RIJKSINSTITUUT VOOR VOLKSGEZONDHEID EN MILIEUHYGIËNE
BILTHOVEN

Rapportnr 441110 001

HET LINEAIR GEMENGD MODEL MET DE SAS PROCEDURE MIXED

C. de Lezenne Coulander

maart 1995

Dit onderzoek werd verricht in opdracht en ten laste van het ministerie van Volksgezondheid, Welzijn en Sport, Directie Preventie, Algemene Gezondheidszorg en Opleidingen.

VERZENDLIJST

1-3	Directie Preventie, Algemene Gezondheidszorg en Opleidingen van het Ministerie van Welzijn, Volksgezondheid en Cultuur
4	Directeur-Generaal van de Volksgezondheid
5	Hoofinspectie voor de Gezondheidszorg
6-7	Hoofdinspectie voor de preventieve en curatieve gezondheidszorg
8	Depot van Nederlandse publikaties en Nederlandse bibliografie
9	Directie RIVM
10	Prof. dr. ir. D. Kromhout
11	C.B. Ameling
12	Dr. ir. B.P.M. Bloemberg
13	Ir. G. Doornbos
14	Dr. ir. E.J.M. Feskens
15	Ir. P.H. Fischer
16	Drs. A. van der Giessen
17	Dr. S.H. Heisterkamp
18	Ir. J. in 't Hout
19	J.M. Klokman-Houweling
20	Dr. ir. E. Lebret
21	Ir. A.H.P. Luijben
22	Dr. ir. J.C. Seidell
23	Dr. ir. H.A. Smit
24-25	Auteur
26	Hoofd Bureau Voorlichting en Public Relations
27-28	Bibliotheek RIVM
29	Bureau projecten- en rapportenregistratie
30-50	Reserve exemplaren

INHOUDSOPGAVE

ABSTRACT

The SAS procedure MIXED fits mixed linear models (models with both fixed and random effects). The mixed model analyzes data with several sources of variation instead of just one (as with the General Linear Model used by the SAS procedure GLM). Some different mixed models are accessible, for instance split-plot designs, repeated measures, random coefficients, best linear unbiased predictions (BLUP) and heterogeneous variances. Models are fit using maximum likelihood (ML) or restricted maximum likelihood (REML).

SAMENVATTING

De SAS procedure MIXED schat parameters voor gemengde lineaire modellen (modellen met zowel 'fixed' als 'random' effecten). Het gemengde model analyseert data met meerdere variantie bronnen in plaats van één bron (zoals met het algemene lineaire model, dat door de SAS procedure GLM gebruikt wordt). Een aantal verschillende gemengde modellen zijn toepasbaar, b.v. split-plot modellen, herhaalde metingen, random coëfficinten, BLUP (best linear unbiased predictor) modellen en heterogene varianties. Model parameters worden geschat met behulp van 'maximum likelihood' (ML) of 'restricted maximum likelihood' (REML).
Dit rapport wordt uitgebracht in het kader van de taak die de afdeling IMA voor het RIVM vervult.

RIJKSINSTITUUT VOOR VOLKSGEZONDHEID EN MILIEUHYGIËNE
BILTHOVEN

Rapportnr 441111 007

HET LINEAIR GEMENGD MODEL MET DE SAS PROCEDURE MIXED

C. de Lezenne Coulander

maart 1995

Dit onderzoek werd verricht in opdracht en ten laste van het ministerie van Volksgezondheid, Welzijn en Sport, Directie Preventie, Algemene Gezondheidszorg en Opleidingen.

VERZENDLIJST

1-3 Directie Preventie, Algemene Gezondheidszorg en Opleidingen van het Ministerie van Welzijn, Volksgezondheid en Cultuur
4 Directeur-Generaal van de Volksgezondheid
5 Plv Directeur-Generaal van de Volksgezondheid, tevens Hoofddirecteur Financiering en Planning
6 Hoofddirecteur Gezondheidsbescherming
7-8 Geneeskundig Hoofdinspectie van de Volksgezondheid
9 Depot van Nederlandse publikaties en Nederlandse bibliografie
10 Directie RIVM
11 Prof. dr. ir. D. Kromhout
12 Mw. C.B. Ameling
13 Dr. ir. B.P.M. Bloemberg
14 Mw. ir. G. Doornbos
15 Mw. dr. ir. E.J.M. Feskens
16 Ir. P.H. Fischer
17 Drs. A. van der Giessen
18 Drs. S.H. Heisterkamp
19 Mw. ir. J. in 't Hout
20 Mw. J.M. Klokman-Houweling
21 Dr. ir. E. Lebret
22 Ir. A.H.P. Luijben
23 Dr. ir. J.C. Seidell
24 Mw. dr. ir. H.A. Smit
25-26 Auteur
27 Hoofd Bureau Voorlichting en Public Relations
28-29 Bibliotheek RIVM
30 Bureau projecten- en rapportenregistratie
31-50 Reserve exemplaren

INHOUDSOPGAVE

ABSTRACT

The SAS procedure MIXED fits mixed linear models (models with both fixed and random effects). The mixed model analyzes data with several sources of variation instead of just one (as with the General Linear Model used by the SAS procedure GLM). Some different mixed models are accessible, for instance split-plot designs, repeated measures, random coefficients, best linear unbiased predictions (BLUP) and heterogeneous variances. Models are fit using maximum likelihood (ML) or restricted maximum likelihood (REML).

SAMENVATTING

De SAS procedure MIXED schat parameters voor gemengde lineaire modellen (modellen met zowel 'fixed' als 'random' effecten). Het gemengde model analyseert data met meerdere variantie bronnen in plaats van één bron (zoals met het algemene lineaire model, dat door de SAS procedure GLM gebruikt wordt). Een aantal verschillende gemengde modellen zijn toepasbaar, b.v. split-plot modellen, herhaalde metingen, random coëfficienten, BLUP (best linear unbiased predictor) modellen en heterogene varianties. Model parameters worden geschat met behulp van 'maximum likelihood' (ML) of 'restricted maximum likelihood' (REML).
Dit rapport wordt uitgebracht in het kader van de taak die de afdeling IMA voor het RIVM vervult.

1. INLEIDING.

De SAS procedure PROC MIXED is bedoeld voor het analyseren van data aan de hand van een 'Linear Mixed Model' (LMM), een lineair gemengd model. Het woord 'mixed' slaat op de aanwezigheid van zowel 'fixed effect' variabelen als 'random effect' variabelen in het model. Deze termen zullen nader worden toegelicht (zie par. 2.1). Een model wordt gespecificeerd als betrekking tussen een afhankelijke variabele en een aantal onafhankelijke variabelen met coëfficiënten, die parameters of ook wel effecten worden genoemd. Een model heet lineair als het model linair in de parameters is. Zo is b.v. het model $y = a + b_1x + b_2x^2 + b_3x^3$ een lineair model, omdat de parameters b_i allen van de eerste graad zijn, terwijl de variabelen x, x^2en x^3 van hogere graad kunnen zijn.
Het preciese onderscheid tussen 'mixed' en 'random' is moeilijk aan te geven. In wezen hangt dit sterk af van het doel van het onderzoek. Wil men voorspellingen doen in situaties, die algemener zijn dan de condities van het onderzoek of experiment? Zo ja, dan zou men sommige regressie variabelen als random kunnen interpreteren. Soms heeft een zelfde onderzoek meer dan één toepassingen!

Het LMM volgt uit het gewone 'Linear Model' (LM) door het toelaten van onafhankelijke random variabelen in het model. Een andere bestaande uitbreiding van het LM is het zgn. 'Generalized Linear Model' (GLM). Hierbij mag de afhankelijke variabele een exponentiële verdeling volgen i.p.v. een normale en bovendien mag het verband tussen de afhankelijke variabele en de onafhankelijke variabelen met een continue functie (de zgn. 'link' functie) worden beschreven. Een derde uitbreiding van het LM is een model met de combinatie van beide uitbreidingen, dus random variabelen, exponentiële verdeling en een 'link' functie. De naam voor deze derde mogelijkheid is 'Generalized Linear Mixed Model' (GLMM). Schematisch kunnen we deze vier modellen als volgt weergeven:

LINEAR MODEL (LM)	→	LINEAR MIXED MODEL (LMM)
onafhankelijk normaal verdeeld 'fixed' effecten	plus random effecten	afhankelijk normaal verdeeld 'fixed' en 'random' effecten
plus 'link' functie ↓		
GENERALIZED LINEAR MODEL (GLM)		GENERALIZED LINEAR MIXED MODEL (GLMM)
onafhankelijk exponentiële familie 'link' functie 'fixed' effecten		afhankelijk exponentiële familie 'link' functie 'fixed' en 'random' effecten

Het vierde model in dit schema, het GLMM, is een logische derde uitbreiding van het LM, waarbij de eigenschappen van het LMM en het GLM worden gecombineerd. Het GLMM maakt het mogelijk om b.v. een logistische regressie uit te voeren op herhaalde metingen. Het GLMM is al door veel auteurs bediscussieerd. Er bestaat nog geen SAS procedure voor het GLMM, alhoewel men met PROC CATMOD voor discrete variabelen een aardig eind in de richting komt.

Het LMM onderscheidt zich van het LM doordat de eis van onafhankelijke waarnemingen wordt losgelaten. Afhankelijke waarnemingen ontstaan doordat b.v. metingen van één persoon op verschillende tijdstippen van elkaar zullen afhangen. Deze afhankelijkheid komt in de covariantie matrix van de waarnemingen tot uiting als niet-diagonale elementen ongelijk nul (bij onafhankelijke waarnemingen is de covariantie matrix zuiver diagonaal). Een verdere toelichting van het LMM volgt in het tweede hoofdstuk.

De procedure PROC MIXED kan worden gebruikt voor b.v. het split-plot model en voor herhaalde metingen. Vooral dit laatste is een veel gebruikte toepassing. PROC MIXED kan wat herhaalde metingen betreft meer dan PROC GLM. Dit komt doordat in PROC MIXED de vorm van de var-covar matrix van de error variabelen en de random variabelen gekozen kan worden. We zetten de mogelijkheden even op een rijtje:

In PROC GLM zijn de vier belangrijkste benaderingen voor de herhaalde metingen analyse:
- een afzonderlijke univariate analyse binnen elke periode
- een analyse van de data alsof ze afkomstig zijn van een split-unit experiment
- een multivariate variantie analyse, waarbij voor elke periode een afzonderlijke variabele wordt genomen
- een analyse van specifieke contrasten over de perioden.

In PROC MIXED zijn de bovengenoemde vier benaderingen ook mogelijk, maar bovendien heeft de procedure de volgende extra's:
- een nieuwe benadering is mogelijk in de vorm van het random coëfficiënten regressie model
- personen met missing data worden toch meegerekend in de analyse
- tijdsafhankelijke covariabelen
- een continue tijdschaal is mogelijk

Het gebruik van random variabelen in het lineaire model werd in het verleden bemoeilijkt door een gebrek aan daarvoor geschikte software. Bovendien waren de resultaten voor gebalanceerde designs (dit betekent een gelijk aantal personen voor elke mogelijke combinatie van de waarden van de onafhankelijke variabelen; als dus één of meer combinaties niet voorkomen, is het design reeds ongebalanceerd) gelijk aan die van de fixed-effect modellen voor wat betreft de schattingen van de effecten. Alleen de standaardfouten wijzigden zich. Maar voor ongebalanceerde designs is de situatie toch wel anders en hier heeft het wel degelijk zin om het LMM toe te passen. Zeker nu de benodigde software wel voor handen is.

2. HET LINEAIR GEMENGD MODEL.

2.1 *OVER 'FIXED' EN 'RANDOM' EFFECTEN.*

Het lineair gemengd model (LMM) onderscheidt zich van het 'gewone' lineaire model (LM) door de introductie van 'random' effecten. Om het LMM goed te begrijpen is het daarom nodig om eerst te weten wat onder 'fixed' en 'random' effecten wordt verstaan.

Neem b.v. het eenvoudige variantie model met één afhankelijke variabele y en één onafhankelijke factor A (een factor is een variabele met discrete waarden). Het verband tussen y en A kan nu met het volgende model worden beschreven:

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij}, \qquad i = 1,.....,p \text{ (factor A heeft p niveaus)}$$
$$j = 1,.....,n \text{ (aantal personen n)}$$

Met

y_{ij} = de gemeten y waarde voor persoon j en factorwaarde i.

μ = het berekende gemiddelde van alle y waarden.

α_i = het effect van factorwaarde i, d.i. het verschil tussen μ en het gemiddelde van de y waarden met factorwaarde i. De term $\mu + \alpha_i$ is dus het berekende gemiddelde van alle y waarden met factorwaarde i.

ε_{ij} = het verschil tussen de gemeten y_{ij} en de door het model berekende y-waarde (dit verschil wordt 'error' term genoemd).

Op grond van waargenomen y waarden en factorwaarden van A (p niveaus) worden de onbekende parameters μ, α_1,....., α_p (ook wel effecten genoemd) geschat. Als de p niveaus van factor A vaste, door de onderzoeker gekozen waarden hebben, spreken we van een factor A met vaste ('fixed') effecten α_i. De waargenomen y waarden worden opgevat als realisaties van de random variabelen Y_j (j = 1,...,n) met een normale verdeling (gemiddelde μ en gelijke variantie σ^2). De ε_{ij} (j = 1,...,n) hebben eveneens een normale verdeling, maar met gemiddelde nul en variantie σ^2.

Het bovenstaande 'fixed' model komt overeen met het opsplitsen van de hele populatie in p vaste subpopulaties. Maar het is ook denkbaar dat de onderzoeker volkomen willekeurig p verschillende subpopulaties trekt uit een oneindig grote populatie van mogelijke subpopulaties. Binnen elke subpopulatie i worden dan weer n waarnemingen gedaan van de random variabele Y (als we van gebalanceerde data uitgaan); stel subpopulatie i heeft gemiddelde y waarde m_i. Deze gemiddelde waarden m_i (i = 1,.....,p) vertegenwoordigen een random steekproef uit een normaal verdeelde populatie met gemiddelde μ (dit is tevens weer het totale gemiddelde voor alle y waarden) en variantie σ_a^2. We representeren deze p verschillende gemiddelden m_i door middel van een effect $a_i = m_i - \mu$ en schrijven het geheel als een lineair model:

$$y_{ij} = \mu + a_i + \varepsilon_{ij}, \qquad i = 1,.....,p \text{ en } j = 1,.....,n$$

Dit is geheel gelijk aan het bovenstaand model met als enig verschil dat de effecten a_i nu afhangen van de uitkomst van een steekproef. We spreken daarom van 'random' effecten en van een 'random' effect model.

Voorbeeld 1(Fixed effect).

Een landbouw experiment bestaat uit het testen van de efficiëntie van drie kunstmest stoffen (n.l. stikstof, fosfor en kalium) voor de opbrengst van de oogst. Stel er worden 24 planten gekweekt, waarvan er zes kunstmest met stikstof krijgen, zes kunstmest met fosfor,

zes kunstmest met kalium en zes helemaal geen kunstmest (controle groep). Een geschikt analyse model voor dit experiment zou zijn

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

met i = 1,.....,4 (aantal experimenten)

j = 1,.....,6 (aantal planten per experiment)

De bedoeling is om het verschil in opbrengst per experiment te testen. De belangstelling gaat hierbij uit naar alleen de vier genoemde zeer specifieke behandelingen, aan eventuele andere kunstmest stoffen wordt niet gedacht. Daarom spreken we hier over vaste effecten α_i en omdat er buiten de error termen hier geen andere effecten in het model zijn heet het model 'fixed'.

Voorbeeld 2(Random effect).

Een laboratorium experiment voor de bepaling van de vruchtbaarheid bij muizen gebruikt het gewicht van een tien dagen oude worp als een maat voor de vruchtbaarheid. Vier muizen, die elk zes worpen hadden, leveren de data. Het model kan geschreven worden als:

$$y_{ij} = \mu + \delta_i + \varepsilon_{ij}$$

met i = 1,.....,4 (aantal muizen met zes worpen)

j = 1,.....,6 (aantal worpen per muis)

Het model ziet er precies eender uit als het vorige, maar is toch anders. De primaire interesse gaat nu niet uit naar de vier specifieke muizen met zes worpen, maar in feite naar alle vrouwelijke muizen in de populatie. De vier specifieke muizen worden hier opgevat als een random steekproef uit de populatie van alle vrouwelijke muizen. De effecten δ_i zijn hier random met een gemiddelde nul en een variantie ongelijk nul. Omdat het enige effect in dit model buiten de error term en de niveau parameter om random is, spreken we hier van een 'random effects' model.

De keuze tussen een 'fixed effect' of een 'random effect' is niet altijd eenduidig te maken. Neem b.v. het geval van jaar effecten bij de bestudering van graan oogsten: zijn de effecten van de jaren op de oogst te beschouwen als 'fixed' of als 'random' ? De jaren zelf zijn niet random, want meestal worden achtereenvolgende jaren beschouwd. Maar de effecten op de oogst zouden wel als random kunnen worden genomen, tenzij men geïnteresseerd is in een aantal specifieke jaren. De consequenties van deze keuze zullen in het vervolg blijken. Het hangt dus sterk van het doel af, en dezelfde gegevens zouden (met andere doelen) verschillend geanalyseerd kunnen worden.

2.2 *DE WISKUNDIGE BESCHRIJVING VAN HET LINEAIR GEMENGD MODEL.*

We gaan nu over tot het beschrijven van het lineair gemengd model (LMM). Dit model bevat, zoals met de naam al aangegeven wordt, zowel 'fixed' als 'random' effecten en deze worden ook in de formule apart weergegeven.

$$\mathbf{Y} = \mathbf{X}\,\beta + \mathbf{Z}\,\nu + \varepsilon$$

Hierin is:

Y een vector van Y_i variabelen (continue random variabelen)

X is een model matrix voor de 'fixed' effect variabelen (overeenkomend met de x variabelen in een regressie model)

β een vector van 'fixed' effecten

Z een model matrix voor de 'random' effect variabelen (overeenkomend met matrix

X, maar dan alleen voor de 'random' effect variabelen)
ν een vector van 'random' effecten
ε een vector van 'error' termen

Uitgeschreven in matrix notatie ziet het bovenstaande er als volgt uit:

$$\begin{pmatrix} y_1 \\ . \\ . \\ . \\ . \\ y_n \end{pmatrix} = \begin{pmatrix} x_{11} & . & . & x_{1p} \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ x_{n1} & . & . & x_{np} \end{pmatrix} \begin{pmatrix} \beta_1 \\ . \\ . \\ \beta_p \end{pmatrix} + \begin{pmatrix} z_{11} & . & z_{1q} \\ . & . & . \\ . & . & . \\ . & . & . \\ . & . & . \\ z_{n1} & . & z_{nq} \end{pmatrix} \begin{pmatrix} \nu_1 \\ . \\ \nu_q \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ . \\ . \\ . \\ . \\ \epsilon_n \end{pmatrix}$$

met p = aantal te schatten fixed effect parameters
en q = aantal te schatten random effect parameters
Verondersteld wordt, dat ν en ε onafhankelijk van elkaar zijn en dat ze allebei de verwachtingswaarde nul hebben en var-covar matrices **G** en **R** respectievelijk. De verwachtingswaarde van Y is dan gelijk aan:

$$E(\mathbf{Y}) = \mathbf{X}\ \beta$$

De var-covar matrix **V** van **Y** (ook wel genoteerd als var(**Y**) = **V**) wordt gevonden met de betrekking:

$$\text{var}(\mathbf{Y}) = \mathbf{V} = E\ [\ (\mathbf{Y} - E(\mathbf{Y}))\ (\mathbf{Y} - E(\mathbf{Y}))'\ ]$$

$$\mathbf{V} = E[(\mathbf{Y} - \mathbf{X}\ \beta)\ (\mathbf{Y} - \mathbf{X}\ \beta)'] = E[(\mathbf{Z}\ \nu + \varepsilon)\ (\mathbf{Z}\ \nu + \varepsilon)'] = E[(\mathbf{Z}\ \nu + \varepsilon)\ (\nu'\ \mathbf{Z}' + \varepsilon')]$$

$$\mathbf{V} = E[\ \mathbf{Z}\ \nu\ \nu'\ \mathbf{Z}'] + E[\mathbf{Z}\ \nu\ \varepsilon'] + E[\varepsilon\ \nu'\ \mathbf{Z}'] + E[\varepsilon\varepsilon']$$

$$\mathbf{V} = \mathbf{Z}\ E[\nu\ \nu']\ \mathbf{Z}' + \mathbf{Z}\ E[\nu\ \varepsilon'] + E[\varepsilon\ \nu']\ \mathbf{Z}' + E[\varepsilon\varepsilon']$$

Omdat ν en ε onafhankelijk zijn van elkaar is $E[\nu\ \varepsilon'] = 0$ en $E[\varepsilon\ \nu'] = 0$; voorts geldt dat $E[\nu\ \nu'] = \mathbf{G}$ en $E[\varepsilon\ \varepsilon'] = \mathbf{R}$, zodat:

$$\mathbf{V} = \mathbf{Z}\ \mathbf{G}\ \mathbf{Z}' + \mathbf{R}$$

Voor het fixed effect model was de var-covar matrix van **Y** een diagonaal matrix (d.w.z alleen de termen op de hoofddiagonaal ongelijk nul en alle termen daarbuiten gelijk aan nul) en de var-covar matrix van de error termen was daaraan gelijk. Voor het random effect model zien we dat matrix **V** over het algemeen niet diagonaal zal zijn. Zelfs als matrices **G** en **R** diagonaal zijn, dan hangt het nog van de structuur van **Z** af hoe matrix **V** er uit zal zien. Een niet diagonale **V** betekent, dat er correlaties bestaan tussen verschillende y_i 's, d.w.z. één of meer metingen zijn afhankelijk van elkaar. Dergelijke omstandigheden komen b.v. voor bij herhaalde metingen aan dezelfde personen. Matrix **V** wordt gemoduleerd door de twee matrices **G** en **R**, zodat het mogelijk is om zowel via een keuze van **G** als van **R** hetzelfde resultaat voor **V** te bereiken. We zullen in de volgende paragraaf de samenhang tussen de drie genoemde matrices wat nader bekijken.

Samenvattend kan voor het mixed model gezegd worden, dat het toevoegen van een

random effect aan een fixed effect model tot gevolg heeft, dat de var-covar matrix van de waarnemingen Y_i verandert en dat de betrouwbaarheidsgrenzen voor de schatting van de effecten zich wijzigen, terwijl de schattingen van de effecten zelf gelijk blijven (want E(**Y**) = **X** β en deze vergelijkingen bevatten niet de random effecten).

2.3 *NADERE BESCHOUWING VAN DE VARIANTIE MATRICES G EN R.*

In deze paragraaf staat de hierboven gegeven betrekking tussen matrices **V**, **G** en **R** centraal:

$$\mathbf{V} = \mathbf{Z}\,\mathbf{G}\,\mathbf{Z}' + \mathbf{R}$$

of uitgeschreven in matrix notatie:

$$\begin{pmatrix} v_{11} & \cdot & \cdot & \cdot & v_{1n} \\ \cdot & & & & \cdot \\ \cdot & & & & \cdot \\ \cdot & & & & \cdot \\ v_{n1} & \cdot & \cdot & \cdot & v_{nn} \end{pmatrix} = \begin{pmatrix} z_{11} & \cdot & \cdot & z_{1q} \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ z_{n1} & \cdot & \cdot & z_{nq} \end{pmatrix} \begin{pmatrix} g_{11} & \cdot & \cdot & g_{1q} \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ g_{q1} & \cdot & \cdot & g_{qq} \end{pmatrix} \begin{pmatrix} z_{11} & \cdot & \cdot & \cdot & z_{n1} \\ \cdot & & & & \cdot \\ \cdot & & & & \cdot \\ z_{1q} & \cdot & \cdot & \cdot & z_{nq} \end{pmatrix} + \begin{pmatrix} r_{11} & \cdot & \cdot & \cdot & r_{1n} \\ \cdot & & & & \cdot \\ \cdot & & & & \cdot \\ \cdot & & & & \cdot \\ r_{n1} & \cdot & \cdot & \cdot & r_{nn} \end{pmatrix}$$

De **Z** matrix is (evenals de **X** matrix) gegeven met de keuze van de onafhankelijke variabelen en bevat nullen en énen voor een discrete variabele en hogere waarden voor een continue variabele. De variantie matrices **G** en **R** worden uit de data geschat. Dit kan een omvangrijke klus zijn, want matrix **R** bevat n × n onbekende termen. Het aantal te schatten termen kan worden verminderd door te veronderstellen, dat de matrices bepaalde structuren aannemen. PROC MIXED biedt naast de ongestructureerde matrices nog een aantal verschillende mogelijkheden. Zo kunnen voor **G** en **R** afzonderlijk de volgende structuren worden gekozen:

- diagonaal (ook wel simpel genoemd)
- compound symmetrie
- banden structuur
- eerste orde autoregressief
- Toeplitz structuur
- ruimtelijke structuur

(zie SAS Technical Report P-229: blz 311-313 voor de betekenis van deze termen)

We gaan na hoe de var-covar matrix van y er uit gaat zien bij een bepaalde keuze van de matrices **G** en **R**.

In alle volgende voorbeelden binnen deze paragraaf zullen we de hiernaast afgebeelde **Z** matrix gebruiken:

In dit voorbeeld zijn er vijf verschillende scores voor het random effect (vijf kolommen) en er zijn steeds drie metingen per persoon (in totaal vijf personen), dus 15 rijen.

$$\boldsymbol{Z} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

Keuze 1: matrix **G** diagonaal en matrix **R** diagonaal.

$$\boldsymbol{G} = \begin{pmatrix} 7 & 0 & 0 & 0 & 0 \\ 0 & 7 & 0 & 0 & 0 \\ 0 & 0 & 7 & 0 & 0 \\ 0 & 0 & 0 & 7 & 0 \\ 0 & 0 & 0 & 0 & 7 \end{pmatrix}$$

$$\boldsymbol{R} = \begin{pmatrix} 3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 3 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 3 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 3 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 3 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 3 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 3 \end{pmatrix}$$

Dan geeft $\mathbf{V} = \mathbf{Z}\,\mathbf{G}\,\mathbf{Z}' + \mathbf{R}$ het volgende resultaat:

$$
\boldsymbol{V} = \begin{pmatrix}
10 & 7 & 7 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
7 & 10 & 7 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
7 & 7 & 10 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 10 & 7 & 7 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 7 & 10 & 7 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 7 & 7 & 10 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 10 & 7 & 7 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 7 & 10 & 7 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 7 & 7 & 10 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 10 & 7 & 7 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 7 & 10 & 7 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 7 & 7 & 10 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 10 & 7 & 7 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 7 & 10 & 7 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 7 & 7 & 10
\end{pmatrix}
$$

Matrix **V** blijkt een zgn. blok-diagonale vorm aan te nemen met vijf gelijke 3 × 3 blokken langs de hoofddiagonaal. Per antwoordcategorie van het random effect wordt dus een vierkant blok gevormd in de **V** matrix en de afmeting van een vierkant blok wordt bepaald door het aantal observaties per antwoordcategorie van het random effect.
Om de gedachte te bepalen kan men b.v. uitgaan van een herhaalde meting van een variabele, waarbij steeds bij dezelfde vijf personen op drie verschillende tijdstippen deze variabele wordt gemeten. De vijf personen worden gezien als een random steekproef uit een grote populatie, zodat deze personen als een random effect in het model kunnen worden opgenomen. Dit model leidt tot een var-covar matrix voor de waarnemingen als hierboven aangegeven. We zien hieraan, dat voor elk persoon (= vierkante blok in **V**) dezelfde covariantie structuur geldt, n.l. tussen elk tweetal tijdstippen steeds dezelfde covariantie van 7. De matrixvorm, die we hier aantreffen voor elk blok, staat bekend als 'compound symmetry', terwijl we matrix **V** als geheel omschrijven als een blok-diagonale matrix.
Ondertussen is deze benadering voor een herhaalde metingen analyse in het algemeen niet voldoende. Vaak zal het voorkomen, dat de covariantie tussen de metingen afneemt naarmate de tijd verstrijkt. Dit kan in matrix **V** tot uitdrukking komen als elk blok in matrix **R** van het type AR(1) (d.i. eerste orde autoregressief) wordt gekozen.

Keuze 2: matrix **G** diagonaal en matrix **R** van type AR(1).

$$
\boldsymbol{G} = \begin{pmatrix}
7 & 0 & 0 & 0 & 0 \\
0 & 7 & 0 & 0 & 0 \\
0 & 0 & 7 & 0 & 0 \\
0 & 0 & 0 & 7 & 0 \\
0 & 0 & 0 & 0 & 7
\end{pmatrix}
$$

$$
\boldsymbol{R} = \begin{pmatrix}
100 & 90 & 81 & 73 & 66 & 59 & 53 & 48 & 43 & 39 & 35 & 31 & 28 & 25 & 23 \\
90 & 100 & 90 & 81 & 73 & 66 & 59 & 53 & 48 & 43 & 39 & 35 & 31 & 28 & 25 \\
81 & 90 & 100 & 90 & 81 & 73 & 66 & 59 & 53 & 48 & 43 & 39 & 35 & 31 & 28 \\
73 & 81 & 90 & 100 & 90 & 81 & 73 & 66 & 59 & 53 & 48 & 43 & 39 & 35 & 31 \\
66 & 73 & 81 & 90 & 100 & 90 & 81 & 73 & 66 & 59 & 53 & 48 & 43 & 39 & 35 \\
59 & 66 & 73 & 81 & 90 & 100 & 90 & 81 & 73 & 66 & 59 & 53 & 48 & 43 & 39 \\
53 & 59 & 66 & 73 & 81 & 90 & 100 & 90 & 81 & 73 & 66 & 59 & 53 & 48 & 43 \\
48 & 53 & 59 & 66 & 73 & 81 & 90 & 100 & 90 & 81 & 73 & 66 & 59 & 53 & 48 \\
43 & 48 & 53 & 59 & 66 & 73 & 81 & 90 & 100 & 90 & 81 & 73 & 66 & 59 & 53 \\
39 & 43 & 48 & 53 & 59 & 66 & 73 & 81 & 90 & 100 & 90 & 81 & 73 & 66 & 59 \\
35 & 39 & 43 & 48 & 53 & 59 & 66 & 73 & 81 & 90 & 100 & 90 & 81 & 73 & 66 \\
31 & 35 & 39 & 43 & 48 & 53 & 59 & 66 & 73 & 81 & 90 & 100 & 90 & 81 & 73 \\
28 & 31 & 35 & 39 & 43 & 48 & 53 & 59 & 66 & 73 & 81 & 90 & 100 & 90 & 81 \\
25 & 28 & 31 & 35 & 39 & 43 & 48 & 53 & 59 & 66 & 73 & 81 & 90 & 100 & 90 \\
23 & 25 & 28 & 31 & 35 & 39 & 43 & 48 & 53 & 59 & 66 & 73 & 81 & 90 & 100
\end{pmatrix}
$$

De hieruit volgende **V** matrix is:

$$
\boldsymbol{V} = \begin{pmatrix}
107 & 97 & 88 & 73 & 66 & 59 & 53 & 48 & 43 & 39 & 35 & 31 & 28 & 25 & 23 \\
97 & 107 & 97 & 81 & 73 & 66 & 59 & 53 & 48 & 43 & 39 & 35 & 31 & 28 & 25 \\
88 & 97 & 107 & 90 & 81 & 73 & 66 & 59 & 53 & 48 & 43 & 39 & 35 & 31 & 28 \\
73 & 81 & 90 & 107 & 97 & 88 & 73 & 66 & 59 & 53 & 48 & 43 & 39 & 35 & 31 \\
66 & 73 & 81 & 97 & 107 & 97 & 81 & 73 & 66 & 59 & 53 & 48 & 43 & 39 & 35 \\
59 & 66 & 73 & 88 & 97 & 107 & 90 & 81 & 73 & 66 & 59 & 53 & 48 & 43 & 39 \\
53 & 59 & 66 & 73 & 81 & 90 & 107 & 97 & 88 & 73 & 66 & 59 & 53 & 48 & 43 \\
48 & 53 & 59 & 66 & 73 & 81 & 97 & 107 & 97 & 81 & 73 & 66 & 59 & 53 & 48 \\
43 & 48 & 53 & 59 & 66 & 73 & 88 & 97 & 107 & 90 & 81 & 73 & 66 & 59 & 53 \\
39 & 43 & 48 & 53 & 59 & 66 & 73 & 81 & 90 & 107 & 97 & 88 & 73 & 66 & 59 \\
35 & 39 & 43 & 48 & 53 & 59 & 66 & 73 & 81 & 97 & 107 & 97 & 81 & 73 & 66 \\
31 & 35 & 39 & 43 & 48 & 53 & 59 & 66 & 73 & 88 & 97 & 107 & 90 & 81 & 73 \\
28 & 31 & 35 & 39 & 43 & 48 & 53 & 59 & 66 & 73 & 81 & 90 & 107 & 97 & 88 \\
25 & 28 & 31 & 35 & 39 & 43 & 48 & 53 & 59 & 66 & 73 & 81 & 97 & 107 & 97 \\
23 & 25 & 28 & 31 & 35 & 39 & 43 & 48 & 53 & 59 & 66 & 73 & 88 & 97 & 107
\end{pmatrix}
$$

Deze matrix is bijna gelijk aan de **R** matrix. Het verschil is, dat in de vijf blokken (drie bij drie) langs de hoofddiagonaal alle elementen met zeven vermeerderd zijn. De **V** matrix wijkt hier niet veel af van het type AR(1). We zien per persoon echter blokken van het type AR(1), zodat inderdaad tussen de metingen op het eerste en het derde tijdstip een

zwakkere covariantie bestaat dan tussen de metingen op het eerste en het tweede tijdstip. We zien tevens, dat de covarianties afnemen voor metingen tussen persoon 1 en persoon 2 t.o.v. metingen tussen persoon 1 en persoon 3 enz. Dit betekent, dat er hier een rangorde bestaat tussen de personen. Dit zal in werkelijkheid meestal niet zo zijn. Uiteindelijk kan men het beste twee ongestructureerde matrices **G** en **R** opgeven, maar door het grote aantal parameters, dat dan geschat moeten worden, kan het in de praktijk gauw een onoplosbaar probleem worden. Omdat het illustratief is voor het begrip van de betrekking **V = Z G Z' + R** geven we hieronder nog de voorbeelden met een ongestructureerde **G** of een ongestructureerde **R**.

Keuze 3: matrix **G** ongestructureerd en matrix **R** diagonaal.

$$G = \begin{pmatrix} 1 & 3 & 2 & 6 & 8 \\ 3 & 3 & 4 & 4 & 6 \\ 5 & 4 & 9 & 8 & 5 \\ 3 & 7 & 6 & 2 & 7 \\ 1 & 4 & 3 & 5 & 7 \end{pmatrix}$$

Matrix **R** heeft allemaal drieën op de hoofddiagonaal staan en bevat verder nullen (15 × 15). Dan is matrix **V** als volgt:

$$V = \begin{pmatrix} 4 & 1 & 1 & 3 & 3 & 3 & 2 & 2 & 2 & 6 & 6 & 6 & 8 & 8 & 8 \\ 1 & 4 & 1 & 3 & 3 & 3 & 2 & 2 & 2 & 6 & 6 & 6 & 8 & 8 & 8 \\ 1 & 1 & 4 & 3 & 3 & 3 & 2 & 2 & 2 & 6 & 6 & 6 & 8 & 8 & 8 \\ 3 & 3 & 3 & 6 & 3 & 3 & 4 & 4 & 4 & 4 & 4 & 4 & 6 & 6 & 6 \\ 3 & 3 & 3 & 3 & 6 & 3 & 4 & 4 & 4 & 4 & 4 & 4 & 6 & 6 & 6 \\ 3 & 3 & 3 & 3 & 3 & 6 & 4 & 4 & 4 & 4 & 4 & 4 & 6 & 6 & 6 \\ 5 & 5 & 5 & 4 & 4 & 4 & 12 & 9 & 9 & 8 & 8 & 8 & 5 & 5 & 5 \\ 5 & 5 & 5 & 4 & 4 & 4 & 9 & 12 & 9 & 8 & 8 & 8 & 5 & 5 & 5 \\ 5 & 5 & 5 & 4 & 4 & 4 & 9 & 9 & 12 & 8 & 8 & 8 & 5 & 5 & 5 \\ 3 & 3 & 3 & 7 & 7 & 7 & 6 & 6 & 6 & 5 & 2 & 2 & 7 & 7 & 7 \\ 3 & 3 & 3 & 7 & 7 & 7 & 6 & 6 & 6 & 2 & 5 & 2 & 7 & 7 & 7 \\ 3 & 3 & 3 & 7 & 7 & 7 & 6 & 6 & 6 & 2 & 2 & 5 & 7 & 7 & 7 \\ 1 & 1 & 1 & 4 & 4 & 4 & 3 & 3 & 3 & 5 & 5 & 5 & 10 & 7 & 7 \\ 1 & 1 & 1 & 4 & 4 & 4 & 3 & 3 & 3 & 5 & 5 & 5 & 7 & 10 & 7 \\ 1 & 1 & 1 & 4 & 4 & 4 & 3 & 3 & 3 & 5 & 5 & 5 & 7 & 7 & 10 \end{pmatrix}$$

We kunnen in deze matrix weer blokken van 3 × 3 herkennen en wel zodanig, dat elk element uit matrix **G** is uitgegroeid tot een 3 × 3 matrix met negen keer datzelfde element, waarbij dan matrix **R** nog eens wordt opgeteld (vermeerdering van de elementen op de hoofddiagonaal met het cijfer drie).

Keuze 4: matrix **G** diagonaal en matrix **R** ongestructureerd.
Matrix **G** heeft allemaal zevens op de hoofddiagonaal en bevat verder nullen (5 × 5).

$$\boldsymbol{R} = \begin{pmatrix}
8 & 1 & 3 & 4 & 2 & 3 & 1 & 1 & 1 & 1 & 7 & 8 & 5 & 2 & 2 \\
3 & 8 & 3 & 3 & 2 & 2 & 8 & 7 & 9 & 0 & 1 & 1 & 2 & 1 & 1 \\
2 & 2 & 8 & 1 & 2 & 3 & 5 & 4 & 2 & 2 & 1 & 1 & 3 & 4 & 6 \\
4 & 3 & 6 & 8 & 3 & 3 & 4 & 5 & 2 & 5 & 6 & 1 & 1 & 2 & 3 \\
3 & 7 & 6 & 7 & 8 & 1 & 1 & 6 & 7 & 6 & 6 & 5 & 2 & 8 & 8 \\
3 & 7 & 7 & 6 & 3 & 8 & 4 & 4 & 5 & 6 & 7 & 3 & 7 & 4 & 2 \\
1 & 1 & 2 & 1 & 3 & 5 & 8 & 4 & 5 & 6 & 5 & 4 & 4 & 6 & 7 \\
3 & 4 & 2 & 3 & 4 & 6 & 7 & 8 & 5 & 5 & 4 & 3 & 2 & 2 & 1 \\
2 & 3 & 4 & 5 & 4 & 3 & 4 & 2 & 8 & 4 & 5 & 6 & 7 & 3 & 3 \\
5 & 5 & 5 & 3 & 6 & 6 & 6 & 4 & 3 & 8 & 6 & 5 & 5 & 7 & 5 \\
8 & 8 & 6 & 5 & 5 & 7 & 8 & 4 & 3 & 3 & 8 & 4 & 3 & 1 & 1 \\
5 & 4 & 4 & 5 & 7 & 3 & 2 & 1 & 1 & 2 & 1 & 8 & 3 & 4 & 7 \\
6 & 6 & 6 & 7 & 8 & 9 & 9 & 5 & 4 & 5 & 2 & 3 & 8 & 5 & 2 \\
4 & 3 & 4 & 4 & 2 & 2 & 3 & 1 & 5 & 5 & 4 & 3 & 2 & 8 & 2 \\
3 & 4 & 5 & 6 & 3 & 4 & 3 & 2 & 5 & 3 & 6 & 6 & 5 & 5 & 8
\end{pmatrix}$$

Uit deze **G** en **R** matrices krijgt men de volgende **V** matrix:

$$\boldsymbol{V} = \begin{pmatrix}
15 & 8 & 10 & 4 & 2 & 3 & 1 & 1 & 1 & 1 & 7 & 8 & 5 & 2 & 2 \\
10 & 15 & 10 & 3 & 2 & 2 & 8 & 7 & 9 & 0 & 1 & 1 & 2 & 1 & 1 \\
9 & 9 & 15 & 1 & 2 & 3 & 5 & 4 & 2 & 2 & 1 & 1 & 3 & 4 & 6 \\
4 & 3 & 6 & 15 & 10 & 10 & 4 & 5 & 2 & 5 & 6 & 1 & 1 & 2 & 3 \\
3 & 7 & 6 & 14 & 15 & 8 & 1 & 6 & 7 & 6 & 6 & 5 & 2 & 8 & 8 \\
3 & 7 & 7 & 13 & 10 & 15 & 4 & 4 & 5 & 6 & 7 & 3 & 7 & 4 & 2 \\
1 & 1 & 2 & 1 & 3 & 5 & 15 & 11 & 12 & 6 & 5 & 4 & 4 & 6 & 7 \\
3 & 4 & 2 & 3 & 4 & 6 & 14 & 15 & 12 & 5 & 4 & 3 & 2 & 2 & 1 \\
2 & 3 & 4 & 5 & 4 & 3 & 11 & 9 & 15 & 4 & 5 & 6 & 7 & 3 & 3 \\
5 & 5 & 5 & 3 & 6 & 6 & 6 & 4 & 3 & 15 & 13 & 12 & 5 & 7 & 5 \\
8 & 8 & 6 & 5 & 5 & 7 & 8 & 4 & 3 & 10 & 15 & 11 & 3 & 1 & 1 \\
5 & 4 & 4 & 5 & 7 & 3 & 2 & 1 & 1 & 9 & 8 & 15 & 3 & 4 & 7 \\
6 & 6 & 6 & 7 & 8 & 9 & 9 & 5 & 4 & 5 & 2 & 3 & 15 & 12 & 9 \\
4 & 3 & 4 & 4 & 2 & 2 & 3 & 1 & 5 & 5 & 4 & 3 & 9 & 15 & 9 \\
3 & 4 & 5 & 6 & 3 & 4 & 3 & 2 & 5 & 3 & 6 & 6 & 12 & 12 & 15
\end{pmatrix}$$

Deze matrix **V** wordt opgebouwd uit matrix **R** vermeerderd met een even grote matrix, die weer uit 3 × 3 blokken bestaat, waarbij elk blok negen gelijke elementen bevat overeenkomstig de elementen uit matrix **G**.

Aan deze twee laatste voorbeelden valt duidelijk af te lezen hoe matrix **V** afhangt van de matrices **G** en **R**. Overigens mag de afhankelijkheid van matrix **Z** niet uit het oog worden

verloren. Als matrix **Z** elementen groter dan één bevat dan groeit een element van matrix **G** uit tot een blok met verschillende elementen (n.l. produkt termen van **Z**). Bovendien bepalen het aantal gelijke rijen in de **Z** matrix de afmetingen van het overeenkomstige blok. Een theoretisch voorbeeld van dit laatste volgt hieronder:

Ga uit van de volgende matrices:

$$\boldsymbol{Z} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 2 \\ 0 & 0 & 2 \\ 0 & 0 & 2 \end{pmatrix} \qquad \boldsymbol{G} = \begin{pmatrix} 5 & 4 & 1 \\ 3 & 6 & 2 \\ 8 & 5 & 4 \end{pmatrix}$$

Dan is

$$\boldsymbol{ZGZ'} = \begin{pmatrix} 5 & 5 & 12 & 2 & 2 & 2 \\ 5 & 5 & 12 & 2 & 2 & 2 \\ 9 & 9 & 54 & 12 & 12 & 12 \\ 16 & 16 & 30 & 16 & 16 & 16 \\ 16 & 16 & 30 & 16 & 16 & 16 \\ 16 & 16 & 30 & 16 & 16 & 16 \end{pmatrix}$$

Deze matrix verschilt nog slechts met een optelling van matrix **R** van matrix **V**. Langs de hoofddiagonaal liggen drie blokken van verschillende afmetingen, n.l. een 2 × 2 blok (met vier elementen van 5), een 1 × 1 blok (met één element van 54) en een 3 × 3 blok (met 9 elementen van 16). De buiten de hoofddiagonaal liggende blokken zijn niet meer vierkant. De elementen binnen een blok van matrix **ZGZ'** zijn alleen verschillend als er per meting verschillende antwoorden mogelijk zijn (verschillende elementen in een kolom van **Z**). Een matrix **V** van het type AR(1) kan dus alleen ontstaan als matrix **R** precies de juiste 'correcties' geeft. Maar de enige manier om zeker te zijn van een bepaalde structuur van **V** is door alleen een **R** matrix met die structuur te definiëren en geen random effect in het model op te nemen.

Stel nu dat we een analyse willen doen op een databestand, dat gegevens bevat over herhaalde metingen van een bepaalde variabele bij een groep personen. We nemen aan dat metingen afkomstig van verschillende personen onafhankelijk van elkaar zijn en dat bij dezelfde persoon de herhaalde metingen een afnemende covariantie in de tijd vertonen. Dit wordt beschreven met een matrix **V**, die langs de hoofddiagonaal blokken heeft met een AR(1) structuur en buiten de diagonaal allemaal nullen. Hoe moeten we **G** en **R** kiezen om dit te verwezenlijken?
Het antwoord volgt al min of meer uit het bovenstaande. Er zijn twee mogelijkheden:
1) Geen random effect kiezen en voor matrix **R** een blokdiagonaal met AR(1) structuur nemen.
2) Matrix **G** diagonaal kiezen en voor matrix **R** een blokdiagonaal met AR(1) structuur. (Als we voor matrix **G** een AR(1) structuur kiezen, dan is matrix **V** namelijk niet meer blokdiagonaal!). Het verschil met de eerste mogelijke instelling is dat hier een deel van de te verklaren variantie door het random effect wordt verklaard en de rest door de error

term.
Hoe we bovengenoemde keuzen in PROC MIXED kunnen instellen wordt in hoofdstuk drie aangegeven.

2.4 *SCHATTING VAN DE PARAMETERS VIA REML.*

Het schatten van de parameters in een gemengd model met 'fixed' en 'random' variabelen c.q. effecten is een lastiger opgave dan in een model met alleen 'fixed' effecten. Dit komt doordat de introductie van 'random' variabelen tot gevolg heeft dat de var-covar matrix van **Y** niet meer diagonaal is. De oplossing volgens de kleinste kwadraten methode is dan niet meer toepasbaar. Wel is het mogelijk om voor de bepaling van de 'fixed' effecten de zgn. gewogen kleinste kwadraten methode toe te passen. Dit houdt in dat er gewichtsfactoren worden ingevoerd in de bekende 'normaal vergelijkingen' waarmee de var-covar matrix van **Y** als het ware wordt getransformeerd naar een diagonaal matrix. Het probleem kan daarna weer met de gewone kleinste kwadraten methode worden opgelost. Even een toelichting met formules:

LM: $\mathbf{Y} = \mathbf{X}\beta + \varepsilon$ met $E(\mathbf{Y}) = \mathbf{X}\beta$ en $var(\mathbf{Y}) = \sigma^2 \mathbf{I}$

Dan zijn de geschatte waarden van β gelijk aan:
$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}' \mathbf{Y}$ (de normaal vergelijkingen).

LMM: $\mathbf{Y} = \mathbf{X}\beta + \mathbf{Z}\nu + \varepsilon$ met $E(\mathbf{Y}) = \mathbf{X}\beta$ en $var(\mathbf{Y}) = \mathbf{V} = \mathbf{ZGZ}' + \mathbf{R}$

Voor wat de 'fixed' effecten betreft kan het LMM opgevat worden als een LM met een $var(\mathbf{Y}) = \mathbf{V}$ matrix, die niet meer diagonaal is en dit probleem is op te lossen met de gewogen kleinste kwadraten methode. Deze methode gaat als volgt:
De oorspronkelijke model vergelijking $\mathbf{Y} = \mathbf{X}\beta + \eta$ met $\eta = \mathbf{Z}\nu + \varepsilon$ wordt links en rechts vermenigvuldigd met de inverse van de symmetrische matrix **P**, waarvoor geldt $\mathbf{P}'\mathbf{P} = \mathbf{P}\mathbf{P} = \mathbf{V}$. Deze transformatie heeft tot gevolg dat er een nieuwe model vergelijking ontstaat $\mathbf{A} = \mathbf{Q}\beta + \tau$ (met $\mathbf{A} = \mathbf{P}^{-1}\mathbf{Y}$, $\mathbf{Q} = \mathbf{P}^{-1}\mathbf{X}$ en $\tau = \mathbf{P}^{-1}\eta$ zodanig dat var(**A**) wel een diagonaal matrix is. De gewone kleinste kwadraten methode geeft dan als oplossing:
$\mathbf{b} = (\mathbf{Q}'\mathbf{Q})^{-1} \mathbf{Q}' \mathbf{A}$ ofwel $\mathbf{b} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1} \mathbf{X}' \mathbf{V}^{-1}\mathbf{Y}$.

Het bovenstaande houdt in, dat de 'fixed' effecten zouden kunnen worden geschat met behulp van de gewogen kleinste kwadraten methode, waarbij als gewichtsfactoren de matrix $\mathbf{V}^{-1}$ moet worden genomen. Maar matrix $\mathbf{V} = \mathbf{ZGZ}' + \mathbf{R}$ is niet bekend, want **G** en **R** moeten hier ook geschat worden. Alleen in speciale gevallen waarin de matrix **V** een eenvoudige structuur heeft met slechts enkele parameters (zoals b.v. bij compound symmetrie) is het mogelijk om **V** uit de waarnemingen te schatten, waarna de 'fixed' effecten en vervolgens de 'random' effecten' kunnen worden gevonden. Een algemene oplossing voor het LMM wordt geboden door de maximum likelihood schattingsmethode (ML), maar dan meestal met restricties t.a.v. de te schatten parameters. Een deel van de parameters wordt 'constant' gehouden, zodanig dat het overige aantal parameters schatbaar is. Deze methode heet daarom ook de REML schattingsmethode (restricted maximum likelihood). Om dit verder te verduidelijken zullen we nu eerst de ML schattingsmethode

wat meer toelichten om vervolgens de REML schattingsmethode beter te kunnen behandelen.

De ML schattingsmethode voor de effecten in een GLM (generalized linear model) maakt gebruik van een iteratie proces. Begonnen wordt met het kiezen van startwaarden voor de te schatten parameters $\mathbf{b} = (b_1, \ldots\ldots\ldots, b_p)'$; noem deze waarde $\mathbf{b}_0$. Deze parameter waarden geven met behulp van de gegeven x-waarden schattingen voor de y-waarden, b.v. $\mathbf{y}_0$, die in het algemeen nog duidelijk zullen afwijken van de gemeten y-waarden, aangegeven met **y**. Dan wordt er een nieuwe waarde van **b** gezocht, noem deze $\mathbf{b}_1$, zodanig dat er een lineaire extrapolatie van $\mathbf{y}_0$ in de richting van de werkelijke waarde **y** wordt gedaan. Het vinden van deze nieuwe waarde $\mathbf{b}_1$ blijkt nu neer te komen op het toepassen van een gewogen kleinste kwadraten methode (ook weer op getransformeerde y-waarden). Bij de nieuwe $\mathbf{b}_1$ waarde hoort weer een nieuwe $\mathbf{y}_1$ waarde, die dichter bij de werkelijke waarde **y** zal liggen. Dit iteratie proces wordt voortgezet totdat de werkelijke waarde **y** voldoende dicht benaderd is.
Deze GLM schattingsmethode is in principe bruikbaar voor het vinden van de parameters in het LMM. Maar als er weinig structuur in de **V** matrix is, kan het aantal te schatten parameters wel erg groot worden (matrix **V** is immers van de orde $n \times n$!). Vaak zal het daarom nodig zijn om van de zgn. 'restricted maximum likelihood' (REML) uit te gaan. Deze methode brengt een reductie aan in het aantal te schatten parameters door een zodanige lineaire transformatie op het oorspronkelijke model uit te voeren, dat er loglikelihood functies overblijven met alleen de gewenste parameters erin. Bovendien wordt er voor gewaakt dat het verlies aan informatie zo klein mogelijk is. Ook hier houdt men weer modellen over , die met de gewogen kleinste kwadraten methode kunnen worden opgelost. Het is b.v. mogelijk om via een bepaalde transformatie de 'fixed' effecten uit het model te 'verwijderen' (de term $\mathbf{X}\beta$), waarna het mogelijk wordt om de 'random' effecten te schatten. Zijn deze bekend, dan kunnen vervolgens de 'fixed' effecten apart worden geschat, waarna het proces kan worden herhaald totdat een bevredigende oplossing is bereikt.

3. DE SAS PROCEDURE PROC MIXED.

3.1 *DE STATEMENTS MET HUN OPTIES.*

De MIXED procedure van SAS kent de volgende 'statements' :

```
PROC MIXED  <options>;
        BY  variables;
        CLASS  variables;
        ID  variables;
        MODEL  dependent = <fixed-effects> </ options >;
        RANDOM  random-effects </ options >;
        REPEATED  <repeated-effects> </ options>;
        PARMS (value) ...</ options>;
        CONTRAST  'label'  <fixed-effect values ...> </ random-effect values ...>,
        ... </ options>;
        ESTIMATE  'label'  <fixed-effect values ...> </ random-effect values ...>
        </ options>;
```

LSMEANS fixed-effects </ options>;
MAKE 'table' OUT = SAS-data-set;

Met PROC MIXED wordt de SAS procedure gestart en tevens kunnen een aantal opties worden opgegeven; we geven een overzicht van de opties:

Opties bij PROC MIXED:

alpha = number
vraagt om de schatting van een normaal betrouwbaarheidsinterval voor de covariantie parameter schattingen met een betrouwbaarheidsniveau van 1- α; de defaultwaarde voor number is 0.05.
asycov
vraagt om het printen van de asymptotische covariantie matrix van de covariantie parameters.
cl
vraagt naar de betrouwbaarheidsgrenzen van de covariantie parameters.
data = SAS-data-set
maxiter = number
specificeert het maximum aantal iteraties; de defaultwaarde is 50.
method = REML, ML of MIVQUE0
voor de keuze van de oplossingsmethode.
order = DATA, FORMATTED, FREQ of INTERNAL
voor het specificeren van de volgorde van de antwoordcategoriën van de discrete variabelen. (Toelichting: met DATA worden de categoriën gerangschikt in de volgorde, waarin ze bij de dataset worden aangetroffen. Met FORMATTED (tevens de defaultwaarde) wordt de volgorde van de categoriën bepaald door het externe format; dit komt in de praktijk neer op de alfabetische volgorde. Met FREQ wordt de categorie met de hoogste frequentie als eerste genomen etc. Met INTERNAL wordt de interne machine representatie gevolgd en dit zal vaak ook de alfabetische volgorde zijn).

Met het MODEL statement worden de afhankelijke variabele en de 'fixed-effect' variabelen (de X matrix) van het LMM opgegeven.
Met het RANDOM statement worden de random variabelen (de G matrix) opgegeven.
Met het REPEATED statement wordt de R matrix opgegeven.
Deze drie statements vormen de ruggegraat van PROC MIXED. Elk statement kan vergezeld gaan van een aantal nuttige opties. We bespreken er een aantal:

Opties bij MODEL:
alpha = number
vraagt om het construeren van een t-type betrouwbaarheidsinterval voor de 'fixed effect' parameters met betrouwbaarheidsniveau 1- α; de defaultwaarde is .05.
chisq
vraagt om het type III χ^2 test voor alle opgegeven effecten, naast de al gegeven type III F-test.
cl
vraagt om het t-type betrouwbaarheidsgrenzen voor alle 'fixed effect' parameters.

noint
vraagt om geen intercept in het model op te nemen; zonder deze optie wordt altijd een intercept in het model opgenomen.
predicted
vraagt om een tabel met o.a. de door het model voorspelde y-waarde per observatie.
solution
vraagt om de oplossing voor de 'fixed effecten' te printen.

Opties bij RANDOM:
alpha = number
vraagt om het construeren van een t-type betrouwbaarheidsinterval voor de 'fixed effect' parameters met betrouwbaarheidsniveau 1- α; de defaultwaarde is .05.
cl
vraagt om het t-type betrouwbaarheidsgrenzen voor alle 'fixed effect' parameters.
g
vraagt om de **G** matrix te printen.
gi
vraagt om de inverse van de **G** matrix te printen.
group = effect
definieert heterogeniteit in de **G** matrix d.m.v. een 'effect'; alle observaties met een zelfde waarde van het 'effect' hebben dezelfde covariantie parameters; alle observaties met een volgende waarde van 'effect' hebben onderling weer dezelfde covariantie parameters, maar deze verschillen van de eerste groep, etc. Een oorspronkelijk diagonale **G** matrix met b.v. 10 keer de waarde 23 op de hoofddiagonaal kan met behulp van de *group* optie veranderd worden in een diagonale matrix met op de hoofddiagonaal b.v. achtereenvolgens drie keer 12, vijf keer 18 en twee keer 22.
solution
vraagt om de oplossing voor de 'fixed effecten' te printen.
subject = effect
identificeert de subjecten in een dataset, waarbij een volledige onafhankelijkheid tussen de subjecten wordt verondersteld. De **G** matrix krijgt hierdoor een blokdiagonale structuur met identieke blokken. In feite wordt door een *subject* optie bereikt, dat alle onder de RANDOM statement opgegeven effecten genest worden binnen het *subject* effect.
type = sim, cs, un, un(q), ar(1), toep, toep(q) of sp(sptype)(coordinates)
hiermee wordt het type van de **G** matrix opgegeven, waarbij de afkortingen het volgende betekenen (zie voor voorbeelden SAS Technical Report P-229, blz 311-313):

sim	simple	1 parameter
cs	compound symmetry	1 of 2 parameters
un	unstructured	n(n+1)/2 parameters
un(q)	banded	(2n-q+1)q/2 parameter
ar(1)	autoregressive	1 of 2 parameters
sp(sph)(c)	spatial(spherical)	1 of 2 parameters
sp(pow)(c)	spatial(power)	1 of 2 parameters
sp(exp)(c)	spatial(exponential)	1 of 2 parameters
sp(gau)(c)	spatial(Gaussian)	1 of 2 parameters
sp(lin)(c)	spatial(linear)	1 of 2 parameters
sp(linl)(c)	spatial(linear log)	1 of 2 parameters

Opties bij REPEATED:

local

vraagt om het toevoegen van σ^2 **I** aan de **R** matrix; deze optie wordt gebruikt om een waarnemingsfout toe te voegen aan een tijdreeks structuur.

group = effect

definieert heterogeniteit in de **R** matrix d.m.v. een 'effect'; alle observaties met een zelfde waarde van het 'effect' hebben dezelfde covariantie parameters; alle observaties met een volgende waarde van 'effect' hebben onderling weer dezelfde covariantie parameters, maar deze verschillen van de eerste groep, etc. Een oorspronkelijk diagonale **R** matrix met b.v. 10 keer de waarde 23 op de hoofddiagonaal kan met behulp van de *group* optie veranderd worden in een diagonale matrix met op de hoofddiagonaal b.v. achtereenvolgens drie keer 12, vijf keer 18 en twee keer 22.

r

vraagt om de **R** matrix te printen.

ri

vraagt om de inverse van de **R** matrix te printen.

subject = effect

identificeert de subjecten in een dataset, waarbij een volledige onafhankelijkheid tussen de subjecten wordt verondersteld. De **R** matrix krijgt hierdoor een blokdiagonale structuur met identieke blokken. In feite wordt door een *subject* optie bereikt, dat alle onder de RANDOM statement opgegeven effecten genest worden binnen het *subject* effect.

type = sim, cs, un, un(q), ar(1), toep, toep(q) of sp(sptype)(coordinates)

hiermee wordt het type van de **R** matrix opgegeven, waarbij de afkortingen het volgende betekenen (zie voor voorbeelden SAS Technical Report P-229, blz 311-313):

sim	simple	1 parameter
cs	compound symmetry	1 of 2 parameters
un	unstructured	n(n+1)/2 parameters
un(q)	banded	(2n-q+1)q/2 parameter
ar(1)	autoregressive	1 of 2 parameters
sp(sph)(c)	spatial(spherical)	1 of 2 parameters
sp(pow)(c)	spatial(power)	1 of 2 parameters
sp(exp)(c)	spatial(exponential)	1 of 2 parameters
sp(gau)(c)	spatial(Gaussian)	1 of 2 parameters
sp(lin)(c)	spatial(linear)	1 of 2 parameters
sp(linl)(c)	spatial(linear log)	1 of 2 parameters

Naast de voorgaande drie kern statements van de MIXED procedure zijn er drie nuttige statements, die ook bij andere procedures vaak voorkomen, n.l.

BY variables	Dit leidt tot gescheiden analyses op groepen uit de dataset met gelijke waarden op de 'variables'.
CLASS variables	De bij dit statement opgegeven variabelen worden in de analyse opgevat als discrete variabelen.
ID variables	De bij dit statement opgegeven variabelen worden gebruikt voor identificatie van de afzonderlijke records bij het weergeven van voorspelde waarden in tabellen.

Dan is er het statement PARMS (value)...</options>, waarmee beginwaarden voor de covariantie parameters kunnen worden opgegeven. Dit zijn de waarden waarmee het iteratie proces voor de schatting van de parameters van start gaat. Er zijn twee opties bij PARMS n.l.

noiter
geeft aan dat er geen iteraties gedaan moeten worden, d.w.z. de startwaarden worden gebruikt als schattingen voor de parameters.

ratios
geeft aan dat ratio's t.o.v. de residuele variantie worden gespecificeerd in plaats van de covariantie parameters zelf.

Vervolgens zijn daar drie statements om hypothesen te toetsen, n.l. CONTRAST, ESTIMATE en LSMEANS. Met CONTRAST kunnen contrasten van alle parameters, 'fixed' of 'random', getest worden met gebruikmaking van de F-test. Er kan met een contrast matrix gewerkt worden, d.w.z. verschillende contrasten (de rijvectoren van de contrast matrix) kunnen gelijktijdig getest worden. Het ESTIMATE statement lijkt heel veel op het CONTRAST statement; ook hiermee worden contrasten gemeten, maar hier kan maar voor één contrast tegelijk getest worden. LSMEANS tenslotte geeft schattingen voor de gemiddelde waarden van de 'fixed effect' variabelen zoals ze voor een gebalanceerd design zouden zijn en gecorrigeerd voor de overige variabelen. De bijbehorende standaardvarianties zijn aangepast aan de aanwezige random effecten in het model. PROC MIXED geeft dus de juiste schattingen van de standaardvarianties en daarmee ook de juiste p-waarden, dit in tegenstelling tot de overeenkomstige schattingen bij PROC GLM.

Opties bij CONTRAST:

chisq
vraagt om ook een χ^2-test uit te voeren naast de F-test.

df = number
specificeert de vrijheidsgraden voor de noemer bij de F-test; de default waarde is N - rang(**X Z**), waarin N de steekproef grootte is.

e
vraagt om de **L** matrix te printen (dit is de matrix met alle parameters, 'fixed' en 'random')

singular = number
'number' is een getal tussen 0 en 1 (default 1E-4), dat de singulariteit bepaalt bij het testen van de schatbaarheid van een opgegeven contrast.

Opties bij ESTIMATE:

alpha = number
vraagt om een de constructie van t-type betrouwbaarheidsintervallen met betrouwbaarheidsniveau 1 - α; de default waarde is .05.

cl
vraagt om de constructie van t-type betrouwbaarheidsgrenzen; het betrouwbaarheidsniveau is default 0.95.

df = number
specificeert het aantal vrijheidsgraden voor de t-test; de default waarde is N - rang(**X Z**), waarin N de steekproef grootte is.

divisor = number
specificeert een waarde waardoor alle coëfficiënten gedeeld moeten worden, zodat gebroken getallen met gehele getallen kunnen worden opgegeven.

e
vraagt om de **L** matrix te printen (dit is de matrix met alle parameters, 'fixed' en 'random')
singular = number
'number' is een getal tussen 0 en 1 (default 1E-4), dat de singulariteit bepaalt bij het testen van de schatbaarheid van een opgegeven contrast.

Opties bij LSMEANS:
alpha = number
vraagt om een de constructie van t-type betrouwbaarheidsintervallen met betrouwbaarheidsniveau 1 - α; de default waarde is .05.
cl
vraagt om de constructie van t-type betrouwbaarheidsgrenzen; het betrouwbaarheidsniveau is default 0.95.
df = number
specificeert het aantal vrijheidsgraden voor de t-test; de default waarde is N - rang(**X Z**), waarin N de steekproef grootte is.
e
vraagt om de **L** matrix te printen (dit is de matrix met alle parameters, 'fixed' en 'random')
singular = number
'number' is een getal tussen 0 en 1 (default 1E-4), dat de singulariteit bepaalt bij het testen van de schatbaarheid van een opgegeven contrast.

Tenslotte rest nog het MAKE statement, waarmee het mogelijk is om elke tabel, die door PROC MIXED gemaakt wordt, weg te schrijven naar een SAS dataset zodat deze tabellen gebruikt kunnen worden in verdere analyses. Deze mogelijkheid is betrekkelijk nieuw binnen SAS en komt nog maar bij enkele procedures voor. De syntax is als volgt:

MAKE 'table' OUT = SAS dataset;

In 'table' kunnen een aantal vastgelegde namen worden opgegeven, n.l.

Table Name	Description
AsyCov	asymptotic covariance matrix of covariance parameters
ClassLevels	level information from the CLASS statement
Coefficients#	**L** matrix coefficients
Contrast	results from the CONTRAST statement(s)
CovParms	estimated covariance parameters
Estimate	results from the ESTIMATE statement(s)
Fitting	model fitting information
G	**G** matrix
GI	inverse of the **G** matrix
LSMeans	results from the LSMEANS statement(s)

ML	ML estimation iteration history
MMEQ	mixed model equations
MMEQSOL	mixed model equations solution
Parms	results from the PARMS statement
Predicted	predicted values
R	first block of the **R** matrix
REML	REML estimation iteration history
RI	inverse of the first block of the **R** matrix
SolutionF	fixed effects solution vector
SolutionR	random effects solution vector
Tests	tests of effects
XPVIX	$(\mathbf{X}\ \mathbf{y})' \mathbf{V}^{-1} (\mathbf{X}\ \mathbf{y})$
XPVIXI	$(\mathbf{X}' \mathbf{V}^{-1}\mathbf{X})^{-}$ with **y** border

De dataset, waar de gegevens naar toe geschreven moeten worden, wordt opgegeven met de optie OUT = SAS dataset. Voor uitgebreide toelichting zij verwezen naar de SAS OUTPUT procedure (SAS Technical Report P-229, blz 409 e.v.).

3.2 *TOELICHTING OP HET GEBRUIK VAN DE STATEMENTS.*

We zullen nu aan de hand van een eenvoudige dataset nagaan hoe de boven beschreven statements toegepast kunnen worden bij het analyseren van een gemengd model, aangegeven met de modelvergelijking $\mathbf{Y} = \mathbf{X}\beta + \mathbf{Z}\nu + \varepsilon$.

Ga uit van de volgende dataset:

```
data sp;
   input block a b y @@;
   cards;
   1 1 1  56  1 1 2  41
   1 2 1  50  1 2 2  36
   1 3 1  39  1 3 2  35
   2 1 1  30  2 1 2  25
   2 2 1  36  2 2 2  28
   2 3 1  33  2 3 2  30
   3 1 1  32  3 1 2  24
   3 2 1  31  3 2 2  27
   3 3 1  15  3 3 2  19
   4 1 1  30  4 1 2  25
   4 2 1  35  4 2 2  30
   4 3 1  17  4 3 2  18
;
```

3.2.1 De statements MODEL en RANDOM.

```
proc mixed data=sp;
   class a b block ;
   model y = a|b/solution;
   random block/g solution;
run;
```

Het MODEL statement definieert de variabelen a, b en a*b als 'fixed' effecten(de notatie a|b is een verkorte schrijfwijze voor het verzadigd model a b a*b).
Het RANDOM statement definieert de variabele block als 'random effect'. Er is geen type opgegeven voor de bijbehorende **G** matrix, zodat een simpele diagonaal matrix zal worden geschat. Hetzelfde geldt voor de niet opgegeven, maar altijd aanwezige **R** matrix (de 'error' matrix). Met de optie 'solution' in het MODEL statement en het RANDOM statement wordt een parameter schatting gevraagd. De optie 'g' in het RANDOM statement vraagt om een afdruk van de **G** matrix. De random variabele 'block' heeft vier antwoord-categoriën. De **G** matrix heeft dus 4 rijen en 4 kolommen met 4 gelijke waarden op de hoofddiagonaal en alle overige termen nul. De **R** matrix is 24 × 24 en heeft allemaal dezelfde waarden op de hoofddiagonaal en de overige termen nul. We kijken naar de output:

The SAS System

The MIXED Procedure

Class Level Information

Class	Levels	Values
A	3	1 2 3
B	2	1 2
BLOCK	4	1 2 3 4

The MIXED Procedure

REML Estimation Iteration History

Iteration	Evaluations	Objective	Criterion
0	1	106.73282502	
1	1	90.53560981	0.00000000

Convergence criteria met.

The MIXED Procedure

G Matrix

Parameter	Row	COL1	COL2	COL3	COL4
BLOCK 1	1	65.47222222			
BLOCK 2	2		65.47222222		
BLOCK 3	3			65.47222222	
BLOCK 4	4				65.47222222

The MIXED Procedure

Covariance Parameter Estimates (REML)

Cov Parm	Ratio	Estimate	Std Error	Z	Pr > |Z|
BLOCK	3.02179487	65.47222222	56.42171550	1.16	0.2459
Residual	1.00000000	21.66666667	7.91154805	2.74	0.0062

Uit de hierboven staande tabel kan de **G** matrix en de **R** matrix worden afgeleid. Vervolgens kan dan met kennis van de **Z** matrix (volgt uit de opgegeven data structuur) de var-covar matrix **V** van de y-waarden worden bepaald. We redeneren als volgt: de **G** matrix en de **R** matrix zijn beide diagonaal met gelijke waarden langs de hoofddiagonaal. Per matrix is dus slechts één parameter nodig en deze staat in de bovenstaande tabel. De variabele 'block' heeft vier antwoordcategoriën, dus matrix **G** is 4 × 4 met de waarden 65.47222222 op de hoofddiagonaal. Matrix **R** is 24 × 24 (aantal records) en heeft de waarden 21.66666667 op de hoofddiagonaal. Als de **G** of **R** matrix niet diagonaal is moeten er meer parameters per matrix geschat worden en verschijnen er dus ook meer waarden in de bovenstaande tabel. Met meer verschillende parameters wordt het lastiger om de structuur van de matrices uit deze tabel te bepalen en is het handig om van de g optie van het RANDOM statement of de 'r' optie van het REPEATED statement gebruik te maken. Is **G** bekend, dan moet vervolgens **ZGZ'** uitgerekend worden. Uit de hier gegeven dataset zien we dat er per categorie van de variabele 'block' zes records zijn; de matrix vermenigvuldiging **ZGZ'** heeft dan als resultaat, dat elk element uit **G** opgeblazen wordt tot een 6 × 6 matrix met 36 gelijke cellen. Het resultaat is dus een 24 × 24 matrix met vier blokken langs de hoofddiagonaal, waarbij elk blok weer een 6 × 6 matrix is met allemaal dezelfde celwaarden 65.47222222. Tellen we bij deze matrix de **R** matrix op dan worden dus uitsluitend de waarden op de hoofddiagonaal vermeerderd met 21.66666667. Hiermee is de gevraagde **V** matrix afgeleid. De gevonden structuur van **V** betekent, dat de 24 oorspronkelijke metingen zo samenhangen, dat steeds zes metingen onderling even sterk met elkaar correleren en helemaal niet met de metingen uit één van de andere drie blokken.

The MIXED Procedure

Model Fitting Information for Y

Description	Value
Observations	24.0000
Variance Estimate	21.6667
Standard Deviation Estimate	4.6547
REML Log Likelihood	-61.8087
Akaike's Information Criterion	-63.8087
Schwartz's Bayesian Criterion	-64.6991
-2 REML Log Likelihood	123.6174
Null Model LRT Chi-Square	16.1972
Null Model LRT DF	1.0000
Null Model LRT P-Value	0.0001

Betekenis van deze 'model fitting information':
Observations geeft het aantal bij de analyse gebruikte records.
Variance Estimate is een schatting van de residuele variantie σ^2en *Standard Deviation Estimate* is de wortel daaruit.
REML Log Likelihood; is dit een zeer groot negatief getal, dan wordt de **R** matrix door SAS als singulier beoordeeld en moeten alle verdere resultaten gewantrouwd worden. Als het ene covariantie model het submodel is van een ander model, dan kan men die modellen vergelijken met behulp van de likelihood ratio test (-2 maal het verschil tussen de log likelihoods). Deze ratio test volgt een χ^2 verdeling met als vrijheidsgraad het verschil tussen de aantal parameters in de twee modellen.
Akaike's Information Criterion (AIC) wordt gebruikt om modellen met dezelfde 'fixed effects' maar verschillende 'random effects' te vergelijken. Het model met het grootste AIC is het beste. Voor *Schwartz's Bayesian Criterion* (SBC) geldt hetzelfde als voor AIC.
Null Model LRT Chi-Square is de likelihood ratio test voor het gegeven model t.o.v. het

nulmodel, d.i. het model met alleen 'fixed effects' en dus zonder de 'random effects'.

De overige tabellen van deze output spreken voor zich.

The MIXED Procedure

Solution for Fixed Effects

Parameter	Estimate	Std Error	DDF	T	Pr > |T|
INTERCEPT	25.50000000	4.66741065	15	5.46	0.0001
A 1	3.25000000	3.29140294	15	0.99	0.3391
A 2	4.75000000	3.29140294	15	1.44	0.1695
A 3	0.00000000	.	.	.	.
B 1	0.50000000	3.29140294	15	0.15	0.8813
B 2	0.00000000	.	.	.	.
A*B 1 1	7.75000000	4.65474668	15	1.66	0.1167
A*B 1 2	0.00000000	.	.	.	.
A*B 2 1	7.25000000	4.65474668	15	1.56	0.1402
A*B 2 2	0.00000000	.	.	.	.
A*B 3 1	0.00000000	.	.	.	.
A*B 3 2	0.00000000	.	.	.	.

The MIXED Procedure

Solution for Random Effects

Parameter	Estimate	Std Error	DDF	T	Pr > |T|
BLOCK 1	11.29376089	4.35141579	15	2.60	0.0203
BLOCK 2	-0.55284144	4.35141579	15	-0.13	0.9006
BLOCK 3	-5.92330117	4.35141579	15	-1.36	0.1935
BLOCK 4	-4.81761828	4.35141579	15	-1.11	0.2857

The MIXED Procedure

Tests of Fixed Effects

Source	NDF	DDF	Type III F	Pr > F
A	2	15	7.54	0.0054
B	1	15	8.38	0.0111
A*B	2	15	1.74	0.2097

Het zal duidelijk zijn, dat bij wat uitgebreidere modellen en bij grote aantallen records de matrix **V** niet meer met de hand kan worden uitgerekend. Toch kan uit deze beschrijving van een eenvoudig model wel een zeker intuïtie worden verkregen voor wat er gebeurt bij ingewikkelder modellen. Vaak zal een blokstructuur van de matrices daarbij helpen.

Een belangrijke optie bij zowel het RANDOM statement als het REPEATED statement is de mogelijkheid om een blokstructuur op te leggen aan de **G** - en **R** matrix. Dit gaat met de optie 'subject'; hiermee wordt een variabele opgegeven, waarvan de antwoordcategoriën de verschillende blokken bepalen. Alle records met een zelfde waarde van de opgegeven variabele behoren tot hetzelfde blok. Er wordt in PROC MIXED van uit gegaan, dat er tussen de blokken geen covariantie bestaat, dus alle elementen van de **G** - en **R** matrix, die buiten de blokken vallen zijn nul; de blokken zijn gegroepeerd langs de hoofddiagonaal. Hierdoor worden de berekeningen met de matrices aanzienlijk efficiënter. Het hierboven gegeven programma kan b.v. ook als volgt worden geformuleerd:

```
proc mixed data=sp;
   class a b block;
   model y = a|b/ solution;
   random intercept / subject=block g solution;
run;
```

Intercept is hier een SAS variabele met voor ieder record de waarde 1. Met RANDOM worden dus een variabele opgegeven, intercept, die 'geblokt' wordt door de variabele 'block'. Dit komt er op neer dat in feite de variabele 'intercept*block' als random variabele wordt gedefinieerd. Aangezien 'intercept*block' hetzelfde oplevert als 'block' staat er in het laatste programma dus hetzelfde als in het eerste programma. De naamgeving in de tabel 'Covariance Parameter Estimates' is op het eerste gezicht wat verwarrend, maar zal met deze toelichting duidelijk zijn.

The MIXED Procedure

Covariance Parameter Estimates (REML)

Cov Parm	Ratio	Estimate	Std Error	Z	Pr > |Z|
INTERCEPT	3.02179487	65.47222222	56.42171550	1.16	0.2459
Residual	1.00000000	21.66666667	7.91154805	2.74	0.0062

In deze tabel moet dus voor 'intercept' gelezen worden 'block'. Residual geeft als gewoonlijk de 'error' aan en staat in **R**. Nog een voorbeeld maar dan met twee random variabelen: Het programma:

```
proc mixed data=sp;
   class a b block;
   model y = a|b/ solution;
   random block a*block/ g solution;
run;
```

kan ook geschreven worden als:

```
proc mixed data=sp;
   class a b block;
   model y = a|b/ solution;
   random intercept a / subject=block g solution;
run;
```

In de Covariance Parameter Estimates (REML) tabel komt dan onder Cov Parm te staan: INTERCEPT, A en Residual. Deze parameters moeten dan gelezen worden als 'block', 'a*block' en 'Residual'.

3.2.2 De statements MODEL en REPEATED.

De volgende opgave, die we ons stellen, is hoe het oorspronkelijke programma geschreven kan worden met behulp van het REPEATED statement en zonder het RANDOM statement. Dit komt er op neer, dat matrix **R** nu gelijk is aan matrix **V**. Matrix **V** was een 24 × 24 matrix met vier gelijke blokken langs de hoofddiagonaal. Elke blok is een 6 × 6 matrix met de getallen 87.14 op de hoofddiagonaal en overal elders de getallen 65.47 (deze structuur heet 'compound symmetry'). Om met het statement REPEATED een blok diagonale matrix te definiëren moet gebruik worden gemaakt van de optie *subject*. Als we geen *subject* optie gebruiken wordt door de procedure verondersteld, dat elk record een

blok is. Dit betekent dat er van een zuiver diagonale matrix (alle termen buiten de diagonaal nul) wordt uitgegaan, hetgeen meestal niet gewenst zal zijn. Hier moeten we voor *subject* = 'block' kiezen, want block heeft vier waarden en zes records per blok, zodat er vier blokken van 6 × 6 langs de hoofddiagonaal ontstaan. Maar de blokken moeten op zich ook nog een 'compound symmetry' hebben, dus moet ook nog worden opgegeven *type* = cs. Omdat verder per blok de records steeds in dezelfde volgorde voor wat betreft de variabelen a en b staan, hoeft er bij REPEATED geen variabele worden opgegeven. Het programma gaat er dus als volgt uitzien:

```
proc mixed data=sp;
   class a b block ;
   model y = a|b/ solution;
   repeated / subject = block r type=cs;
run;
```

De output van dit programma is als volgt:

The SAS System

The MIXED Procedure

Class Level Information

Class	Levels	Values
A	3	1 2 3
B	2	1 2
BLOCK	4	1 2 3 4

The MIXED Procedure

REML Estimation Iteration History

Iteration	Evaluations	Objective	Criterion
0	1	106.73282502	
1	1	90.53560981	0.00000000

Convergence criteria met.

The MIXED Procedure

R Matrix for BLOCK 1

Row	COL1	COL2	COL3	COL4	COL5
1	87.13888889	65.47222222	65.47222222	65.47222222	65.47222222
2	65.47222222	87.13888889	65.47222222	65.47222222	65.47222222
3	65.47222222	65.47222222	87.13888889	65.47222222	65.47222222
4	65.47222222	65.47222222	65.47222222	87.13888889	65.47222222
5	65.47222222	65.47222222	65.47222222	65.47222222	87.13888889
6	65.47222222	65.47222222	65.47222222	65.47222222	65.47222222

The MIXED Procedure

R Matrix for BLOCK 1

COL6
65.47222222

65.47222222
65.47222222
65.47222222
65.47222222
87.13888889

The SAS System

The MIXED Procedure

Covariance Parameter Estimates (REML)

Cov Parm	Ratio	Estimate	Std Error	Z	Pr > |Z|
DIAG CS	3.02179487	65.47222222	56.42171550	1.16	0.2459
Residual	1.00000000	21.66666667	7.91154805	2.74	0.0062

The MIXED Procedure

Model Fitting Information for Y

Description	Value
Observations	24.0000
Variance Estimate	21.6667
Standard Deviation Estimate	4.6547
REML Log Likelihood	-61.8087
Akaike's Information Criterion	-63.8087
Schwartz's Bayesian Criterion	-64.6991
-2 REML Log Likelihood	123.6174
Null Model LRT Chi-Square	16.1972
Null Model LRT DF	1.0000
Null Model LRT P-Value	0.0001

The MIXED Procedure

Solution for Fixed Effects

Parameter	Estimate	Std Error	DDF	T	Pr > |T|
INTERCEPT	25.50000000	4.66741065	18	5.46	0.0000
A 1	3.25000000	3.29140294	18	0.99	0.3365
A 2	4.75000000	3.29140294	18	1.44	0.1662
A 3	0.00000000	.	.	.	.
B 1	0.50000000	3.29140294	18	0.15	0.8809
B 2	0.00000000	.	.	.	.
A*B 1 1	7.75000000	4.65474668	18	1.66	0.1132
A*B 1 2	0.00000000	.	.	.	.
A*B 2 1	7.25000000	4.65474668	18	1.56	0.1367
A*B 2 2	0.00000000	.	.	.	.
A*B 3 1	0.00000000	.	.	.	.
A*B 3 2	0.00000000	.	.	.	.

The MIXED Procedure

Tests of Fixed Effects

Source	NDF	DDF	Type III F	Pr > F
A	2	18	7.54	0.0042
B	1	18	8.38	0.0097
A*B	2	18	1.74	0.2044

De output is vrijwel gelijk aan het programma met het RANDOM statement. Omdat hier geen variabele was opgegeven bij het REPEATED statement staat in de 'covariance

parameter estimates' tabel de naam 'diag'. Bij het testen van de 'fixed effects' zien we dat het aantal vrijheidsgraden voor de noemer (DDF) 18 is, terwijl dat bij gebruik van het RANDOM statement 15 was. Dit komt doordat het aantal vrijheidsgraden van de 'random effects' variabele (hier 'block' met drie vrijheidsgraden) meegeteld wordt met het aantal vrijheidsgraden van de 'fixed effects' (hier 6). Bij gebruik van het RANDOM statement wordt DDF dus 24 - 6 - 3 = 15 en bij gebruik van het REPEATED statement 24 - 6 = 18 (24 is het aantal records). Het aantal vrijheidsgraden voor de teller (NDF) is direct gelijk aan het aantal vrijheidsgraden van het betreffende effect.
De vraag komt nu naar boven voor welk statement (RANDOM of REPEATED) het best kan worden gekozen in een bepaalde situatie. Het antwoord daarop is dat het in rekenkundig opzicht efficiënter zal zijn om een variabele met veel antwoordcategoriën bij het RANDOM statement te gebruiken, omdat de **Z** matrix veel kleiner is dan de **R** matrix.

3.2.3 Het CONTRAST statement.

Met het CONTRAST statement kunnen zelf gedefinieerde contrasten (lineaire combinaties van de in het model gedefinieerde effecten, waarbij de coëfficiënten tot nul optellen) ingesteld worden. Zowel 'fixed effects' als 'random effects' zijn toegestaan, maar ze worden op een verschillende manier toegepast. Het contrast werkt in feite op de 'fixed effects', terwijl de 'random effects' het referentie kader vormen t.o. waarvan getest gaat worden. Als er geen 'random effects' worden opgegeven, dan vormt de totale populatie, waaruit de 'random effects' worden geselecteerd, de referentie en we spreken dan over de brede referentie ruimte. Worden er wel 'random effects' opgegeven, dan vormen de waargenomen 'random effects' het referentie kader en we spreken over de smalle referentie ruimte. Er kunnen ook meerdere contrasten tegelijk getest worden, d.w.z. het contrast kan als een contrastmatrix worden opgegeven. Er volgt weer een voorbeeld met de reeds eerder gebruikte dataset.

```
proc mixed data=sp;
   class a b block ;
   model y = a|b/ solution;
   random block/g solution;
   contrast 'a broad'
            a 1 -1  0 a*b .5 .5 -.5 -.5   0   0,
            a 1  0 -1 a*b .5 .5   0   0 -.5 -.5;
   contrast 'b broad'
            b 1 -1 a*b .3333333 -.3333333 .3333333 -.3333333 .3333333
-.3333333;
   contrast 'a*b broad'
            a*b 1 -1 -1  1  0  0,
            a*b 1 -1  0  0 -1  1;
run;
```

We geven een toelichting op het CONTRAST statement. De syntax is als volgt:

CONTRAST 'label' <fixed-effect values...> < | random-effect values...> , ... < / options> ;

Het 'label' is verplicht en bestaat uit een tekst van hoogstens 20 karakters tussen enkele quotes. Op de plaats van Fixed-effect en random-effect kunnen de variabelen worden genoemd, die respectievelijk bij het MODEL statement en het RANDOM statement zijn opgegeven. De waarden achter de effecten bepalen het gewenste contrast. In het eerstgenoemde contrast van het bovenstaande voorbeeld moet het effect van a worden getest

tegen de brede referentie ruimte (dus geen 'fixed efects'). Dit gebeurt door het eerste niveau van a tegen het tweede niveau van a te testen (met voor de overige 'fixed effects' b en a*b gemiddelde waarden) en tevens het eerste niveau van a tegen het derde niveau van a. De waarden moeten per effect tot nul optellen, want het gaat immers om contrasten. We kiezen dus voor a de drie waarden van a 1 -1 0; b heeft twee waarden, zodat voor het gemiddelde 0.5 moet worden gekozen; a*b heeft dan de waarden a1b1, a1b2, a2b1, a2b2, a3b1 en a3b2 ofwel .5 .5 -.5 -.5 0 0. Het effect b hoeft zelf niet in het contrast worden opgenomen. Op analoge wijze wordt voor het tweede contrast gekozen: a 1 0 -1 a*b .5 .5 0 0 -.5 -.5. Het totale contrast moet hetzelfde opleveren als de waarden in de tabel 'Tests of Fixed Effects'.
Bij het contrast 'b broad' moet voor b worden gekozen 1 -1 (een andere mogelijkheid is er niet met de twee niveaus van b, dus hier is maar één contrast nodig); voor a nemen de gemiddelde waarde 1/3 (a heeft immers drie niveaus), zodat a*b (de b index loopt het hardst omdat dit de tweede variabele in het CLASS statement is) dan de volgende waarden heeft: 1/3 -1/3 1/3 -1/3 1/3 -1/3.
Het contrast 'a*b broad' tenslotte alleen waarden voor a*b (en niet voor a en b afzonderlijk). We moeten hier de waarden 1 en -1 kiezen, omdat niet wordt gemiddeld. Nemen voor a 1 -1 0 in gedachten en voor b 1 -1, dan worden de waarden voor a*b 1 -1 -1 1 0 0. Voor het tweede contrast nemen we voor a 1 0 -1 en voor b 1 -1, zodat a*b dan de waarden 1 -1 0 0 -1 1 aanneemt.
We vergelijken nu deze drie contrasten met de resultaten van de tabel 'Test of Fixed Effects' in de volgende programma output:

The MIXED Procedure

Tests of Fixed Effects

Source	NDF	DDF	Type III F	Pr > F
A	2	15	7.54	0.0054
B	1	15	8.38	0.0111
A*B	2	15	1.74	0.2097

The MIXED Procedure

CONTRAST Statement Results

Source	NDF	DDF	F	Pr > F
a broad	2	15	7.54	0.0054
b broad	1	15	8.38	0.0111
a*b broad	2	15	1.74	0.2097

We zien dat de resultaten inderdaad hetzelfde zijn.
Vervolgens geven we nog een voorbeeld van een contrast met een smalle referentie ruimte. We kiezen hiervoor weer dezelfde dataset, maar met een iets uitgebreider model:

```
options nodate nonumber;
proc mixed data=sp;
   class a b block ;
   model y = a|b/solution;
   random block a*block/g solution;
   contrast 'a narrow'
      a 1 -1 0  a*b .5 .5 -.5 -.5 0 0|
      a*block .25 .25 .25 .25 -.25 -.25 -.25 -.25 0 0 0 0,
      a 1 0 -1  a*b .5 .5 0 0 -.5 -.5|
      a*block .25 .25 .25 .25 0 0 0 0 -.25 -.25 -.25 -.25/df=6;
```

```
run;
```

In de output vergelijken we de waarden van de tabel 'Test of Fixed Effects' met de resultaten van de contrast test voor 'a narrow':

The MIXED Procedure

Tests of Fixed Effects

Source	NDF	DDF	Type III F	Pr > F
A	2	6	4.07	0.0764
B	1	9	19.39	0.0017
A*B	2	9	4.02	0.0566

The MIXED Procedure

CONTRAST Statement Results

Source	NDF	DDF	F	Pr > F
a narrow	2	6	17.44	0.0032

In de tabel 'Tests of Fixed Effects' is de test gelijk aan die voor een brede referentie ruimte. De test voor 'a narrow' heeft duidelijk een hogere F-waarde, zoals te verwachten was. Maar ondanks die hoge F-waarde zal in de praktijk het contrast met de brede referentie ruimte het meest geschikt zijn.

3.2.4 Het ESTIMATE statement.

Het ESTIMATE statement lijkt in veel opzichten op het CONTRAST statement. De syntax is hetzelfde en de opties komen grotendeels overeen. Maar er zijn toch ook verschillen te noemen. Zo is het bij het ESTIMATE statement alleen mogelijk om één effect tegelijk te schatten, dus geen 'contrast' matrix , maar een 'contrast' rij vector. Bovendien hoeven de coëfficiënten niet tot nul te sommeren, zoals bij een echt contrast. De significantie wordt nu met een t-test gedaan i.p.v. een F-test. Er volgt een voorbeeld met de al eerder gebruikte dataset. Van a wordt de gemiddelde waarde van het eerste niveau geschat en wel tegen een smalle referentie ruimte, een wat bredere referentie ruimte en een brede referentie ruimte.

```
proc mixed data=sp;
   class a b block ;
   model y = a|b;
   random block a*block;
   estimate 'a1 mean narrow'
            intercept 1
            a 1
            b .5 .5
            a*b .5  .5
          | block .25 .25 .25 .25
            a*block .25 .25 .25 .25;

   estimate 'a1 mean intermediate'
            intercept 1
            a 1
            b .5 .5
            a*b .5  .5
```

```
          | block .25 .25 .25 .25;

   estimate 'a1 mean broad'
            intercept 1
            a 1
            b .5 .5
            a*b .5 .5;
run;
```

De waarden voor de effecten worden op dezelfde manier toegekend als bij het CONTRAST statement, alle niet genoemde parameters zijn nul. Dus in het eerste 'estimate' statement staat:

intercept 1 a 1 0 0 b .5 .5 a*b .5 .5 0 0 0 0 |
block .25 .25 .25 .25 a*block .25 .25 .25 .25 0 0 0 0 0 0 0 0 0;

Let op dat de coëfficiënten hier niet tot nul optellen!
Het uitkomsten van dit programma zijn:

The MIXED Procedure

Tests of Fixed Effects

Source	NDF	DDF	Type III F	Pr > F
A	2	6	4.07	0.0764
B	1	9	19.39	0.0017
A*B	2	9	4.02	0.0566

The MIXED Procedure

ESTIMATE Statement Results

Parameter	Estimate	Std Error	DDF	T	Pr > |T|
a1 mean narrow	32.87500000	1.08172958	9	30.39	0.0000
a1 mean intermediate	32.87500000	2.23955911	9	14.68	0.0000
a1 mean broad	32.87500000	4.54032855	9	7.24	0.0000

Om de grote overeenkomst tussen CONTRAST en ESTIMATE te tonen, kiezen we het volgende programma:

```
options nodate nonumber;
proc mixed data=sp;
   class a b block ;
   model y = a|b;
   random block a*block;
   contrast 'b mean broad'
            b 1 -1 a*b .33333333 -.33333333 .33333333 -.33333333
            .33333333 -.33333333/e;
   estimate 'b1 mean broad'
            b 1 -1 a*b .33333333 -.33333333 .33333333 -.33333333
            .3333333 -.33333333/e;
run;
```

Let op! Het CONTRAST statement moet het eerst genoemd worden. Bij de statements is nog de optie 'e' meegegeven om een overzicht van de gekozen coëfficiënten te krijgen. De uitkomst is als volgt:

The MIXED Procedure

Coefficients for b mean broad

Parameter	Row 1
INTERCEPT	0
A 1	0
A 2	0
A 3	0
B 1	1
B 2	-1
A*B 1 1	0.33333333
A*B 1 2	-0.33333333
A*B 2 1	0.33333333
A*B 2 2	-0.33333333
A*B 3 1	0.33333333
A*B 3 2	-0.33333333

The MIXED Procedure

Coefficients for b1 mean broad

Parameter	Row 1
INTERCEPT	0
A 1	0
A 2	0
A 3	0
B 1	1
B 2	-1
A*B 1 1	0.33333333
A*B 1 2	-0.33333333
A*B 2 1	0.33333333
A*B 2 2	-0.33333333
A*B 3 1	0.3333333
A*B 3 2	-0.33333333

The SAS System

The MIXED Procedure

ESTIMATE Statement Results

Parameter	Estimate	Std Error	DDF	T	Pr > |T|
b1 mean broad	5.49999995	1.24907373	9	4.40	0.0017

The MIXED Procedure

CONTRAST Statement Results

Source	NDF	DDF	F	Pr > F
b mean broad	1	9	19.39	0.0017

De coëfficiënten voor de beide statements zijn hetzelfde en ook de significantie testen komen redelijk overeen, aangezien $\sqrt{19.39} \approx 4.40$.

3.2.5 Het LSMEANS statement.

Het statement LSMEANS geeft schattingen voor de gemiddelde waarden van alle antwoordcategoriën van de opgegeven variabelen ('fixed effects'). De schattingen zijn dan gecorrigeerd voor de aanwezigheid van alle andere 'fixed effects' uit het model en bij het

testen wordt uitgegaan van de brede referentie ruimte. Dit is allemaal ook met het ESTIMATE statement in te stellen, maar bij LSMEANS hoeft dan niet nagedacht te worden over de op te geven waarden voor de coëfficiënten. Als men alleen in de schattingen voor de gemiddelde waarden geïnteresseerd is, dan is het LSMEANS statement dus veel handiger. Ter toelichting weer een voorbeeld:

```
proc mixed data=sp;
   class a b block ;
   model y = a|b;
   random block a*block;
   estimate 'a1 mean broad'
            intercept 1
            a 1
            b .5 .5
            a*b .5 .5;
   estimate 'a2 mean broad'
            intercept 1
            a 0 1
            b .5 .5
            a*b 0 0 .5 .5;
   estimate 'a3 mean broad'
            intercept 1
            a 0 0 1
            b .5 .5
            a*b 0 0 0 0 .5 .5;
   estimate 'b1 mean broad'
            intercept 1
            a .33333333 .33333333 .33333333
            b 1
            a*b .33333333 0 .33333333 0 .33333333 0;
   estimate 'b2 mean broad'
            intercept 1
            a .33333333 .33333333 .33333333
            b 0 1
            a*b 0 .33333333 0 .33333333 0 .33333333;
   lsmeans a b;
run;
```

We vergelijken de uitkomsten van bovenstaande ESTIMATE statements met die voor LSMEANS:

The MIXED Procedure

ESTIMATE Statement Results

Parameter	Estimate	Std Error	DDF	T	Pr > |T|
a1 mean broad	32.87500000	4.54032855	9	7.24	0.0000
a2 mean broad	34.12500000	4.54032855	9	7.52	0.0000
a3 mean broad	25.75000000	4.54032855	9	5.67	0.0003
b1 mean broad	33.66666659	4.20248494	9	8.01	0.0000
b2 mean broad	28.16666664	4.20248494	9	6.70	0.0001

The MIXED Procedure

Least Squares Means

Level	LSMEAN	Std Error	DDF	T	Pr > |T|
A 1	32.87500000	4.54032855	9	7.24	0.0000
A 2	34.12500000	4.54032855	9	7.52	0.0000
A 3	25.75000000	4.54032855	9	5.67	0.0003
B 1	33.66666667	4.20248494	9	8.01	0.0000
B 2	28.16666667	4.20248494	9	6.70	0.0001

De uitkomsten zijn inderdaad gelijk.

3.2.6 Het MAKE statement.

Met het MAKE statement is het mogelijk geworden om tabellen, die in de uitvoer geprint worden op te slaan in SAS systeem files. De kolommen in de tabel worden variabelen in deze output systeem files. Dit biedt de mogelijkheid om de resultaten van een mixed model analyse te gebruiken bij eventuele verdere analyses. Er zijn twee verschillende manieren om de output systeem files te produceren, n.l.

1) via het MAKE statement in PROC MIXED of
2) via het ODS (Output Delivery System) en de procedure OUTPUT.

Bij methode 1) maken we gebruik van het MAKE statement. De syntax hiervan is simpel:

```
MAKE  'table'  out = SAS dataset;
```

Voor 'table' kan één van de omschrijvingen, die al hiervoor in een tabel zijn opgenomen, gebruikt worden. Is het gewenst om meer dan één van de output tabellen naar een systeem file te schrijven, dan moet dat met meer dan één MAKE statements in hetzelfde programma. Bij elk MAKE statement moet ook een aparte output systeem file opgegeven worden. De geproduceerde output files kunnen naderhand worden bekeken met b.v. PROC PRINT.

Bij methode 2) moeten we vóórdat het SAS programma gestart wordt, eerst twee zgn. globale macro variabelen definiëren, als volgt:

```
%global _DISK_;
%let _DISK_ = ON;
%global _PRINT_;
%let _PRINT_ = ON;  ( OFF, als geen print uitvoer naar scherm gewenst is)
```

Deze regels moeten gewoon 'gesubmit' worden met de PROGRAM EDITOR en blijven dan geldig voor de duur van de sessie.
Vervolgens wordt een programma met PROC MIXED uitgevoerd, zonder dat daarin MAKE statements hoeven worden opgenomen. Is dit programma klaar, dan kan de procedure PROC OUTPUT worden 'gesubmit', als volgt:

```
proc output;
run;
```

Er verschijnt een window, waarin een lijst staat met alle beschikbare output tabellen, behorend bij de zo juist uitgevoerde analyse. Met de muis kan een van de tabellen worden uitgekozen. Na het indrukken van de rechter muisknop wordt er een pop-up menu geactiveerd, waarin een aantal bewerkingen staan, die op de uitgekozen tabel kan worden toegepast. Kiezen we b.v. 'make', dan kan een naam van een output file worden opgegeven. Kiezen we 'print' dan wordt de uitgekozen tabel op het OUTPUT window getoond. Is de gevraagde actie uitgevoerd, dan gaat de cursor weer terug naar de oorspronkelijke lijst met output tabellen en kan er weer een andere tabel worden uitgekozen.

4. ENKELE UITGEWERKTE VOORBEELDEN.

4.1 *HERHAALDE METINGEN.*

De data van dit voorbeeld zijn afkomstig van Pothoff en Roy (1964) en bestaan uit groeimetingen bij 11 meisjes en 16 jongens; per persoon wordt op vier verschillende tijdstippen gemeten, n.l. op de leeftijd van 8, 10, 12 en 14 jaar. Jennrich en Schluchter (1986) analyseren deze data met gebruikmaking van verschillende covariantie structuren. Het SAS programma ziet er als volgt uit:

```
data pr;
   input person sex$ y1 y2 y3 y4;
   y=y1; age=8;  output;
   y=y2; age=10; output;
   y=y3; age=12; output;
   y=y4; age=14; output;
   drop y1-y4;
   cards;
   1    F   21.0    20.0    21.5    23.0
   2    F   21.0    21.5    24.0    25.5
   3    F   20.5    24.0    24.5    26.0
   4    F   23.5    24.5    25.0    26.5
   5    F   21.5    23.0    22.5    23.5
   6    F   20.0    21.0    21.0    22.5
   7    F   21.5    22.5    23.0    25.0
   8    F   23.0    23.0    23.5    24.0
   9    F   20.0    21.0    22.0    21.5
  10    F   16.5    19.0    19.0    19.5
  11    F   24.5    25.0    28.0    28.0
  12    M   26.0    25.0    29.0    31.0
  13    M   21.5    22.5    23.0    26.5
  14    M   23.0    22.5    24.0    27.5
  15    M   25.5    27.5    26.5    27.0
  16    M   20.0    23.5    22.5    26.0
  17    M   24.5    25.5    27.0    28.5
  18    M   22.0    22.0    24.5    26.5
  19    M   24.0    21.5    24.5    25.5
  20    M   23.0    20.5    31.0    26.0
  21    M   27.5    28.0    31.0    31.5
  22    M   23.0    23.0    23.5    25.0
  23    M   21.5    23.5    24.0    28.0
  24    M   17.0    24.5    26.0    29.5
  25    M   22.5    25.5    25.5    26.0
  26    M   23.0    24.5    26.0    30.0
  27    M   22.0    21.5    23.5    25.0
;
```

Het data bestand 'work.pr' bestaat uit vier records per persoon met de variabelen person (persoonsnummer), sex (geslacht), y (groeimeting) en age (leeftijd); in totaal zijn er 108 records. De vraag, die nu onderzocht kan worden, is in hoeverre de gemeten groei afhankelijk is van de leeftijd en of deze afhankelijkheid verschillend is voor meisjes of jongens. Het antwoord op deze vraag wordt bemoeilijkt, doordat de metingen niet allemaal onafhankelijk zijn. De metingen op de vier verschillende tijdstippen zijn aan elkaar gecorreleerd doordat bij dezelfde personen gemeten wordt. De var-covar matrix van de y-waarden geeft daardoor per vier metingen (van één persoon) covarianties te zien. Bij het formuleren van het model (in hoeverre hangt y af van sex en age) moet dus tevens de

juiste var-covar matrix opgesteld worden. Dit kan met PROC MIXED op twee manieren, n.l. via het REPEATED statement (moduleren van de R-matrix) of via het RANDOM statement (moduleren van de G-matrix). We beginnen met het volgende SAS programma:

```
proc mixed data=pr method=ml scoring;
  class person sex;
  model y = sex age(sex) / noint s;
  repeated / type=un sub=person r;
run;
```

Toelichting:
De gekozen oplossingsmethode is maximum likelihood met gebruikmaking van de Fisher's scoring methode. De variabelen 'person' en 'sex' worden als discrete variabelen opgevat. In het model worden hier opgenomen 'sex' als een hoofdeffect en 'age' als een binnen de variabele 'sex' genest effect. Nesting gedraagt zich hetzelfde als interactie met dit verschil, dat het geneste effect nooit tevens als hoofdeffect zal voorkomen. De variabele 'sex' heeft twee niveau's en omdat bij het MODEL statement als optie 'noint' is meegegeven zullen er twee effecten worden geschat, die de gemiddelde y-waarde voor elk niveau van 'sex' weergeven. De variabele 'age' wordt opgevat als een continue variabele, zodat dan één effect wordt geschat, dat de helling aangeeft van de regressielijn van 'y' op 'age'. En omdat hier 'age(sex)' is gekozen worden er voor meisjes en jongens apart een helling geschat. De optie 's' zorgt er voor dat de gevraagde effecten ook worden afgedrukt.
Vervolgens moet het model gecorrigeerd worden voor de afhankelijkheid in de herhaalde metingen aan personen. In dit voorbeeld is gekozen voor het REPEATED statement. Het effect, dat hiermee kan worden opgegeven moet voor de verschillende records binnen ieder subject een andere waarde hebben. Zijn de records zo gerangschikt, dat per subject de volgorde van het repeated effect dezelfde is (zoals hier het geval is, want het repeated effect is in dit voorbeeld 'age'), dan hoeft er geen effect te worden opgegeven. Met de optie 'subject' wordt vervolgens aangegeven welke records bij één persoon (of een andere meetgrootheid) horen. Wordt er geen subject variabele opgegeven, dan gaat de procedure er van uit dat elk record voor één aparte persoon geldt. Dit leidt tot een diagonale R matrix. Wordt er wel een subject variabele opgegeven, dan geeft dit aanleiding tot het 'blokken' van de R matrix. In het hier gegeven voorbeeld is 'person' de subject variabele; er zijn 27 personen, dus de R matrix krijgt hiermee 27 blokken, langs de hoofddiagonaal, met elk blok bestaande uit een 4 × 4 matrix. De subject variabele vertegenwoordigt het random effect, de repeated variabele (die overigens altijd discreet moet zijn) geeft de verschillende meet tijdstippen.
Met de optie type=un wordt aangegeven, dat er een ongestructureerde R matrix moet worden geschat. Aangezien de R matrix in dit geval bestaat uit 27 gelijke blokken van 4 × 4 symmetrische matrices, moeten er in totaal 10 parameters geschat worden.
We bekijken nu eerst de output, die door het opgegeven programma wordt geproduceerd:

```
                    The MIXED Procedure

                  Class Level Information

Class      Levels  Values

PERSON         27  1 2 3 4 5 6 7 8 9 10 11 12 13
                   14 15 16 17 18 19 20 21 22 23
                   24 25 26 27
SEX             2  F M
```

The MIXED Procedure

ML Estimation Iteration History

Iteration	Evaluations	Objective	Criterion
0	1	279.75103669	
1	2	220.98663078	0.00000248
2	1	220.98632822	0.00000003
3	1	220.98632494	0.00000000

Scoring stopped after iteration 1.

Convergence criteria met.

The MIXED Procedure

R Matrix for PERSON 1

Row	COL1	COL2	COL3	COL4
1	5.11919889	2.44090159	3.61051034	2.52224364
2	2.44090159	3.92794750	2.71751364	3.06234944
3	3.61051034	2.71751364	5.97979819	3.82346068
4	2.52224364	3.06234944	3.82346068	4.61798404

The MIXED Procedure

Covariance Parameter Estimates (MLE)

Cov Parm		Estimate	Std Error	Z	Pr > |Z|
DIAG	UN(1,1)	5.11919889	1.41686560	3.61	0.0003
	UN(2,1)	2.44090159	0.98353460	2.48	0.0131
	UN(2,2)	3.92794750	1.08245774	3.63	0.0003
	UN(3,1)	3.61051034	1.27665783	2.83	0.0047
	UN(3,2)	2.71751364	1.07398146	2.53	0.0114
	UN(3,3)	5.97979819	1.62787950	3.67	0.0002
	UN(4,1)	2.52224364	1.06485905	2.37	0.0179
	UN(4,2)	3.06234944	1.01346595	3.02	0.0025
	UN(4,3)	3.82346068	1.25076893	3.06	0.0022
	UN(4,4)	4.61798404	1.25732801	3.67	0.0002
Residual		1.00000013	.	.	.

The MIXED Procedure

Model Fitting Information for Y

Description	Value
Observations	108.0000
Variance Estimate	1.0000
Standard Deviation Estimate	1.0000
Log Likelihood	-209.739
Akaike's Information Criterion	-219.739
Schwartz's Bayesian Criterion	-233.149
-2 Log Likelihood	419.4770
Null Model LRT Chi-Square	58.7647
Null Model LRT DF	9.0000
Null Model LRT P-Value	0.0000

The MIXED Procedure

Solution for Fixed Effects

Parameter	Estimate	Std Error	DDF	T	Pr > |T|
SEX F	17.42536849	1.12838047	104	15.44	0.0000
SEX M	15.84228933	0.93560366	104	16.93	0.0000
AGE(SEX) F	0.47636470	0.09541519	104	4.99	0.0000
AGE(SEX) M	0.82680330	0.07911409	104	10.45	0.0000

The MIXED Procedure

Tests of Fixed Effects

Source	NDF	DDF	Type III F	Pr > F
SEX	2	104	262.60	0.0000
AGE(SEX)	2	104	67.07	0.0000

De oplossingen voor de vaste effecten laten zien, dat voor meisjes de beginwaarde van de groei variabele hoger is dan die voor jongens, maar dat de groeisnelheid van de meisjes bijna de helft is van die voor jongens.
De oplossingen voor de R matrix laten zien, dat de niet diagonale effecten variëren van 2.4 tot 3.8 en de diagonale effecten van 3.9 tot 5.9; deze variaties zijn niet zo groot en het is dus de moeite waard om de wat eenvoudiger compound symmetrie te proberen. Het vorige programma behoeft slechts een kleine wijziging om dit uit te voeren; type=un moet veranderd worden in type=cs.

```
proc mixed data=pr method=ml scoring;
  class person sex;
  model y = sex age(sex) / noint s;
  repeated / type=cs sub=person r;
run;
```

De output die dit programma oplevert en voor zover verschillend van de vorige, is als volgt:

ML Estimation Iteration History

Iteration	Evaluations	Objective	Criterion
0	1	279.75103669	
1	1	230.14833485	0.00000000

Scoring stopped after iteration 1.

Convergence criteria met.

The MIXED Procedure

R Matrix for PERSON 1

Row	COL1	COL2	COL3	COL4
1	4.90515835	3.03056169	3.03056169	3.03056169
2	3.03056169	4.90515835	3.03056169	3.03056169
3	3.03056169	3.03056169	4.90515835	3.03056169
4	3.03056169	3.03056169	3.03056169	4.90515835

The MIXED Procedure

Covariance Parameter Estimates (MLE)

Cov Parm	Ratio	Estimate	Std Error	Z	Pr > |Z|

DIAG CS	1.61664733	3.03056169	0.95520746	3.17	0.0015
Residual	1.00000000	1.87459666	0.29456445	6.36	0.0000

The MIXED Procedure

Model Fitting Information for Y

Description	Value
Observations	108.0000
Variance Estimate	1.8746
Standard Deviation Estimate	1.3692
Log Likelihood	-214.320
Akaike's Information Criterion	-216.320
Schwartz's Bayesian Criterion	-219.002
-2 Log Likelihood	428.6391
Null Model LRT Chi-Square	49.6027
Null Model LRT DF	1.0000
Null Model LRT P-Value	0.0000

The MIXED Procedure

Solution for Fixed Effects

Parameter	Estimate	Std Error	DDF	T	Pr > |T|
SEX F	17.37272727	1.16152410	104	14.96	0.0000
SEX M	16.34062500	0.96308491	104	16.97	0.0000
AGE(SEX) F	0.47954545	0.09230869	104	5.20	0.0000
AGE(SEX) M	0.78437500	0.07653832	104	10.25	0.0000

The MIXED Procedure

Tests of Fixed Effects

Source	NDF	DDF	Type III F	Pr > F
SEX	2	104	255.79	0.0000
AGE(SEX)	2	104	66.01	0.0000

The SAS System

We zien dat met de invoering van een compound symmetrie de R matrix twee verschillende waarden bevat, n.l. 4.905 op de hoofddiagonaal en 3.031 overal daarbuiten. De AIC (Akaike's Information Criterion) score voor dit model is -216.320, terwijl de AIC score voor het vorige model met ongestructureerde R matrix -219.739 bedroeg. Omdat in beide gevallen de 'fixed' effecten dezelfde zijn mag geconcludeerd worden , dat het model met de hoogste AIC score het beste is. Dat is hier dus het model met een compound symmetrie in de R matrix. De schattingen voor de 'fixed' effecten verschillen nauwelijks van de waarden in het vorige model.

Alvorens nu verder te gaan met de analyse zullen we nu eerst laten zien dat het voorgaande programma met een compound symmetrie R matrix ook opgesteld kan worden met een RANDOM statement. Zoals al eerder is opgemerkt vertegenwoordigt de variabele 'person' het random effect, dus we kunnen het volgende programma proberen:

```
proc mixed data=pr method=ml scoring;
   class person sex;
   model y = sex age(sex) / noint s;
   random person;
run;
```

Dit programma produceert de volgende output:

The MIXED Procedure

Class Level Information

Class	Levels	Values
PERSON	27	1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27
SEX	2	F M

The MIXED Procedure

ML Estimation Iteration History

Iteration	Evaluations	Objective	Criterion
0	1	279.75103669	
1	1	230.14833485	0.00000000

Scoring stopped after iteration 1.

Convergence criteria met.

The MIXED Procedure

Covariance Parameter Estimates (MLE)

Cov Parm	Ratio	Estimate	Std Error	Z	Pr > |Z|
PERSON	1.61664733	3.03056169	0.95520746	3.17	0.0015
Residual	1.00000000	1.87459666	0.29456445	6.36	0.0000

The MIXED Procedure

Model Fitting Information for Y

Description	Value
Observations	108.0000
Variance Estimate	1.8746
Standard Deviation Estimate	1.3692
Log Likelihood	-214.320
Akaike's Information Criterion	-216.320
Schwartz's Bayesian Criterion	-219.002
-2 Log Likelihood	428.6391
Null Model LRT Chi-Square	49.6027
Null Model LRT DF	1.0000
Null Model LRT P-Value	0.0000

The MIXED Procedure

Solution for Fixed Effects

Parameter	Estimate	Std Error	DDF	T	Pr > |T|
SEX F	17.37272727	1.16152410	79	14.96	0.0000
SEX M	16.34062500	0.96308491	79	16.97	0.0000
AGE(SEX) F	0.47954545	0.09230869	79	5.20	0.0000
AGE(SEX) M	0.78437500	0.07653832	79	10.25	0.0000

The MIXED Procedure

Tests of Fixed Effects

Source	NDF	DDF	Type III F	Pr > F
SEX	2	79	255.79	0.0000
AGE(SEX)	2	79	66.01	0.0000

Het RANDOM statement wijst de variabele 'person' als een random effect aan. Deze variabele heeft 27 discrete waarden (27 persoonsnummers). Er wordt dus een symmetrische G matrix gevormd van 27 × 27 en wel diagonaal, want dat is het default type. Alle waarden langs die hoofddiagonaal zijn gelijk. De bijdrage aan de var-covar matrix van de y-waarden wordt gegeven door het matrixprodukt ZGZ', waarbij de Z matrix hier een 108 × 27 matrix is met steeds vier dezelfde rijen per persoon. Zo'n matrix produkt resulteert in een 108 × 108 matrix met 27 gelijke blokken langs de hoofddiagonaal en in elk 4 × 4 blok 16 gelijke waarden (zoals in paragraaf 2.3 is beschreven). Als hier dan nog de diagonale R matrix wordt opgeteld ontstaat weer een totale var-covar met 27 blokken langs de hoofddiagonaal en elk 4 × 4 blok met een compound symmetrie, precies gelijk aan de var-covar matrix bij het programma met het REPEATED statement en type = cs. De resultaten zijn ook precies gelijk aan die van de vorige output. Omdat de G matrix hier zo groot is hebben we maar geen optie meegegeven om de G matrix af te drukken.

Er bestaat nog een derde manier om de compound symmetrie structuur te definiëren, n.l. als volgt:

```
proc mixed data=pr method=ml scoring;
   class person sex;
   model y = sex age(sex) / noint s;
   repeated intercept diag / sub=person r;
run;
```

Zowel 'intercept' als 'diag' zijn hier speciale sleutelwoorden van SAS, die er voor zorgen dat onafhankelijke random effecten (opgegeven met sub =) met een gemeenschappelijke covariantie en een verhoging van de waarden langs de hoofddiagonaal zullen worden geschat.

We gaan weer verder met de analyse. Met de compound symmetrie R matrix hoeven slechts twee waarden geschat te worden. Het zou echter kunnen zijn dat de varianties en covarianties anders zijn voor jongens dan voor meisjes. Om dit te kunnen onderzoeken bestaat er bij het REPEATED en het RANDOM statement een optie 'group ='. Per niveau van de met 'group' op te geven variabele worden er waarden geschat van de covariantie parameters. Hieronder volgen programma en output:

```
proc mixed data=pr method=ml scoring;
   class person sex;
   model y = sex age(sex) / noint s;
   repeated intercept diag / sub=person group=sex;
run;
```

The SAS System

The MIXED Procedure

Class Level Information

Class	Levels	Values

PERSON	27	1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27
SEX	2	F M

The MIXED Procedure

ML Estimation Iteration History

Iteration	Evaluations	Objective	Criterion
0	1	279.75103669	
1	1	210.32224911	0.00000000

Scoring stopped after iteration 1.

Convergence criteria met.

The MIXED Procedure

Covariance Parameter Estimates (MLE)

Cov Parm	Estimate	Std Error	Z	Pr > |Z|
INTERCEPT SEX F	3.88038912	1.71788299	2.26	0.0239
INTERCEPT SEX M	2.44630534	1.11754866	2.19	0.0286
DIAG SEX F	0.59001377	0.14525135	4.06	0.0000
DIAG SEX M	2.75774740	0.56292283	4.90	0.0000
Residual	1.00000000	.	.	.

The MIXED Procedure

Model Fitting Information for Y

Description	Value
Observations	108.0000
Variance Estimate	1.0000
Standard Deviation Estimate	1.0000
Log Likelihood	-204.406
Akaike's Information Criterion	-208.406
Schwartz's Bayesian Criterion	-213.771
-2 Log Likelihood	408.8130
Null Model LRT Chi-Square	69.4288
Null Model LRT DF	3.0000
Null Model LRT P-Value	0.0000

The SAS System

The MIXED Procedure

Solution for Fixed Effects

Parameter	Estimate	Std Error	DDF	T	Pr > |T|
SEX F	17.37272727	0.83107137	104	20.90	0.0000
SEX M	16.34062500	1.11299466	104	14.68	0.0000
AGE(SEX) F	0.47954545	0.05178688	104	9.26	0.0000
AGE(SEX) M	0.78437500	0.09283297	104	8.45	0.0000

The MIXED Procedure

Tests of Fixed Effects

Source	NDF	DDF	Type III F	Pr > F
SEX	2	104	326.26	0.0000
AGE(SEX)	2	104	78.57	0.0000

Uit de tabel van de covariantie parameter schattingen blijkt, dat er duidelijke verschillen zijn tussen de covarianties bij meisjes of bij jongens. De AIC score en de SBC score voor dit model (-208.4 en -213.8) zijn allebei hoger dan die voor het vorige model (-216.3 en -219.0). Bovendien is het verschil in de -2log likelihood scores voor beide modellen (428.64 - 408.81 = 19.83) met een verschil van twee vrijheidsgraden ook zeer significant ($p < 0.0001$). Het model met de 'group' optie fit beter dan zonder 'group' optie. De 'fixed' effecten blijven in beide gevallen dezelfde, alleen de standaard errors veranderen.

Hiermee sluiten we deze analyse af. Maar er is met PROC MIXED nog meer mogelijk dan de gegeven analyses. In de nu volgende paragraaf zullen we het zgn. random coëfficiënten model aan de orde stellen.

4.2 *RANDOM COËFFICIËNTEN.*

We gebruiken hetzelfde databestand als in paragraaf 4.1 en ook nu weer zijn we geïnteresseerd in de betrekking tussen de leeftijd en de groeivariabele voor meisjes en jongens. De personen worden als een random steekproef gedacht uit een populatie, zodat de variabele 'person' (persoonsnummer) als een random variabele wordt beschouwd. Door herhaalde metingen aan dezelfde personen wordt een afhankelijkheid tussen waarnemingen van de groeivariabele geïntroduceerd. Deze afhankelijkheid komt tot uitdrukking in een var-covar matrix van y-waarden, die nu ook niet-diagonaal elementen ongelijk nul heeft. In paragraaf 4.1 hebben we een blokdiagonale matrix bekeken met compound symmetrie voor de blokken, wat erop neer kwam dat tussen alle herhaalde metingen eenzelfde covariantie wordt verondersteld. We schrijven nog even het desbetreffende programma op:

```
proc mixed data=pr method=ml scoring;
   class person sex;
   model y = sex age(sex) / noint s;
   random intercept / type=sim sub=person g;
run;
```

In paragraaf 4.1 staat in feite "random person", maar dit is hetzelfde als hierboven staat, want "random intercept/ sub=person" is gelijk aan "random intercept*person" en dit is weer gelijk aan "random person". Dit model beschrijft, zoals reeds was uitgelegd, de regressie van y (groeivariabele) op 'age' en dat voor meisjes en jongens apart. Als we ons zouden beperken tot alleen het MODEL statement (dus zonder het RANDOM statement) dan hebben te maken met een klassiek regressie model, waarbij twee regressie lijnen worden berekend, ieder met een eigen intercept en een eigen richtingscoëfficiënt (β-waarde). Door het toevoegen van het bovenstaand RANDOM statement wordt dit model uitgebreid tot een stelsel van in dit geval 27 regressie lijnen, met dezelfde richtingscoëfficiënt voor jongens, maar allemaal verschillende intercepts, en een andere richtingscoëfficiënt voor meisjes, ook met verschillende intercepts.

Het RANDOM COËFFICIËNTEN model, gaat nu nog een stap verder en veronderstelt, dat ook alle richtingscoëfficiënten per persoon verschillend zullen zijn. Hiermee wordt een tweede random effect in het model gehaald. Omdat 'age' hier de variabele is, waarvan de

coëfficiënt gelijk is aan de te schatten β-waarde, wordt deze variabele aan de RANDOM statement toegevoegd. De variabele 'age' komt daardoor twee keer in het programma voor, een keer in het MODEL statement als 'fixed' effect en een keer in het RANDOM statement als 'random effect'! Dit model beschrijft dan dus een stelsel van 27 regressie lijnen met onderling verschillende intercepts en verschillende richtingscoëfficiënten. Het programma is als volgt:

```
proc mixed data=pr method=ml scoring;
   class person sex;
   model y = sex age(sex) / noint s;
   random intercept age / type=un sub=person;
run;
```

Intercept is een variabele met één kolom in de G matrix en de continue variabele 'age' heeft eveneens maar één kolom, zodat hier een 2 × 2 matrix gedefinieerd wordt, maar wel per persoon. De totale G matrix bestaat dus uit 54 rijen en kolommen met 27 blokken van gelijke 2 × 2 matrices langs de hoofddiagonaal. Er is hier gekozen voor een ongestructureerde matrix, zodat het element un(1,1) de variantie schatting voor de intercepts geeft, het element un(2,2) de variantie schatting voor de richtingscoëfficiënten en het element un(2,1) de covariantie schatting daartussen. Dit laatste betekent dat een verandering in de richtingscoëfficiënt ook een verandering in de intercept tot gevolg heeft. Toegespitst op het gebruikte voorbeeld kan dit gelezen worden als: de groeisnelheid houdt verband met de aanvangswaarde van de groeivariabele. We geven nu eerst de resultaten van het bovenstaande programma:

The SAS System

The MIXED Procedure

Class Level Information

Class	Levels	Values
PERSON	27	1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27
SEX	2	F M

The MIXED Procedure

ML Estimation Iteration History

Iteration	Evaluations	Objective	Criterion
0	1	279.75103669	
1	1	229.31522763	0.00000000

Scoring stopped after iteration 1.

Convergence criteria met.

The MIXED Procedure

G Matrix

Parameter	Subject	Row	COL1	COL2
INTERCEPT	PERSON 1	1	4.55691340	-0.19825389
AGE	PERSON 1	2	-0.19825389	0.02375894

The MIXED Procedure

Covariance Parameter Estimates (MLE)

Cov Parm		Ratio	Estimate	Std Error	Z	Pr > |Z|
INTERCEPT	UN(1,1)	2.65522874	4.55691340	4.67187514	0.98	0.3294
	UN(2,1)	-0.11551886	-0.19825389	0.37905734	-0.52	0.6010
	UN(2,2)	0.01384389	0.02375894	0.03408822	0.70	0.4858
Residual		1.00000000	1.71620370	0.33028356	5.20	0.0000

The MIXED Procedure

Model Fitting Information for Y

Description	Value
Observations	108.0000
Variance Estimate	1.7162
Standard Deviation Estimate	1.3100
Log Likelihood	-213.903
Akaike's Information Criterion	-217.903
Schwartz's Bayesian Criterion	-223.267
-2 Log Likelihood	427.8060
Null Model LRT Chi-Square	50.4358
Null Model LRT DF	3.0000
Null Model LRT P-Value	0.0000

The MIXED Procedure

Solution for Fixed Effects

Parameter	Estimate	Std Error	DDF	T	Pr > |T|
SEX F	17.37272727	1.18202433	54	14.70	0.0000
SEX M	16.34062500	0.98008280	54	16.67	0.0000
AGE(SEX) F	0.47954545	0.09980396	54	4.80	0.0000
AGE(SEX) M	0.78437500	0.08275307	54	9.48	0.0000

The MIXED Procedure

Tests of Fixed Effects

Source	NDF	DDF	Type III F	Pr > F
SEX	2	54	247.00	0.0000
AGE(SEX)	2	54	56.46	0.0000

De schattingen voor de covariantie parameters laten wel variantie zien voor de intercepts (un(1,1) = 4.557), maar nauwelijks voor de richtingscoëfficiënten (un(2,2) = 0.024). De correlatie tussen beide schattingen kunnen we uitrekenen door de covariantie (un(2,1) = -0.198) te delen door de standaarddeviaties van de beide schattingen. De berekening gaat

dus als volgt: $-0.198/(\sqrt{4.557} \times \sqrt{0.024}) = -0.60$. Er blijkt dus wel sprake te zijn van enige correlatie tussen intercepts en richtingscoëfficiënten, wat hier ook wel te verwachten was. Als het draaipunt van de rechte namelijk ver naar links ligt, zoals hier het geval is (leeftijd nul, groeivariabele nul), dan zal een verandering in de helling ook een duidelijke verandering in de intercept (waarde van groeivariabele bij aanvang onderzoek) te zien geven. Maar veel belangrijker dan het vinden van deze correlatie is de constatering , dat geen van de geschatte parameters voor de var-covar matrix significant zijn. Kijken we naar de informatie over de model fitting, dan zien we dat ten opzichte van het ongestructureerde model wel sprake is van enige verbetering, maar niet ten opzichte van het compound symmetrie model (afgekort cs model). De log-likelihood (LL) score -213.903 verschilt niet significant van de LL score -214.320 voor het cs model. De AIC en SBC scores van het random coëfficiënten model (resp. -217.903 en -223.267) zijn zelfs lager dan die voor het cs model (resp. -216.320 en -219.002). De conclusie is dat het random coëfficiënten model voor de gegeven data geen verbetering geeft ten opzichte van het cs model. De varianties in de richtingscoëfficiënten zijn hier kennelijk niet van belang, d.w.z. ongeacht de grootte van de groeivariabele bij acht jarigen is de groeisnelheid voor meisjes (van acht tot veertien jaar) onderling vrijwel gelijk en ook voor jongens (van acht tot veertien) onderling gelijk, maar de groeisnelheid is voor jongens wel duidelijk groter dan voor meisjes.

Tot zo ver het random coëfficiënten model. Dit was een voorbeeld van een nieuwe analyse mogelijkheid, die door PROC MIXED gegeven wordt.

==========