KM 22(1986)
pag 35 - 59

ANALYSE VAN EEN MIXED MODEL

Bas Engel(*), Marion van den Bol(*), Pieter Vereijken(**)

0 INLEIDING

In dit artikel wordt de analyse van een (ongebalanceerd) mixed model beschreven. De nadruk ligt vooral op het schatten van de variantiecomponenten met behulp van de volgende methoden:
Henderson methode III, MINQUE, MIVQUE, REML en ML. De bedoeling van dit artikel is om een beknopt en eenvoudig overzicht te geven van deze methoden en hun onderlinge samenhang. Daarnaast wordt aandacht besteed aan het schatten van de fixed effecten en het voorspellen van de randomeffecten. Voor uitvoerige informatie over de vele details die onbesproken blijven verwijzen we naar Harville (1975, 1977) en Searle (1979). Ter illustratie van de methoden is een dataset uit Dempster et al. (1984) geanalyseerd. De voor deze illustratie ontwikkelde computerprogramma's zijn geschreven in GENSTAT en opgenomen in Engel et al. (1985).

(*) ITI-TNO
Afdeling Statistiek
Postbus 214
2600 AE Delft
tel.: 015-569330

(**) ITI-TNO
Afdeling Statistiek-Landbouw
Postbus 100
6700 AC Wageningen
tel.: 08370-19100

1 HET MIXED MODEL

Het mixed model voor een vector $\vec{y}$ van waarnemingen formuleren we als volgt:

$$\vec{y} = X\vec{\alpha} + Z_1\vec{b}_1 + \ldots + Z_c\vec{b}_c + \vec{e} \qquad (1)$$

De vector $\vec{\alpha}$ bevat de onbekende fixed effecten. De vectoren $\vec{b}_1, \ldots \vec{b}_c$ bevatten toevalsbijdragen van c verschillende bronnen van variatie, de zogenaamde random effecten . De vector $\vec{e}$ is de gebruikelijke vector van restbijdragen. De matrices X, $Z_1, \ldots, Z_c$ zijn bekend. Vaak veronderstellen we dat $\vec{b}_1, \ldots, \vec{b}_c, \vec{e}$ onafhankelijk zijn met

$$\vec{b}_i \sim N(\vec{0}, I\sigma_i^2) \qquad i = 1, 2, \ldots, c \qquad (2)$$

$$\vec{e} \sim N(\vec{0}, I\sigma_e^2)$$

De varianties $\sigma_e^2, \sigma_1^2, \ldots, \sigma_c^2$ zijn de variantiecomponenten in het model. Het kan zijn dat het ons voornamelijk om de fixed effecten te doen is. Dit is bijvoorbeeld het geval wanneer $\vec{\alpha}$ contrasten tussen verschillende behandelingen bevat en de randomeffecten $\vec{b}_1, \ldots, \vec{b}_c$ zijn opgenomen om (positieve) correlaties tussen de waarnemingen te modelleren. Het kan ook zijn dat $\vec{\alpha}$ effecten bevat die mogelijk een rol spelen en die daarom in het model zijn opgenomen, maar dat het ons meer om voorspellingen voor de randomeffecten gaat.

Stel de vector $\vec{y}$ bevat melkproduktiegegevens van dochters van verschillende stiervaders. De genetische bijdragen van de stiervaders aan hun nakomelingen zijn opgeborgen in $\vec{b}_1$. De vector $\vec{\alpha}$ bevat termen die verschillen tussen bedrijven, jaren en seizoenen vertegenwoordigen, de zogenaamde HYS- (heard-year-season)effecten. Op basis van een voorspelling voor $\vec{b}_1$ selecteren we de beste stieren ten aanzien van melkproduktie voor verdere opfok om zo in de loop der jaren het produktieniveau van de populatie te verhogen.

Soms zijn $\vec{b}_1, \ldots, \vec{b}_c$ zelf niet interessant maar willen we weten hoe de variatiebronnen ten opzichte van elkaar en ten opzichte van de restbijdragen aan de totale variatie in de waarnemingen bijdragen. In dat geval zijn we voornamelijk geïnteresseerd in de variantiecomponenten $\sigma_e^2, \sigma_1^2, \ldots, \sigma_c^2$.

Stel de vector $\vec{y}$ bevat spekdiktemetingen verricht met een meetpistool gehanteerd door verschillende personen aan aselect gekozen karkassen van varkens. We beschouwen deze personen als representatief voor een populatie van personen en representeren de persoonsbijdragen tot de waarnemingen door de vector $\vec{b}_1$. Verrichten we bovendien herhaalde waarnemingen aan hetzelfde karkas, dan representeren we de karkasbijdragen door de vector $\vec{b}_2$.

Waarnemingen verzameld door dezelfde persoon aan hetzelfde karkas hebben de persoons- en karkasbijdrage gemeen. De correlatie tussen twee van die waarnemingen is $(\sigma_1^2 + \sigma_2^2)/(\sigma_0^2 + \sigma_1^2 + \sigma_2^2)$. Waarnemingen van verschillende personen aan hetzelfde karkas hebben alleen de karkasbijdrage gemeen, de correlatie is dan $\sigma_2^2/(\sigma_0^2 + \sigma_1^2 + \sigma_2^2)$. In dit geval zouden we graag zien dat σ_2^2 groot is t.o.v. σ_0^2 en σ_1^2. Immers, in dat geval lijken metingen aan hetzelfde karkas veel op elkaar. Begrippen die in deze context vaak gehanteerd worden zijn "herhaalbaarheid" (repeatability) en "reproduceerbaarheid" (reproducibility).

Tenslotte nog enige notatie. De matrices Z_i en de vectoren $\vec{b}_i$ vatten we samen in de matrix Z en de vector $\vec{b}$:

$$Z = (Z_1, Z_2, \ldots, Z_c) \quad \text{en} \quad \vec{b}' = (\vec{b}_1', \vec{b}_2', \ldots, \vec{b}_c').$$

Soms zullen we geen onderscheid tussen $\vec{b}_1, \ldots, \vec{b}_c$ en $\vec{e}$ maken. In dat geval is $Z_0 = I$ en $\vec{b}_0 = \vec{e}$ en nummeren we van 0 tot c in plaats van 1 tot c:

$$\dot{Z} = (Z_0, Z_1, \ldots, Z_c) \text{ en } \dot{\vec{b}}' = (\vec{b}_0', \vec{b}_1', \ldots, \vec{b}_c').$$

We nemen aan dat $\vec{y}$ lengte N heeft, $\vec{\alpha}$ lengte p, $\vec{b}_i$ lengte q_i, $\vec{b}$ lengte q en dat rang (X) = p.
We kunnen het model dus ook op de volgende twee manieren formuleren:

$$\vec{y} = X\vec{\alpha} + Z\vec{b} + \vec{e} \quad \text{of} \quad \vec{y} = X\vec{\alpha} + \dot{Z}\dot{\vec{b}} .$$

De variantiecomponenten $\sigma_0^2, \sigma_1^2, \ldots, \sigma_c^2$ nemen we samen in de vector

$$\vec{\sigma}^2 = (\sigma_0^2, \sigma_1^2, \ldots, \sigma_c^2)'.$$

2 DE MIXED MODEL VERGELIJKINGEN

In de volgende paragrafen schetsen we verschillende methoden voor het schatten van de variantiecomponenten. In deze paragraaf gaat onze interesse uit naar de vectoren $\vec{\alpha}$, $\vec{b}$ en $\vec{e}$ van fixed effecten, randomeffecten en restbijdragen. In eerste instantie nemen we aan dat de variantiecomponenten, en dus de varianties en covarianties in $V = \text{var}(\vec{y})$, bekend zijn. Later komen we hierop terug.

De fixed effecten in het model schatten we met behulp van de gegeneraliseerde kleinste kwadratenmethode:

$$\tilde{\vec{\alpha}} = (X'V^{-1}X)^{-1}X'V^{-1}\vec{y} \qquad (3)$$

Deze schatter is BLUE (Best Linear Unbiased Estimator) en tevens de maximum likelihood schatter onder normaliteit.

De randomeffecten $\vec{b}$ voorspellen we met behulp van regressie van $\vec{b}$ op de waarnemingsvector $\vec{y}$:

$$\tilde{\vec{b}} = DZ'V^{-1}(\vec{y}-X\tilde{\vec{a}}), \tag{4}$$

waarbij we $\vec{a}$ hebben vervangen door $\tilde{\vec{a}}$ uit (3) en D de matrix van varianties en covarianties van $\vec{b}$ is.(*) Deze voorspeller is BLUP (Best Linear Unbiased Predictor).
De schatter uit (3) en de voorspeller uit (4) voldoen aan de zogenaamde mixed model vergelijkingen:

$$\begin{pmatrix} X'X & X'Z \\ Z'X & Z'Z + \sigma_e^2 D^{-1} \end{pmatrix} \begin{pmatrix} \tilde{\vec{a}} \\ \tilde{\vec{b}} \end{pmatrix} = \begin{pmatrix} X'\vec{y} \\ Z'\vec{y} \end{pmatrix} \tag{5}$$

Merk op dat deze vergelijkingen numeriek aantrekkelijker zijn dan (3) en (4). In (3) en (4) moeten we immers de N × N matrix V inverteren, de coëfficiëntenmatrix in (5) is echter (p+q)×(p+q) en in het algemeen veel kleiner. Wanneer we de fixed effecten uit (5) elimineren vinden we:

$$\tilde{\vec{b}} = (Z'MZ + \sigma_e^2 D^{-1})^{-1} Z'M\vec{y} \tag{6}$$

met $M = I - X(X'X)^{-1}X'$ de projectie-matrix op het orthogonale complement van de kolommenruimte van X. Verder:

$$\tilde{\vec{a}} = (X'X)^{-1} X'(\vec{y} - Z\tilde{\vec{b}}) \tag{7}$$

Nu behoeven we dus slechts matrices van omvang p×p en q×q te inverteren. Een voor de hand liggende voorspelling voor de vector van restbijdragen $\vec{e}$ is:

$$\tilde{\vec{e}} = \vec{y} - X\tilde{\vec{a}} - Z\tilde{\vec{b}} = M(\vec{y} - Z\tilde{\vec{b}}) \tag{8}$$

In de praktijk zijn de variantiecomponenten en de elementen van V natuurlijk niet bekend. We zouden kunnen proberen om met behulp van (6) en (8) iteratief de variantiecomponenten te schatten. Een dergelijk iteratieproces zou er als volgt uit kunnen zien:
(i) kies startwaarden voor de variantiecomponenten,
(ii) bepaal de voorspellingen uit (6) en (8),

(*) Meestal is D een diagonaal matrix. Covariantie-termen kunnen echter ook worden opgenomen. Wanneer $\vec{y}$ melkproduktiegegevens van dochters van stieren en $\vec{b}$ de genetische bijdragen van de stiervaders bevat kunnen middels covariantie-termen in D familieverbanden tussen de stieren in het model worden opgenomen.

(iii) bepaal nieuwe waarden voor de variantiecomponenten, bijvoorbeeld op basis van $\tilde{b}_i'\tilde{b}_i$ $i = 1, \ldots, c$ en $\tilde{e}'\tilde{e}$,
(iv) stop als opeenvolgende iteratiestappen resultaten geven die veel op elkaar lijken, ga anders terug naar (ii).

Alle iteratieve procedures in dit artikel kunnen in termen van de oplossingen van de mixed model vergelijkingen worden geformuleerd. In 7.3 schetsen we een algoritme dat nauw bij het bovenstaande schema aansluit.

2.1 Opmerkingen

Stel dat C de inverse van de coëfficiënten matrix in (5) is en als volgt is gepartitioneerd:

$$C = \begin{pmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{pmatrix} \text{ met } C_{11}\ p\times p,\ C_{12} = C_{21}'\ p\times q \text{ en } C_{22}\ q\times q$$

Dan is voor V bekend, $\text{var}(\tilde{\alpha}) = C_{11}\sigma_e^2 = (X'V^{-1}X)^{-1}$ en $\text{var}(\tilde{b} - \vec{b}) = C_{22}\sigma_e^2$.

Soms kunnen we nog verder bezuinigen op de omvang van de te inverteren matrices. Wanneer bijvoorbeeld $D = I\sigma_i^2$ kunnen we de inverse in (6) schrijven in termen van de inverse van een diagonaal matrix van omvang q×q en de inverse van een matrix van omvang p×p. Dit is handig wanneer q groot is en p klein.

3 HENDERSONS METHODE III

Hendersons methode III is de derde in een reeks van methoden voor het schatten van variantiecomponenten afkomstig van C.R. Henderson. Deze methode wordt ook de 'fitting constants method' genoemd.
Hendersons methode III is een zeer populaire methode voor het schatten van variantiecomponenten niet in het minst omdat deze methode de basis vormt van het LSML76-programma (Mixed Model Least Squares and Maximum Likelihood Computer Program) van W.R. Harvey (1977).

3.1 Rationale

De methode maakt gebruik van de volgende eigenschap:
We doen even alsof de vector $\vec{b}$ niet een vector van random effecten is maar een vector van fixed effecten. We passen nu twee modellen aan, namelijk het volledige model (model I) en het model zonder de bijdragen uit $\vec{b}$ (model II). Noem de bijbehorende restkwadraatsommen RSS_I en RSS_{II}.

Definieer de kwadraatsom van $\vec{b}$ als

$$SS_b = RSS_{II} - RSS_I$$

Nu herinneren we ons dat $\vec{b}$ random bijdragen bevat en we bepalen de verwachtingswaarde van SS_b.

Nu geldt:

(i) $E(SS_b)$ hangt niet af van de vector van fixed effecten $\vec{\alpha}$,
(ii) $E(SS_b)$ is een lineaire combinatie, met bekende coëfficiënten, van de variantiecomponenten corresponderende met $\vec{b}$ en σ_δ^2.

Met behulp van deze eigenschap, die ook geldt voor deelvectoren van $\vec{b}$, bepalen we c kwadraatsommen. De kwadraatsom RSS_I voegen we aan deze set toe. We schrijven nu c+1 vergelijkingen op van de vorm E(kwadraatsom) = lineaire combinatie van variantiecomponenten en vervangen het linkerlid door de bijbehorende realisatie. Het resulterende stelsel van c+1 lineaire vergelijkingen met c+1 onbekenden lossen we vervolgens op.

3.2 Nadere uitwerking

Veronderstel dat het model drie random effecten bevat. Bijvoorbeeld $\vec{b}_1$ en $\vec{b}_2$ zijn hoofdeffecten en $\vec{b}_3$ de bijbehorende interaktie. Verder nemen we voor het gemak aan dat het model behalve het algemeen gemiddelde μ geen fixed effecten bevat.

We passen nu de volgende modellen aan:
I : model met μ, b_1, b_2, b_3.
II : model met μ, b_1, b_2.
III: model met μ, b_1.
IV : model met μ.

We gebruiken de volgende kwadraatsommen:
$SS_{b_3} = RSS_{II} - RSS_I$, $SS_{b_2\,3} = RSS_{III} - RSS_I$, $SS_{b_1\,2\,3} = RSS_{IV} - RSS_I$ en RSS_I.
We vinden hetzelfde stelsel vergelijkingen met:
SS_{b_3}, $SS_{b_2} = RSS_{III} - RSS_{II}$, $SS_{b_1} = RSS_{IV} - RSS_{III}$ en RSS_I.

3.3 Opmerkingen

In bovenstaand voorbeeld kunnen we model III vervangen door het model met μ en b_2.
In dat geval vinden we een ander stelsel vergelijkingen met andere oplossingen als schattingen voor de variantiecomponenten! Dit is een slechte eigenschap van de methode: soms kun je meerdere stelsels met verschillende oplossingen samenstellen.

Welk stelsel we ook kiezen de schatters die uit het stelsel volgen zijn zuivere schatters. Dit is overigens niet helemaal waar daar de schattingen negatief kunnen zijn en we negatieve uitkomsten doorgaans door nul vervangen.

4 MINQUE

MINQUE is een afkorting van "Minimum Norm Quadratic Unbiased Estimator (of soms: Estimation)" en afkomstig van C.R. Rao.

4.1 Rationale

Veronderstel dat we een lineaire combinatie $\sum_i p_i \sigma_i^2$ willen schatten. Als schatter neemt Rao een kwadratische vorm $\vec{y}'A\vec{y}$ met A een symmetrische matrix. Om te beginnen eisen we nu dat de schatter zuiver en translatie invariant is, dat wil zeggen:

$$E(\vec{y}'A\vec{y}) = \sum_i p_i \sigma_i^2$$

$$(\vec{y} + X\vec{\alpha}_o)' A(\vec{y} + X\vec{\alpha}_o) = \vec{y}'A\vec{y} \text{ voor alle } \vec{\alpha}_o.$$

Nodige en voldoende voorwaarden hiervoor zijn:

$$AX = 0 \tag{9}$$

$$\text{spoor } (AZ_iZ_i') = p_i \quad i = 0, 1, \ldots, c \tag{10}$$

Wanneer we $\dot{\vec{b}}$ kennen is $\sum_i p_i \frac{\dot{\vec{b}}_i'\dot{\vec{b}}_i}{q_i}$ een "natuurlijke" schatter voor $\sum_i p_i \sigma_i^2$. Voor een geschikt gedefinieerde matrix Δ kunnen we deze "natuurlijke" schatter schrijven als

$$\dot{\vec{b}}'\Delta\dot{\vec{b}}$$

Uit (9) volgt dat we $\vec{y}'A\vec{y}$ kunnen schrijven als

$$\ddot{\vec{b}}Z'AZ\ddot{\vec{b}}$$

We bepalen nu een matrix A die aan (9) en (10) voldoet en waarvoor $\dot{Z}'A\dot{Z}$ op Δ "lijkt" in die zin dat we de norm

$$\|\dot{Z}'A\dot{Z} - \Delta\|$$

minimaliseren, waarbij $\|B\|^2 \overset{\text{def}}{=} \text{spoor } (BB')$, dit is de som van de kwadraten van de elementen van de matrix B.
Algemener kunnen we ook een "gewogen" norm minimaliseren:

$$\|W^{\frac{1}{2}} (\dot{Z}'A\dot{Z} - \Delta)W^{\frac{1}{2}}\| \tag{11}$$

met $W = \text{diag } (w_0^2 I_{q_0}, \ldots, w_c^2 I_{q_c})$ en I_{q_i} de $q_i \times q_i$ eenheidsmatrix.

Dit resultaat kunnen we ook bereiken door het model als volgt te formuleren:

$$\vec{y} = X\vec{\alpha} + \sum_{i=0}^{c} (w_iZ_i)(\vec{b}_i/w_i) = X\vec{\alpha} + \sum_{i=0}^{c} Z_i^*\vec{b}_i^* = X\vec{\alpha} + \dot{Z}^*\dot{\vec{b}}^*$$

$$\sum_i p_i \sigma_i^2 = \sum_i (p_i w_i^2) \frac{\sigma_i^2}{w_i^2} = \sum p_i^* \sigma_i^{2*}$$

en de ongewogen procedure toe te passen.

We kunnen de gewichten w_i^2 als apriori waarden voor de variantiecomponenten σ_i^2 beschouwen.

4.2 Nadere uitwerking

Door $(p_o \ldots p_c)$ achtereenvolgens gelijk te kiezen aan de rijen van de eenheidsmatrix vinden we schatters voor $\sigma_0^2, \ldots, \sigma_c^2$. Deze schatters, samengevat in de vector $\tilde{\vec{\sigma}}^2$, voldoen aan het volgende stelsel vergelijkingen:

$$S_w \tilde{\vec{\sigma}}^2 = \vec{u}_w \qquad (12)$$

met $S_w = (s_{ij})_{i,j} \quad \vec{u}_w' = (u_i)_i$

$$s_{ij} = \text{spoor } (P_w Z_i Z_i' P_w Z_j Z_j')$$

$$u_i = \vec{y}' P_w Z_i Z_i' P_w \vec{y}$$

en $P_w = V_w^{-1} - V_w^{-1} X (X' V_w^{-1} X)^{-1} X' V_w^{-1}$

$$V_w = \overset{\cdot\cdot}{Z} W Z'$$

Merk op dat V_w gelijk is aan $\text{var}(\vec{y}) = V$ indien de apriori waarden w_i^2 en de variantiecomponenten σ_i^2 overeenstemmen. Daar de MINQUE schattingen negatief kunnen zijn, moeten we ook hier de zuiverheid met een korreltje zout nemen wanneer we negatieve uitkomsten door nul vervangen.

4.3 MINQUE-O

Voor MINQUE-O kiezen we de gewichten w_i^2 i = 0, ..., c als volgt:

$$w_0^2 = 1, \quad w_1^2 = \ldots = w_c^2 = 0.$$

Dit brengt een aanzienlijke vereenvoudiging in (12) met zich mee, daar in dat geval geldt: $V_w = I$.

4.4 I-MINQUE

I-MINQUE is een afkorting voor "Iterated-MINQUE".
In dit geval gebruiken we de uitkomsten van (12) als nieuwe apriori waarden (gewichten) en berekenen we opnieuw de MINQUE schatters. Dit herhalen we, in theorie tot in het oneindige, in praktijk tot de resultaten van opeenvolgende iteraties voldoende op elkaar lijken. Merk op dat bij het itereren de zuiverheidsvoorwaarde zeker verloren gaat.

Wanneer het iteratieproces convergeert voldoen de I-MINQUE oplossingen aan het volgende stelsel vergelijkingen:

$$S\,\vec{\tilde{\sigma}}^2 = \vec{u} \qquad (13)$$

met $S = (s_{ij})_{i,j}$ $\vec{u}' = (u_i)_i$

s_{ij} = spoor $(\bar{P}Z_iZ_i'\bar{P}Z_jZ_j')$

$u_i = \vec{y}'\bar{P}Z_iZ_i'\bar{P}\vec{y}$

$\bar{P} = \bar{V}^{-1} - \bar{V}^{-1}X(X'\bar{V}^{-1}X)^{-1}X'\bar{V}^{-1}$

Dit stelsel volgt uit (12) door de gewichten en oplossingen aan elkaar gelijk te nemen, $\bar{V} = \sum_{i=0}^{c} Z_iZ_i'\bar{\sigma}_i^2$.

4.5 Opmerkingen

Wanneer een iteratiestap van I-MINQUE een negatieve schatting voor een variantiecomponent geeft vervangt men deze in praktijk doorgaans door nul en vervolgt het iteratieproces. Het blijft mogelijk dat het iteratieproces niet convergeert.
MINQUE en I-MINQUE kunnen ook in termen van oplossingen van de mixed model vergelijkingen worden geformuleerd. We gaan hier niet verder op in.

5 MIVQUE

MIVQUE is een afkorting van "Minimum Variance Quadratic Unbiased Estimator (of soms: estimation)" en afkomstig van L.R. LaMotte.

5.1 Rationale

Veronderstel dat we een lineaire combinatie $\Sigma p_i\sigma_i^2$ willen schatten met behulp van een kwadratische vorm $\vec{y}'A\vec{y}$. We nemen aan dat de waarnemingsvector een multivariaat normale verdeling volgt. We proberen nu een symmetrische matrix A te bepalen die aan (9) en (10) voldoet (schatter is zuiver en translatie invariant) en die bovendien de variantie $\mathrm{Var}(\vec{y}'A\vec{y})$ minimaliseert.

De oplossing voor A blijkt niet uitsluitend van de waarnemingen maar ook van de onbekende variantiecomponenten af te hangen. Dit wordt "opgelost" door de variantiecomponenten in A door apriori waarden te vervangen. Het resultaat is dus alleen "minimum variance" indien apriori waarden en werkelijke waarden voor de variantiecomponenten overeenstemmen. Daar we de variantiecomponenten nu juist willen schatten worden we daar niet veel wijzer van. Doorgaans spreekt men van "locally minimum variance".

5.2 Nadere uitwerking

De MIVQUE schatter met apriori waarden $w_0^2, \ldots, w_c^2$ blijkt dezelfde te zijn als de MINQUE oplossing van (11). Analoog aan 4.3 en 4.4 kunnen we MIVQUE-0 en I-MIVQUE definiëren, deze stemmen onder normaliteit overeen met respectievelijk MINQUE-0 en I-MINQUE.

6 MAXIMUM LIKELIHOOD

6.1 Rationale

We veronderstellen dat de waarnemingsvector een multivariaat normale verdeling volgt. We schatten de onbekende parameters in het model door die waarden waarvoor de likelihood of, omdat dat makkelijker werkt, de logaritme van de likelihood L maximaal is. Omdat we bij het maximaliseren van L alleen waarden voor de parameters toelaten die tot de afsluiting van de parameterruimte behoren zijn de ML-schattingen voor de variantiecomponenten nooit negatief. Het is echter niet uitgesloten dat L zijn maximum als het ware in het oneindige bereikt.

6.2 Nadere uitwerking

Wanneer L zijn maximum aanneemt in het inwendige van de parameterruimte dan voldoen de ML-schatters aan de volgende vergelijkingen:

$$X'\bar{V}^{-1}X\tilde{\vec{\alpha}} = X'\bar{V}^{-1}\vec{y} \tag{14}$$

$$\text{spoor } (\bar{V}^{-1}Z_iZ_i') = (\vec{y} - X\tilde{\vec{\alpha}})'\,\bar{V}^{-1}Z_iZ_i'\bar{V}^{-1}\,(\vec{y} - X\tilde{\vec{\alpha}}), \quad i = 0, 1, \ldots, c, \tag{15}$$

waarin $\bar{V} = \sum_{i=0}^{c} Z_iZ_i'\,\tilde{\sigma}_i^2$.

Deze vergelijkingen verkrijgen we door de partiële afgeleiden van L ten aanzien van de onbekende parameters gelijk aan nul te stellen. Daar alle stationaire punten van L aan (14) en (15) voldoen is niet elke oplossing van dit stelsel noodzakelijk gelijk aan de ML-schatter.
Door (14) in te vullen in (15) volgt dat het rechterlid van (15) gelijk is aan $\vec{y}'\bar{P}Z_iZ_i'\bar{P}\vec{y}$. Het linkerlid kunnen we als volgt herschrijven:

$$\text{spoor } (\bar{V}^{-1}Z_iZ_i') = \text{spoor } (\bar{V}^{-1}Z_iZ_i'\,\bar{V}^{-1}\bar{V}) = \text{spoor } (\bar{V}^{-1}Z_iZ_i'\bar{V}^{-1}\sum_j \tilde{\sigma}_j^2 Z_jZ_j')$$
$$= \sum_j \text{spoor } (\bar{V}^{-1}Z_iZ_i'\bar{V}^{-1}Z_jZ_j')\,\tilde{\sigma}_j^2$$

Zo vinden we voor (15) de uitdrukking:

$$S\,\tilde{\vec{\sigma}}^2 = \vec{u} \tag{16}$$

met $S = (s_{ij})_{i,j}$ $\quad \vec{u}' = (u_i)_i$

$$s_{ij} = \text{spoor } (\bar{V}^{-1}Z_iZ_i'\,\bar{V}^{-1}Z_jZ_j')$$

$$u_i = \vec{y}'\bar{P}Z_iZ_i'\,\bar{P}\vec{y}$$

$$\bar{P} = \bar{V}^{-1} - \bar{V}^{-1}\,X(X'\bar{V}^{-1}X)^{-1}X'\bar{V}^{-1}$$

$$\bar{V} = \sum_{i=0}^{c} Z_iZ_i'\,\tilde{\sigma}_i^2$$

Merk op dat (16) erg op (13) lijkt. We krijgen (16) uit (13) door $\vec{P}$ door $\vec{V}^{-1}$ te vervangen in de uitdrukking voor s_{ij}.

6.3 Opmerkingen

Er kan worden aangetoond dat de ML-schattingen voor de variantiecomponenten uitsluitend van $\vec{y}$ afhangen via een vector $K\vec{y}$ waarbij de rijen van K een willekeurig, maar onafhankelijk, N-p tal rijen van de matrix $M = I - X(X'X)^{-1}X'$ zijn. Daar $E(K\vec{y}) = \vec{0}$ volgt dat de ML-schattingen voor de variantiecomponenten invariant zijn.

7 REML

REML is een afkorting voor Restricted Maximum Likelihood en wordt meestal toegeschreven aan het duo H.D. Patterson en R. Thompson.

7.1 Rationale

Wanneer we voor het eenvoudige geval van een aselecte steekproef $y_1, y_2, \ldots, y_N$ uit een $N(\mu, \sigma^2)$ verdeling, met onbekende μ en σ^2, de maximum likelihood schatter voor σ^2 bepalen, vinden we:

$$\frac{1}{N} \sum_i (y_i - \bar{y}_.)^2$$

Deze schatter heeft verwachtingswaarde $\frac{N-1}{N} \sigma^2$ en is onzuiver. Brengen we voor het schatten van μ door het gemiddelde $\bar{y}_.$ een verlies van 1 vrijheidsgraad in rekening dan vinden we:

$$\frac{1}{N-1} \sum_i (y_i - \bar{y}_.)^2.$$

Deze schatter is zuiver.

In het veel complexere mixed model doet iets dergelijks zich ook voor. Bij het schatten van de variantiecomponenten met behulp van ML wordt geen rekening gehouden met het verlies aan graden van vrijheid voortkomende uit het schatten van de vector van fixed parameters $\vec{\alpha}$. De REML methode probeert hier iets aan te doen.

We bepalen N-p lineaire functies $\vec{k}_i'\vec{y}$ zodat $E(\vec{k}_i'\vec{y}) = 0$, i = 1, 2, ...,N-p, en de vectoren van coëfficiënten $\vec{k}_1, \ldots, \vec{k}_{N-p}$ lineair onafhankelijk zijn. Stel nu dat $\vec{k}_i'$ de i-de rij is van de (N-p) × N matrix K.
Voor de REML methode passen we nu de ML methode toe, echter niet op de vector van waarnemingen $\vec{y}$, maar op de vector van zogenaamde error-contrasten $K\vec{y}$.

7.2 Nadere uitwerking

Merk op dat elk N-p tal onafhankelijke rijen van de projectiematrix $M = I - X(X'X)^{-1}X'$ als rijen voor de matrix K kunnen worden gebruikt. Welke collectie van N-p onafhankelijke rijen we ook kiezen, de bijbehorende log-likelihood L_1 van $K\vec{y}$ is behoudens een constante altijd dezelfde.

Wanneer L_1 zijn maximum in het inwendige van de parameterruimte aanneemt, voldoen de REML-schattingen aan het volgende stelsel vergelijkingen (partiële afgeleiden van L_1 gelijk aan nul stellen):

$$S\,\vec{\tilde{\sigma}}^2 = \vec{u} \qquad (17)$$

met $S = (s_{ij})_{i,j}$ $\vec{u} = (u_i)_i$

$$s_{ij} = \text{spoor}\,(\bar{P}Z_iZ_i'\,\bar{P}Z_jZ_j')$$

$$u_i = \vec{y}'\,\bar{P}Z_iZ_i'\,\bar{P}\vec{y}$$

$$\bar{P} = \bar{V}^{-1} - \bar{V}^{-1}X(X'\bar{V}^{-1}X)^{-1}X'\bar{V}^{-1}$$

$$\bar{V} = \sum_i Z_iZ_i'\,\tilde{\sigma}_i^2$$

Vergelijken we dit stelsel met (16), het stelsel voor de ML schatters, dan zien we dat (17) uit (16) volgt door de matrix $\bar{V}^{-1}$ in de uitdrukking voor de s_{ij} door $\bar{P}$ te vervangen. Verder zien we dat (12) en (17) overeenkomen. Dus, wanneer de oplossing van REML positief is dan stemt deze oplossing overeen met de oplossing van I-MINQUE en I-MIVQUE onder normaliteit.

7.3 Een andere afleiding van REML

De REML-schattingen kunnen ook worden afgeleid als uitkomsten van een zogenaamd EM-algoritme (Dempster et al. (1984)).
Aan het begin van de t-de stap van dit iteratieve proces veronderstellen we dat de variantiecomponenten bekend zijn, zeg $\sigma^2_{0(t-1)}$, $\sigma^2_{1(t-1)}$, ..., $\sigma^2_{c(t-1)}$. Voor $\vec{\alpha}$, $\vec{b}_i$ i = 1, 2, ..., c en $\vec{e}$ nemen we de volgende apriori verdelingen:

$$\vec{\alpha} \sim N(\vec{0},\ I\tau^2)$$

$$\vec{b}_i \sim N(\vec{0},\ I\sigma^2_{i(t-1)}) \qquad i = 1, 2, \ldots, c$$

$$\vec{e} \sim N(\vec{0},\ I\sigma^2_{0(t-1)})$$

We veronderstellen dat $\vec{\alpha}$, $\vec{b}_i$, i = 1,2,...,c en $\vec{e}$ onafhankelijk zijn verdeeld en dat τ^2 erg groot is.
Nu geldt:

$$\text{Var}(\vec{y}) = XX'\tau^2 + V_{(t-1)}$$

met $V_{(t-1)} = \sum_{i=1}^{c} Z_iZ_i'\,\sigma^2_{i(t-1)} + I\sigma^2_{0(t-1)}$

Verder:

$$\lim_{\tau \to \infty} \mathrm{Var}(\vec{y})^{-1} = V_{(t-1)}^{-1} - V_{(t-1)}^{-1}X(X'V_{(t-1)}^{-1}X)^{-1}X'V_{(t-1)}^{-1} \overset{\text{def}}{=} P_{(t-1)}.$$

Merk op dat $P_{(t-1)}$ vergelijkbaar is met $\bar{P}$ uit (13), (16) en (17). We "schatten" nu $\vec{\alpha}$, $\vec{b}$ en $\vec{e}$ met de conditionele verwachtingen $E(\vec{\alpha}|\vec{y})$, $E(\vec{b}|\vec{y})$ en $E(\vec{e}|\vec{y})$. Dit levert uitkomsten $\tilde{\vec{\alpha}}$, $\tilde{\vec{b}}$ en $\tilde{\vec{e}}$, waarbij $\tilde{\vec{\alpha}}$ en $\tilde{\vec{b}}$ oplossingen zijn van de mixed model vergelijkingen en

$$\tilde{\vec{e}} = \vec{y} - X\tilde{\vec{\alpha}} - Z\tilde{\vec{b}}.$$

Als nieuwe waarden voor de variantiecomponenten nemen we

$$\begin{aligned} \sigma^2_{i\,(t)} &= E(\vec{b}_i'\vec{b}_i|\vec{y})/q_i, \quad i = 1, 2, \ldots, c \\ \sigma^2_{0\,(t)} &= E(\vec{e}'\vec{e}|\vec{y})/N \end{aligned} \tag{18}$$

Bij nadere uitwerking blijkt dit een iteratieve procedure voor het oplossen van (17) te zijn.

7.4 Opmerkingen

Uitdrukking (18) kan ook worden afgeleid als iteratief proces door de vergelijkingen (17) te manipuleren. In dat geval wordt het proces Hendersons iteratieve procedure genoemd. De procedure convergeert langzaam maar heeft één sterk pluspunt: er kan worden aangetoond dat de uitkomsten voor de variantiecomponenten verkregen aan het einde van iedere iteratiestap niet-negatief zijn.

Wanneer we de REML oplossingen bepalen met behulp van I-MINQUE of I-MIVQUE dan wordt het iteratieproces ook wel Anderson's iteratieve procedure genoemd. Deze procedure is verwant aan de Fisher scoring methode, en kan ook direct uit (17) worden afgeleid. Anderson's procedure convergeert doorgaans sneller dan Hendersons procedure, negatieve uitkomsten voor de variantiecomponenten zijn echter niet uitgesloten.

Voor ML kunnen equivalenten van Hendersons en Andersons procedure uit (14) en (15) worden afgeleid met behoud van bovengenoemde eigenschappen.

In 6.3 merkten we op dat de ML-schattingen voor de variantiecomponenten uitsluitend afhangen van een vector van error-contrasten $K\vec{y}$. ML en REML gebruiken dus ten aanzien van het schatten van de variantiecomponenten dezelfde informatie, zij het op een verschillende manier. Uiteraard zijn de REML-schattingen ook invariant.

8 RELATIES TUSSEN DE SCHATTINGSMETHODEN

In de voorafgaande paragrafen merkten we reeds op dat de beschreven methoden voor het schatten van de variantiecomponenten (gelukkig) niet los van elkaar staan.
Andersons iteratieve procedure voor REML is gelijk aan I-MINQUE. Wanneer REML positieve schattingen geeft zijn dit tevens de I-MINQUE-schattingen.

REML en ML zijn beide maximum likelihood methoden, echter REML werkt op de vector van error-contrasten $K\vec{y}$ in plaats van op de waarnemingsvector $\vec{y}$. In gebalanceerde modellen stemmen Henderson III en REML overeen met de gebruikelijke ANOVA aanpak, mits de schattingen positief zijn. Onder normaliteit geven MINQUE en MIVQUE dezelfde schatters.

We kunnen nog wat meer over de relaties tussen de methoden zeggen aan de hand van het zogenaamde "dispersion-mean model".
Uitgangspunt is een vector van error-contrasten $K\vec{y}$. We bepalen de matrix van kwadraten en produkten van elementen van $K\vec{y}$. Dit is

$$U = (K\vec{y})(K\vec{y})' = K\vec{y}\vec{y}'K'$$

We doorlopen de matrix U nu rij-gewijs van linksboven naar rechtsonder en plaatsen de elementen die we tegenkomen in de kolomvector vec(U). Voor de verwachtingswaarde van de vector vec(U) geldt:

$$E[\mathrm{vec}(U)] = F\,\vec{\sigma}^2 \qquad (19)$$

waarin F een bekende matrix is die kan worden afgeleid uit de matrices X en Z. Dit is het dispersion-mean model. Het is een lineair model in de variantiecomponenten σ_0^2, σ_1^2, ..., σ_c^2.

Met enig kunst en vliegwerk kan nu worden afgeleid dat de "gewone" kleinste kwadraten methode toegepast op (19) MINQUE-0 geeft en de gegeneraliseerde kleinste kwadraten methode MINQUE oplevert, waarbij we apriori waarden nodig hebben voor de constructie van Var(vec(U)). Wanneer we eerst $\vec{\alpha}$ bekend veronderstellen en in plaats van $K\vec{y}$ de vector $y-X\vec{\alpha}$ nemen en achteraf $\vec{\alpha}$ door $\tilde{\vec{\alpha}}$ uit (3) vervangen, vinden we met behulp van de gegeneraliseerde kleinste kwadraten methode de ML-procedure.

De relaties tussen de verschillende methoden zijn weergegeven in figuur 4 in de appendix.

9 SIMULATIE

We zouden graag weten hoe de schatters voor de variantiecomponenten uit Henderson III, MINQUE-, MIVQUE-, REML- en ML-verschillen met betrekking tot onzuiverheid, variantie en mean-square-error.

Praktijksituaties kunnen erg verschillen zowel ten aanzien van de verhouding tussen de variantiecomponenten als ten aanzien van de mate van ongebalanceerdheid.

Om het rekenwerk bij de simulaties binnen de perken te houden is het aantal schema's wat in een simulatie wordt doorgerekend beperkt. Per schema is het aantal factoren en instellingen van deze factoren gering. Bij het trekken van conclusies uit simulatieresultaten moeten we dus voorzichtig zijn.

Voor het 1-weg random model met ongebalanceerde data hebben Swallow en Monahan (1984) een Monte Carlo studie verricht. De schatters onder Henderson III, REML, ML, MIVQUE-0 en MIVQUE(A) (dit is MIVQUE met als apriori-waarden de Henderson III-schattingen) zijn vergeleken ten aanzien van onzuiverheid en mean-square-error. Bij alle methoden is rekening gehouden met het feit dat de schatters niet negatief dienen te zijn. Henderson III is goed behalve ingeval de data erg ongebalanceerd zijn en tevens de variantie van de random factor groter is dan de residuele variantie. In geval de variantie van de random factor minder dan de helft van de residuele variantie bedraagt, verdient ML de voorkeur vanwege de kleine onzuiverheid en de lage mean-square-error.
MIVQUE-0 komt als duidelijke verliezer naar voren vanwege de hoge mean-square-error ingeval de variantiecomponent groter is dan de residuele variantie. MIVQUE(A) is goed en verder itereren naar de REML-oplossing lijkt niet nodig. Giesbrecht en Burns (1985) stellen eveneens dat twee stappen MINQUE afdoende is voor de REML-oplossing.

Lee en Kapadia (1984) hebben een simulatiestudie verricht voor het 2-weg mixed model zonder interaktie en met één waarneming per cel. ML- en REML-schatters voor de variantiecomponenten zijn vergeleken op onzuiverheid en mean-square-error. Bij de schattingsmethoden is de niet negativiteits-voorwaarde voor de schatters meegenomen.
REML geeft een verwaarloosbare onzuiverheid. De relatieve onderschatting bij ML is ongeveer omgekeerd evenredig met het aantal niveaus van de random factor; Corbeil en Searle (1976) komen tot eenzelfde onderschatting, alleen hebben deze geen rekening gehouden met de niet-negativiteitsvoorwaarde. Qua mean-square-error verschillen REML en ML weinig.

Voor het 2-weg random model met interaktie met gebalanceerde data is door Miller (1979) een Monte Carlo studie verricht. Voor kleine steekproeven is de onzuiverheid bij REML aanzienlijk kleiner dan bij ML, doch dit gaat ten koste van een hogere mean-square-error. Voor grote steekproeven verschilt de mean-square-error bij REML niet veel van die bij ML.

Op grond van de simulatiestudies is er geen duidelijke winnaar aan te wijzen. MINQUE-0 kan beter niet gebruikt worden, tenzij men er zeker van is dat de variantiecomponenten erg klein zijn ten opzichte van de residuele variantie. REML geeft minder onzuivere schatters dan ML, doch dit gaat gepaard met een hogere mean-square-error. Waarschijnlijk geven twee stappen met MINQUE reeds een goede benadering voor de REML-oplossing.

10 EEN UITGEWERKT VOORBEELD

We analyseren een dataset uit Dempster et al. (1984). De waarnemingen zijn gewichten van jonge ratten afkomstig uit verschillende nesten. We onderscheiden twee bronnen van toevalsbijdragen:

(i) nestbijdragen, hieronder rekenen we alle invloeden die ratten uit hetzelfde nest gemeenschappelijk hebben, zoals bijvoorbeeld de genetische bijdrage van hun vader en moeder,

(ii) restbijdragen, variatie tussen ratten binnen hetzelfde nest.

De moeders zijn verdeeld in drie groepen: een controlegroep en twee groepen die respectievelijk een lage en een hoge dosis van een experimenteel mengsel toegediend hebben gekregen.
Andere factoren die een rol spelen zijn sexe en omvang van het nest. De waarnemingen zijn geanalyseerd volgens het split-plot model:

$$y_{ijkl} = \mu + d_i + \beta l_{ik} + b_{ik} + s_l + e_{ijkl}$$

$$b_{ik},\ e_{ijkl} \text{ onafhankelijk}$$

$$b_{ik} \sim N(0,\ \sigma_1^2) \text{ en } e_{ijkl} \sim N(0,\ \sigma_0^2)$$

waarin:

y_{ijkl} = gewicht in grammen van de j-de rat van sexe l van het k-de nest waarvan de moeder dosis i heeft ontvangen,
μ = basisniveau,
d_i = dosisbijdrage (i=0 is controle, i = 1, 2 zijn resp. lage en hoge dosis),
l_{ik} = omvang van het k-de nest waarvan de moeder dosis i heeft ontvangen met bijbehorende regressiecoëfficiënt β,
b_{ik} = nestbijdrage (whole-plot error),
s_l = sexebijdrage (l=1 is een mannetje, l=2 is een vrouwtje),
e_{ijkl} = restbijdrage (sub-plot error).

De dataset bevat waarnemingen aan 322 jonge ratten, 171 mannetjes en 151 vrouwtjes, verdeeld over 27 nesten van verschillende moeders. De doses zijn over de moeders verloot. Voor de controle, lage en hoge dosis zijn de aantallen respectievelijk 10, 10 en 7.

De nestomvang varieert van 2 tot 18 en is gemiddeld 11.93. De waarnemingen variëren van 3.68 tot 8.33 met een gemiddelde van 6.08.

In tabel 1 zijn de schattingen voor de variantiecomponenten gegeven.
Startwaarden voor I-MINQUE, REML en ML waren $\sigma_0^2 = 1$ en $\sigma_1^2 = 1$.
In alle iteratieve procedures is het volgende stopcriterium gebruikt:

stop als

$$[(\tilde{\sigma}^2_{1\,nieuw} - \tilde{\sigma}^2_{1\,oud})^2 + (\tilde{\sigma}^2_{0\,nieuw} - \tilde{\sigma}^2_{0\,oud})^2]^{1/2} <$$

$$0.00001 * [(\tilde{\sigma}^2_{1\,oud} + \tilde{\sigma}^2_{0\,oud})^{1/2} + 0.001]$$

51

Tabel 1

	methode						
	Henderson III	MINQUE	MINQUE-0	I-MINQUE = REML Anderson	REML Henderson	ML Anderson	ML Henderson
	aantal iteraties						
	-	-	-	6	7	4	8
$\tilde{\sigma}_I^2$	0.1025	0.1037	0.1185	0.0974	0.0974	0.0815	0.0815
$\tilde{\sigma}_\delta^2$	0.1630	0.1615	0.1492	0.1628	0.1628	0.1621	0.1621
$\tilde{\sigma}_I^2/\tilde{\sigma}_\delta^2$	0.629	0.642	0.794	0.598	0.598	0.503	0.503

De gebruikte computerprogramma's zijn geschreven in GENSTAT (Alvey et al. (1977)) en opgenomen in Engel et al. (1985).

Het rekenwerk per iteratieslag is voor de iteratieve processen vergelijkbaar, in alle gevallen worden de mixed model vergelijkingen opgelost waarna er nog wat sporen van matrices worden berekend. Bij REML wint Andersons methode (dit is tevens I-MINQUE) met een neuslengte van Hendersons iteratieve procedure ten aanzien van het aantal iteraties. Bij ML is het verschil groter. In het algemeen is Hendersons methode veel trager dan Andersons methode. Merk op dat de eerste stap van Andersons algoritme voor REML (dit is MINQUE in kolom 2 van de tabel) al dicht in de buurt van de eindoplossing ligt. Dit is vaak het geval. We gaan nu nog wat nader op de verschillende methoden in.

Bij Henderson methode III passen we twee modellen aan namelijk het volledige model met nestbijdragen fixed (deze zijn dan gestrengeld met de factor dosis en de variabele nestomvang) en het model waarin nestbijdragen niet zijn opgenomen. Uit de restkwadraatsommen RSS_I en RSS_{II} en hun vrijheidsgraden df_I en df_{II} bepalen we de kwadraatsom voor nestbijdragen $SS_{nest} = RSS_{II} - RSS_I$ met $d = df_{II} - df_I$ vrijheidsgraden. Nu geldt:

$$E(SS_{nest}/d) = \sigma_\delta^2 + \phi\,\sigma_I^2$$

$$E(RSS_I/df_I) = \sigma_\delta^2,$$

hierin is de constante ϕ het spoor van een matrix. Hier geldt $\phi = 11.839$. Vaak is het gemiddeld aantal waarnemingen per whole-plot een redelijke benadering voor ϕ (de gemiddelde omvang van een nest is 11.926). Door in het linkerlid verwachtingen te vervangen door de bijbehorende gemiddelde kwadraatsommen verkrijgen we een stelsel vergelijkingen waaruit we de schattingen $\tilde{\sigma}_\delta^2$ en $\tilde{\sigma}_I^2$ kunnen oplossen.

Op basis van $\tilde{\sigma}^2_\delta - \sigma^2_\delta \chi^2_{df_I}/df_I$ kunnen we eenvoudig een betrouwbaarheidsinterval voor σ^2_δ afleiden. Voor σ^2_f is dit moeilijker, maar het is wel betrekkelijk eenvoudig (zie Seely en El-Bassiouni (1983) of Harville en Fenech (1985)) om een interval voor $\sigma^2_f/\sigma^2_\delta$ te construeren.
De 0.95 intervallen voor σ^2_δ en $\sigma^2_f/\sigma^2_\delta$ zijn respectievelijk (0.14; 0.19) en (0.32; 1.24). Door de eindpunten van de intervallen te vermenigvuldigen volgt een conservatief 0.90 interval voor σ^2_f: (0.05; 0.25).
Voor alle methoden liggen de uitkomsten in de betrouwbaarheidsintervallen. Voor deze dataset ontlopen de methoden elkaar niet veel.

SS_{nest} en RSS_I zijn onafhankelijke kwadratische vormen waarvan we de variantie betrekkelijk eenvoudig kunnen bepalen onder normaliteit (zie Searle (1971), p. 54 e.v.). Schattingen voor de varianties van de oplossingen van Henderson III kunnen we hieruit eenvoudig afleiden.

Voor ML kunnen onder zekere voorwaarden (zie Miller (1977)) de grote steekproefvarianties aan de Fisher informatiematrix worden ontleend. Andersons algoritme gebruikt deze matrix zodat de grote steekproefbenaderingen voor varianties en covarianties direct beschikbaar zijn.
In onderstaande tabel zijn de benaderingen voor varianties en covarianties van de schatters voor Henderson III, REML en ML weergegeven.

Tabel 2

	Henderson III	REML	ML
$\widehat{Var}\begin{pmatrix}\tilde{\sigma}^2_\delta \\ \tilde{\sigma}^2_f\end{pmatrix} \times 1000 =$	$\begin{pmatrix}0.18 & -0.015 \\ -0.015 & 1.20\end{pmatrix}$	$\begin{pmatrix}0.18 & -0.026 \\ -0.026 & 1.20\end{pmatrix}$	$\begin{pmatrix}0.18 & -0.017 \\ -0.017 & 0.70\end{pmatrix}$

Henderson III en REML lijken op elkaar. Daar de ML-schatters erg onzuiver kunnen zijn is de uitkomst 0.70 voor $Var(\tilde{\sigma}^2_f)$ bij de ML-methode waarschijnlijk niet erg bruikbaar voor een dataset van deze omvang.

In de iteratieve procedures in de laatste vier kolommen van tabel 1 worden in iedere iteratieslag de mixed model vergelijkingen opgelost. De voorspellingen $\tilde{b}_{ik}$ en $\tilde{e}_{ijkl}$ uit de laatste iteratieslag kunnen we gebruiken voor eenvoudige controles op uitbijters en normaliteit. In figuur 1 zijn de gestandaardiseerde voorspellingen

$$\tilde{b}_{ik}/\sqrt{\left(\sum_{i,k}\tilde{b}^2_{ik}/27\right)}$$

uitgezet tegen de normal scores. Onder normaliteit moet het resultaat op een rechte lijn lijken.

Aan de hand van de resultaten van REML gaan we nu nader op de schattingen voor de fixed effecten in.
We bepalen schattingen voor de fixed effecten volgens (3). De bijbehorende varianties en covarianties volgen uit

$$\widehat{Var}(\tilde{\vec{\alpha}}) = (X'V^{-1}X)^{-1}$$

Voor V nemen we de benadering $\tilde{V} = ZZ' \tilde{\sigma}_l^2 + I \tilde{\sigma}_\delta^2$. De schattingen voor de fixed effecten zijn zuiver, zie Kackar en Harville (1981).

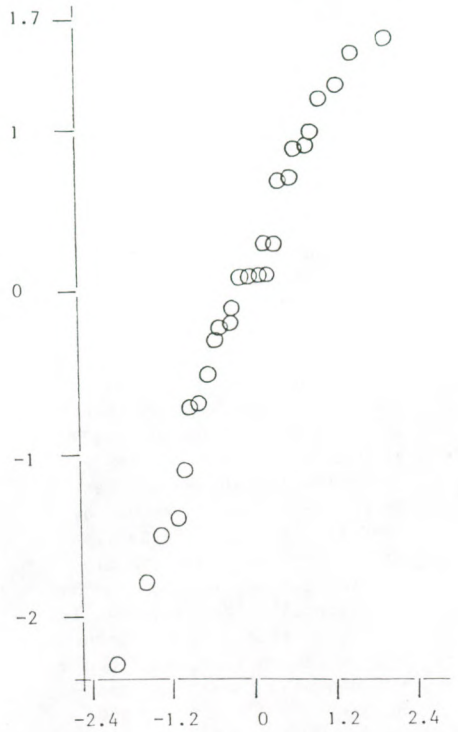


Fig. 1: De gestandaardiseerde voorspelde nestbijdragen uitgezet tegen de normal scores

In de eerste kolom van tabel 3 zijn de schattingen voor de fixed effecten en de bijbehorende standaardafwijkingen gegeven.

Op basis van de normale benadering zijn alle effecten significant. Voor het dosiseffect kunnen we ook een chi-kwadraattoets afleiden door de met dosis corresponderende contrasten in de vorm van een vector van lengte 2, zeg $\vec{x}$, uit $\tilde{\vec{\alpha}}$ te nemen en de bijbehorende matrix $\mathrm{Var}(\vec{x})$ uit $\mathrm{Var}(\tilde{\vec{\alpha}})$ te halen:

$$\vec{x}' \, \mathrm{Var}(\vec{x})^{-1} \, \vec{x}$$

Het resultaat is 23.19 met 2 vrijheidsgraden.

Bij het bekijken van de fixed effecten hebben we schattingen voor de variantiecomponenten gebruikt. Om het effect hiervan op onze conclusies na te gaan zijn de schattingen en bijbehorende standaardafwijkingen ook bepaald voor vier combinaties ontleend aan de eindpunten van de 0.95-betrouwbaarheidsintervallen voor σ_δ^2 en $\sigma_l^2/\sigma_\delta^2$.

Tabel 3: Schattingen en bijbehorende standaardafwijkingen (tussen haakjes) voor de fixed effecten.

$\sigma_l^2/\sigma_\delta^2$ =	0.598	0.32	1.24	0.32	1.24
σ_δ^2 =	0.1628	0.14	0.14	0.19	0.19
Regressie-coëfficiënt voor nest omvang	-0.13(0.019)	-0.13(0.014)	-0.13(0.023)	-0.13(0.016)	-0.13(0.027)
Sexeverschil mannetjes - vrouwtjes	-0.36(0.047)	-0.35(0.044)	-0.37(0.044)	-0.35(0.051)	-0.37(0.051)
Dosis					
laag-controle	-0.43(0.150)	-0.43(0.107)	-0.43(0.191)	-0.43(0.125)	-0.43(0.225)
hoog-controle	-0.86(0.182)	-0.87(0.132)	-0.85(0.232)	-0.87(0.153)	-0.85(0.270)

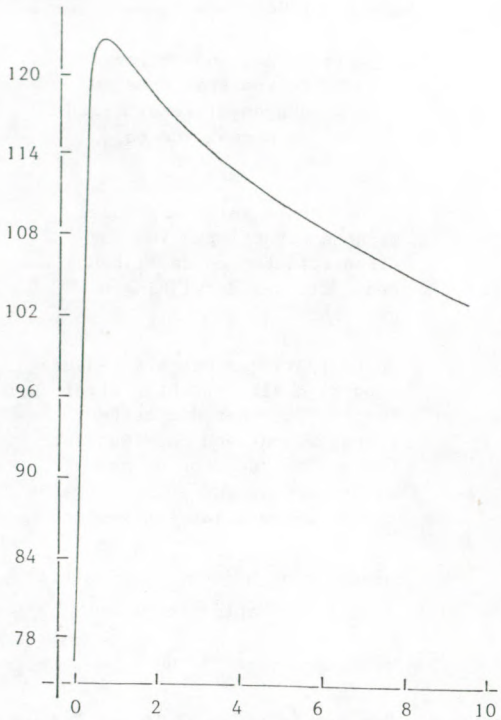


Fig. 2: De loglikelihood L_1 van REML uitgezet tegen het quotiënt $\sigma_\gamma^2/\sigma_\delta^2$

De resultaten zijn weergegeven in de kolommen 2, 3, 4 en 5 van tabel 3. De schattingen voor de effecten lopen niet veel uiteen, de bijbehorende standaardafwijkingen wèl. Ten aanzien van de conclusies is er in dit geval weinig reden voor ongerustheid.

Giesbrecht en Burns (1985) gebruiken een chi-kwadraat verdeling als benadering voor de verdeling van de schatting van de variantie van een contrast. In dat geval werken we met de verdeling van Student in plaats van de Normale verdeling bij het toetsen en het construeren van betrouwbaarheidsintervallen voor contrasten. De p-waarden voor effecten in tabel 3 bedragen dan (van boven naar beneden): 0.00, 0.00, 0.01 en 0.00.

Kackar en Harville (1984) geven schattingen voor standaardafwijkingen behorende bij schattingen voor de fixed effecten waarin de extra spreiding die volgt uit het schatten van de variantiecomponenten is doorberekend.

Wanneer dat redelijk mogelijk lijkt dienen we altijd te controleren of de REML of ML oplossing correspondeert met een absoluut maximum. We doen dit hier voor REML. Gegeven het quotiënt $\sigma_\gamma^2/\sigma_\delta^2$ is het betrekkelijk eenvoudig om analytisch de waarde voor σ_δ^2 te bepalen waarvoor de loglikelihood L_1 maximaal is. Deze optimale waarde voor L_1 kunnen we met GENSTAT eenvoudig bepalen (Barendregt (1985)). In figuur 2 is dit maximum uitgezet tegen $\sigma_\gamma^2/\sigma_\delta^2$.

In figuur 3 zien we een hoogtekaart van de loglikelihood L_1 als functie van σ_δ^2 en σ_γ^2. We zien dat er maar één maximum is en dat maximum hebben we gevonden.

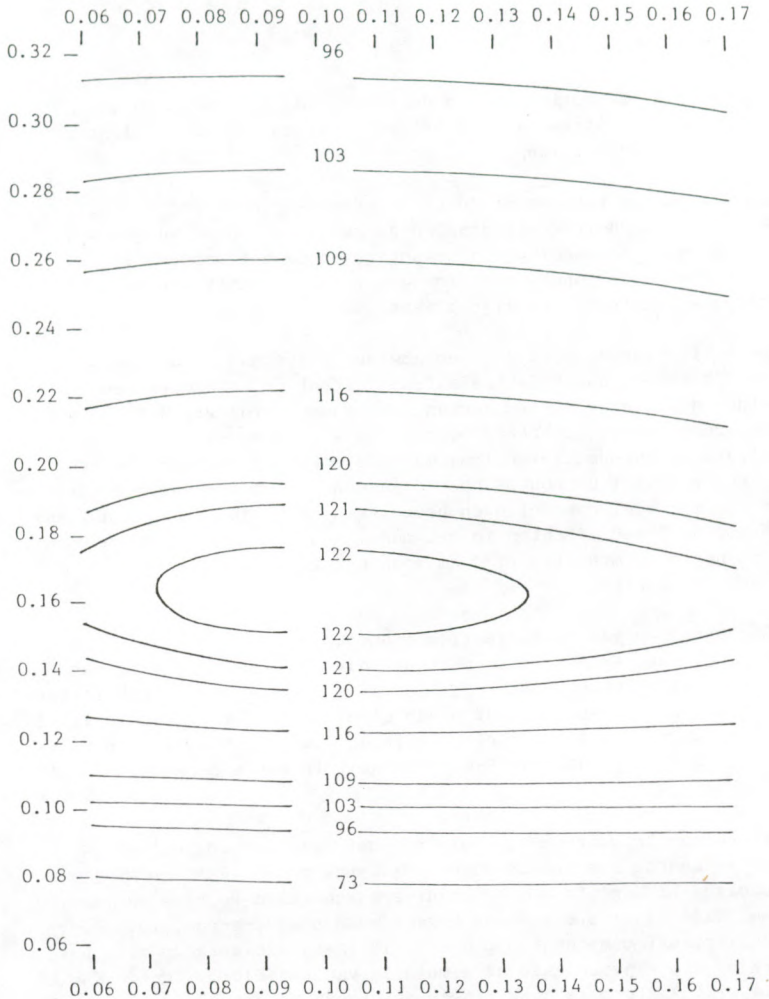


Fig. 3: Een contourplot van L_1 als functie van de variantiecomponenten σ_δ^2 (verticaal) en σ_τ^2 (horizontaal).

11 DISCUSSIE

Wanneer we over schattingen voor de variantiecomponenten beschikken kunnen we hiermee schattingen voor de fixed effecten en voorspellingen voor de randomeffecten afleiden.

Voor het uitvoeren van toetsen op de fixed effecten (rasverschillen, sexeverschillen, behandelingseffecten en dergelijke) kunnen we gebruik maken van de normale benadering zoals aangegeven in het voorbeeld. Een controle in hoeverre de conclusies afhangen van de schattingen voor de variantiecomponenten is dan op zijn plaats.

Wanneer een variantiecomponent zeer onnauwkeurig geschat is kunnen we soms door het bijbehorende effect als fixed effect te beschouwen een toets afleiden die niet van deze variantiecomponent afhangt. Het sexeverschil in 10 zouden we bijvoorbeeld binnen nesten kunnen schatten. Dit betekent uiteraard informatieverlies, maar de toets op sexeverschillen binnen moeders zal veel betrouwbaarder zijn wanneer we over betrekkelijk weinig moeders met veel nakomelingen beschikken. Er zijn dus verschillende manieren om de fixed effecten in het model aan te pakken, over het algemeen zal het niet moeilijk zijn om voor een gegeven situatie een geschikte aanpak te kiezen.

Waar het het schatten van de variantiecomponenten betreft ligt de situatie anders. Onder de verschillende methoden die we hebben bekeken, namelijk Henderson III, MINQUE, MINQUE-0, REML en ML, is noch op theoretische gronden, noch op grond van resultaten van simulatiestudies een duidelijke winnaar aan te wijzen. Wel is er een verliezer namelijk MINQUE-0. Het treft daarom ongelukkig dat MINQUE-0 de default in het SAS-pakket is (SAS (1982)).

REML lijkt een aardig compromis in die zin dat deze methode duidelijk gerelateerd is aan de overige methoden. De keuze tussen REML en ML is geen eenvoudige. REML geeft minder onzuivere schatters voor de variantiecomponenten, maar ten koste van een hogere mean-square-error. Voor functies van de variantiecomponenten behoeft dit laatste overigens niet het geval te zijn, zie bijvoorbeeld de simulatie van een BIB-design in Kackar en Harville (1984). Toch is de kleinere onzuiverheid van REML wel wat waard; in het voorbeeld in 10 is het maar de vraag wat we aan de uitkomst 0.70 voor de variantie van $\tilde{\sigma}_f^2$ bij ML hebben. Misschien is REML door zijn sterke verbondenheid met I-MINQUE robuuster tegen afwijkingen van normaliteit dan ML, hoewel ook ML, zij het op een enigszins getruceerde wijze, uit het dispersion mean model kan worden afgeleid. Als benadering voor de REML- oplossing zouden we voor grote datasets 2 MINQUE-stappen kunnen nemen, zie Giesbrecht en Burns (1985).

Henderson III is theoretisch de minst elegante methode. Dat de methode in sommige gevallen niet uniek is, zal de gemoedsrust van de gebruiker, die de collectie van kwadraatsommen moet kiezen, geen goed doen. Aan de andere kant is de methode vrij recht toe recht aan en zeker voor split-plot en split-split-plot designs, waar geen problemen liggen wat betreft het

kiezen van de kwadraatsommen, erg aantrekkelijk. Bovendien is de methode voor een aantal standaard situaties geprogrammeerd in het LSML76 programma (Harvey (1977))wat onlangs, na een oppervlakkige face-lift, aan het SAS-pakket is toegevoegd.

Een toets die nagaat of een variantiecomponent significant van nul verschilt is betrekkelijk eenvoudig wanneer we, zoals in 10, naast de restbijdragen maar één randomeffect hebben. In het algemeen ligt het niet altijd zo eenvoudig. De lezer hoede zich in dit verband voor de likelihood-ratio-test zoals die bijvoorbeeld in BMDP (BMDP (1981)) is opgenomen, de toetsingsgrootheid is niet noodzakelijk asymptotisch chi-kwadraat verdeeld, zie Miller (1977).

Alvey, N. et al. (1977). "GENSTAT, a General Statistical Program". Rothamsted Experimental Station, Harpenden, England.

Barendregt, Leo G. (1985). Persoonlijke mededeling.

BMDP Statistical software 1981. W.J. Dixon, chief editor, University of California press.

Corbeil, R.R. en Searle, S.R. (1976). "A Comparison of Variance Component Estimators". Biometrics 32, p. 779-791.

Dempster, Arthur P., Selwyn, Murray R., Patel, Chandu M., Roth, Arthur J.(1984). "Statistical and Computational Aspects of Mixed Model Analysis". Appl. Stats. 33, p. 203-214.

Engel, B., Van den Bol, M.E., Vereijken, P.F.G. (1985). "Analyse van een mixed model", rapport IWIS-TNO (thans ITI-TNO) A 85 ST 59

Giesbrecht, F.G., Burns, J.C. (1985). "Two-Stage Analysis Based on a Mixed Model: Large-Sample Asymptotic Theory and Small-Sample Simulation Results". Biometrics 41, p. 477-486.

Harvey, Walter R. (1977). "User's guide for LSML 76", Ohio State University.

Harville, David A. en Fenech, A.P. (1985). "Confidence Intervals for a Variance Ratio, or for Heritability, in an Unbalanced Mixed Linear Model". Biometrics 41, p. 137-152.

Harville, David A. (1975). "Maximum Likelihood Approaches to Variance Component Estimation and to Related Problems". Technical Report nr. 75-01175, Aerospace Research Laboratories. Wright-Patterson AFB, Ohio.

Harville, David A. (1977). "Maximum Likelihood Approaches to Variance Component Estimation and to Related Problems". JASA 72, p. 320-340.

Kackar Raghu N. en Harville David A. (1981). "Unbiasedness of two-stage estimation and prediction procedures for mixed linear models". Commun. Statist. Theor. Meth., A10 (13) p. 1249-1261.

Kackar Raghu N. en Harville David A. (1984). "Approximations for standard errors of estimates of fixed and random effects in mixed linear models". JASA 79, p. 853-862.

Lee, Kwan R. en Kapadia, C.H. (1984). "Variance Components Estimators for the Balanced Two-Way Mixed Model". Biometrics 40, p. 507-512.

Miller, John J. (1977). "Asymptotic Properties of Maximum Likelihood Estimates in the Mixed Model of the Analysis of Variance". The Annals of Stats., 5, p. 746-762.

Miller, John J. (1979). "Maximum Likelihood Estimation of Variance Components - A Monte Carlo Study". Journal of Statistical Computation and Simulation 8, p. 175-190.

SAS User's Guide: Statistics, 1982 Edition, Cary, NC: SAS Institute Inc.

Searle, S.R. (1971). "Linear Models", John Wiley.

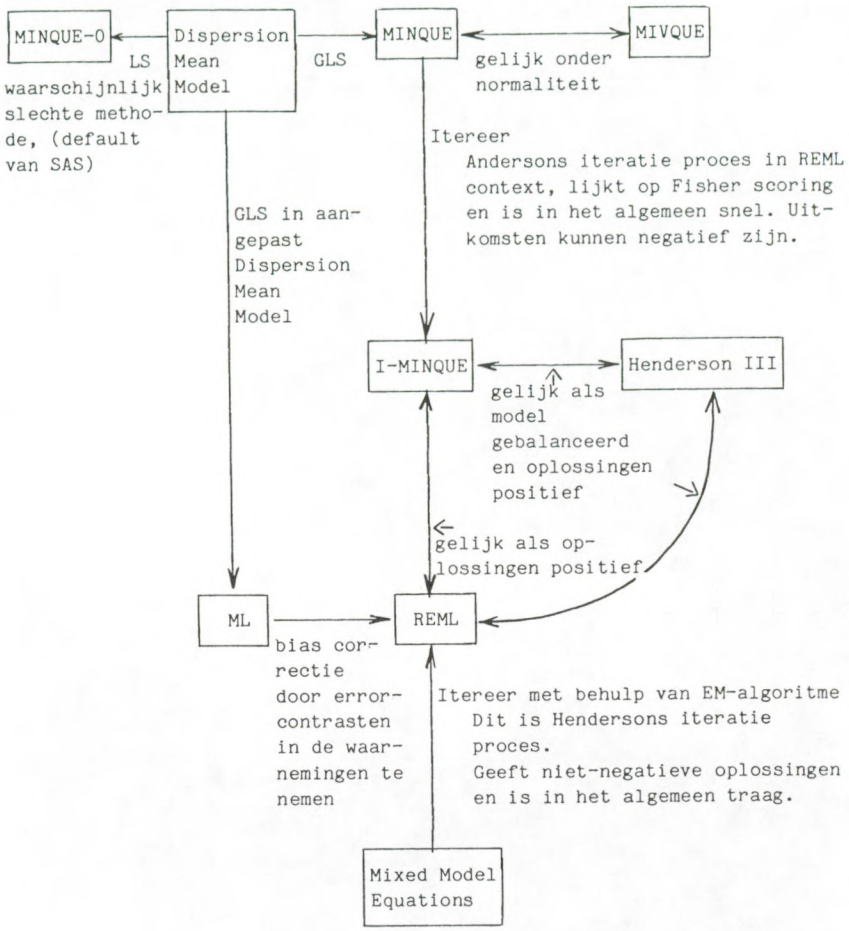
Searle, S.R. (1979). "Notes on Variance Component Estimation: a detailed account of maximum likelihood and kindred methodology". Biometrics Unit, Warren Hall, Cornell University, Ithaca, New York, paper BU-673-M.

Seely, Justus F. en El-Bassiouni, Yahia (1983). "Applying Wald's Variance Component test." The Annals of Stats. 11, p. 197-201.

Swallow, William H. en Monahan, John F. (1984). "Monte Carlo Comparison of ANOVA, MIVQUE, REML and ML Estimators of Variance Components". Technometrics 26, p. 47-57.

AO - OVERZICHT VAN DE SCHATTINGSMETHODEN

Fig. 4



Afkortingen:

LS	= Least Squares = kleinste kwadraten methode
GLS	= Generalized Least Squares = Gegeneraliseerde kleinste kwadraten methode
MINQUE	= Minimum Norm Quadratic Unbiased Estimation
MIVQUE	= Minimum Variance Quadratic Unbiased Estimation
ML	= Maximum Likelihood
REML	= Restricted Maximum Likelihood