VARIANCE COMPONENTS FOR SIBGROUPS

JAN DE LEEUW

ABSTRACT. We outline the computation of both FIML and REML estimates in twin studies of a single quantitative characteristic, in the case in which there are no covariates. A program in R is provided.

1. MODEL AND NOTATION

Suppose we have $j = 1, \cdots, m$ groups of pairs of individuals, with group j having n_j pairs. For each of the $i = 1, \cdots, n_j$ pairs of individuals in group j we have

$$(\underline{x}_i, \underline{y}_i) \sim \mathcal{N}\left(\begin{bmatrix}\mu_j\\ \mu_j\end{bmatrix}, \begin{bmatrix}\sigma_j^2 & \omega_j^2\\ \omega_j^2 & \sigma_j^2\end{bmatrix}\right).$$

The interpretation is that individuals within pairs are exchangeable, in the sense that there is no "first" and no "second" member of the pair, and the labeling of an individual as either $\underline{x}_i$ or $\underline{y}_i$ is arbitrary. The different groups of pairs can be thought of as having a different level of "relatedness". The groups can be, for example, monozygotic twins, dizygotic twins, non-twin siblings, and so on. The treatment is easily extended from pairs to sibgroups of more than two individuals.

When calculating within a model we underline the hypothetical random variables whose realizations we observe. The realizations

Date: April 14, 2007.
Key words and phrases. Variance Components, Mixed Linear Models, Maximum Likelihood, REML, Twin Studies.

themselves use the same symbol as we use for the random variable, but without the underlining. In addition we use overlining for means of vectors of observed quantities, and tildes for the elements of the vector in deviations from the mean.

2. Single Group Full Information Maximum Likelihood or FIML

Consider the situation in which we have only a single group of n pairs (or in which we analyze each group of pairs separately). Before we give the likelihood, we transform to a new set of variables.

$$\underline{u}_i = \frac{1}{2}(\underline{x}_i - \underline{y}_i),$$
$$\underline{v}_i = \frac{1}{2}(\underline{x}_i + \underline{y}_i).$$

All $2n$ transformed random varables are independent. The $\underline{u}_i$ have expectation zero and variance $\frac{1}{2}(\sigma^2 - \omega^2)$, and the $\underline{v}_i$ have expectation μ and variance $\frac{1}{2}(\sigma^2 + \omega^2)$.

The negative log-likelihood becomes

$$\mathcal{D}(\sigma^2, \omega^2) = n\sum_{i=1}^{n} \log\frac{1}{2}(\sigma^2 + \omega^2) + n\sum_{i=1}^{n} \log\frac{1}{2}(\sigma^2 - \omega^2) +$$
$$+ \frac{\sum_{i=1}^{n} u_i^2}{\frac{1}{2}(\sigma^2 - \omega^2)} + \frac{\sum_{i=1}^{n}(v_i - \mu)^2}{\frac{1}{2}(\sigma^2 + \omega^2)}.$$

Thus the FIML estimate of μ is

$$\hat{\mu} = \overline{v} = \frac{1}{2}(\overline{x} + \overline{y}),$$

which is the mean of all $2n$ measurements.

The unrestricted FIML estimates of the variance components can be computed from

$$\frac{1}{2}\widehat{(\sigma^2 + \omega^2)} = \frac{1}{n}\sum_{i=1}^{n} \tilde{v}_i^2,$$

and

$$\frac{1}{2}\widehat{(\sigma^2-\omega^2)}=\frac{1}{n}\sum_{i=1}^{n}u_i^2.$$

The non-negative ML estimate of the intra-class correlation is

$$\begin{aligned}\hat{\rho}&=\max\left(0,\frac{\hat{\omega}^2}{\hat{\sigma}^2}\right)=\\&=\max\left(0,\frac{\sum_{i=1}^{n}\tilde{v}_i^2-\sum_{i=1}^{n}u_i^2}{\sum_{i=1}^{n}\tilde{v}_i^2+\sum_{i=1}^{n}u_i^2}\right).\end{aligned}$$

To translate this into the usual ANOVA terminology, observe that the sums of squares within and between groups are

$$\begin{aligned}SS_W&=2\sum_{i=1}^{n}u_i^2,\\SS_B&=2\sum_{i=1}^{n}\tilde{v}_i^2.\end{aligned}$$

Thus the mean squares are

$$\begin{aligned}MS_W&=\frac{2}{n}\sum_{i=1}^{n}u_i^2,\\MS_B&=\frac{2}{n-1}\sum_{i=1}^{n}\tilde{v}_i^2,\end{aligned}$$

and

$$\hat{\rho}=\max\left(0,\frac{(n-1)MS_B-nMS_W}{(n-1)MS_B+nMS_W}\right).$$

3. Single Group Restricted Maximum Likelihood or REML

Here we use a different transformation of the variables. Suppose k_0 is a vector of length n with all elements equal to $\frac{1}{\sqrt{n}}$. Moreover $k_1,\cdots,k_{n-1}$ are vectors of length one, orthogonal to each other, and orthogonal to k_0. Define for $i = 1, \cdots, n-1$

$$\begin{aligned}\underline{a}_i &= \frac{k_i'(\underline{x}-\underline{z})}{\sqrt{2}},\\ \underline{b}_i &= \frac{k_i'(\underline{x}+\underline{z})}{\sqrt{2}},\end{aligned}$$

as well as

$$\underline{c} = \frac{k_0'(\underline{x}-\underline{z})}{\sqrt{2}}.$$

All $2n-1$ new variables have expectation zero, and are independent. The variance of the $\underline{a}_i$ and of $\underline{c}$ is $\sigma^2 - \omega^2$, while that of the $\underline{b}_i$ is $\sigma^2 + \omega^2$.

Thus for REML

$$\frac{1}{2}(\widehat{\sigma^2+\omega^2}) = \frac{1}{2(n-1)}\sum_{i=1}^{n-1} b_i^2 = \frac{1}{n-1}\sum_{i=1}^{n} \tilde{v}_i^2$$

while

$$\frac{1}{2}(\widehat{\sigma^2-\omega^2}) = \frac{1}{2n}(c^2+\sum_{i=1}^{n-1} a_i^2) = \frac{1}{n}\sum_{i=1}^{n} u_i^2.$$

The non-negative REML estimate of the intra-class correlation is

$$\hat{\rho} = \max\left(0, \frac{\frac{n}{n-1}\sum_{i=1}^n \tilde{v}_i^2 - \sum_{i=1}^n u_i^2}{\frac{n}{n-1}\sum_{i=1}^n \tilde{v}_i^2 + \sum_{i=1}^n u_i^2}\right) = \max\left(0, \frac{MS_B - MS_W}{MS_B + MS_W}\right).$$

For small n the REML and ML estimates can differ.

4. Multiple Group FIML

Suppose we have n MZ twins and m DZ twins. If the number of pairs is small, it may be useful to fit a more restrictive model with a smaller number of parameters.

So suppose both types of twins have the same means μ, the same variances σ^2, but different covariances ω_{MZ}^2 and ω_{DZ}^2. Transform to independent variables as before. The negative log-likelihood becomes

$$\begin{aligned}\mathcal{D}(\mu,\sigma^2,\omega^2_{MZ},\omega^2_{DZ}) &= n\log\frac{1}{2}(\sigma^2+\omega^2_{MZ}) + n\log\frac{1}{2}(\sigma^2-\omega^2_{MZ})+\\ &+ m\log\frac{1}{2}(\sigma^2+\omega^2_{DZ}) + m\log\frac{1}{2}(\sigma^2-\omega^2_{DZ})+\\ &+ \frac{\sum_{i=1}^n u_i^2}{\frac{1}{2}(\sigma^2-\omega^2_{MZ})} + \frac{\sum_{i=1}^n (v_i-\mu)^2}{\frac{1}{2}(\sigma^2+\omega^2_{MZ})}+\\ &+ \frac{\sum_{p=n+1}^{n+m} u_i^2}{\frac{1}{2}(\sigma^2-\omega^2_{DX})} + \frac{\sum_{p=n+1}^{n+m} (v_i-\mu)^2}{\frac{1}{2}(\sigma^2+\omega^2_{DZ})}.\end{aligned}$$

This must be minimized over $\sigma^2 \geq \omega^2_{MZ} \geq \omega^2_{DZ}$. It requires a (simple) iterative algorithm using either scoring or Newton. Obvious initial estimates are

$$\begin{aligned}\hat{\mu} &= \frac{1}{n+m}\sum_{i=1}^{n+m} v_i = \overline{v},\\ \hat{\sigma}^2 &= \frac{1}{n+m}\left(\sum_{i=1}^{n+m} u_i^2 + \sum_{i=1}^{n+m} \tilde{v}_i^2\right),\\ \hat{\omega}^2_{MZ} &= \frac{1}{n}\left(\sum_{i=1}^{n} u_i^2 - \sum_{i=1}^{n} \tilde{v}_i^2\right),\\ \hat{\omega}^2_{DZ} &= \frac{1}{m}\left(\sum_{i=n+1}^{n+m} u_i^2 - \sum_{i=n+1}^{n+m} \tilde{v}_i^2\right).\end{aligned}$$

We can reparametrize by letting

$$\begin{aligned}\theta^2 &= \omega^2_{DZ},\\ \xi^2 &= \omega^2_{MZ} - \omega^2_{DZ},\\ \eta^2 &= \sigma^2 - \omega^2_{MZ},\end{aligned}$$

and then maximize the likelihood over these parameters, requiring that they are all non-negative.

The reparametrized negative log-likelihood is

$$\begin{aligned}\mathcal{D}(\mu,\eta^2,\xi^2,\theta^2) &= n\log(\eta^2+2\xi^2+2\theta^2)+n\log\eta^2+\\ &+m\log(\eta^2+\xi^2+2\theta^2)+m\log(\eta^2+\xi^2)\\ &+\frac{\sum_{i=1}^n u_i^2}{\eta^2}+\frac{\sum_{i=1}^n(v_i-\mu)^2}{\eta^2+2\xi^2+2\theta^2}+\\ &+\frac{\sum_{i=n+1}^{n+m} u_i^2}{\eta^2+\xi^2}+\frac{\sum_{i=n+1}^{n+m}(v_i-\mu)^2}{\eta^2+\xi^2+2\theta^2}.\end{aligned}$$

It probably makes sense in the FIML case to alternate minimization over μ for fixed variance components and minimization over the variance components for fixed μ. Of course the minimum over μ for fixed variance components is attained at

$$\hat{\mu}=\frac{\frac{\sum_{i=1}^n v_i}{\eta^2+2\xi^2+2\theta^2}+\frac{\sum_{i=n+1}^{n+m} v_i}{\eta^2+\xi^2+2\theta^2}}{\frac{n}{\eta^2+2\xi^2+2\theta^2}+\frac{m}{\eta^2+\xi^2+2\theta^2}}.$$

5. Multiple Group REML

One can repeat the previous arguments to find REML estimates in the multiple group case too. We just give the final result.

$$\begin{aligned}\mathcal{D}(\eta^2,\xi^2,\theta^2) &= (n-1)\log(\eta^2+2\xi^2+2\theta^2)+n\log\eta^2+\\ &+(m-1)\log(\eta^2+\xi^2+2\theta^2)+m\log(\eta^2+\xi^2)+\\ &+\log(\eta^2+\frac{3}{2}\xi^2+2\theta^2)+\\ &+\frac{\sum_{i=1}^n u_i^2}{\eta^2}+\frac{\sum_{i=1}^n \tilde{v}_i^2}{\eta^2+2\xi^2+2\theta^2}+\\ &+\frac{\sum_{i=n+1}^{n+m} u_i^2}{\eta^2+\xi^2}+\frac{\sum_{i=n+1}^{n+m}\tilde{v}_i^2}{\eta^2+\xi^2+2\theta^2}+\frac{c^2}{\eta^2+\frac{3}{2}\xi^2+2\theta^2}\end{aligned}$$

Here the $\tilde{v}_i$ are in deviations from the corresponding group means $\overline{v}_{MZ}$ and $\overline{v}_{DZ}$. Also

$$c=\frac{\sqrt{n}\overline{v}_{MZ}-\sqrt{m}\overline{v}_{DZ}}{\sqrt{2}}.$$

6. Newton and Scoring

We illustrate the Newton and Scoring algorithms using the REML likelihood. For FIML the situation is very similar.

Define the five column vectors

$$\begin{aligned} a_1' &= \begin{bmatrix} 1 & 2 & 2 \end{bmatrix}, \\ a_2' &= \begin{bmatrix} 1 & 0 & 0 \end{bmatrix}, \\ a_3' &= \begin{bmatrix} 1 & 1 & 2 \end{bmatrix}, \\ a_4' &= \begin{bmatrix} 1 & 1 & 0 \end{bmatrix}, \\ a_5' &= \begin{bmatrix} 1 & \frac{3}{2} & 2 \end{bmatrix}, \end{aligned}$$

and the five sums-of-squares

$$\begin{aligned} S_1 &= \sum_{i=1}^{n} u_i^2, \\ S_2 &= \sum_{i=n+1}^{n+m} u_i^2, \\ T_1 &= \sum_{i=1}^{n} \tilde{v}_i^2, \\ T_2 &= \sum_{i=n+1}^{n+m} \tilde{v}_i^2, \\ U &= c^2 \end{aligned}$$

Also define $\lambda = \begin{bmatrix} \eta^2 & \xi^2 & \theta^2 \end{bmatrix}$. Then the REML log-likelihood is

$$\begin{aligned}
\mathcal{D} = {} & (n-1)\log a_1'\lambda + \frac{T_1}{a_1'\lambda} + \\
& + n\log a_2'\lambda + \frac{S_1}{a_2'\lambda} + \\
& + (m-1)\log a_3'\lambda + \frac{T_2}{a_3'\lambda} + \\
& + m\log a_4'\lambda + \frac{S_2}{a_4'\lambda} + \\
& + \log a_i'\lambda + \frac{U}{a_5'\lambda}.
\end{aligned}$$

First derivatives are

$$\begin{aligned}
\frac{\partial\mathcal{D}}{\partial\lambda} = {} & \frac{(n-1)a_1'\lambda - T_1}{a_1'\lambda} a_1 + \\
& + \frac{na_2'\lambda - S_1}{a_2'\lambda} a_2 + \\
& + \frac{(m-1)a_3'\lambda - T_2}{a_3'\lambda} a_3 + \\
& + \frac{ma_4'\lambda - S_2}{a_4'\lambda} a_4 + \\
& + \frac{a_5'\lambda - U}{a_5'\lambda} a_5.
\end{aligned}$$

Second derivatives are

$$\begin{aligned}
\frac{\partial^2\mathcal{D}}{\partial\lambda^2} = {} & \frac{2(n-1)a_1'\lambda - T_1}{(a_1'\lambda)^2} a_1 a_1' + \\
& + \frac{2na_2'\lambda - S_1}{(a_2'\lambda)^2} a_2 a_2' + \\
& + \frac{2(m-1)a_3'\lambda - T_2}{(a_3'\lambda)^2} a_3 a_3' + \\
& + \frac{2ma_4'\lambda - S_2}{(a_4'\lambda)^2} a_4 a_4' + \\
& + \frac{2a_5'\lambda - U}{(a_5'\lambda)^2} a_5 a_5',
\end{aligned}$$

and the expected values of the second derivatives are

$$\begin{aligned}\mathbf{E}\left(\frac{\partial^2\mathcal{D}}{\partial\lambda^2}\right) = & \frac{(n-1)}{a_1'\lambda}a_1a_1'+\\ & +\frac{n}{a_2'\lambda}a_2a_2'+\\ & +\frac{m-1}{a_3'\lambda}a_3a_3'+\\ & +\frac{m}{a_4'\lambda}a_4a_4'+\\ & +\frac{1}{a_5'\lambda}a_5a_5',\end{aligned}$$

7. Confidence Intervals

From the formulas given above it is fairly easy to derive asymptotic standard errors (which work for large n and do not depend on the normality assumption) or exact standard errors, which do depend on normality.

Department of Statistics, University of California, Los Angeles, CA 90095-1554

E-mail address, Jan de Leeuw: `deleeuw@stat.ucla.edu`

URL, Jan de Leeuw: `http://gifi.stat.ucla.edu`