn verschillende alternatieve werkwijzen om betrouwbaarheidsintervallen te berekenen. Die alternatieven zijn in sommige situaties ook beter, bijvoorbeeld bij (zeer) hoge of lage percentages of bij kleine steekproeven (zie Newcombe, 1998). Maar die andere berekeningswijzen vallen buiten het bestek van deze tekst. Voor de opbouw van onze redenering is het ook niet noodzakelijk erbij stil te staan.

We berekenen dezelfde tevredenheidspercentages op basis van de SCV-survey 2009. Een belangrijk aandachtspunt daarbij is het uniformiseren van de populaties. Vanaf 2009 werden de nationaliteitsvoorwaarde en de bovengrens voor leeftijd (85 jaar) losgelaten. Om de populaties vergelijkbaar te maken, moeten we dus de niet-Belgen en de 85-plussers verwijderen uit het bestand van 2009.

Tabel 3 Tevredenheid met de woning in 2009

		Aantal	Percentage
heel ontevreden		12	0,8
ontevreden		78	5,4
tevreden		753	52,3
heel tevreden		597	41,5
	Totaal	1.440	100,0

[1] SPSS heeft – tot in versie 16 – geen eenvoudige manier om betrouwbaarheidsintervallen op te vragen voor categorische variabelen. Een omweg is mogelijk door de dichotome categorische variabele te hercoderen tot een 0/1-variabele, die variabele als metrisch te beschouwen en er "*Descriptive Statistices*" voor op te vragen via "*Explore*".

Tabel 4 Tevredenheid met de woning in 2009 bij de 18- tot 85-jarige Belgen

		Aantal	Percentage
heel ontevreden		9	0,7
ontevreden		71	5,2
tevreden		708	52,1
heel tevreden		570	42,0
	Totaal	1.358	100,0

Ook voor de data van de SCV-survey van 2009 dichotomiseren we de variabele en berekenen we betrouwbaarheidsintervallen.

Tabel 5 Tevredenheid met de woning in 2009, gedichotomiseerd en met betrouwbaarheidsintervallen (bij 18- tot 85-jarige Belgen)

		Aantal	Percentage	95%-betr.interv.		
niet heel tevreden		788	58,0	55,4	-	60,6
heel tevreden		570	42,0	39,4	-	44,6
	Totaal	1.358	100,0			

Tabel 2 bis Tevredenheid met de woning in 2008, gedichotomiseerd en met betrouwbaarheidsintervallen (herhaling)

		Aantal	Percentage	95%-betr.interv.		
niet heel tevreden		799	54,2	51,6	-	56,7
heel tevreden		676	45,8	43,3	-	48,4
	Totaal	1.475	100,0			

Als we tabel 5 en tabel 2 (die we voor het overzicht hier opnieuw weergegeven hebben) vergelijken, zouden we verschillende conclusies kunnen trekken over een al dan niet significant verschil tussen beide jaargangen. We zouden ten eerste kunnen besluiten dat er geen significant verschil is, omdat de betrouwbaarheidsintervallen overlap vertonen. Een alternatieve conclusie zou kunnen zijn dat er wel een significant verschil is omdat het percentage van 2009 niet in het interval van 2008 zit en vice versa. Of misschien kunnen we op basis van deze tabellen niets besluiten over een al dan niet significant verschil?

We nemen er nog een tweede voorbeeld bij. In dat tweede voorbeeld kijken we naar de tevredenheid met de levensstandaard in 2008 en 2010. In tabellen 6 en 7 hebben we die tevredenheid ook gedichotomiseerd én hebben we betrouwbaarheidsintervallen rond de percentages berekend.

Tabel 6 Tevredenheid met de levensstandaard in 2008, gedichotomiseerd en met betrouwbaarheidsintervallen

		Aantal	Percentage	95%-betr.interv.		
(heel) ontevreden		129	8,76	7,32	-	10,21
(heel) tevreden		1.343	91,24	89,79	-	92,68
	Totaal	1.472	100,00			

Tabel 7 Tevredenheid met de levensstandaard in 2010, gedichotomiseerd en met betrouwbaarheidsintervallen (bij 18- tot 85-jarige Belgen)

		Aantal	Percentage	95%-betr.interv.		
(heel) ontevreden		94	7,26	5,85	-	8,67
(heel) tevreden		1.201	92,74	91,33	-	94,16
	Totaal	1.295	100,00			

Ook nu zijn verschillende conclusies mogelijk. De betrouwbaarheidsintervallen vertonen overlap, waaruit we zouden kunnen besluiten dat er geen significant verschil is. Het percentage van 2010 zit niet vervat in het interval van 2008 (let op het extra cijfer achter de komma), waaruit we wel significantie zouden kunnen afleiden. Of misschien moeten we opnieuw erkennen dat we op basis van deze twee tabellen niet tot een besluit over een al dan niet significant verschil kunnen komen.

De correcte conclusie voor beide voorbeelden is dat we op basis van een vergelijking van de afzonderlijke tabellen voor de twee jaargangen geen besluiten mogen trekken over de significantie van het verschil, zelfs niet als we ook de betrouwbaarheidsintervallen berekenen. Er bestaat daarentegen een specifieke *z*-toets voor het verschil tussen proporties in twee onafhankelijke steekproeven[2]. De *z*-waarde voor die toets wordt berekend als volgt:

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1}+\frac{1}{n_2})}}$$

Waarbij

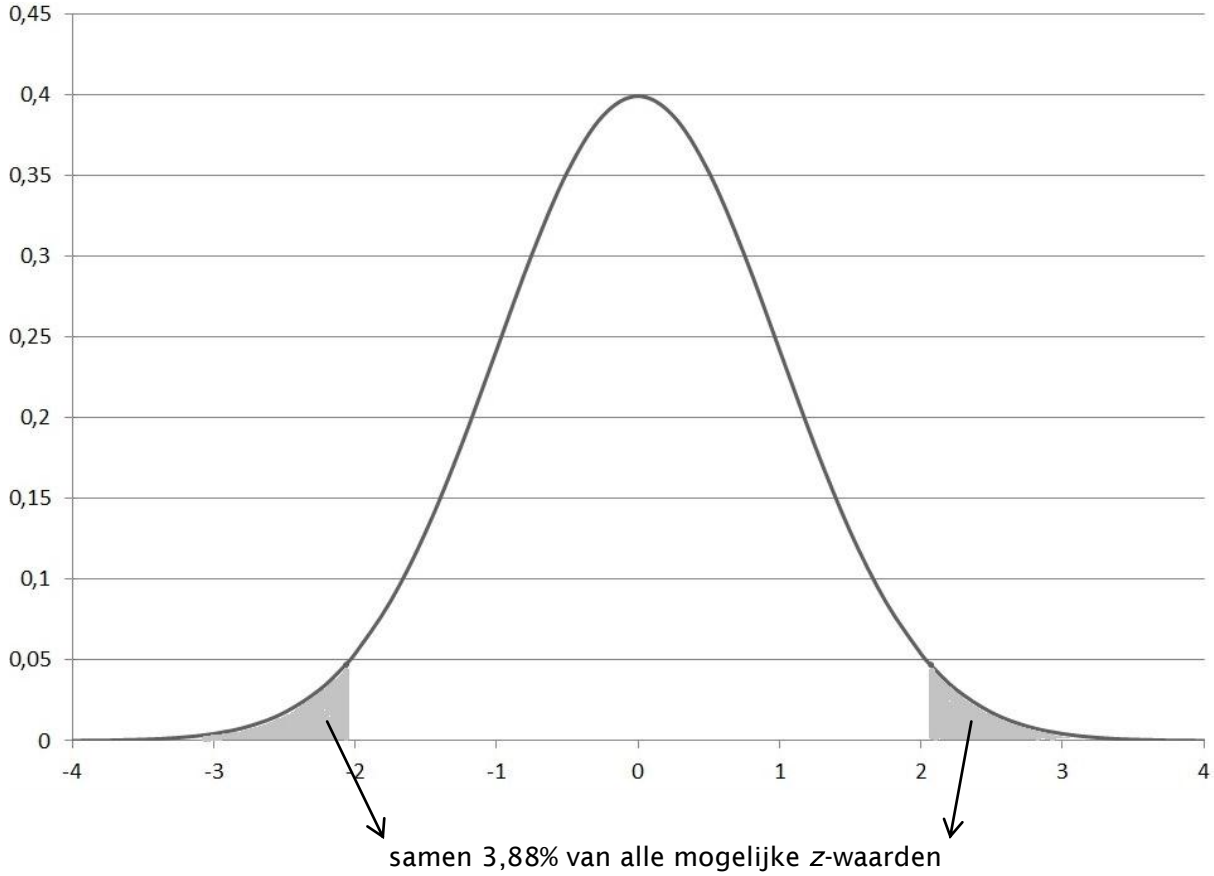
z = *z*-waarde op de standaardnormaalverdeling
$\hat{p}_1$ = steekproefschatter voor de populatieproportie π_1
= p_1, de waargenomen steekproefproportie in steekproef 1
$\hat{p}_2$ = steekproefschatter voor de populatieproportie π_2
= p_2, de waargenomen steekproefproportie in steekproef 2
$\hat{p}$ = de geschatte proportie als beide steekproeven samengevoegd worden
= de gezamenlijke schatter van π
n_1 = de steekproefomvang van steekproef 1
n_2 = de steekproefomvang van steekproef 2

Voor het eerste voorbeeld (tevredenheid met de woning) is de berekende *z*-waarde gelijk aan 2,07. Uit de Standaardnormaalverdeling kunnen we afleiden hoe waarschijnlijk het is om een *z*-waarde te bekomen die zo groot is of nog groter in absolute waarde.

Figuur 1 toont dat slechts 3,88% van alle *z*-waarden kleiner zijn dan -2,07 of groter dan 2,07. De interpretatie luidt dat als er in werkelijkheid geen verschil zou zijn in tevredenheid met de woning tussen 2008 en 2009, we slechts 3,88% kans hebben om in steekproeven die even groot zijn als de onze, minstens het verschil aan te treffen dat wij vaststelden. Die kans is klein, kleiner dan de conventionele 5% (α = 0,05). Daarom verwerpen we de nulhypothese dat er geen verschil is tussen 2008 en 2009 en besluiten we dus tot een significant verschil.

[2] Dit is de meest eenvoudige hypothesetoets. Net als bij de betrouwbaarheidsintervallen, zijn er alternatieven, die in verschillende situaties (zeer hoge of zeer lage proporties, kleine steekproeven) beter zijn. De geïnteresseerde lezer kan zich hiervoor wenden tot Fleiss e.a. (2003, 50-63).

Figuur 1 Standaardnormaalverdeling, met aanduiding van de waarschijnlijkheid om *z*-waarden te bekomen die in absolute waarde groter dan of gelijk zijn aan 2,07



Dezelfde toets toegepast op het tweede voorbeeld levert een *z*-waarde op die gelijk is aan 1,45. De kans om een dergelijke *z*-waarde te bekomen bedraagt 14,68%. Voor het tweede voorbeeld kunnen we dus niet besluiten dat er een significant verschil is tussen 2008 en 2010 in tevredenheid met de levensstandaard.

De twee voorbeelden tonen aan dat de betrouwbaarheidsintervallen voor afzonderlijke jaargangen geen goede basis vormen om al dan niet tot significantie van het verschil te besluiten. Bij voorbeeld 1 is er overlap tussen de betrouwbaarheidsintervallen, maar het percentage van 2009 zit niet vervat in het betrouwbaarheidsinterval van 2008. De test wijst uiteindelijk uit dat er een significant verschil is. De vergelijking van de twee afzonderlijke tabellen bij voorbeeld 2 volgt hetzelfde patroon: overlap bij de betrouwbaarheidsintervallen, maar het percentage van het ene jaar zit niet vervat in het interval van het andere jaar. Hier blijkt uit de test dat er geen significant verschil is.

Bemerk dat dit principe niet alleen geldt als je verschillende jaargangen van een survey vergelijkt, maar ook voor verschillen tussen steden (bijvoorbeeld bij de stadsmonitor) of verschillen volgens een bepaald kenmerk (bijvoorbeeld tussen mannen en vrouwen). Op basis van het betrouwbaarheidsinterval bij de mannen en dat bij de vrouwen kan je niet concluderen of het verschil tussen beide significant is. Het idee dat er een significant verschil is als de betrouwbaarheidsintervallen van twee jaargangen/groepen geen overlap vertonen, is misschien ruim ingeburgerd maar niet correct. Om het verschil na te gaan heb je gewoon een andere statistische toets nodig, in casu een verschiltoets. Ook de optie om te kijken of het cijfer voor jaargang X zich al dan niet in het betrouwbaarheidsinterval van jaargang Y bevindt, is niet correct. Bij die redenering misken je de onzekerheid die geldt voor het cijfer van jaargang X.

Dit principe geldt natuurlijk ook als de afhankelijke variabele metrisch is. De gemiddelden van één numerieke variabele in twee jaargangen van een survey met bijhorende betrouwbaarheidsintervallen vormen op zich een onvoldoende basis om de significantie van het verschil tussen beide jaargangen na te gaan. Daarvoor moet er teruggegrepen worden naar een specifieke statistische verschiltoets, zoals de *t*-toets voor het verschil tussen twee gemiddelden van onafhankelijke steekproeven.

Hoewel het toetsen van verschillen tussen proporties of percentages op de hierboven beschreven manier nu ook weer niet zo complex is, vergt de uitvoering toch enig eigen rekenwerk – iets wat de meeste onderzoekers eigenlijk niet meer gewoon zijn. Er is een alternatief waarbij we binnen een vertrouwde statistische software-omgeving kunnen blijven. Dat alternatief gaat uit van het samenvoegen van de databestanden. Hieronder illustreren we dat alternatief voor beide voorbeelden.

Als we de data van de SCV-surveys van 2008 en 2009 samennemen, hebben we 2.833 mensen waarbij we naar de tevredenheid met de woning gevraagd hebben. Surveyjaar en tevredenheid (dichotoom) kunnen we daarbij beschouwen als 2 afzonderlijke variabelen (zie tabel 8 en tabel 9).

Tabel 8 Surveyjaar in de samengevoegde dataset voor tevredenheid met de woning

		Aantal	Percentage
2008		1.475	52,1
2009		1.358	47,9
	Totaal	2.833	100,0

Tabel 9 Tevredenheid met de woning in de samengevoegde dataset

		Aantal	Percentage
niet heel tevreden		1.587	56,0
heel tevreden		1.246	44,0
	Totaal	2.833	100,0

We hebben nu één databestand met twee (categorische) variabelen. Of er een samenhang is tussen die twee variabelen kunnen we nagaan met een statistische toets die ongetwijfeld bekender is dan de hierboven beschreven *z*-toets: de Chi-kwadraattoets voor afhankelijkheid van twee categorische kenmerken. Die toets vergelijkt de verwachte en geobserveerde frequenties. De verwachte frequenties worden berekend op basis van de globale verdelingen van beide variabelen (ook wel marginale verdelingen genoemd). Details zijn in alle statistiekhandboeken te vinden. Die toets is eenvoudig toe te passen op (de kruistabel uit) het samengevoegde databestand.

Tabel 10 Tevredenheid met de woning volgens surveyjaar

			2008	2009	Totaal
niet heel tevreden	aantal		799	788	1.587
	kolom%		54,2	58,0	
heel tevreden	aantal		676	570	1.246
	kolom%		45,8	42,0	
		Totaal	1.475	1.358	2.833

De berekende Chi-kwadraatwaarde voor deze tabel is gelijk aan 4,269; bij 1 vrijheidsgraad[3] is de bijhorende kans gelijk aan 0,0388. De kans om dergelijke percentages te vinden in een steekproef van deze omvang als er in de populatie geen samenhang zou zijn tussen beide kenmerken (onafhankelijkheid) is dus kleiner dan 4%. Omdat die kans zo klein is, verwerpen we de nulhypothese van onafhankelijkheid en kunnen we besluiten dat er wel een samenhang is (en dus een verschil in tevredenheid tussen 2008 en 2009).

We passen dezelfde methode toe op het tweede voorbeeld: data samenvoegen in één bestand, kruistabel berekenen en Chi-kwadraattoets uitvoeren. De tabel en de resultaten van de toets worden hieronder weergegeven.

Tabel 11 Tevredenheid met de levensstandaard volgens surveyjaar

			2008	2010	Totaal
(heel) ontevreden	aantal		129	94	223
	kolom%		8,8	7,3	
(heel) tevreden	aantal		1.343	1.201	2.544
	kolom%		91,2	92,7	
		Totaal	1.472	1.295	2.767

$\chi^2 = 2{,}106$; df = 1, $p = 0{,}147$

De aandachtige lezer heeft al gemerkt dat de *p*-waarden van de Chi-kwadraattoetsen gelijk zijn aan de *p*-waarden van de eerder beschreven *z*-toetsen. Bovendien is de χ^2-waarde ook telkens gelijk aan het kwadraat van de bekomen *z*-waarde. Beide toetsen zijn bijgevolg volledig equivalent. Het samenvoegen ("poolen") van datasets en het uitvoeren van een toets op het geheel is dus een correcte manier om significantie van de verschillen tussen twee jaargangen na te gaan.

De strategie van het samenvoegen van databestanden is natuurlijk ook toepasbaar als de afhankelijke variabele metrisch is. Op het samengevoegde bestand voeren we dan een variantieanalyse of *t*-toets voor een verschil tussen gemiddelden uit.

3. Twee surveys, (ruw) verschil tussen percentages – gewogen analyse

De techniek van het samenvoegen van databestanden biedt een bijkomend voordeel. Zo is het relatief eenvoudig om ook gewogen analyses uit te voeren. Een gewogen variant van de *z*-toets met correcte statistische inferentie bestaat waarschijnlijk wel, maar lijkt alvast niet zo evident. We hebben ook geen referenties gevonden van een dergelijke gewogen *z*-toets. Hedendaagse statistische software biedt daarentegen de mogelijkheden voor een correcte gewogen analyse van de afhankelijkheid van twee categorische variabelen in één databestand. In SPSS is dat mogelijk met Complex Samples.

In eerste instantie tonen we hieronder een gewogen Chi-kwadraattoets, waarbij we de gewichten gebruiken zoals SPSS dat default doet[4]. Dat default gebruik van gewichten is niet correct, maar vormt wel een goede opstap naar de Complex Samples analyse nadien.

Om een gewogen analyse te kunnen uitvoeren, moet ook het gewicht van de twee jaargangen in het samengevoegde databestand opgenomen worden. Voor beide jaargangen

[3] Het aantal vrijheidsgraden wordt als volgt berekend: (aantal kolommen – 1) x (aantal rijen – 1).

[4] Met *default* gebruik van gewichten in SPSS bedoelen we eerst een gewicht definiëren via "Weight Cases" en nadien de analyses laten lopen via "Analyze", bijvoorbeeld "Crosstabs", "Regression"...

nemen we het (cross-sectionele) gewicht[5], voegen dat als nieuwe variabele toe aan het databestand en gebruiken die nieuwe variabele als gewicht bij de analyse. De op die manier gewogen kruistabel met bijhorende Chi-kwadraattoets wordt getoond in tabel 12.

Tabel 12 Tevredenheid met de woning volgens surveyjaar, gewogen

		2008	2009	Totaal
niet heel tevreden	aantal	799	771	1.570
	kolom%	54,5	57,2	
heel tevreden	aantal	666	577	1.243
	kolom%	45,5	42,8	
	Totaal	1.465	1.348	2.813

χ^2 = 2,009; df = 1, p = 0,156

In tabel 12 valt als eerste op dat de gewogen percentages minder sterk van elkaar verschillen (45,5% versus 42,8% heel tevredenen) dan de ongewogen percentages (45,8% versus 42,0% heel tevredenen). Hiermee samenhangend wijst de χ^2-waarde en bijhorende probabiliteit erop dat het verschil tussen 2008 en 2009 niet significant is. Merk ook op dat het aantal cases gedaald lijkt. De totale omvang is nu gelijk aan 2.813 respondenten, terwijl dat er in tabel 10 nog 2.833 waren. Dit is mede een gevolg van het feit dat de respondenten die uit de analyse weggelaten werden (de niet-Belgen en de 85-plussers) over het algemeen een hoger gewicht hadden dan de respondenten die wel in de analyse opgenomen zijn. Eigenlijk doen die gewogen aantallen er niet toe. Alle aantallen in tabel 12 zijn immers "virtueel". Het *aantal tevredenen in het databestand* is gelijk aan het ongewogen aantal tevredenen. Het aantal tevredenen in tabel 12 zou geïnterpreteerd kunnen worden als het geschatte aantal in een steekproef van 2.813 personen met de populatieverdeling die gebruikt werd bij de weging. Dat velen voorbijgaan aan die interpretatie van de gewogen aantallen is op zich niet zo problematisch. Problematischer is wel dat ook de Chi-kwadraattoets op die virtuele aantallen gebaseerd is. Daarom zijn hypothesetoetsen die op een correctere wijze omgaan met gewichten noodzakelijk. We hebben dezelfde gewogen toets voor onafhankelijkheid van de variabelen surveyjaar en tevredenheid met de woning uitgevoerd in SPSS Complex Samples. Die procedure houdt rekening met de werkelijke steekproefaantallen en ook met het feit dat ongelijke gewichten het statistische onderscheidingsvermogen doen dalen. In SPSS Complex Samples ziet de tabel er hetzelfde uit als tabel 12 met identieke gewogen percentages maar ook (indien gevraagd) ongewogen aantallen. Maar de waarschijnlijkheid van de toetswaarde[6] bedraagt 0,189. De conclusie blijft dat er geen significant verschil is, maar deze kans is toch nog duidelijk groter dan de hierboven bekomen p-waarde.

Meer uitleg over de correcte omgang met gewichten (in SPSS) is te vinden in Pickery (2010). Dit voorbeeld moest vooral aantonen dat het goed mogelijk is om rekening te houden met gewichten als de werkwijze van het samenvoegen van databestanden gevolgd wordt. Bij die gewogen analyses moet dan adequate software gebruikt worden. Als de afhankelijke variabele metrisch is, zijn andere toetsen van toepassing, maar ook dan is adequaat gebruik van gewichten in samengevoegde databestanden mogelijk en nodig. In de laatste sectie van

[5] Tot voor enkele jaren werd bij de SCV-survey zowel een cross-sectioneel als een longitudinaal gewicht voorzien. Voor dat longitudinale gewicht (de zogenaamde referentieweging) werd niet de (geschatte) populatieverdeling van het betreffende jaar gebruikt, maar wel de (geschatte) populatieverdeling van het jaar 2000. De bekomen percentages moeten bij gebruik van dat gewicht geïnterpreteerd worden als een soort gestandaardiseerde percentages, bijvoorbeeld "het aantal tevredenen als de verdeling volgens geslacht, leeftijd en opleidingsniveau dezelfde zou zijn als in 2000". Omdat dit voor verwarring zorgde, wordt dat referentiegewicht in de laatste jaargangen niet meer aangeboden. Voor de probleemstelling hier zou het trouwens onlogisch zijn om dat longitudinale gewicht te gebruiken. De eenvoudige vraagstelling is immers "Is er een verschil tussen beide jaargangen" zonder controle voor welke variabelen dan ook.

[6] De toets voor onafhankelijkheid van twee dichotome variabelen in SPSS Complex Samples is een Adjusted F-toets.

deze tekst komen we nog terug op het gebruik van gewichten in samengevoegde databestanden.

4. Twee surveys, verschil tussen percentages onder controle van andere variabelen

Tot nu toe hebben we ons beperkt tot ruwe verschillen tussen surveyjaargangen. Soms rijst de vraag of die ruwe verschillen ook blijven gelden onder controle van een aantal kenmerken. Tevredenheid met de woning hangt bijvoorbeeld samen met huishoudtype. Misschien is er een significant verschil in tevredenheid tussen 2008 en 2009 omdat de verdeling over de verschillende huishoudtypes anders is in beide surveys (en al dan niet ook veranderd is in de samenleving)? De eenvoudigste manier om zulke vraag te beantwoorden is overstappen naar een regressiemodel. In deze sectie tonen wij dat regressiemodel in eerste instantie ongewogen omdat de berekeningen zo makkelijker te volgen zijn, daarna volgt de gewogen variant.

In een regressiemodel wordt tevredenheid met de woning de afhankelijke variabele. Omdat dat hier een dichotome categorische variabele is, grijpen we terug naar een binaire logistische regressie. Meer uitleg hierover is te vinden in Hosmer & Lemeshow (2000). Hiervoor maken we van 'tevredenheid' een dummy (0/1-variabele). In een eerste model tonen we de resultaten van de logistische regressie met alleen surveyjaar als onafhankelijke variabele. Ook daarvan hebben we een dummy gemaakt, zoals tabel 13 toont. Voor die dummy-variabele (survey2009) krijgen alle respondenten van 2008 waarde 0 en alle respondenten van 2009 waarde 1.

Tabel 13 Dummy – survey2009

	Aantal	Percentage
0	1.475	52,1
1	1.358	47,9
Totaal	2.833	100,0

Een ongewogen logistische regressie met tevredenheid als afhankelijke variabele en de dummy survey2009 als onafhankelijke variabele, geeft volgende resultaten:

Tabel 14 Logistische regressie van tevredenheid met de woning en alleen surveyjaar als onafhankelijke variabele

	b	st.fout	sign.	exp(b)
intercept	-0,167	0,052	0,001	0,846
survey2009	-0,158	0,076	0,039	0,855

Eigenlijk bevat deze logistische regressie grotendeels dezelfde informatie als tabel 10 (ongewogen kruistabel). De *p*-waarde voor het "effect" van survey2009 is exact dezelfde als deze van de Chi-kwadraattoets. Verder is 0,855 een oddsratio die ook uit die tabel berekend kan worden:

$$\frac{570/788}{676/799} = 0{,}855$$

De logistische regressie geeft dus dezelfde resultaten als de kruistabel, maar laat ook toe om bijkomende onafhankelijke variabelen op te nemen in het model. In ons tweede model nemen we leeftijd (in 3 categorieën), huishoudtype (6 categorieën) en urbanisatiegraad van de woonplaats (6 categorieën) op. De univariate frequentieverdelingen van die drie variabelen in het samengevoegde databestand worden weergegeven in tabellen 15 tot 17.

Tabel 15 Leeftijd

	Aantal	Percentage
18-40	1.020	36,0
41-60	1.039	36,7
61-85	774	27,3
Totaal	2.833	100,0

Tabel 16 Huishoudtype

	Aantal	Percentage
woont bij ouders	394	13,9
woont alleen	361	12,7
woont niet met partner wel met kind(eren)	91	3,2
woont met partner	993	35,1
woont met partner en kind(eren)	913	32,2
andere	81	2,9
Totaal	2.833	100,0

Tabel 17 Urbanisatiegraad van de woonplaats

	Aantal	Percentage
grootsteden	333	11,8
centrumsteden	361	12,7
stedelijke rand	435	15,4
kleinere steden	579	20,4
overgangsgebied	657	23,2
platteland	468	16,5
Totaal	2.833	100,0

In onze logistische regressie nemen we deze 3 variabelen op als onafhankelijke, dummy-gecodeerde variabelen, naast de dummy survey2009. De resultaten van dat model worden weergegeven in tabel 18.

Tabel 18 toont dat de drie bijkomende variabelen duidelijk een impact hebben op de tevredenheid met de woning. Ouderen zijn meer tevreden dan jongeren (zie de positieve b-parameters en de oddsratio's die groter zijn dan 1). Verder zijn diegenen die bij de ouders wonen het meest tevreden en mensen die met kinderen maar zonder partner wonen het minst tevreden. Tenslotte is de tevredenheid nergens lager dan in de grootsteden; alleen met de stedelijke rand is het verschil niet significant.

Tabel 18 Logistische regressie van tevredenheid met de woning en leeftijd, urbanisatiegraad, huishoudtype en surveyjaar als onafhankelijke variabelen (ongewogen)

	b	st.fout	sign.	exp(b)
intercept	-0,480	0,158	0,002	0,619
leeftijd			0,000	
18-40 (referentie)				
41-60	0,428	0,103	0,000	1,535
61-85	0,762	0,119	0,000	2,142
huishoudtype			0,000	
woont bij ouders (refer.)				
woont alleen	-0,639	0,169	0,000	0,528
woont niet met partner wel met kind(eren)	-1,534	0,289	0,000	0,216
woont met partner	-0,438	0,144	0,002	0,645
woont met partner en kind(eren)	-0,477	0,137	0,000	0,621
andere	-0,520	0,258	0,044	0,594
urbanisatie			0,003	
grootsteden (referentie)				
centrumsteden	0,408	0,160	0,011	1,503
stedelijke rand	0,268	0,154	0,081	1,307
kleinere steden	0,513	0,145	0,000	1,670
overgangsgebied	0,411	0,143	0,004	1,509
platteland	0,560	0,151	0,000	1,750
survey2009	-0,153	0,077	0,048	0,858

Het belangrijkste voor ons voorbeeld is het resultaat voor survey2009. De parameterschatting is vrijwel gelijk aan deze in het model zonder de 3 bijkomende variabelen en het effect blijft (net) significant op niveau $\alpha = 0,05$. De conclusie is dus dat er ook na controle voor leeftijd, urbanisatiegraad en huishoudtype een significant verschil is in tevredenheid met de woning tussen 2008 en 2009; in 2009 ligt de tevredenheid lager.

Ook bij deze logistische regressies is het mogelijk om rekening te houden met gewichten. Voor de volledigheid geven we daarom ook nog de resultaten mee van de cross-sectioneel gewogen logistische regressie (schatting met SPSS Complex Samples). In de gewogen analyse blijkt het verschil tussen 2008 en 2009 niet significant, maar dat was ook al zo als er niet gecontroleerd werd voor bijkomende variabelen.

Tabel 19 Logistische regressie van tevredenheid met de woning en leeftijd, urbanisatiegraad, huishoudtype en surveyjaar als onafhankelijke variabelen (gewogen)

	b	st.fout	sign.	exp(b)
intercept	-0,619	0,169	0,000	0,538
leeftijd				
18-40 (referentie)				
41-60	0,433	0,113	0,000	1,542
61-85	0,734	0,131	0,001	2,084
huishoudtype				
woont bij ouders (refer.)				
woont alleen	-0,578	0,181	0,001	0,561
woont niet met partner				
wel met kind(eren)	-1,555	0,309	0,000	0,211
woont met partner	-0,368	0,160	0,022	0,692
woont met partner en				
kind(eren)	-0,488	0,148	0,001	0,614
andere	-0,313	0,278	0,261	0,732
urbanisatie				
grootsteden (referentie)				
centrumsteden	0,513	0,165	0,002	1,670
stedelijke rand	0,378	0,162	0,019	1,459
kleinere steden	0,551	0,152	0,000	1,735
overgangsgebied	0,547	0,153	0,000	1,727
platteland	0,696	0,159	0,000	2,006
survey2009	-0,106	0,082	0,198	0,858

Deze sectie toonde hoe de significantie van een verschil onderzocht kan worden onder controle van andere variabelen, via een overstap naar een regressiemodel. Dit principe geldt natuurlijk ook als de afhankelijke variabele metrisch is. In dat geval, valt de keuze op een lineaire regressie.

5. Meerdere surveys, verschillende dummy's

Doordat een regressiemodel meerdere onafhankelijke variabelen kan bevatten, laat het ook toe om meer dan twee jaargangen van een survey in de analyse te betrekken. Eén manier om dat te doen, is werken met verschillende dummy's. In het volgende voorbeeld bekijken we het vertrouwen in de Vlaamse Regering gemeten in vijf jaargangen van de SCV-survey (2004 tot en met 2008). In al die surveys werd dat vertrouwen gemeten met een vijfpuntenschaal. Wij hebben terug gedichotomiseerd waarbij we zeer veel vertrouwen en veel vertrouwen afzetten tegen de drie andere categorieën. Tabel 20 toont dat er toch wel wat variatie is in dat vertrouwen. In 2004 had slechts 18% van de Vlamingen (zeer) veel vertrouwen in de Vlaamse Regering, van 2006 tot 2008 was dat om en bij de 30%. Bemerk dat we vanaf deze sectie alleen nog gewogen percentages en gewogen analyses rapporteren. De steekproefomvang (N) wordt echter niet gewogen.

Tabel 20 (Zeer) veel vertrouwen in de Vlaamse Regering volgens surveyjaargang

	Gewogen %	Ongewogen N
2004	18,0	1.535
2005	24,9	1.511
2006	30,7	1.526
2007	31,1	1.406
2008	29,5	1.460

Een eerste manier om deze data te analyseren is het aanmaken van een dummy voor de verschillende surveyjaargangen. Zo heeft de dummy survey2004 bijvoorbeeld waarde 1 voor de respondenten van 2004 en waarde 0 voor alle andere surveyjaren. Die dummy's kunnen dan opgenomen worden als onafhankelijke variabelen van de (logistische) regressie. Het aantal dummy's in de regressie moet altijd kleiner zijn dan het aantal jaargangen. De niet opgenomen dummy bepaalt dan het referentiejaar. In tabellen 21 en 22 nemen we dus telkens vier dummy's op. In tabel 21 is 2004 het referentiejaar, in tabel 22 is dat 2008.

Tabel 21 Logistische regressie van vertrouwen in de Vlaamse Regering met de verschillende dummy's voor vier surveyjaren als onafhankelijke variabelen (2004 is referentie)

	b	st.fout	sign.	exp(b)
intercept	-1,516	0,071	0,000	0,220
survey2005	0,414	0,096	0,000	1,513
survey2006	0,706	0,093	0,000	2,025
survey2007	0,721	0,094	0,000	2,056
survey2008	0,647	0,093	0,000	1,910

Nagelkerke R^2: 0,019

Tabel 22 Logistische regressie van vertrouwen in de Vlaamse Regering met de verschillende dummy's voor vier surveyjaren als onafhankelijke variabelen (2008 is referentie)

	b	st.fout	sign.	exp(b)
intercept	-0,869	0,060	0,000	0,419
survey2004	-0,647	0,093	0,000	0,524
survey2005	-0,233	0,088	0,008	0,792
survey2006	0,059	0,084	0,487	1,060
survey2007	0,074	0,086	0,388	1,077

Nagelkerke R^2: 0,019

De regressiemodellen in tabellen 21 en 22 zijn equivalent. Dat blijkt bijvoorbeeld uit de pseudo-verklaarde variantie, hier weergeven door de *Nagelkerke R^2*, die dezelfde is. Maar de modellen verschillen natuurlijk ook. In tabel 21 geldt het intercept voor 2004, de jaargang waarvan de dummy niet is opgenomen in de regressie. Zo kan uit exp(b) het percentage uit tabel 20 worden afgeleid: 0,220/(1+0,220) = 0,18 of 18%. In tabel 22 geldt het intercept voor 2008. De dummy's en de significanties daarvan hebben betrekking op het verschil tussen dat referentiejaar en de andere jaargangen. Zo blijkt uit tabel 21 dat in alle onderzochte jaren er significant meer Vlamingen (zeer) veel vertrouwen hadden in de Vlaamse Regering dan in 2004. Uit tabel 22 blijk dat het vertrouwen in 2004 en 2005 lager lag dan in 2008, maar 2006 en 2007 verschilden niet significant van dat referentiejaar.

Doordat survey2006 en survey2007 in het laatste model niet significant zijn, zou overwogen kunnen worden om die weg te laten uit het model. Het model zal hierdoor niet significant verslechteren. Maar het resultaat zou ook zijn dat de referentiegroep drie surveys omvat (2006, 2007 en 2008) en het intercept minder gemakkelijk te interpreteren is. Bovendien zouden beide modellen niet meer volledig equivalent zijn. De vergelijking van tabel 21 en

tabel 22 toont dat het perfect mogelijk is dat er twee equivalente regressiemodellen zijn, waarvan er één vier significante effecten bevat en het andere slechts twee. Welk van beide de voorkeur verdient, kan niet bepaald worden op basis van statistische argumenten. Inhoudelijke overwegingen moeten de keuze voor één bepaald referentiejaar en bijgevolg voor één van beide modellen bepalen. Die inhoudelijke argumenten kunnen afgeleid worden uit voorgaand onderzoek, beleidsdocumenten…

6. Meerdere surveys, een tijdsvariabele (en machtsverheffingen daarvan)

Een andere manier om met verschillende jaargangen om te gaan is het surveyjaar te modelleren als een tijdsvariabele. Eigenlijk hebben we zo één variabele (surveyjaar) waarvan we de waarden (2004, 2005…) gewoon als metrisch beschouwen. We zullen die variabele opnemen als onafhankelijke variabele van onze logistische regressie, samen met eventuele machtstransformaties ervan. Als we uitsluitend de oorspronkelijke tijdsvariabele opnemen, modelleren we een monotoon stijgende of dalende trend. Door opname van machtstransformaties kunnen we complexere evoluties modelleren. Het voorbeeld in deze sectie zal dat illustreren. Omwille van een aantal redenen (interpretatie intercept, minder multicollineariteit bij de machtsverheffingen) kan het echter interessant zijn om die variabele te centreren, dat wil zeggen het gemiddelde aftrekken van de oorspronkelijke waarde. Tabellen 23 en 24 tonen het resultaat van het berekenen van die deviatiescores.

Tabel 23 Surveyjaar in de samengevoegde dataset voor vertrouwen in de Vlaamse Regering

		Aantal	Percentage
2004		1.554	20,6
2005		1.522	20,2
2006		1.540	20,4
2007		1.449	19,2
2008		1.475	19,6
	Totaal	7.540	100,0

We hebben vijf jaargangen, waarvan 2006 de middelste is. Dat is dus ook het (afgeronde) gemiddelde. De nieuwe variabele (deviatiescore) gaat bijgevolg van -2 tot 2.

Tabel 24 Deviatiescore van surveyjaar in de samengevoegde dataset voor vertrouwen in de Vlaamse Regering

		Aantal	Percentage
-2		1.554	20,6
-1		1.522	20,2
0		1.540	20,4
1		1.449	19,2
2		1.475	19,6
	Totaal	7.540	100,0

We beschouwen dit als een metrische variabele. Verder kunnen we in het logistische regressiemodel niet alleen de gewone deviatiescore opnemen als onafhankelijke variabele, maar ook machtsverheffingen ervan. Tabel 25 toont de beschrijvende statistieken tot en met de 4de macht.

Tabel 25 Beschrijvende statistieken van de verschillende machtsverheffingen van de deviatiescore van surveyjaar

	Minimum	Maximum	Gemiddelde	N
deviatiescore surveyjaar	-2,0	2,0	-0,0	7.540
deviatiescore surveyjaar tot 2de macht	0,0	4,0	2,0	7.540
deviatiescore surveyjaar tot 3de macht	-8,0	8,0	-0,1	7.540
deviatiescore surveyjaar tot 4de macht	0,0	16,0	6,8	7.540

We nemen nu stap voor stap deze onafhankelijke variabelen op in onze logistische regressie. In het eerste model alleen de deviatiescore van surveyjaar, in het tweede model ook het kwadraat daarvan enzoverder. De resultaten van die verschillende modellen worden gerapporteerd in tabel 26.

Tabel 26 Logistische regressies van vertrouwen in de Vlaamse Regering met surveyjaar metrisch gemodelleerd

	b	st.fout	sign.	exp(b)
Model 1 – lineair				
intercept	-1,011	0,028	0,000	0,364
deviatiescore surveyjaar	0,151	0,019	0,000	1,163
Nagelkerke R^2: 0,013				
Model 2 – tot 2de macht				
intercept	-0,839	0,043	0,000	0,432
deviatiescore surveyjaar	0,160	0,021	0,000	1,174
deviatiescore surveyjaar tot 2de macht	-0,090	0,017	0,000	0,914
Nagelkerke R^2: 0,019				
Model 3 – tot 3de macht				
intercept	-0,838	0,043	0,000	0,432
deviatiescore survey	0,150	0,059	0,011	1,162
deviatiescore surveyjaar tot 2de macht	-0,090	0,017	0,000	0,914
deviatiescore surveyjaar tot 3de macht	0,003	0,017	0,852	1,003
Nagelkerke R^2: 0,019				
Model 4 – tot 4de macht				
intercept	-0,811	0,059	0,000	0,445
deviatiescore survey	0,151	0,060	0,012	1,163
deviatiescore surveyjaar tot 2de macht	-0,152	0,095	0,109	0,859
deviatiescore surveyjaar tot 3de macht	0,003	0,017	0,869	1,003
deviatiescore surveyjaar tot 4de macht	0,014	0,021	0,505	1,014
Nagelkerke R^2: 0,019				

Het eerste model wordt een lineair model genoemd omdat alleen de gewone deviatiescore wordt opgenomen als onafhankelijke variabele. Uit dit model valt af te leiden dat er inderdaad een lineaire toename is van de ***logit*** (zeer) veel vertrouwen in de Vlaamse Regering. Die logit stijgt met 0,151 per surveyjaar. Dat wil niet zeggen dat volgens dit model het percentage Vlamingen met (zeer) veel vertrouwen ook lineair stijgt. De afhankelijke variabele van de logistische regressie is immers niet de kans op (zeer) veel vertrouwen, maar de logit-transformatie van die kans[7]. In een volgende stap nemen we ook het kwadraat op

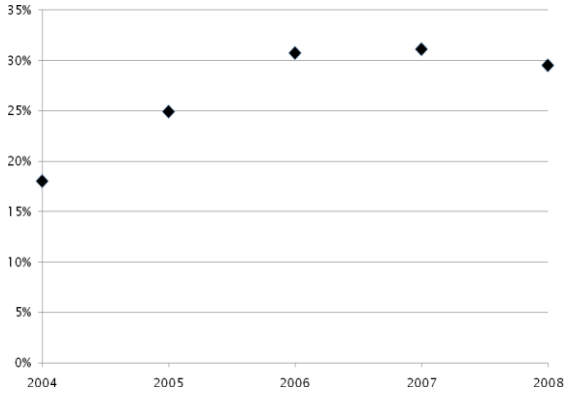
[7] Meer uitleg over de logit-transformatie is bijvoorbeeld te vinden in Pampel (2000). Hier kunnen we ons beperken tot de vaststelling dat een lineair effect op de logit overeenkomt met een meer S-vormige curve voor het effect op de kans.

van de deviatiescore. Ook die variabele blijkt een significant effect te hebben op de kans om (zeer) veel vertrouwen te hebben. De derdemachtsterm en vierdemachtsterm verbeteren het model echter niet meer. Die variabelen blijken niet significant en een likelihood-ratiotoets (niet gerapporteerd) wijst uit dat opname ervan geen significante verbetering van het model met zich meebrengt.

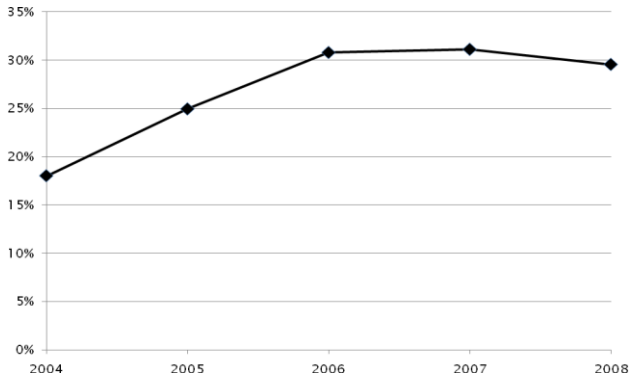
Het is niet zo eenvoudig om de verschillende parameters van deze modellen te interpreteren. Daarom wordt er vaak teruggegrepen naar grafieken met de voorspelde waarden (en curves). Op de volgende pagina tonen we eerst de geobserveerde waarden, nadien een grafiek met de voorspellingen op basis van het model met dummy's en tenslotte een grafiek met de voorspellingen op basis van de vier modellen in tabel 26. We gebruiken die grafieken ook als basis om de verschillende modellen te vergelijken.

Figuur 2 toont de geobserveerde waarden, figuur 3 de voorspellingen op basis van het model met dummy's. De achterliggende idee van dat model is eigenlijk een aparte schatting per surveyjaar. De voorspellingen vertonen dus ook geen vloeiend patroon. Eigenlijk wordt er helemaal geen evolutie gemodelleerd en je kan je afvragen of het opportuun is om de punten te verbinden.

Figuur 2 Geobserveerd percentage Vlamingen met (zeer) veel vertrouwen in de Vlaamse Regering

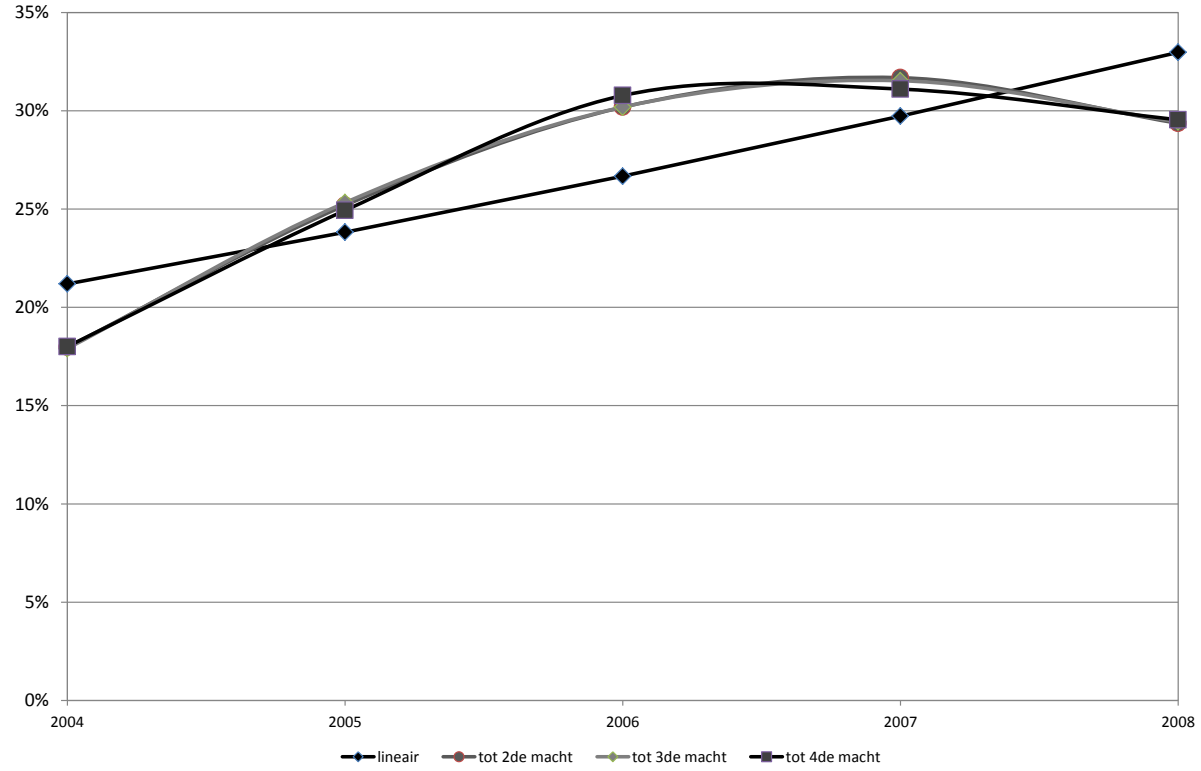


Figuur 3 Voorspeld percentage Vlamingen met (zeer) veel vertrouwen in de Vlaamse Regering op basis van het model met 4 dummy's



Dat is heel anders bij figuur 4. Omdat de invloed van het surveyjaar metrisch gemodelleerd werd, zijn er niet alleen voorspellingen voor de verschillende jaren, maar ook voor alle tussenliggende punten. Bij die modellen (waarvan de parameters zich dus bevinden in tabel 26) is het veel logischer om ook de curves te tekenen. Net zoals tabel 26 toont ook figuur 4 dat er weinig verschillen zijn tussen de modellen 2 tot en met 4. De voorspelde waarden zijn vrijwel gelijk en tonen een stijgend vertrouwen tussen 2004 en 2006, waarna het min of meer gelijk blijft. Er is wel een groot verschil met het lineaire model, dat een continue stijging toont, die voor de periode 2004-2006 wel minder sterk is dan die van de modellen met meer parameters.

Figuur 4 Voorspeld percentage Vlamingen met (zeer) veel vertrouwen in de Vlaamse Regering waarbij de tijdsevolutie metrisch gemodelleerd werd



Als we figuren 2, 3 en 4 op elkaar zouden leggen (met dezelfde schaal), zou blijken dat de voorspelde waarden van het model met dummy's en van model 4 (met modellering van surveyjaar tot 4de macht) gelijk zijn aan de geobserveerde waarden. In beide modellen zijn er vier parameters voor vijf tijdsmomenten. Dat is ook het maximum. In een model waar tijd metrisch gemodelleerd wordt, kan je ook niet verder gaan dan een macht gelijk aan het aantal tijdsmomenten min 1. Model 4 is daarom ook equivalent aan het model met vier dummy's. De filosofie is anders: in het ene geval modelleer je elk tijdsmoment afzonderlijk; in het andere geval modelleer je een evolutie. Maar het resultaat, afgemeten aan de voorspelde waarden of pseudo-verklaarde variantie is identiek. De *Nagelkerke* R^2 bedraagt eveneens 0,019. Als we meer cijfers achter de komma zouden meegeven, zou blijken dat alleen voor model 4 die waarde gelijk is aan deze van het model met dummy's. De *Nagelkerke* R^2 van modellen 2 en 3 is een beetje kleiner.

De keuze tussen beide vormen van modellering (met dummy's of metrisch) wordt bepaald door verschillende afwegingen. In tabel 26 krijgen we geen directe significantietoets voor de verschillen tussen twee afzonderlijke jaargangen. We krijgen wel een indicatie van de significantie van een algemene trend, maar als het de expliciete bedoeling is jaargangen te vergelijken, is het logischer te kiezen voor een model met dummy's. Een model met een metrische modellering is niet aangewezen als er zeer weinig meetpunten zijn of als de data alleen maar "bokkesprongen" vertonen. Het sluit daarentegen beter aan bij de realiteit als er een evolutie waar te nemen lijkt (en die moet niet noodzakelijk monotoon stijgend of dalend zijn). Als er (zeer) veel tijdsmomenten zijn, vraagt een model met dummy's snel veel parameters. Een belangrijk voordeel van de metrische modellering is dan dat het logischer is om onnodige parameters weg te laten. Bij het voorbeeld van het vertrouwen in de Vlaamse Regering volstaan we duidelijk met een modellering tot de 2de macht (twee onafhankelijke variabelen). We kunnen dus een spaarzamer model toepassen dat de realiteit toch voldoende goed weergeeft. Als we in het model met dummy's niet-significante variabelen weglaten (tabel 22), krijgen we een minder logische en moeilijker te interpreteren referentiegroep. Een nadeel van de metrische modellering tot slot is dat de parameters op zich niet altijd

makkelijk interpreteerbaar zijn. Een grafiek met de curve met de voorspelde waarden, is dan op zijn plaats.

Bemerk dat de figuren "voorspelde waarden" geven voor gemeten surveyjaren. Het zijn dus voorspellingen van geobserveerde variabelen op basis van een model. Ook al is het aanlokkelijk, het model en de figuur hebben niet de expliciete bedoeling om voorspellingen te doen over de toekomst. Bij variabelen die aan grote schommelingen onderhevig zijn, zoals vertrouwen, blijken die voorspellingen vaak ook niet correct.

7. Meerdere surveys, tijdsvariabele en controle voor bijkomende variabelen

In een laatste voorbeeld brengen we de elementen van sectie 6 (modellering met een tijdsvariabele) en sectie 4 (controle voor bijkomende variabelen) samen.

In de verschillende SCV-surveys werd ook de deelname aan kerkelijke erediensten bevraagd. De vraag met zeven antwoordcategorieën in de surveys luidt: *"Mensen nemen wel eens deel aan kerkelijke of religieuze plechtigheden naar aanleiding van een huwelijk, begrafenis en dergelijke. Als we deze NIET meetellen, hoe vaak neemt u dan deel aan kerkelijke of godsdienstige erediensten?"*. Ook hier maken we een dichotome variabele van. We kijken naar de groep die minstens wekelijks een eredienst bijwoont en dat voor de jaren 2000 tot en met 2007. Tabel 27 toont dat de groep regelmatige bezoekers van erediensten van 2000 tot 2007 stelselmatig kleiner geworden is.

Tabel 27 Minstens wekelijks bijwonen van kerkelijke of godsdienstige erediensten

	Gewogen %	Ongewogen N
2000	12,3	1.283
2001	11,3	1.445
2002	10,8	1.475
2003	8,8	1.434
2004	8,9	1.553
2005	10,8	1.519
2006	8,0	1.540
2007	7,1	1.448

Onze analyse omvat acht surveyjaren. We hebben een deviatiescore berekend, waarbij we gecentreerd hebben rond 2003. In een logistische regressie met wekelijks bezoek als afhankelijke variabele nemen we die deviatiescore en enkele machtsverhoudingen ervan op. Uiteindelijk blijkt de toevoeging van een tweede-, derde- of vierdemachtsterm geen verbetering van het model op te leveren en behouden we een eenvoudig model met een lineaire afname van de logit van het minstens wekelijks bijwonen van erediensten. Tabel 28 toont de parameters van dat model.

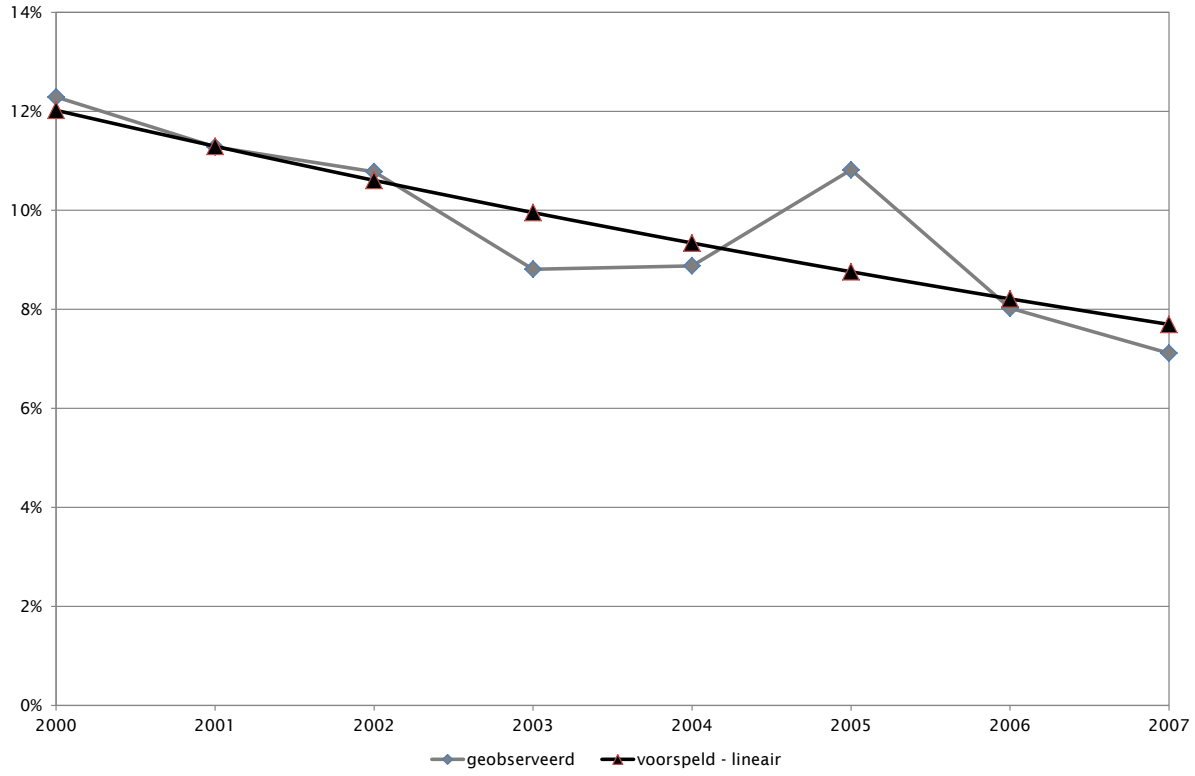
Tabel 28 Logistische regressie van het minstens wekelijks bijwonen van erediensten met surveyjaar metrisch gemodelleerd

	b	st.fout	sign.	exp(b)
intercept	-2,202	0,034	0,000	0,111
deviatiescore surveyjaar	-0,071	0,015	0,000	0,932

Nagelkerke R^2: 0,005

Figuur 5 vergelijkt de door dit model voorspelde waarden met de geobserveerde percentages. Voor de meeste jaren blijkt de voorspelling redelijk accuraat. Alleen voor 2003 is er een duidelijke overschatting en voor 2005 een duidelijke onderschatting. Misschien is het model hier wel beter dan de schatting door beide surveys. Het lijkt alvast niet onwaarschijnlijk dat de afwijkende waarden voor die surveys vooral uitingen zijn van steekproeffluctuaties en dat de ware percentages meer aanleunen bij de door het model geschatte percentages. Maar uitsluitsel daarover hebben we natuurlijk niet.

Figuur 5 Geobserveerde en voorspelde percentages van het minstens wekelijks bijwonen van kerkelijke of godsdienstige erediensten volgens surveyjaar



In een volgende stap nemen we in het regressiemodel drie bijkomende onafhankelijke variabelen op: geslacht, leeftijd en opleidingsniveau. De laatste twee variabelen werden tot drie categorieën teruggebracht. Tabel 29 toont de resultaten van die logistische regressie.

Het grootste aandeel regelmatige bezoekers van erediensten vinden we – niet verrassend – bij de oudste leeftijdsgroep. Zowel bij de middengroep als bij de jongste groep is dit aandeel veel kleiner. De verschillen volgens leeftijd zijn zeer groot. Verder zijn er meer vrouwen dan mannen die minstens wekelijks een kerkelijke of religieuze eredienst bijwonen en zijn er ook verschillen volgens opleidingsniveau. Het hoogste percentage vinden we bij de hoogste opleidingscategorie. Maar deze verschillen zijn beduidend kleiner dan de verschillen volgens leeftijd. Onder controle van deze drie bijkomende variabelen, blijft het tijdseffect (het effect van surveyjaar) bestaan. Het is zelfs een beetje groter geworden. Tot slot merken we nog op dat de voorspellende waarde sterk gestegen is door de toevoeging van de bijkomende variabelen (*Nagelkerke* R^2 = 0,160).

Bij dit voorbeeld is de keuze voor een modellering met een tijdsvariabele evident. Een spaarzaam model met slechts één parameter volstaat om een redelijk goed beeld te geven van de evolutie over acht tijdsmomenten. Het model is bovendien eenvoudiger te interpreteren omdat er geen machtsverheffingen van de tijdsvariabele nodig zijn: er is een

significante monotoon dalende trend. Deze trend blijft behouden wanneer gecontroleerd wordt voor leeftijd, geslacht en opleidingsniveau. Omgekeerd kan op basis van dit model gesteld worden dat de verschillen volgens die drie kenmerken blijven bestaan wanneer gecontroleerd wordt voor de evolutie in de tijd.

Tabel 29 Logistische regressies van het minstens wekelijks bijwonen van erediensten met surveyjaar metrisch gemodelleerd

	b	st.fout	sign.	exp(b)
intercept	-0,855	0,101	0,000	0,425
deviatiescore surveyjaar	-0,085	0,016	0,000	0,918
opleidingsniveau				
laag	-0,222	0,097	0,021	0,801
midden	-0,240	0,098	0,014	0,787
hoog (referentie)				
leeftijd				
18-40	-2,392	0,114	0,000	0,091
41-60	-1,482	0,083	0,000	0,227
61-85 (referentie)				
geslacht				
man	-0,152	0,070	0,030	0,859
vrouw (referentie)				

Nagelkerke R^2: 0,160

8. Een belangrijke voetnoot over gewichten

In sectie 3 noemden we de mogelijkheid om rekening te houden met gewichten één van de voordelen van de strategie van het samenvoegen van databestanden. Hoe de gewichten moeten toegevoegd worden aan het samengevoegde bestand, is iets complexer dan daar werd gesuggereerd. In de voorbeelden van sectie 3 en de daarop volgende secties namen we telkens het cross-sectionele gewicht voor de verschillende jaargangen, gaven dit één uniforme naam en namen dit op als nieuwe variabele in de samengevoegde dataset. Dat was voor de getoonde voorbeelden correct, maar is niet altijd de juiste oplossing.

In de literatuur hierover gaat het meestal over gewichten die sommeren tot het populatietotaal. Door die gewichten te gebruiken kan je uitspraken doen over aantallen, bijvoorbeeld schattingen van het aantal Vlamingen met (zeer) veel vertrouwen in de Vlaamse Regering of het aantal Vlamingen dat minstens wekelijks kerkelijke of religieuze erediensten bijwoont (in plaats van het percentage). Het is evident dat, als je databestanden samenvoegt, die gewichten behoudt, en één schatting voor het samengevoegde bestand maakt, je die aantallen zal overschatten. Als je twee surveys met dergelijke gewichten samenvoegt, schat je bijvoorbeeld het *aantal Vlamingen* dat minstens wekelijks kerkelijke of religieuze erediensten bijwoont, ongeveer dubbel zo hoog in als het werkelijke aantal. Merk op dat de interpretatie van één schatting voor het samengevoegde bestand sowieso wat moeilijker is omdat er geen eenduidige populatie is, eerder een gemiddelde populatie. Het probleem stelt zich ook niet als je proporties of percentages rapporteert. Toch kan een eenvoudige stelregel zijn om mogelijke problemen te vermijden, in dit geval de oorspronkelijke gewichten te delen door het aantal surveys dat wordt samengevoegd (Thomas & Wannel, 2009). Een meer ingewikkelde berekening om de impact van de gewichten op de variantie van de schatters te minimaliseren is ook mogelijk. Daarbij wordt aan sommige surveys een grotere rol toebedeeld bij het bepalen van het uiteindelijke gewicht voor het samengevoegde bestand dan aan andere surveys (Chu e.a., 1999). Het probleem hierbij is dat deze strategie andere

gewichten kan opleveren naargelang de variabele waarin de onderzoeker geïnteresseerd is. Dat kan verwarring creëren en is daarom niet wenselijk.

Iets complexer wordt het als één bestand gewichten bevat die sommeren tot het populatietotaal en een ander bestand gewichten die sommeren tot de steekproefomvang (en dus een gemiddelde gelijk aan 1 hebben). Voor de SCV-surveys werd bijvoorbeeld tot 2009 uitsluitend een gewicht voorzien waarbij de ongewogen en gewogen steekproefomvangen gelijk zijn. Sinds 2010 worden ook gewichten voorzien die sommeren tot het populatietotaal (+/- 5 miljoen Vlamingen). De combinatie in één samengevoegd databestand van een gewicht dat sommeert tot de steekproefomvang en een ander gewicht dat sommeert tot de populatieomvang, is problematisch en het gebruik van aangepaste software lost dat probleem niet op. De stelregel bij adequate software zoals SPSS Complex Samples is dat de absolute omvang van de gewichten onbelangrijk is, maar de relatieve verhouding is natuurlijk wel belangrijk. Door een bestand met gewichten die sommeren tot het populatietotaal samen te voegen met een bestand met gewichten die sommeren tot het steekproeftotaal, verandert natuurlijk de onderlinge verhouding van de gewichten. De meest eenvoudige oplossing in dit geval is eerst de gewichten herschalen voor elke survey afzonderlijk, zodanig dat het gemiddelde gewicht binnen elke survey gelijk is aan 1. Nadien kunnen de surveybestanden samengevoegd worden, de herschaalde gewichten een uniforme naam gegeven en toegevoegd aan het databestand als nieuwe variabele. Let wel, ook na deze herschaling blijft het gebruik van adequate software-toepassingen nodig.

Uitleiding

Deze tekst probeerde duidelijk te maken hoe uitspraken gedaan kunnen worden over evoluties in de tijd door het samenvoegen van (resultaten van) cross-sectionele surveys. Het samenvoegen van cross-secties (“pooled cross-sections”) wordt traditioneel afgezet tegen panelanalyse, waarbij één groep gevolgd wordt in de tijd. Panelsurveys kunnen een ander en vaak correcter beeld geven van een veranderingsdynamiek dan samengevoegde cross-secties. Die laatste hebben dan weer het voordeel dat er geen speciale statistische modellen vereist zijn. Bij paneldata moeten er speciale technieken toegepast worden omdat de verschillende observaties gecorreleerd zijn (metingen bij dezelfde personen). Bovendien verloopt de dataverzameling vaak heel wat moeizamer bij panelsurveys. Deze nota beperkte zich tot technieken voor samengevoegde cross-secties en wilde vooral praktijkgericht tonen hoe de analyses op zulke bestanden kunnen gebeuren.

Referenties

Chu, A., Brick J.M. & Kalton, G. (1999). Weights For Combining Surveys Across Time Or Space. In: *Bulletin of the International Statistical Institute: 52nd Session, Contributed Papers, Book 2*, pp. 103-104.

Fleiss, J.L., Levin, B. & Paik, M.C. (2003). *Statistical Methods for Rates & Proportions, Third Edition.* New York: Wiley.

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

Newcombe, R. G. (1998). Two-Sided Confidence Intervals For The Single Proportion: Comparison of Seven Methods. In: *Statistics in Medicine*, 17, 857-872.

Pampel, F. C. (2000). *Logistic Regression*. A Primer. Londen: Sage University Papers.

Pickery, J. (2010). *Aanmaak en gebruik van gewichten voor surveydata. Met toepassing in SPSS. SVR-Methoden & Technieken 2010/3*. Brussel: Vlaamse overheid/Studiedienst van de Vlaamse Regering.

Thomas, S. & Wannel, B. (2009). Combining Cycles of the Canadian Community Health Survey. In: *Health Reports*, 20, 1, 1-6.