Multiniveau modellen voor panel data. Een vergelijking van multiniveau regressie- en structurele modellen met gesimuleerde gegevens

J. J. Hox

Samenvatting

Voor de analyse van panel data wordt in toenemende mate gebruik gemaakt van multiniveau regressie modellen. Eén van de voordelen daarvan is, dat het niet nodig is om personen uit de dataset te verwijderen bij wie van een van de meetpunten de gegevens ontbreken. Een recente ontwikkeling bij structurele vergelijking modellen is het zogenaamde latente curve model. Dit artikel laat zien dat zowel multiniveau regressie modellen als latente curve modellen effectief gebruikt kunnen worden bij de analyse van panel gegevens. Bij eenvoudige modellen komen de verschillen neer op een verschil in parametrisering. Een vergelijking van beide modellen bij een gesimuleerde dataset laat zien dat de overeenkomstige schattingen dan ook vrijwel identiek zijn, zowel bij volledige als onvolledige gegevens.

J.J. Hox, Faculteit POW, Universiteit van Amsterdam, Wibautstraat 4, 1091 GM Amsterdam, hox@educ.uva.nl.
Met dank aan David Rogosa voor het beschikbaar stellen van de gesimuleerde datasets, en een anonieme reviewer voor commentaar.

1. Inleiding

Panel studies zijn studies waarin een steekproef een aantal malen wordt bevraagd, met de bedoeling om daardoor trends in de tijd vast te stellen. Voor de analyse van panel data zijn een groot aantal modellen ontwikkeld. Afhankelijk van het aantal meetpunten en variabelen kunnen onderzoekers kiezen uit variantie analyse voor herhaalde metingen, tijdreeksanalyse, structurele modellen, latente klasse analyse en Markovmodellen. Wanneer het aantal variabelen en het aantal meetpunten niet al te groot is, is een variantie analyse voor herhaalde metingen (ANOVA of MANOVA) een aantrekkelijke optie, onder andere omdat het een goed bekende techniek is die in standaard software zoals SPSS beschikbaar is. Bij een betrekkelijk klein aantal metingen en meerdere variabelen wordt veel gebruik gemaakt van causale modellen, die geanalyseerd worden met structurele modellen (vgl. Jöreskog & Sörbom, 1977). Een overzicht van deze en andere klassieke modellen voor panel data wordt gegeven door Plewis (1985) en Engel & Reinecke (1994).

Sinds enige tijd zijn de analysemogelijkheden uitgebreid met verschillende multiniveau modellen. In het meest simpele geval worden de individuen beschouwd als de eenheid van analyse op het tweede (hogere) niveau, en de metingen binnen de individuen als de eenheid van analyse op het eerste (lagere) niveau. Wanneer alle individuen op gelijke momenten gemeten zijn (hetgeen bij panelonderzoek doorgaans het geval is) wordt wel gesproken van een fixed occasion of gefixeerde meetmomenten model. Wanneer de individuen op verschillende tijdstippen gemeten zijn, wordt wel gesproken van een groeicurve model. Het gefixeerde meetmomenten model kan beschouwd worden als een speciaal geval van het groeicurve model. Voor multiniveau regressie analyse is het onderscheid niet van groot belang. De meeste andere modellen voor panel analyse gaan uit van gefixeerde meetmomenten. De betreffende multiniveau modellen zijn besproken door o.a. Raudenbush & Bryk (1987), Bryk & Raudenbush (1987, 1992) en Goldstein (1987, 1989, 1995). Kenmerkend is dat doorgaans aangenomen wordt dat de regressiecoëfficiënten voor het verloop over de tijd random zijn, dat wil zeggen dat verschillende individuen gekenmerkt worden door verschillende groeicurven.

Multiniveau regressie analyse voor panel data is met name aantrekkelijk wanneer er sprake is van één of enkele afhankelijke variabelen en een betrekkelijk groot aantal achtereenvolgende metingen. Bij variantie analyse en de gebruikelijke structurele modellen wordt voor elke variabele op elk tijdstip doorgaans een individuele parameter geschat, terwijl in het multini-

veau model voor elke regressiecoëfficiënt een verdeling wordt aangenomen waarvan het gemiddelde en de variantie geschat wordt. Bij een groter aantal metingen is het multiniveau model daardoor veel spaarzamer.[1] Daarnaast heeft het multiniveau model duidelijke voordelen wanneer er ontbrekende data zijn doordat niet alle individuen op alle meetmomenten beschikbaar zijn. Bij klassieke modellen voor panel data zoals variantie analyse moeten dergelijke gevallen geheel uit de analyse verwijderd worden (listwise deletion), hetgeen tot een drastische reductie van de hoeveelheid gegevens kan leiden. Bij multiniveau modellen vervallen alleen de gegevens die betrekking hebben op dat specifieke tijdstip.

Toepassing van het multiniveau regressiemodel op panel data heeft ook voordelen. In het groeicurve model is een belangrijk voordeel dat niet ieder individu op dezelfde tijdstippen gemeten hoeft te zijn. In onderwijskundig panel onderzoek is dit voordeel niet doorslaggevend, omdat daar doorgaans op vaste momenten (bijvoorbeeld aan het begin of eind van ieder schooljaar) gemeten wordt. Voordelen die ook van toepassing zijn op het gefixeerde meetmomenten model zijn de reeds genoemde eenvoudige inpassing van personen die op een enkel meetmoment afwezig zijn geweest, en de direct aanwezige mogelijkheid om de multiniveau structuur uit te breiden met de gebruikelijke klas- en schoolniveaus.

Een nadeel van het multiniveau model voor panel data is dat het gefixeerde (niet-stochastische) deel van het model neerkomt op een standaard multipele regressieanalyse. Het meer gecompliceerde structurele vergelijkingen model (Structural Equations Model ofwel SEM, ook wel aangeduid als covariantie structuur analyse) laat zich met de standaard multiniveau software niet schatten.[2] Muthén (1989, 1994) en McDonald (1994) beschrijven een structureel model voor twee-niveau data (zie ook Hox, 1995). Dit model kan gebruikt worden om onderscheiden structurele modellen te schatten voor de individuele en de groepsgegevens. Een belangrijke beperking van dit model is dat het beperkt is tot over de groepen variërende intercepts; met andere woorden, variaties in de regressiehellingen kunnen niet gemodelleerd worden. Het modelleren van cross-level interacties tussen verklarende variabelen van verschillende niveaus is bij deze benadering een gecompliceerde zaak, vooral

[1]De meeste modellen kennen ook covarianties tussen geschatte parameters. Ook wanneer daar rekening mee gehouden wordt, is het multiniveau model doorgaans spaarzamer. Al bij drie meetpunten kan voor het multiniveau model een spaarzamere formulering gekozen worden (vgl. Snijders & Maas, 1996).

[2]Bryk & Raudenbush (1992) en Goldstein (1995) beschrijven manieren om multivariate modellen te schatten, maar deze zijn het best op te vatten als een multiniveau analogon van MANOVA, en zeker niet als een padmodel.

wanneer het gaat om interacties tussen latente variabelen (vgl. Jöreskog & Yang, 1996). Het resultaat is dat de toepassing van multiniveau structurele modellen op panel data beperkt is: enerzijds kunnen meer gecompliceerde padmodellen worden gespecificeerd, anderzijds is de random error structuur beperkter.

Een interessant structureel model voor panel data met gefixeerde meetmomenten is het Latente Curve Model (LCM). In het latente curve model, dat tot nu toe vooral is toegepast op ontwikkelingsdata (het staat ook wel bekend als het latente groei model), wordt de tijdsdimensie rechtstreeks opgenomen in de definitie van de (latente) afhankelijke variabelen in het structurele model (vgl. Meredith & Tisak, 1990; Muthén, 1991; Willett & Sayer, 1994). Het LCM modelleert de opeenvolgende metingen door een latente variabele voor de intercept en een latente variabele voor de helling van de regressie van de metingen op de tijdvariabele. De regressie van de metingen op de tijdvariabele is een random coëfficiënten model, hetgeen inhoudt dat zowel de intercepts als de regressiehellingen van de individuele curven kunnen verschillen. Het LCM kan geschat worden met standaard SEM software. In de context van structurele modellen bestaan geavanceerde oplossingen voor het analyseren van datasets met ontbrekende gegevens (Arbucle, 1996), waardoor het voordeel van multiniveau regressie wat dit betreft minder relevant is.

2. Probleemstelling

In dit artikel wordt een vergelijking gemaakt tussen multiniveau regressie modellen en latente curve modellen voor panel data. Er zijn drie onderscheiden sub-probleemstellingen:

1) In hoeverre kunnen parameters in beide modellen in elkaar vertaald worden, en welke parameters zijn uniek voor het betreffende model?
2) In hoeverre leiden de beide analysemethoden tot gelijke schattingen?
3) Hoe effectief kunnen gegevens met ontbrekende data worden geanalyseerd?

Beide modellen worden vergeleken op gesimuleerde data met bekende kenmerken, waarbij ter vergelijking ook een klassieke variantie analyse voor herhaalde metingen wordt uitgevoerd.

3. Het multiniveau regressie model voor panel data

In het multiniveau regressiemodel wordt de meting van persoon j (j=1,...J) op T meetmomenten i (i=1,...T) aangeduid als Y_{ij}. De score op de responsvariabele Y wordt doorgaans voorspeld uit een polynoomfunctie van het tijdsverloop t op meetmoment i (andere functies zoals splines zijn uiteraard ook mogelijk, maar worden in de praktijk weinig gebruikt) sinds een bepaalde gebeurtenis, bijvoorbeeld geboorte, begin van een schooljaar, of eenvoudig vanaf de eerste meting. Het model op het eerste niveau is:

$$Y_{ij} = ß_{0j} + ß_{1j}\, t_{ij} + ß_{2j}\, t_{ij}^2 + \ldots ß_{tj}\, t_{ij}^T + e_{ij} \qquad (1)$$

Dit model is een Random Coëfficiënt model, waarin ieder individu een eigen groeicurve heeft, gerepresenteerd door de individu-specifieke regressiecoëfficiënten $ß_{pj}$. Goldstein (1995) beschrijft een aantal uitbreidingen van dit model. Zo is het mogelijk verklarende variabelen toe te voegen op beide niveaus (meetmoment en individu). Voorbeelden van verklarende variabelen op meetmoment niveau (d.w.z. binnen individuen variërend over verschillende tijdstippen) zijn treatment-variabelen (bijvoorbeeld in een longitudinaal N=1 design); voorbeelden van verklarende variabelen op individueel niveau zijn sekse en SES. Zo leidt toevoeging van een verklarende variabele X_{ij} op het eerste niveau en een verklarende variabele Z_j op het tweede niveau tot het volgende model:

$$Y_{ij} = \beta_{0j} + \beta_{1j}\, t_{ij} + \beta_{2j}\, t_{ij}^2 + \ldots \beta_{Tj}\, t_{ij}^T + \beta_{T+1,j}\, Xij + e_{ij} \qquad (2)$$

waarbij

$$\beta_{pj} = \gamma_{p0} + \gamma_{p1} Z_j + u_{pj},\ p=1,\ldots T \qquad (3)$$

Substitutie van (2) in (1) leidt tot:

$$\begin{aligned} Y_{ij} &= \gamma_{00} + \gamma_{10}\, t_{ij} + \gamma_{20}\, t_{ij}^2 + \ldots + \gamma_{T0}\, t_{ij}^T + \gamma_{T+1,0}\, X_{ij} + \\ &+ \gamma_{01}\, Z_j + \gamma_{11}\, Z_j\, t_{ij} + \gamma_{21}\, Z_j\, t_{ij}^2 + \ldots + \gamma_{T1}\, Z_j\, t_{ij}^T + \gamma_{T+1,1}\, Z_j\, X_{ij} + \\ &+ u_{0j} + u_{1j}\, t_{ij} + u_{2j}\, t_{ij}^2 + \ldots + u_{Tj}\, t_{ij}^T + u_{T+1,j}\, X_{ij} + e_{ij} \end{aligned} \qquad (4)$$

In dit model worden individuele verschillen in groeitrajecten gemodelleerd door een interactie met de individuele eigenschap Z. De variatie tussen de individuen is een functie van de tijd, en de responsen binnen één individu zijn gecorreleerd. Het gebruikelijke multiniveau regressiemodel (cf. Bryk & Raudenbush, 1992; Goldstein, 1995) veronderstelt geen specifieke

covariantiestructuur voor de metingen op verschillende tijdstippen, hoewel dit in principe wel mogelijk is (cf. Gibbons et al., 1993; Hedeker & Gibbons, 1995).

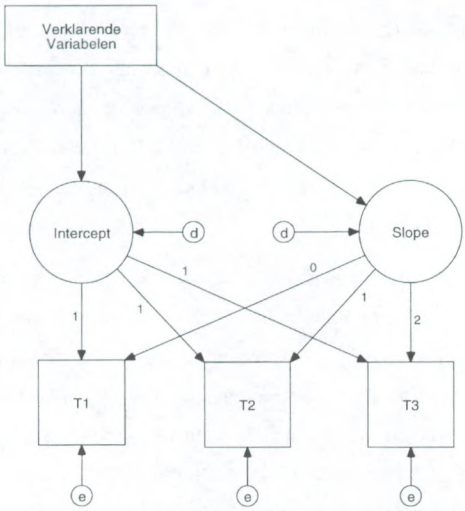
Het fixed-occasions model veronderstelt dat een groep personen j (j=1..J) is gemeten op opeenvolgende vaste meetmomenten i (i=1..T). De T meetmomenten kunnen dan gerepresenteerd worden door T indicatorvariabelen, waarna op het laagste niveau een model wordt gepostuleerd analoog aan model (1) hierboven, maar zonder intercept (cf. Goldstein, 1995). Ook dit model kan worden uitgebreid zoals hierboven in model (2) tot (4) geschetst is. Snijders en Maas (1996) laten zien hoe langs de weg van de multiniveau analyse van data met gefixeerde meetmomenten toetsen kunnen worden geconstrueerd die bij volledige data equivalent zijn aan de gebruikelijke toetsen bij multivariate variantie analyse voor herhaalde metingen. Bij incomplete gegevens treden er verschillen op.

Wanneer een individu op enig meetmoment afwezig is, is de gebruikelijke procedure bij een variantie analyse dat het betreffende individu uit de datamatrix verwijderd wordt. Bij multiniveau regressie analyse is er geen eis dat de aantallen waarnemingen op het laagste niveau steeds gelijk zijn; ongebalanceerde data worden door de schattingsprocedure 'vanzelf' correct geanalyseerd. Dit geldt uiteraard niet voor ontbrekende gegevens bij de verklarende variabelen; wanneer die zich voordoen is de gebruikelijke procedure eveneens het verwijderen van het betreffende individu. Bij dit type ontbrekende gegevens onderscheidt multiniveau analyse zich dus niet van de klassieke variantie analyse.

4. Het latente curve model voor panel data

Figuur 1 bevat het pad diagram van een eenvoudig latente curve model voor panel data met een afhankelijke variabele en drie meetmomenten.[1]

[1]De figuur volgt de SEM notatie van Bentler (1993), waarin de residuele variantie van een latente vareiabele wordt weergegeven als d (disturbance) en de residuele variantie van de geobserveerde variabelen als e (error).



Figuur 1. Pad diagram voor het Latente Curve Model

In het model van figuur 1 zijn T1, T2 en T3 de observaties van de afhankelijke variabele op opeenvolgende meetmomenten. In het latente groei model wordt de verwachte score van een individu weergegeven door een latente *intercept* factor. De intercept is constant over de tijd. De verandering over de tijd wordt weergegeven door een latente *helling* (slope) factor; deze representeert het gemiddelde en de variantie van de individuele regressiehellingen. Uit het pad diagram in figuur 1 is niet zonder meer af te lezen dat het model tevens de gemiddelden en intercepts van zowel de geobserveerde als de latente variabelen bevat. De intercepts van de geobserveerde variabelen worden gefixeerd op nul. Deze restricties leiden ertoe dat de geobserveerde gemiddelden op de opeenvolgende meetmomenten worden vertaald in een factorgemiddelde voor de latente intercept en hellingfactor. Het gemiddelde en de residuele variantie van de intercept factor representeren het gemiddelde en de variantie van de individuele intercepts van de individuele groeicurves. Individuele afwijkingen ten opzichte van het algemeen gemiddelde worden vertaald in de variantie van de intercept factor, en individuele verschillen in de ontwikkeling op de drie meetmomenten worden vertaald in de variantie van

de hellingfactor. Uit de gefixeerde ladingen van de slope factor (0,1,2) blijkt dat de gemiddelde groeicurve hier als een rechte lijn wordt gemodelleerd. Wijzigingen van de gefixeerde ladingen van de slope factor impliceren een verandering in de tijdschaal of non-lineaire groeicurves. Zowel de intercept als de slope factor kunnen op hun beurt gemodelleerd worden door verklarende variabelen op individueel niveau. Het LCM is een volledig random coëfficiënten model, dat wil zeggen dat zowel de intercepts als de hellingen van de individuele groeicurven kunnen verschillen.

Uit figuur 1 valt eenvoudig af te leiden dat het ontbreken van een individu op een van de meetmomenten leidt tot het ontbreken van de gegevens op een van de variabelen. Bij de klassieke variantie analyse leidt dat tot verwijdering van die persoon uit de datamatrix. Voor structurele modellen zijn krachtiger methoden beschikbaar. Eén mogelijkheid is het toepassen van het EM algoritme om de covariantiematrix en de gemiddelden volgens een maximum likelihood methode te schatten (vgl. Little & Rubin, 1989), en deze vervolgens in een standaard SEM programma in te voeren. Een efficiëntere methode is het gebruik van gefactoriseerde likelihood (Little & Rubin, 1989) om de parameters van het model direct te schatten op basis van de beschikbare data (Arbucle, 1996). Tenminste twee SEM-programma's kennen deze mogelijkheid: Amos (Arbucle, 1995) en Mx (Neale, 1994).

5. Multiniveau regressie en het latente curve model, bij gesimuleerde data

De basis voor de vergelijking is een gesimuleerde dataset die door Rogosa en Saner (1995) eveneens gebruikt is om verschillende analysetechnieken met elkaar te vergelijken. De dataset heeft de volgende structuur: 1 afhankelijke variabele (Y), 5 gefixeerde tijdpunten (t=0,...4), 200 individuen, en 1 tijdonafhankelijke covariaat Z op individueel niveau. De data zijn gegenereerd volgens een lineair groeimodel met meetfouten in de afhankelijke variabele. Naast een volledige dataset is ook een dataset gecreëerd waarin 7% van de datapunten at random is verwijderd. Voor het genereren van deze data is gebruik gemaakt van het programma TPSIM (Rogosa & Ghandour, 1986). Verschillende kenmerken van deze gesimuleerde dataset worden besproken door Rogosa en Saner (1995). De belangrijkste worden hieronder kort samengevat.

Parameters van gesimuleerde data

De latente variabele h wordt gegenereerd volgens het regressiemodel

$$\eta_{ti} = \beta_{0i} + \beta_{1i}t$$

met geobserveerde variabelen

$$Y_{ti} = \eta_{ti} + \varepsilon_{ti}$$

en

$$\beta_{0i} = \gamma_{00} + \gamma_{01}Z_i + u_0$$
$$\beta_{1i} = \gamma_{10} + \gamma_{11}Z_i + u_1$$

zodat

$$\eta_{ti} = \gamma_{00} + \gamma_{10}t + \gamma_{01}Z_i + \gamma_{11}tZ_i + u_0 + u_1t$$

en

$$Y_{ti} = \gamma_{00} + \gamma_{10}t + \gamma_{01}Z_i + \gamma_{11}tZ_i + u_0 + u_1t + \varepsilon_{ti}$$

Hierbij zijn η_0~N(44,52), η_1~N(49,47), η_2~N(54,52), η_3~N(59,67), η_4~N(64,52), Z~N(10,4), ε~N(0,12). β_0~N(5,5), $r_{\beta 0Z} = r_{\beta 1Z} = 0.60$.

Het voor ons belangrijkste kenmerk is dat het verband tussen het tijdstip en de ongemeten voorspelde variabele lineair is met regressiecoëfficiënt 5, en dat de geobserveerde variabele een normale meetfout heeft waarvan de variantie een functie is van de ongeobserveerde waarde.

5.1 Variantie analyse voor herhaalde metingen

Op de Rogosa/Saner data is een variantieanalyse uitgevoerd met de variabele Z als covariaat. Het programma MANOVA van SPSS eist in zo'n geval dat van de covariaat Z in totaal vijf

identieke kopieën gemaakt worden, voor ieder tijdstip één.[1] De resultaten van de analyses worden hieronder weergegeven in tabel 1.

Tabel 1. Resultaten ANOVA op Rogosa/Saner data (volledige gegevens)

tijdstip	0	1	2	3	4
gemiddelde	44.5	48.4	53.7	59.2	63.9
st. afw.	8.0	8.1	8.7	8.8	10.4
Effecten					
tijdstip	$F_{4,196}$=218.3,		p=0.00		
covariaat	$F_{1,198}$=229.8,		p=0.00	b=2.7	ß=0.73

Toetsingen voor polynome trends levert op dat de lineaire trend significant is (t_{795}=6.16, p=0.00), terwijl de hogere trends alle niet significant zijn. De Tukey test voor non-additiviteit is eveneens significant ($F_{1,795}$=37.2, p=0.00). Het gemiddelde verschil tussen twee opeenvolgende tijdstippen is 4.9; vrijwel gelijk aan de bekende regressiehelling van 5.

De resultaten bij ontbrekende data zijn vrijwel identiek. Kenmerkend voor 'listwise deletion' is wel het grote verlies aan data. Bij de onvolledige data is 7% van de data at random vervangen door een `missing value' code. SPSS' MANOVA verwijdert alle personen met een of meer missing values, waarna er nog maar 126 personen overblijven voor de analyse, ofwel 63% van de oorspronkelijke 200 personen. De resultaten van de analyses op de onvolledige data worden weergegeven in tabel 2.

Tabel 2. Resultaten ANOVA op Rogosa/Saner data (onvolledige gegevens)

tijdstip	0	1	2	3	4
gemiddelde	44.5	48.2	53.8	59.1	64.1
st. afw.	7.6	7.6	8.6	8.6	10.3
Effecten					
tijdstip	$F_{4,122}$=144.2,		p=0.00		
covariaat	$F_{1,124}$=158.2,		p=0.00	b=2.7	ß=0.75

[1]Deze omslachtige procedure maakt in ieder geval op heldere wijze duidelijk wat bedoeld wordt met een 'tijdonafhankelijke covariaat.'

Toetsingen voor polynome trends levert op dat de lineaire trend significant is (t_{500}=4.99, p=0.00), terwijl de hogere trends alle niet significant zijn. De Tukey test voor non-additiviteit is eveneens significant ($F_{1,543}$=28.2, p=0.00). Het gemiddelde verschil tussen twee opeenvolgende tijdstippen is 4.9, gelijk aan het overeenkomstige resultaat bij de volledige gegevens en vrijwel gelijk aan de bekende regressiehelling van 5.

5.2 Multiniveau regressie analyse

Bij de multiniveau regressie analyse is een sequentie van modellen geschat. Tabel 3 bevat het eindmodel voor beide datasets.

Tabel 3. Resultaten multiniveau regressie (Z gecentreerd, Mln, IGLS)

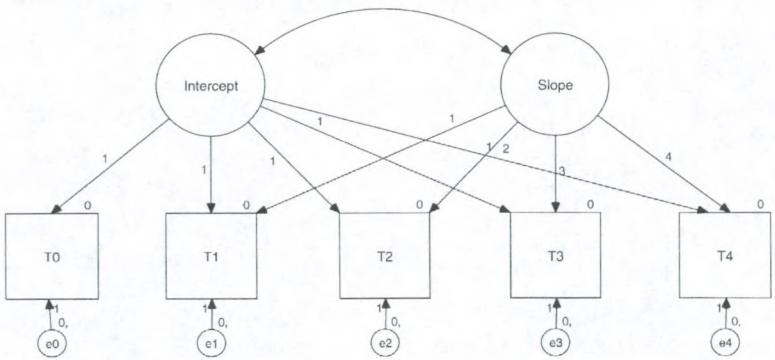
	volledige data	onvolledige data	populatie
interc	44.0 (.51)	44.2 (.52)	44
tijd	5.0 (.14)	4.9 (.14)	5
Z	1.5 (.25)	1.4 (.24)	1.4
Z*tijd	.62 (.07)	.61 (.07)	.67
s^2_e	11.9 (.69)	12.4 (.79)	12
s^2_{int}	45.3 (5.26)	40.8 (5.16)	47
s^2_{tijd}	2.7 (.40)	2.5 (.41)	3.2
s_{i*t}	-7.7 (1.25)	-6.6 (1.24)	-9.0
r_{i*t}	-.70	-.66	-.73
verklaarde variantie (vergeleken met intercept-only model)			
s^2_{int}	20%	19%	
s^2_{tijd}	39%	39%	

Bij de onvolledige dataset worden personen met een ontbrekende waarde op de covariaat Z altijd verwijderd. Bij personen die een ontbrekende meting hebben op een bepaald tijdstip, worden alleen de gegevens voor dat specifieke tijdstip verwijderd. Dit leidt tot een dataset met 866 observaties voor in totaal 186 personen, dat is 73% van de oorspronkelijke 1200 observaties voor 200 personen. Dat is beter dan de 635 bij MANOVA, maar nog steeds worden 27% van de waarnemingen niet gebruikt terwijl er slechts 7% ontbreken. Bij de gevolgde methode, die neerkomt op 'listwise deletion' voor personen met ontbrekende

waarden voor Z, is dit verlies onvermijdelijk. Ook hier geldt dat de resultaten voor beide datasets sterk op elkaar lijken. De verschillen zijn overal kleiner dan de standaardfout. Opvallend is dat bij de onvolledige data de standaardfouten nauwelijks verschillen van de standaardfouten bij de volledige dataset. Bij de volledige dataset wordt de regressiehelling van het tijdstip correct geschat, bij de onvolledige data is de afwijking echter klein. De `verklaarde variantie' is berekend door vergelijking van het eindmodel met een model zonder de betreffende termen.

5.3 Het latente curve model

Het pad diagram voor het LCM voor de Rogosa/Saner data is weergegeven in figuur 2.



Figuur 2. Latente Curve Model voor Rogosa/Saner data

De schattingen van de regressiecoëfficiënten en covarianties voor het LCM voor beide datasets zijn weergegeven in tabel 4. De analyses zijn uitgevoerd met het programma Amos, dat bij de onvolledige gegevens de analyses uitvoert op basis van alle 1119 aanwezige waarnemingen, ofwel de volle 93% die beschikbaar is. Het enige nadeel is dat de chi-kwadraatwaarde niet door het programma kan worden berekend, en daarom handmatig berekend moet worden op basis van de likelihood functie voor het getoetste en het verzadigde model (vgl. Arbucle, 1995, 1996). De resultaten voor beide datasets staan in tabel 4.

Tabel 4. Resultaten LCM (Z gecentreerd, Amos, ML)

	volledige data	onvolledige data	populatie
interc	44.1 (.51)	44.2 (.50)	44
tijd	4.9 (.14)	4.9 (.14)	5
Z	1.5 (.24)	1.5 (.24)	1.4
Z*tijd	.61 (.07)	.58 (.07)	.67
gem. s^2_e	12.4	12.4	12
s^2_{int}	45.1 (5.26)	42.0 (5.19)	47
s^2_{time}	2.9 (.41)	2.6 (.42)	3.2
s_{i*t}	-8.0 (1.28)	-7.1 (1.27)	9.0
r_{i*t}	-.71	-.68	-.73
χ^2	33.8	31.4	
df	13	13	
p	.00	.00	
TLI	.99	.99	
SMC			
voor intercept	.18	.20	
voor tijd	.36	.37	

Opnieuw geldt dat de resultaten voor beide datasets sterk op elkaar lijken. De verschillen zijn overal kleiner dan de standaardfout. Ook bij de SEM aanpak blijken de standaardfouten voor de onvolledige dataset nauwelijks te verschillen van de standaardfouten bij de volledige dataset. De bekende regressiecoëfficiënt voor de factor tijd wordt dicht benaderd. Een verschil ten opzichte van de multiniveau benadering is dat we bij structurele modellen informatie krijgen over de passing van het model als zodanig. Volgens de χ^2-test moet het model verworpen worden, maar de Tucker-Lewis goodness of fit index (TLI) is de passing voldoende. Bij de analyse van structurele modellen kan de passing exploratief verbeterd worden op basis van een inspectie van de zogenaamde modificatie-index die aangeeft welke restricties in het model problematisch zijn. Alle modificatie-indexen zijn echter kleiner dan 10, hetgeen aangeeft dat er niet één simpele modificatie van het model is die de passing statistisch sluitend maakt. In het LCM is de meest voor de hand liggende aanpassing het toestaan van covarianties tussen naastgelegen error-termen e_t, waarmee dan aangegeven wordt dat er nog onverklaarde covarianties zijn tussen dicht bij elkaar gelegen tijdstippen. Omdat volgens de TLI

goodness of fit maat de passing op zich bijzonder goed is, is van verdere modelexploratie hier afgezien.

5.4 Een vergelijking van het multiniveau en het latente curve model

Het regressiemodel voor de individuele curven voor de Rogosa/Saner data, met een intercept en een lineair verband over de tijd, wordt geschreven als:

$$Y_{ij} = \beta_{0j} + \beta_{1j}\, t_{ij} + e_{ij}. \qquad (5)$$

Na substitutie van de regressie op persoonsniveau, waarbij voor de eenvoud de covariaat Z wordt weggelaten, levert dit (zie ook vgl. 3) het multiniveau regressiemodel:

$$Y_{ij} = \gamma_{00} + \gamma_{10}\, t_{ij} + u_{0j} + u_{1j}\, t_{ij} + e_{ij} \qquad (6)$$

Voor dit model schatten we de intercept en slope en de varianties van de storingstermen: σ_{u0}^2, σ_{u1}^2 en σ_e^2.

Voor het latente curvemodel in Figuur 2 geldt dat de factormatrix Λ geheel gefixeerd is, en als volgt gespecificeerd:

$$\Lambda = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \end{bmatrix}$$

De vijf rijen van de factormatrix zijn de vijf tijdstippen; eerste kolom representeert de intercept, en de tweede de slope. De specifieke waarden voor de slope-factor impliceren een lineair verloop over de tijd. De scores van individu j op de intercept en slope factor zijn η_{0j} en η_{1j}. Dan kunnen we de vergelijking voor de score Y van persoon j op tijdstip i schrijven als:

$$Y_{ij} = \eta_{0j} + \lambda_{i1}\, \eta_{1j} \qquad (7).$$

In aanmerking genomen dat de factorladingen λ_{i1} in vergelijking (7) gelijk zijn aan de tijdstippen t=0,...4 in vergelijking (5), zal duidelijk zijn dat de modellen vrijwel identiek zijn. Het latente curve model bevat ook de geobserveerde gemiddelden. Wanneer we de intercepts

van de regressies van de geobserveerde variabelen op nul fixeren, is het gevolg dat het gemiddelde van de intercept factor de geschatte intercept γ_{00} en het gemiddelde van de slope factor is de geschatte regressiehelling γ_{10}. De variantie van de intercept factor is gelijk aan σ_{u0}^2 en de variantie van de slope factor is gelijk aan σ_{u1}^2 (voor details zie MacCallum et al., 1997). Het enige verschil in de standaard parametrisering van het multiniveau regressie model en het latente curve model is de individuele errors op de onderscheiden tijdstippen. In het multiniveau regressiemodel wordt doorgaans aangenomen dat deze constant zijn over de tijd, met variantie σ_e^2. In het latente curve model wordt doorgaans aangenomen wordt aangenomen dat deze niet constant zijn over de tijd, met variantie $\sigma_{e(i)}^2$. Dit is echter een triviaal verschil. In het latente curve model is het eenvoudig om een constraint op te leggen die specificeert dat de error-varianties van de geobserveerde variabelen gelijk zijn. Anderzijds, in multiniveau regressie is het mogelijk voor de verschillende tijdstippen verschillende varianties te laten schatten (vgl. Hedeker & Gibbons, 1995). Het kleine verschil in parametrisering is een verschil uit gewoonte, en niet een fundamenteel verschil.

Gegeven dat de verschillen tussen de beide modellen neerkomen op een andere parametrisering van fundamenteel vrijwel identieke modellen, mogen we bij een vergelijking van beide modellen sterk gelijkende uitkomsten verwachten. In tabel 5 worden de overeenkomstige schattingen van het multiniveau regressie model en het latente curve model voor de volledige data naast elkaar gepresenteerd, voorafgegaan door de bekende populatiewaarden van de gesimuleerde dataset.

Tabel 5. Vergelijking MRM en LCM voor volledige data

	populatie	MRM	LCM
interc	44	44.0 (.51)	44.1 (.56)
tijd	5	5.0 (.14)	4.9 (.17)
Z	1.4	1.5 (.25)	1.5 (.24)
Z*tijd	.67	.62 (.07)	.61 (.07)
s^2_e	12	11.9 (.69)	9.0-14.5
s^2_{int}	47	45.3 (5.26)	45.1 (5.26)
s^2_{tijd}	3.2	2.7 (.40)	2.9 (.41)
r_{i*t}	-.73	-.70	-.71

Uit tabel 5 blijkt inderdaad dat de schattingen van het MRM en het LCM dicht bij elkaar liggen, en eveneens de bekende populatiewaarden nauwkeurig schatten.

5.5 Conclusies

De algemene conclusie wat betreft het toepassen van (M)ANOVA op deze gegevens is dat ANOVA op zich correcte resultaten oplevert, maar de aanwezige informatie erg onvolledig weergeeft. ANOVA detecteert dat er sprake is van een pure lineaire trend en een significante covariaat. De interactie tussen de covariaat en de tijd factor wordt door de standaard procedure niet opgespoord; door de aanwezigheid van deze interactie wordt in feite een assumptie van het covariantie analyse model geschonden. De test voor non-additiviteit detecteert dat er een interactie is tussen de personen en de tijd factor, hetgeen wijst op heterogene individuele tijd-curven, maar deze heterogeniteit wordt niet in het model opgenomen. Onvolledigheid van de gegevens leidt niet tot duidelijk andere schattingen, maar wel tot een ernstig verlies aan gegevens.

Het multiniveau regressie model en het latente curve model leveren vrijwel identieke schattingen op, die ook dicht liggen bij de bekende populatiewaarden. In beide modellen kunnen alle gesimuleerde effecten (tijd-effect, effect van covariaat op zowel intercept als regressiehelling, residuele heterogeniteit) efficiënt gemodelleerd worden. Multiniveau regressie modellen zijn duidelijk eleganter wanneer er sprake is van een groot aantal tijdpunten die voor de verschillende personen ongelijk zijn; bij latente curve modellen is elk tijdpunt een aparte variabele, terwijl bij multiniveau regressie modellen elk tijdpunt een nieuwe waarneming is voor een enkele tijdvariabele. Bij latente curve modellen is het echter eenvoudiger om restricties op te leggen, of om het model juist uit te breiden met ingewikkelde padmodellen of multipele groep analyse.

Wanneer het latente groei model wordt gecombineerd met het eerder beschreven multiniveau covariantie structuur model, dan kunnen ook multiniveau effecten van klas en schoolvariabelen in het model worden opgenomen. Multiniveau effecten van schoolvariabelen op de intercept factor zijn te interpreteren als directe effecten van deze variabelen, terwijl effecten op de slope factor zijn te interpreteren als cross-level interacties. Het een en ander leidt tot tamelijk complexe modellen, waarbij sprake kan zijn van multigroep SEM analyses met

verschillende aantallen variabelen (Het programma EQS van Bentler (Bentler, 1993) kan zulke modellen schatten). Het is echter op voorhand niet inzichtelijk in hoeverre dit gecombineerde model in staat is cross-level interacties nauwkeurig weer te geven, en of de betreffende specificatie tot modellen leidt die empirisch geïdentificeerd zijn (het LCM model met drie meetmomenten is juist geïdentificeerd, zodat we mogen verwachten dat variaties op dit model instabiel zullen zijn).

6. Referenties.

Arbucle, J.L. (1995). *Amos user's guide*. Chicago: Smallwaters.

Arbucle, J.L. (1996). Full information estimation in the presence of incomplete data. In: G.A. Marcoulides & R.E. SChumacker (eds.). *Advanced structural equation modeling*. Mahwah, NJ: Lawrence Erlbaum Associates.

Bentler, P.M. (1993). *EQS. Structural equations program manual*. Los Angeles: BMDP Statistical Software Inc.

Bryk, A.S. & Raudenbush, S.W. (1987). *Applying the hierachical linear model to measurement of change problems*. Psychological Bulletin, 101, 147-158.

Bryk, A.S. & Raudenbush, S.W. (1992). *Hierarchical linear models*. Newbury Park, CA: Sage.

Engel, U. & Reinecke, J. (1994). *Panelanalyse. Grundlagen, Techniken, Beispiele*. Berlijn: Walter de Gruyter.

Gibbons, R.D., Hedeker, D. Elkin, I., Waternaux, C., Kraemer, H.C., Greenhouse, J.B., Shea, M.T., Imber, S.D., Sotsky, S.M. & Watkins, J.T. (1993). Some conceptual and statistical issues in analysis of longitudinal psychiatric data. *Archives of General Psychiatry*, 50, 739-750.

Goldstein, H. (1987). *Multilevel models in educational and social research*. London: Griffin.

Goldstein, H. (1989). Models for multilevel response variables with an application to growth curves. In D. Bock (ed.), *Multilevel analysis of educational data*. San Diego, CA: Academic Press.

Goldstein, H. (1995). *Multilevel statistical models*. London: Griffin.

Hedeker, D. & Gibbons, R.D. (1995). MIXREG. A computer program for mixed-effects regression analysis with autocorrelated errors.

Hox, J.J. (1995). *Applied multilevel analysis. 2nd edition*. Amsterdam: TT-Publikaties.

Jöreskog, K.G. & Sörbom, D. (1977). Statistical models and methods for the analysis of longitudinal data. In: D.J. Aigner & A.S. Goldberger (eds.) *Latent variables in socio-*

economic models. Amsterdam: North Holland.

Jöreskog, K.G. & Yang, F. (1996). Nonlinear structural equation models: the Kenny-Judd model with interactions. In: G.A. Marcoulides & R.E. SChumacker (eds.). *Advanced structural equation modeling*. Mahwah, NJ: Lawrence Erlbaum Associates.

Little, R.J.A. & Rubin, D.B. (1989). The treatment of missing data in multivariate analysis. *Sociological Methods & Research*, 18, 292-326.

MacCallum, R.C., Kim, C., Malarkey, W.B. & Kiecolt-Glaser, J.K. (1997). Studying multivariate change using multilevel models and latent curve models. *Multivariate Behavior Research*, 32, 215-253.

McDonald, R.P. (1994). The bilevel reticular action model for path analysis with latent variables. *Sociological Methods & Research*, 22, 399-413.

Meredith, W. & Tisak, J. (1990). Latent curve analysis. *Psychometrika*, 55, 107-122.

Muthén, B. (1989). Latent variable modeling in heterogeneous populations. *Psychometrika*, 54, 557-585.

Muthén, B. (1991). Analysis of longitudinal data using latent variable models with varying parameters. In: L.C.Collins & J.L. Horn (eds.) *Best methods for the analysis of change*. Washington, DC: American Psychological Association.

Muthén, B. (1994). Multilevel covariance structure analysis. *Sociological Methods & Research*, 22, 376-398.

Neale, M. (1994). *Mx: Statistical modeling*. Richmond, VA: Department of Psychiatry, Medical College of Virginia.

Plewis, I. (1985). *Analysing change*. New York: Wiley.

Raudenbush, S.W. & Bryk, A.S. (1987). Examining correlates of diversity. *Journal of Educational Statistics*, 12, 241-269.

Rogosa, D. & Ghandour, G.A. (1986). *TPSIM: A program for generating longitudinal panel data with a known structure*. Stanford, CA: Stanford University.

Rogosa, D. & Saner, H. (1995). Longitudinal data analysis examples with random coefficient models. *Journal of Educational and Behavioral Statistics*, 20, 149-170.

Snijders, T.A.B. & Maas, C. (1996). Apllication: Using MLn for repeated measures with missing data. *Multilevel modelling newsletter*, 8, 2, 7-10.

Willet, J.B. & Sayer, A.G. (1994). Using covariance structure analysis to detect correlates and predictors of change over time. *Psychological Bulletin*, 116, 363-381.

ontvangen: 06-04-1997
geaccepteerd: 15-04-1998