KM 31(1989)
pag 165-170

REML - Een programma voor REstricted Maximum Likelihood

Bas Engel en Paul Goedhart*

1. Restricted Maximum Likelihood

Restricted (of Residual) Maximum Likelihood (REML) is een techniek voor het schatten van variantiecomponenten in gemengde modellen. Dit zijn lineaire modellen van de vorm:

$$\underline{y} = X\alpha + Z_1\underline{b}_1 + Z_2\underline{b}_2 + \ldots + Z_c\underline{b}_c + \underline{e}$$

De vector van n waarnemingen $\underline{y}$ is de som van een aantal systematische bijdragen (fixed effecten) en een aantal toevalsbijdragen (random effecten). De parameter-vector α voor de fixed effecten bestaat uit bijdragen van factoren in de vorm van hoofdeffecten en interacties en eventueel uit bijdragen van co-variabelen. De toevalsbijdragen $\underline{b}_1, \ldots, \underline{b}_c$ en $\underline{e}$ (de gebruikelijke vector van residuele bijdragen) worden onafhankelijk normaal verdeeld verondersteld met varianties respectievelijk $\sigma_1^2, \ldots, \sigma_c^2$ en σ_0^2. Deze varianties worden de variantiecomponenten genoemd.

Bij REML worden de variantiecomponenten geschat met behulp van de maximum likelihood methode toegepast op de likelihood van een vector van contrasten $Q\underline{y}$. De matrix Q is zodanig gekozen dat rang(Q)=n-rang(X) en QX=0. De matrix Q is niet uniek maar dat geeft verder geen problemen.

REML stemt overeen met geïtereerde versies van Rao's MINQUE (Minimum Quadratic Unbiased Estimation) en La Motte's MIVQUE (Minimum Variance Quadratic Unbiased Estimation), wanneer voor de laatste methoden de schattingen voor de variantiecomponenten positief zijn. In gebalanceerde modellen stemmen REML en de gebruikelijke ANOVA aanpak op basis van kwadraatsommen voor de toevalsbijdragen $\underline{b}_1, \ldots, \underline{b}_c$ en $\underline{e}$ overeen, wanneer de ANOVA schattingen positief zijn. REML kan ook afgeleid worden als uitkomst van een EM-algoritme.

---

* Groep LandbouwWiskunde, Postbus 100, 6700 AC Wageningen,
tel. 08370 - 19100.

Gegeven de schattingen voor de variantiecomponenten kunnen schattingen voor de elementen van α en voorspellingen voor de elementen van $\underline{b}_1,\ldots,\underline{b}_c$ en $\underline{e}$ worden afgeleid. Voor details zie Engel, v.d. Bol en Vereijken (1986).

2. Het REML programma

De numerieke problemen die bij gebruik van REML een rol spelen zijn groot. Zo groot, dat aan de mogelijkheid van een algemeen bruikbaar pakket vaak is getwijfeld. Niettemin is het de Scottish Agricultural Unit of Statistics van de Universiteit van Edinburgh gelukt om een 'general purpose' programma te maken waarmee de doorsnee gebruiker een heel eind kan komen. Voor de zeer grote datasets die bijvoorbeeld bij fokkerij onderzoek voorkomen zal het programma waarschijnlijk te beperkt zijn. Echter, Robinson (1987) maakt melding van een dataset van meer dan 9000 waarnemingen, waarbij een der factoren 1400 niveaus telt, die met het programma REML is geanalyseerd. Een dergelijke capaciteit zal voor de meeste gebruikers voldoende zijn.

Voor het optimaliseren van de likelihood van de contrasten $Q\underline{y}$ gebruikt REML Fishers score methode. Door handig gebruik te maken van de structuur van het model kan in vele gevallen de opslag en inversie van (te) grote matrices worden vermeden. De data worden dan meerdere malen per niveau van de factor met het grootste aantal niveaus ingelezen en de berekeningen worden dan in handzame stukjes uitgevoerd. Deze factor wordt de 'absorbing factor' genoemd.

Gebruikers van statistische pakketten als GENSTAT, BMDP en SPSS zullen weinig moeite hebben hun probleem met REML op te lossen. Het programma wordt gestuurd door 'directives' die elk uit een directive naam gevolgd door 1 of meer parameters bestaan. Een REML programma heeft de volgende structuur:

1. Declaratie en inlezen van variabelen en factoren.
2. Modelbeschrijving.
3. Beschrijving van het iteratieproces en uitvoer tijdens dit proces.
4. Gewenste uitvoer na het iteratieproces.

De uitvoer kan bestaan uit schattingen voor variantiecomponenten en fixed effecten en voorspellingen voor de toevalsbijdragen. Standaardafwijkingen van schattingen en voorspellingen zijn ook opvraagbaar. Paarsgewijze vergelijkingen tussen combinaties van factoren en toetsen op andere contrasten kunnen eenvoudig worden uitgevoerd. Het model kan worden gecontroleerd met behulp van enige plots. Afhankelijk van het aantal

waarnemingen, het aantal factoren en het aantal niveaus kan de rekentijd variëren van enige minuten tot een aantal uren.

3. Tekortkomingen

Het programma heeft de volgende nadelen. Er is geen enkele vorm van data-manipulatie mogelijk. Zelfs een eenvoudige log-transformatie moet buiten het pakket uitgevoerd worden. Simultaan toetsen van contrasten (zoals een F-toets in ANOVA) is niet mogelijk. In sommige situaties zijn de plots misleidend. Overigens wordt de Fortran code meegeleverd, zodat een handige programmeur het programma aan zijn eigen wensen kan aanpassen.

4. Beschikbaarheid

Het programma is verkrijgbaar in Fortran-IV en Fortran-77. De prijs bedraagt f100 voor academische instellingen en £200 voor overige instellingen. Deze prijzen zijn echter onder revisie. Voor bestelling van het programma of meer informatie kan men zich wenden tot:

D.L. Robinson, AFRC, Unit of Statistics, University of Edinburgh,
James Clerk Maxwell Building, The King's Buildings,
Mayfield Road, Edinburgh EH9 3JZ, Great Britain.

Bij de Groep Landbouwwiskunde is het programma geïnstalleerd op een VAX computer en is een DCL-commando procedure geschreven die het gebruik van REML vergemakkelijkt.

5. Een voorbeeld

Om een idee te geven van het programma wordt voor het volgende probleem een REML programma plus uitvoer gegeven. Het doel van het volgende fictieve experiment is om de opbrengst van 4 aardappelrassen (A, B, C en D) te vergelijken. Tevens zijn niveaus F1 en F2 van een bemestingsstof in het experiment opgenomen. Voor het experiment zijn 10 blokken met elk 3 plots beschikbaar. De proefopzet en de opbrengst per veldje zijn als volgt:

blok	1	2	3	4	5	6	7	8	9	10
	A 22	B 37	D 32	C 16	B 26	A 46	C 43	D 44	B 30	A 38
	C 21	D 45	C 29	D 16	A 22	D 46	D 40	B 45	A 38	B 23
	B 16	A 48	A 37	B 11	C 24	B 44	A 36	C 49	C 28	D 32
mest	F1	F1	F1	F1	F2	F2	F2	F2	F1	F1

De factor bemesting ligt dus op blokniveau en de factor ras op plotniveau. De blokken 1-4 en 5-8 vomen elk een gebalanceerde incomplete blokkenproef met betrekking tot ras. De behandelingen in blokken 9 en 10 zijn zo gekozen omdat men speciaal geïnteresseerd is in een vergelijking van ras A met ras B bij het lage bemestings niveau. Een model voor de waarnemingen is:

$$\underline{y}_{ijk} = \mu + \text{mest}_i + \underline{\text{blok}}_j + \text{ras}_k + \text{mest.ras}_{ik} + \underline{e}_{ijk}$$

Een REML programma om deze dataset te analyseren kan er als volgt uitzien.

```
'TITLE'                                  Naam voor het programma
  Split-plot proef

'UNIT' 30                                30 experimentele eenheden
'FACTOR'                                 Declaratie van factoren
  Ras    4    A  B  C  D                    Ras, 4 niveaus, namen A,B,C,D
  Mest   2    F1  F2                       Mest, 2 niveaus, namen F1,F2
  Blok  10    *                            Blok, 10 niveaus, 'namen' 1..10
'VARIATE' Opbrengst                      Declaratie van vector Opbrengst

'READFREE'  4                            Inlezen van 4 variabelen/factoren
   Opbrengst Ras Mest Blok                 Gegevens kunnen ook op een
             22   A   F1    1             externe file staan.
             21   C   F1    1
             16   B   F1    1
              .   .    .    .
              .   .    .    .             Gegevens per experimentele
              .   .    .    .             eenheid
             37   A   F1   10
             22   B   F1   10
             31   D   F1   10

'RANDOM' Blok
'FIXED'  Ras + Mest + Ras.Mest           Specificatie van het model
'DEPENDENT' Opbrengst

'MAXITER'  10                            Maximale aantal iteraties is 10
'INITIAL'  -2   70   0.8                 Beginschattingen voor componenten
'PRINT' 1                                Begin uitvoer directives
'DEC'   5                                Uitvoer met 5 decimalen
'SE'    3                                Afdrukken van standaardafwijkingen
'COMPONENTS'                             Afdrukken van componenten

'AT END'                                 Verdere uitvoer na convergentie
'MEAN EFF'  Ras  Mest  Ras.mest          Schattingen voor fixed effecten
'SE'  0                                  Verder geen uitvoer van stan.afw.
'MEAN EFF'  Blok                         Voorspellingen voor Blok bijdragen
'ENDPRINT'                               Einde uitvoer directives

'GO'                                     Einde van het programma
```

Uitvoer van het REML programma:

```
ITERATION NO     1

  Split-plot proef

 ESTIMATED COMPONENTS OF VARIANCE
     Blok       SIGMA
                SQUARED
         95.25       5.36
SE      235.54       2.03
ITERATION NO     2

  Split-plot proef

 ESTIMATED COMPONENTS OF VARIANCE
     Blok       SIGMA
                SQUARED
         95.23       5.36
SE       48.61       2.03
ITERATION NO     3

  Split-plot proef   ITERATIONS HAVE CONVERGED

 ESTIMATED COMPONENTS OF VARIANCE
     Blok       SIGMA
                SQUARED
         95.23       5.36
SE       48.60       2.03


 MEAN EFFECTS (B.L.U.E.'s) of Ras

  A          B          C          D          MARGIN
   35.47      31.22      34.22      34.23      33.78

          Ras
 STANDARD ERRORS OF DIFFERENCES BETWEEN PAIRS
           A          B          C
B            1.26
C            1.30       1.30
D            1.30       1.30       1.34


 MEAN EFFECTS (B.L.U.E.'s) of Mest

  F1         F2         MARGIN
   28.82      38.75      33.78

          Mest
 STANDARD ERRORS OF DIFFERENCES BETWEEN PAIRS
           F1
F2          6.36
```

MEAN EFFECTS (B.L.U.E.'s) of Ras BY Mest

	F1	F2	MARGIN
A	34.10	36.83	35.47
B	23.83	38.62	31.22
C	27.36	41.07	34.22
D	29.98	38.48	34.23
MARGIN	28.82	38.75	33.78

STANDARD ERRORS OF DIFFERENCES BETWEEN PAIRS Ras BY Mest

		A F1	A F2	B F1	B F2	C F1	C F2	D F1
A	F2	6.54						
B	F1	1.51	6.54					
B	F2	6.54	2.00	6.54				
C	F1	1.65	6.57	1.65	6.57			
C	F2	6.54	2.00	6.54	2.00	6.57		
D	F1	1.65	6.57	1.65	6.57	1.79	6.57	
D	F2	6.54	2.00	6.54	2.00	6.57	2.00	6.57

MEAN EFFECTS (B.L.U.P.'s) of Blok

1	2	3	4	5	6	7	8	9	10
25.18	47.56	35.93	21.30	19.22	41.01	34.64	40.27	37.29	35.45

Referenties

Bas Engel, Marion van den Bol, Pieter Vereijken (1986). "Analyse van een mixed model", Kwantitatieve Methoden, 22, p.35-59.

D.L. Robinson (1987). "Estimation and use of variance components". The Statistician, 36, p.3-14.