Rob Eisinga

Van dwarsdoorsnede naar transitie: het schatten van overgangskansen met herhaalde cross sectie data

Inleiding

Een van de belangrijke recente ontwikkelingen in het grootschalig survey onderzoek is de opeenstapeling van nationale en internationale cross-sectionele databestanden. Een voorbeeld daarvan is het onderzoek *Sociaal culturele ontwikkelingen in Nederland* (SOCON) dat in 1979 door Bert Felling samen met Jan Peters en Osmund Schreuder is opgezet. In het kader daarvan wordt elke vijf jaar een nationale enquête gehouden onder de Nederlandse bevolking van 18 jaar en ouder. Aan de basis van dit onderzoek ligt een willekeurig gekozen dwarsdoorsnede van de Nederlandse populatie. Door een aselect getrokken steekproef van Nederlanders te ondervragen, tracht men opinies, houdingen en gedragingen in de totale Nederlandse samenleving vast te stellen.

Om veranderingen in de populatie door de tijd te kunnen registreren, worden de surveys – zoals SOCON – op gezette tijden herhaald. Daartoe worden telkens nieuwe steekproeven genomen. Nu kleeft er aan het aaneenrijgen van periodiek herhaalde cross-sectionele steekproeven voor het longitudinaal onderzoek een beperking. Door telkens verse steekproeven te nemen, zijn temporele veranderingen in individuele eenheden niet direct observeerbaar. Daarvoor is het nodig dezelfde eenheden herhaaldelijk in een panel te observeren. Volgens sommigen is de analyse van herhaalde cross-sectionele surveys daarom geen echt longitudinaal onderzoek; voor anderen zijn cross sectie gegevens slechts een soort tweede keus artikel: alleen het overwegen waard als er geen panel data voorhanden zijn. Dit is mijn inziens niet terecht. Aan panel data kleven eveneens specifieke bezwaren, zoals selectieve sterfte van het panel en conditionering. Bovendien verouderen panels na verloop van tijd en zijn weinig panel onderzoeken zó opgezet dat ze naast longitudinale gegevens permanent een representatief beeld geven van een (voortdurend veranderende) populatie.

In de afgelopen tijd heb ik me samen met collega Ben Pelzer gebogen over de kwestie welke longitudinale informatie er te ontlenen valt aan een reeks van onafhankelijke trekkingen uit een veranderende populatie over de tijd. In dit verband hebben we ons vooral beziggehouden met de vraag of en in hoeverre het mogelijk is herhaalde cross sectie data te gebruiken voor het bepalen van temporele veranderingen in de toestanden waarin individuen verkeren. Voorbeelden daarvan zijn veranderingen in de positie van individuen op de arbeidsmarkt, veranderingen in

stemgedrag, veranderingen in voorkeur voor merken en producten, etc. Natuurlijk hebben we voor het bepalen van dit soort individuele veranderingen bij voorkeur de beschikking over transitiedata, maar die gegevens zijn dikwijls niet voorhanden.

Wanneer we ons beperken tot een dichotome variabele Y, zoals het wel of niet hebben van een betaalde baan, het wel of niet gaan stemmen bij verkiezingen, dan is het vraagstuk dat we bestuderen methodologisch als volgt te omschrijven. We beschikken over een aantal steekproeven, geobserveerd op opeenvolgende tijdstippen t. De observaties Y van steekproef t duiden we aan met het symbool Y_t. Voor twee opeenvolgende steekproeven, die van tijdstip $t-1$ en die van t, kunnen we de gegevens weergeven met een 2×2 overgangstabel van Y_{t-1} tegen Y_t. Er zijn evenveel overgangstabellen als er steekproeven zijn, minus 1. Tevens zijn er voor verschillende groepen onderzoekseenheden aparte 2×2 tabellen te construeren. Voor al deze overgangstabellen geldt echter dat alleen de marginale (kolom en rij) aantallen bekend (geobserveerd) zijn, terwijl de aantallen in de vier cellen van de tabellen onbekend zijn. De vraag luidt nu: kunnen we de onbekende celaantallen (of proporties) op zinvolle wijze schatten?

De afgelopen decennia hebben onderzoekers van uiteenlopende wetenschapsdisciplines zich met dit vraagstuk (en nauw verwante kwesties) beziggehouden. Zo hebben diverse statistici zich gebogen over de vraag welke informatie de marginaal totalen leveren over de interne cellen en voor verschillende situaties (c.q. steekproefopzetten) schattingsprocedures ontwikkeld (Plackett, 1977; Hamdan en Nasro, 1986; Haber, 1989; Kocherlakota en Kocherlakota, 1992; McCullagh en Nelder, 1992). Verder is er de laatste jaren opnieuw een sterke belangstelling voor het schatten van verbanden op individueel niveau aan de hand van geaggregeerde data (ofwel, meer formeel, voor het bepalen van $P(y \mid x)$ op grond van uitsluitend $P(y)$ en $P(x)$). Dit ecologische inferentie probleem komt in uiteenlopende disciplines aan de orde, zoals de politieke wetenschappen (Achen en Shively, 1995; King, 1997; King, Rosen en Tanner, 2003), epidemiologie (Richardson en Montfort, 2000), statistiek (Wakefield, 2003), marketing (Böckenholt en Dillon, 2000) en econometrie (Cross en Manski, 2002) (zie ook het recente themanummer over ecologische analyse van de *Journal of the Royal Statistical Society, Series A*, 2001, volume 164, issue 1). Ook is er in het kader van de bescherming van persoonsgegevens recentelijk veel methodologische belangstelling voor de vraag in hoeverre het mogelijk is uit aggregeerde gegevensbestanden micro data te construeren (Fienberg, 1997; Tebaldi en West, 1998; Dobra, Tebaldi en West, 2003). Tot slot noemen we de toenemende belangstelling voor het gebruik van maximum entropie schattingsprocedures (i.p.v.

bijvoorbeeld maximum likelihood) voor onder meer het reconstrueren van celaantallen in onvolledig geobserveerde kruistabel data (bijv. Golan, Judge en Robinson, 1994; Golan, Judge en Miller, 1996; Judge, Miller en Tam Cho, 2003) (zie ook het recente themanummer over entropie van de *Journal of Econometrics*, 2002, volume 107, issues 1-2).

In het onderstaande geef ik aan de hand van de bestaande literatuur eerst een beknopt overzicht van de statistische achtergrond van het probleem. Vervolgens plaats ik herhaalde cross sectie data in dit kader. Na de presentatie van het dynamische Markov model voor de analyse van deze data illustreer ik het model aan de hand van een empirisch voorbeeld. Tot slot noem ik enkele wensen voor de toekomst van dit onderzoek.

Achtergrond

We gaan eerst uit van de situatie waarin de interne cellen zijn geobserveerd. Stel we trekken (met teruglegging) een steekproef uit een multinomiale populatie en het resultaat daarvan geven we weer in de vorm van een 2×2 kruistabel. Het symbool y_{ij} geeft het aantal observaties weer in rij i en kolom j, waarbij $i, j = 0,1$, en p_{ij} is de onbekende parameter.

	0	1	
0	y_{00}	y_{01}	$n - y_{t-1}$
1	y_{10}	y_{11}	$y_{t-1} = y_{1.}$
	$n - y_t$	$y_t = y_{.1}$	n

	0	1	
0	p_{00}	p_{01}	$1 - p_{t-1}$
1	p_{10}	p_{11}	$p_{t-1} = p_{1.}$
	$1 - p_t$	$p_t = p_{.1}$	1

De kansverdeling van y_{ij} wordt uiteraard bepaald door de steekproefopzet die aan de tabel ten grondslag ligt. Zijn de marginale aantallen niet gefixeerd en de cellen $(Y_{00}, Y_{01}, Y_{10}, Y_{11})$ random variabelen, dan is er sprake van een multinomiaalverdeling met parameters $(p_{00}, p_{01}, p_{10}, p_{11})$. Als de Y_{ij} sommeren tot een (gefixeerd) totaal aantal n, dan is de kans op de tabel weer te geven als

$$P(Y_{00} = y_{00}, \ldots, Y_{11} = y_{11}) = \frac{n!}{\prod_{i,j} y_{ij}!} \prod_{i,j} p_{ij}^{y_{ij}} ,$$

waarbij $p_{ij} \in [0,1]$ en $\sum_{i,j} p_{ij} = 1$. Stel nu dat y_{t-1} en y_t de optellingen zijn van respectievelijk $y_{10} + y_{11}$ en $y_{01} + y_{11}$. De bijbehorende marginale kansen definiëren we als $p_{t-1} = p_{10} + p_{11}$ en $p_t = p_{01} + p_{11}$. De marginale verdelingen van de sommen zijn binomiaal, dwz. $Y_{t-1} \sim B(n, p_{t-1})$ en $Y_t \sim B(n, p_t)$. Deze binomiaalverdeling geldt ook voor de conditionele verdelingen $y_t \mid y_{t-1} = 0,1$. De likelihood functie kunnen we dan als volgt noteren

$$P(Y_{00} = y_{00}, \ldots, Y_{11} = y_{11}) = \binom{n}{y_{t-1}} p_{t-1}^{y_{t-1}} (1-p_{t-1})^{n-y_{t-1}}$$
$$\times \left[\binom{n-y_{t-1}}{y_{01}} \left(\frac{p_{01}}{1-p_{t-1}} \right)^{y_{01}} \left(\frac{p_{00}}{1-p_{t-1}} \right)^{y_{00}} \times \binom{y_{t-1}}{y_{11}} \left(\frac{p_{11}}{p_{t-1}} \right)^{y_{11}} \left(\frac{p_{10}}{p_{t-1}} \right)^{y_{10}} \right].$$

De likelihood functie van de 2×2 tabel valt dus uiteen in een marginale binomiale random variabele voor de rij totalen en twee conditionele binomialen voor de twee rijen (zie Bishop, Fienberg en Holland, 1975).

Veronderstel nu dat de rij en kolom totalen zijn geobserveerd, maar dat de individuele celaantallen onbekend zijn (vanwege steekproef design, databeveiliging, o.i.d.). In dat geval kunnen er, al naar gelang het aantal gefixeerde marginalen (0, 1 of 2), drie steekproefdesigns aan de geobserveerde tabel ten grondslag liggen (Barnard, 1947).

Beschikken we alleen over de marginale aantallen en is geen van de marginaal totalen gefixeerd, dan kunnen we de bivariaat-binomiaalverdeling hanteren voor het bestuderen van de twee-dimensionele random variabele. Nemen we $g = y_{01}$ om aan te geven dat de celaantallen niet zijn geobserveerd, dan is de gezamenlijke kansfunctie van Y_{t-1} en Y_t gegeven door

$$P(Y_{t-1} = y_{t-1}, Y_t = y_t) = \sum_{g=u_0}^{u_1} \binom{n}{g, n-y_{t-1}-g, y_t-g, y_{t-1}-y_t+g}$$
$$\times {p_{01}}^g (1-p_{t-1}-p_{01})^{n-y_{t-1}-g} (p_t - p_{01})^{y_t-g} (p_{t-1} - p_t + p_{01})^{y_{t-1}-y_t+g},$$

waarbij $u_0 = \max(0, y_t - y_{t-1})$ en $u_1 = \min(n - y_{t-1}, y_t)$. Beschikken we over paren van observaties (y_{t-1}, y_t) afkomstig uit een serie van onafhankelijke steekproeven

van omvang n (d.w.z. meerdere tabellen), dan zijn de parameters p_{t-1}, p_t en p_{01} met maximum likelihood te schatten (zie voor details Hamdan en Nasro, 1986; Kocherlakota en Kocherlakota, 1992).

Is slechts één van de marginaal totalen gefixeerd, dan kunnen we de parameters schatten door de analyse te beperken tot de 2×2 tabellen die dezelfde gefixeerde marginale som hebben als de geobserveerde tabel. Stel $\mu = p_{01}/(1-p_{t-1})$ en $\kappa = p_{11}/p_{t-1}$. In een vergelijkende studie (bijvoorbeeld prospectief onderzoek) met twee onafhankelijke steekproeven, nemen we in het meest eenvoudige model $Y_{01} \sim B(n-y_{t-1},\mu)$ en $Y_{11} \sim B(y_{t-1},\kappa)$ als onafhankelijke random variabelen. De kans op het observeren van Y_t is dan de convolutie van twee onafhankelijke binomialen

$$P(Y_t = y_t) = \sum_g \binom{n-y_{t-1}}{g} \mu^g (1-\mu)^{n-y_{t-1}-g} \binom{y_{t-1}}{y_t-g} \kappa^{y_t-g} (1-\kappa)^{y_{t-1}-y_t+g},$$

waarbij de sommatie opnieuw betrekking heeft op de waarden van g die voldoen aan $\max(0, y_t - y_{t-1}) \le g \le \min(n-y_{t-1}, y_t)$. Deze dubbele binomiaalverdeling is, in het kader van 2×2 tabellen met onbekende celaantallen, uitvoerig besproken door Plackett (1977), Haber (1989) en McCullagh en Nelder (1992). McCue (1995) en Wakefield (2003) bestuderen het gebruik van deze convolutie likelihood voor de analyse van ecologische inferentie problemen.

Tot slot is er de mogelijkheid dat we een 2×2 tabel observeren waarvan zowel de rij als de kolom marginalen door de steekproef opzet zijn gefixeerd (een klassiek voorbeeld is het thee-proef experiment van Fisher). De steekproevenverdeling van Y_{01} gegeven de marginaal totalen is dan niet-centraal ('extended') hypergeometrisch (Fisher, 1935; Agresti, 1992; McCullagh en Nelder, 1992; Wakefield, 2003). Stel dat Y_{01} en Y_{11} onafhankelijke binomiale random variabelen zijn met verdelingen $B(n-y_{t-1},\mu)$ en $B(y_{t-1},\kappa)$, respectievelijk, en dat $\psi = \kappa(1-\mu)/\mu(1-\kappa)$ de odds ratio is die de rij-kolom afhankelijkheid weergeeft. De conditionele verdeling van Y_{01} gegeven $Y_{t-1} = y_{t-1}$ is dan niet-centraal hypergeometrisch met parameter ψ, gegeven door

$$P(Y_{01}=y_{01}) = \frac{\psi^{y_{01}}}{\prod_{i,j} y_{ij}!}\left(\sum_{y_{01}} \frac{\psi^{y_{01}}}{\prod_{i,j} y_{ij}!}\right)^{-1}$$

$$= \psi^{y_{01}} \binom{n-y_{t-1}}{y_{01}} \binom{y_{t-1}}{y_t - y_{01}} \left[\sum_{g=u_0}^{u_1} \psi^g \binom{n-y_{t-1}}{g} \binom{y_{t-1}}{y_t - g}\right]^{-1} \qquad y_{01} = u_0, \ldots, u_1,$$

waarbij de sommatie index g weer varieert van $u_0 = \max(0, y_t - y_{t-1})$ tot $u_1 = \min(n - y_{t-1}, y_t)$; de mogelijke waarden van y_{01} gegeven de marginaal totalen.

Herhaalde cross sectie data

Een belangrijke eigenschap van herhaalde cross-sectionele surveys is dat er op ieder meetmoment een nieuwe steekproef wordt getrokken, waardoor het onmogelijk is individuen door de tijd heen te volgen. Stel dat y_{it} de response is op de binaire (0-1) afhankelijke variabele y van individu i op tijdstip t. Bij cross sectie data observeren we wel y_{it}, maar niet y_{it-1}. Dit betekent voor het bovenstaande dat we van de 2×2 tabel slechts één marginaalverdeling observeren, de kolomtotalen bijvoorbeeld, en dat de andere verdeling, van de rijtotalen, en die van de interne celaantallen onbekend is. We gebruiken de notatie $f = y_{t-1}$ om aan te geven dat de rij sommen niet beschikbaar zijn bij herhaalde cross-sectionele surveys. De gezamenlijke kansverdeling van Y_{t-1} en Y_t is dan

$$P(Y_{t-1} = f, Y_t = y_t) = \sum_{f=0}^{n} P(Y_{t-1} = f) \times P(Y_t = y_t \mid Y_{t-1} = f),$$

waarbij $P(Y_t = y_t \mid Y_{t-1} = f)$ de convolutieverdeling is van Y_{01} en Y_{11}. Indien we nu veronderstellen dat Y_{t-1} een binomiaalverdeling volgt, d.w.z. $Y_{t-1} \sim B(n, p_{t-1})$, dan is de marginale verdeling van Y_t

$$P(Y_t = y_t) = \sum_{f=0}^{n_t} \left[\binom{n_t}{f} p_{t-1}{}^{f} (1-p_{t-1})^{n_t - f} \right.$$
$$\left. \times \sum_{g} \left[\binom{n_t - f}{g} \mu^g (1-\mu)^{n_t - f - g} \binom{f}{y_t - g} \kappa^{y_t - g} (1-\kappa)^{f - y_t + g} \right] \right],$$

en dit correspondeert met een binomiaalverdeling. Met andere woorden, als Y_{01}, Y_{11} èn Y_{t-1} binomiaal verdeeld zijn, dan is de marginaal verdeling van Y_t ook binomiaal, met index parameter n en succes kans p_t. Het is in dit verband informatief de conditionele verdeling te presenteren van Y_t gegeven $Y_{t-1} = f$. De kans-genererende-functie (p.g.f.) van de verdeling is $\Pi_{Y_t}(t \mid f) = \left[(1-\mu) + \mu t\right]^{n-f} \left[(1-\kappa) + \kappa t\right]^{f}$, en dit is de p.g.f. van de convolutie van $B(n-f, \mu)$ en $B(f, \kappa)$. Op basis van deze conditionele verdeling, is de regressie van Y_t op f te schrijven als

$$E(Y_t \mid f) = \mu(n-f) + \kappa f = \mu n + (\kappa - \mu) f$$

en die is lineair in f met regressie coëfficiënt $\kappa - \mu$. Voor deze regressie is de conditionele variantie te bepalen als

$$\mathrm{Var}(Y_t \mid f) = \mu(1-\mu)n + [\kappa(1-\kappa) - \mu(1-\mu)]f$$

en die is eveneens lineair in f. Uit de regressie van Y_t op f zijn onder- en bovengrenzen (Fréchet bounds) te bepalen die het bereik van μ en κ aangeven. Uit de bovenstaande vergelijking voor $E(Y_t \mid f)$ volgt namelijk dat

$$\mu = \frac{\mathrm{E}(Y_t \mid f)}{(n-f)} - \frac{f}{(n-f)}\kappa \quad \text{en} \quad \kappa = \frac{\mathrm{E}(Y_t \mid f)}{f} - \frac{(n-f)}{f}\mu .$$

Aangezien μ en κ kansen zijn die binnen het $[0,1]$ interval liggen, volgt hieruit dat $\mu \in (O_\mu, B_\mu)$ en $\kappa \in (O_\kappa, B_\kappa)$. Daarbij zijn de onder- (O) en bovengrenzen (B) gedefinieerd door de min en max operators

$$O_\mu = \max\left[0, \frac{y_t - f}{n - f}\right] \le \mu \le \min\left[\frac{y_t}{n - f}, 1\right] = B_\mu \text{ en}$$

$$O_\kappa = \max\left[0, \frac{y_t - (n - f)}{f}\right] \le \kappa \le \min\left[\frac{y_t}{f}, 1\right] = B_\kappa$$

(zie bijv. King, 1997; Chambers en Steel, 1997). In geval van herhaalde cross sectie data zijn voor het bepalen van de onder- en bovengrenzen minstens twee kwesties van belang. De eerste is dat f niet is geobserveerd. Er zijn echter situaties waarin het redelijk is om te veronderstellen dat de onbekende marginaal y_{t-1} op tijdstip t, gelijk is aan de geobserveerde marginaal van de steekproef op $t-1$. Deze veronderstelling kunnen we implementeren door op tijdstip t de op tijdstip $t-1$ geobserveerde marginale frequenties (of proporties) te imputeren. De tweede kwestie is dat onder- en bovengrenzen alleen deterministische informatie verschaffen over μ en κ, onder de veronderstelling dat de marginale data correct zijn. In herhaalde cross sectie steekproeven zijn y_t en y_{t-1} echter random observaties. De onder- en bovengrenzen zijn daarom niet deterministisch, maar stochastisch van aard. Omdat de grenzen niet met zekerheid zijn vast te stellen, bepalen we – bijvoorbeeld met bootstrap - betrouwbaarheidsintervallen van de onder- en bovengrenzen. Hierbij is het van belang te vermelden dat in de procedure om de overgangskansen te schatten geen (expliciet) gebruik wordt gemaakt van onder- en bovengrenzen.

Een transitiemodel voor herhaalde cross sectie data

Hierboven zijn de data gepresenteerd in de vorm van geaggregeerde 2×2 tabellen. Het dynamisch Markov model voor herhaalde cross sectie data, dat we hieronder presenteren, is ontworpen voor de analyse van zowel geaggregeerde data als voor individuele observaties. We presenteren het model in termen van individuele observaties en nemen opnieuw een 2×2 tabel in gedachten waarin de cellen rijgewijs sommeren tot één. Stel p_{it} is de kans dat $y_{it} = 1$, μ_{it} de kans dat $y_{it} = 1$ gegeven $y_{it-1} = 0$, en κ_{it} de kans dat $y_{it} = 0$ gegeven $y_{it-1} = 1$. De verwachtingswaarde van y_{it} is dan

$$E(y_{it}) = p_{it} = \mu_{it}(1 - p_{it}) + \kappa_{it} p_{it-1}.$$

In woorden: de kans dat individu i op tijdstip t tot de categorie $y_{it}=1$ behoort, is gelijk aan de kans dat hij op $t-1$ tot de categorie '0' behoort maal de kans dat hij van '0' naar '1' transformeert, plus de kans dat hij op $t-1$ tot categorie '1' behoort maal de kans dat hij in categorie '1' blijft. Voor het schatten van dynamische modellen met cross sectie data is het oplossen van de bovenstaande vergelijking fundamenteel, omdat daarin een relatie wordt gelegd tussen de marginale kansen (p_{it} en p_{it-1}) en de intrede (μ_{it}) en uittrede ($\lambda_{it}=1-\kappa_{it}$) kansen. Om de vergelijking toe te passen op herhaalde cross sectie data, vervangen we p_{it} in de bovenstaande vergelijking bij herhaling door μ_{it} en κ_{it}. We krijgen dan het volgende transitiemodel

$$p_{it}=\mu_{it}+\sum_{\tau=1}^{t-1}\left(\mu_{i\tau}\prod_{s=\tau+1}^{t}\eta_{is}\right)+p_{i0}\prod_{\tau=1}^{t}\eta_{it}\ ,$$

waarbij $\eta_{is}=\kappa_{is}-\mu_{is}$. Dit stelsel van vergelijkingen is alleen uniek oplosbaar voor cross sectie gegevens, door aan de transities van individuen i en/of tijdstippen t bepaalde restricties op te leggen. Daarbij zijn diverse typen van restricties mogelijk.

Eén mogelijkheid is *a priori* (d.w.z. voor kennisneming van de data) beperkingen op te leggen aan de niet-geobserveerde μ_{it} en κ_{it} overgangskansen. De parameters van het transitiemodel zijn bijvoorbeeld eenvoudig te identificeren, indien we veronderstellen dat de transitiekansen homogeen zijn voor zowel de individuen i als de tijdstippen t. In dat geval reduceert p_{it} in het model op de lange termijn tot $p_{it}=\mu/\mu+\lambda$. Modellen met dit soort type homogeniteit zijn uitgebreid bestudeerd in de statistische literatuur (Lee, Judge en Zellner, 1970; Firth, 1982; Kalbfleish en Lawless, 1984, 1985; Lawless en McLeish, 1984; Li en Kwok, 1990; Hawkins, Han en Eisenfeld, 1996). Ze zijn ook veelvuldig toegepast in sociologische (Goodman, 1961; Bartholomew, 1996) en economische (Topel, 1983; McCall, 1971) studies. Identificeerbaarheid is ook te forceren door minder stringente assumpties op te leggen, bijvoorbeeld individu-heterogene, maar tijd-homogene (c.q. stationaire) overgangskansen.

De benadering die we in ons model hanteren, is afkomstig van Moffitt (1990 1993). Hij stelt voor om bij het bepalen van p_{it-1} gebruik te maken van de waarden van de covariaten op tijdstip $t-1$. De gedachte daarbij is dat het met herhaalde cross sectie data dikwijls mogelijk is voorgaande waarden van $\mathbf{x}_{it}$ te reconstrueren door terug te gaan in de tijd. Daarvoor is vereist dat de covariaten $\mathbf{x}_{it}$ tijd-constant

zijn (bijv. geboortedatum, geslacht, etniciteit, hoogst voltooide opleiding) of terug te vertalen zijn in de tijd door retropolatie (bijv. leeftijd). Beschikken we over dit soort gegevens, dan kunnen we de covariaten gebruiken om de transitiekansen μ_{it} en κ_{it} te voorspellen en daarmee ook de marginale kansen p_{it}. Voor de relatie tussen de covariaten en de transitiekansen specificeren we $\mu_{it} = F(\mathbf{x}_{it}\beta_t)$ en $\kappa_{it} = F(\mathbf{x}_{it}\beta_t^*)$ - waarbij $F(\cdot)$ een logistische functie is. Wanneer we deze logit functies substitueren in de bovenstaande vergelijking voor p_{it} dan krijgen we een Markov transitiemodel voor het schatten van covariaat-afhankelijke overgangskansen.

Dit standaard Markov model is op diverse punten aan te passen en uit te breiden. We kunnen bijvoorbeeld covariaten opnemen waarvan het verleden onbekend is en rekening houden met niet-geobserveerde heterogeniteit. Voor deze en andere uitbreidingen verwijzen we naar Pelzer, Eisinga en Franses (2001, 2002a, 2004). Daarin staat ook een uiteenzetting van schattingstechnieken (maximum likelihood, MCMC, parametrisch bootstrappen).

Empirische illustratie

Het model is inmiddels op diverse terreinen toegepast, zoals de in- en uittrede uit het arbeidsproces door vrouwen (Pelzer, Eisinga en Franses, 2001), verandering in politiek stemgedrag (Pelzer, Eisinga en Franses, 2002a), de inschatting van de kans op het krijgen van AIDS door het gezamenlijk gebruik van ongebleekte naalden door drugsgebruikers (Pelzer en Eisinga, 2002b) en de aanschaf van een personal computer door Nederlandse huishoudens (Pelzer, Eisinga en Franses, 2004). Het voorbeeld dat we hier bespreken heeft betrekking op de belangstelling van middelbare scholieren voor natuurkunde onderwijs (voor een uitgebreide bespreking, zie Pelzer en Eisinga, 2003). De data zijn ontleend aan Vermunt, Langeheine en Böckenholt (1999) en afkomstig van een Duits panel survey bestaande uit drie, jaarlijks gehouden metingen. De gegevens zijn onder meer verzameld om de invloed na te gaan van 'geslacht' en 'natuurkundecijfer' op de 'interesse in natuurkunde'. De steekproef bestaat uit 541 scholieren met volledige gegevens voor alle (dichotome) variabelen op alle (3) tijdstippen. Het is belangrijk om te vermelden dat we de drie-golf panelgegevens behandelen als drie onafhankelijke steekproeven. Dit wil zeggen dat er geen informatie in de geanalyseerde data aanwezig is over de samenhang tussen opeenvolgende y_{it}. Door panelgegevens te gebruiken kunnen we de modeluitkomsten vergelijken met de observaties in het panel. Geslacht is een tijdconstant en natuurkundecijfer een tijdvariërend covariaat. De variabelen geslacht (jongens, meisjes) en

natuurkundecijfer (laag, hoog) combineren we tot één nieuwe variabele. We duiden de variabelen aan met de letters j (jongen) dan wel m (meisje), l (laag cijfer) dan wel h (hoog cijfer) en een tijd subscript t. Bijvoorbeeld, mh_1 heeft betrekking op meisjes met een hoog cijfer voor natuurkunde op $t=1$. Het volledig, niet-stationair model ziet er als volgt uit:

$$\text{logit}(p_1) = \delta_1 + \delta_2 \text{jl}_1 + \delta_3 \text{gl}_1 + \delta_4 \text{gh}_1$$

$$\text{logit}(\mu_2) = \beta_1 \text{jl}_2 + \beta_2 \text{ml}_2 + \beta_3 \text{jh}_2 + \beta_4 \text{mh}_2 \qquad \text{logit}(\kappa_2) = \beta_1^* \text{jl}_2 + \beta_2^* \text{ml}_2 + \beta_3^* \text{jh}_2 + \beta_4^* \text{mh}_2$$

$$\text{logit}(\mu_3) = \beta_5 \text{jl}_3 + \beta_6 \text{ml}_3 + \beta_7 \text{jh}_3 + \beta_8 \text{mh}_3 \qquad \text{logit}(\kappa_3) = \beta_5^* \text{jl}_3 + \beta_6^* \text{ml}_3 + \beta_7^* \text{jh}_3 + \beta_8^* \text{mh}_3$$

Het intercept voor $\text{logit}(p_1)$ heeft betrekking op de categorie jongens met een hoog natuurkundecijfer op $t=1$. De overige vergelijkingen hebben geen constante, waardoor we de parameters in termen van logits (log odds-ratios) kunnen interpreteren. De analyse bestaat uit het bepalen van een zo zuinig mogelijk transitiemodel dat goed bij de gegevens past. De procedure start met een stationair model met tijdinvariante parameters voor de categorieën met dezelfde waarde voor geslacht-natuurkundecijfer ($\beta_1 = \beta_5, \beta_2 = \beta_6$, et cetera). Vervolgens is het model uitgebreid door tijdvariërende parameters op te nemen. Het resulterende niet-stationaire model is vervolgens gesimplificeerd door parameters op nul te fixeren dan wel gelijk aan elkaar te stellen. De resultaten van deze procedure zijn besproken in Pelzer en Eisinga (2003). Hieronder geven we alleen het eindmodel weer:

$$\text{logit}(p_1) = \delta_1 + \delta_2 \text{jl}_1 + 2\delta_2 \text{ml}_1 + \delta_2 \text{mh}_1$$

$$\text{logit}(\mu_2) = \beta_1 \text{jl}_2 + 5\beta_4 \text{ml}_2 + \beta_4 \text{mh}_2 \qquad \text{logit}(\kappa_2) = \beta_1^* \text{jl}_2 + \beta_1^* \text{ml}_2 + \beta_3^* \text{jh}_2 + \beta_3^* \text{mh}_2$$

$$\text{logit}(\mu_3) = \beta_1 \text{jl}_3 + \beta_4 \text{ml}_3 + \beta_1 \text{jh}_3 + \beta_4 \text{mh}_3 \qquad \text{logit}(\kappa_3) = \beta_1^* \text{jl}_3 + \beta_3^* \text{jh}_3 + \beta_3^* \text{mh}_3$$

De maximum likelihood schattingen van de parameters en (tussen haakjes) standaard deviaties zijn

p_1		μ_t		κ_t	
δ_1	δ_2	β_1	β_4	β_1^*	β_3^*
.599	-1.036	-.941	-2.437	-.636	1.689
(.127)	(.126)	(.368)	(.355)	(.319)	(.242)

De parameters voor p_1 geven aan dat jongens met een hoog cijfer op $t = 1$ in sterkere mate een hoge interesse in natuurkunde hebben dan zowel jongens met een laag cijfer als meisjes. Bij meisjes met lage cijfers is de kans op een hoge interesse het geringst. De resultaten voor μ_t impliceren dat voor jongens met een hoog cijfer de kans op een transitie van een lage naar een hoge interesse het grootst is. De geschatte logits zijn het geringst voor meisjes met een laag cijfer. De parameter schattingen voor κ_t geven aan dat de kans op een 1-1 transitie het grootst is voor scholieren met een hoog natuurkundecijfer en absoluut gesproken laag zijn voor scholieren met een laag cijfer.

Om de kwaliteit van het model te beoordelen, vergelijken we de voorspellingen op grond van het model met de observaties in het panel. Hieronder staan de geschatte totale kansen en de geobserveerde totale proporties voor μ_t en κ_t. Hieronder staan – in de notatie van de 2×2 tabellen die aan het begin van deze bijdrage zijn gepresenteerd – de geschatte proporties p_{01} en p_{11} en de geobserveerde proporties y_{01}/n en y_{11}/n voor $t = 2$ en $t = 3$.

	jongens laag	meisjes laag	jongens hoog	meisjes hoog	jongens laag	meisjes laag	jongens hoog	meisjes hoog
	$t = 2$				$t = 3$			
$\widehat{p}_{01}$	.13	.00	.19	.05	.13	.07	.10	.06
y_{01}/n	.11	.01	.14	.14	.04	.06	.10	.10
$\widehat{p}_{11}$	.19	.09	.52	.29	.19	.09	.54	.26
y_{11}/n	.23	.07	.52	.22	.27	.09	.55	.22

Ofschoon er enkele afwijkingen zijn, is aan deze uitkomsten te zien dat het model goed in staat is de panel observaties te reproduceren.

Tot besluit

Uit de bovenstaande toepassing zijn geen algemene conclusies te trekken over de universele bruikbaarheid van het Markov model. Daarvoor is het nodig het model onder uiteenlopende omstandigheden te testen. Een van de zaken die op de agenda staat, is dan ook een Monte Carlo simulatie studie waarin het model aan een uitvoeriger test wordt onderworpen. Een ander plan is de toepassing van het model op incomplete panel data. In een panel komen dikwijls ontbrekende waarden voor.

Ze kunnen tussentijds ontbreken of er kan sprake zijn van vroegtijdige uitval uit het panel. We willen nagaan of we met het model ontbrekende gegevens geldig kunnen reconstrueren.

Literatuur

Achen, Christopher H. en W. Phillips Shively. 1995. *Cross-level Inference*. Chicago Ill, University of Chicago Press.

Agresti, Alan. 1992. A Survey of Exact Inference for Contingency Tables. *Statistical Science* 7, 131-153.

Barnard, G.A. 1947. Significance Tests for 2×2 Tables. *Biometrika* 34, 123-138.

Bartholomew, David J. 1996. *The Statistical Approach to Social Measurement*. San Diego. Academic Press.

Bishop, Yvonne M.M., Stephen E. Fienberg en Paul W. Holland. 1975. *Discrete Multivariate Analysis: Theory and Practice*. Cambridge MA, MIT Press.

Böckenholt, Ulf en William R. Dillon. 2000. Inferring Latent Brand Dependencies. *Journal of Marketing Research* 37, 72-87.

Cross, Philip J. en Charles F. Manski. 2002. Regressions, Short and Long. *Econometrica* 70, 357-368.

Dobra, Adrian, Claudia Tebaldi en Mike West. 2003. *Bayesian Inference in Incomplete Multi-way Tables*. Institute of Statistics and Decision Sciences, Duke University, Durham NC.

Fienberg, Stephen. 1997. *Confidentiality and Disclosure Limitation Methodology. Challenges for National Statistics and Statistical Research*. Department of Statistics, Carnegie Mellon University, Pittsburgh PA.

Firth, David. 1982. *Estimation of Voter Transition Matrices from Election Data.* M.Sc. Thesis, Department of Mathematics, Imperial College.

Fisher, Ronald Aylmer. 1935. The Logic of Inductive Inference (with discussion). *Journal of the Royal Statistical Society* 98, 39-82.

Goodman, L.A. 1961. Statistical Methods for the Mover-Stayer Model. *Journal of the American Statistical Association* 56, 841-868.

Golan, Amos, George Judge en Douglas Miller. 1996. *Maximum Entropy Econometrics. Robust Estimation with Limited Data.* New York, Wiley.

Golan, Amos, George Judge en Sherman Robinson. 1994. Recovering Information from Incomplete or Partial Multisectoral Economic Data. *Review of Economics and Statistics* 76, 541-549.

Golan, Amos, George Judge en Jeffrey M. Perloff. 1996. A Maximum Entropy Approach to Recovering Information from Multinomial Response Data. *Journal of the American Statistical Association* 91, 841-853.

Haber, Michael. 1989. Do the Marginal Totals of a 2×2 Contingency Table Contain Information Regarding the Table Proportions? *Communications in Statistics - Theory and Methods* 18, 147-156.

Hamdan, M.A. en M.O. Nasro. 1986. Maximum Likelihood Estimation of the Parameters of the Bivariate Binomial Distribution. *Communication in Statistics - Theory and Methods* 15, 747-754.

Hawkins, D.L., C.P. Han en J. Eisenfeld. 1996. Estimating Transition Probabilities from Aggregate Samples Augmented by Haphazard Recaptures. *Biometrics* 5, 625-638.

Judge, George, Douglas Miller en Wendy K. Tam Cho. 2004. An Information Theoretic Approach to Ecological Estimation and Inference. (verschijnt in) Gary King, Ori Rosen en Martin Tanner. 2004. *Ecological Inference. New Methodological Strategies*. New York, Cambridge University Press.

Kalbfleish, J.D. en J.F. Lawless. 1984. Least Squares Estimation of Transition Probabilities from Aggregate Data. *Canadian Journal of Statistics* 12, 169-182.

Kalbfleish, J.D. en J.F. Lawless. 1985. The Analysis of Panel Data under a Markovian Assumption. *Journal of the American Statistical Association* 80, 863-81.

King, Gary. 1997. *A Solution to the Ecological Inference Problem. Reconstructing Individual Behavior from Aggregate Data.* Cambridge MA, Cambridge University Press.

King, Gary, Ori Rosen en Martin Tanner. 2004. *Ecological Inference. New Methodologial Strategies*. New York, Cambridge University Press.

Kocherlakota, Subrahmaniam en Kathleen Kocherlakota. 1992. *Bivariate Discrete Distributions*. New York, Marcel Dekker.

Lawless, J.F. en D.L. McLeish. 1984. The Information in Aggregate Data from Markov Chains. *Biometrika* 71, 419-430.

Lee, T.C., G.G. Judge en A. Zellner. 1970. *Estimating the Parameters of the Markov Probability Model from Aggregate Time Series Data*. Amsterdam. North-Holland.

Li, W.K. en Michael C.O. Kwok. 1990. Some Results on the Estimation of a Higher Order Markov Chain. *Communications in Statistics. Part B. Simulation and Computation* 19, 363-380.

McCall, John J. 1971. A Markovian Model of Income Dynamics. *Journal of the American Statistical Association* 66, 439-447.

McCue, Kenneth F. 1995. *Individual Choice and Ecological Analysis*. California Institute of Technology, Pasadena CA.

McCullagh, P. en J.A. Nelder. 1992. *Generalized Linear Models (2nd ed.)*. London, Chapman and Hall.

Moffitt, Robert. 1990. The Effect of the U.S. Welfare System on Marital Status. *Journal of Public Economics* 41, 101-124.

Moffitt, Robert. 1993. Identification and Estimation of Dynamic Models with a Time Series of Repeated Cross-sections. *Journal of Econometrics* 59, 99-123.

Pelzer, Ben, Rob Eisinga en Philip Hans Franses. 2001. Estimating Transition Probabilities from a Time Series of Repeated Cross Sections. *Statistica Neerlandica* 55, 248-261

Pelzer, Ben, Rob Eisinga en Philip Hans Franses. 2002a. Inferring Transition Probabilities from Repeated Cross Sections. *Political Analysis* 10, 113-133.

Pelzer, Ben en Rob Eisinga. 2002b. Bayesian Estimation of Transition Probabilities from Repeated Cross Sections. *Statistica Neerlandica* 56, 23-33.

Pelzer, Ben, Rob Eisinga en Philip Hans Franses. 2004. Ecological Panel Inference from Repeated Cross Sections. (verschijnt in) Gary King, Ori Rosen en Martin Tanner. *Ecological Inference. New Methodological Strategies*. New York, Cambridge University Press.

Pelzer, Ben en Rob Eisinga. 2003. *Recovering Transitions from Repeated Cross Sections*. University of Nijmegen.

Plackett, R.L. 1977. The Marginal Totals of a Table. *Biometrika* 64, 37-42.

Richardson, S. en C. Montfort. 2000. Ecological Correlation Studies. Pp. 205-220 In *Spatial Epidemiology. Methods and Applications*, P. Elliott, J.C. Wakefield, N.G. Best en D.J. Briggs (eds.). Oxford, Oxford University Press.

Tebaldi, Claudia en Mike West. 1998. *Reconstruction of Contingency Tables with Missing Data*. Institute of Statistics and Decision Sciences, Duke University, Durham NC.

Topel, Robert H. 1983. On Layoffs and Unemployment Insurance. *American Economic Review* 73, 541-559.

Vermunt, Jeroen K., Rolf Langeheine en Ulf Böckenholt. 1999. Discrete-time Discrete-state Latent Markov Models with Time-constant and Time-varying Covariates. *Journal of Educational and Behavioral Statistics* 24, 179-207.

Wakefield, Jonathan. 2001. *Ecological Inference for* 2×2 *Tables*. Working paper no. 12. Department of Statistics and Biostatistics, University of Washington, Seattle WA.

Wakefield, Jonathan. 2004. Prior and Likelihood Choices in the Analysis of Ecological Data. (verschijnt in) Gary King, Ori Rosen en Martin Tanner. 2004. *Ecological Inference. New Methodological Strategies*. New York, Cambridge University Press.