CS-BIGS 4(1): 44-58
http://www.bentley.edu/csbigs/documents/onyango.pdf
© 2009 CS-BIGS

On the Linear Mixed Effects Regression (lmer) R Function for Nested Animal Breeding Data

Nelson Owuor Onyango
Technical University of Munich

This work highlights aspects of the R ***lmer*** *function for a case where the dataset is nested, highly unbalanced, involves mixed effects and repeated measurements. The lmer function is part of the lme4 package of the statistical software R. The dataset used in the study is simulated from a survey of cow milk off takes from a group of Herds in Uganda, Africa. The purpose of the survey was to identify quality breeds of African Indigenous cattle for purposes of genetic breeding following the difficulties involved in implantation of foreign breeds of cattle in Africa. The work highlights the use of mixed model analysis in the context of animal breed selection. The exposition is accessible to readers with an intermediate background in statistics. Some previous exposure to R is helpful as well as some familiarity with mixed models.*

Key Words: *Mixed Models, Repeated Measures, lmer function in the R Statistical Software, Best Linear Unbiased Predictor (BLUP)*

1. INTRODUCTION

Multilevel data structure is often associated with many studies from medical, agricultural and social sciences, as scientists can capture a variety of factors at different levels of aggregation. Progress is being made day by day to capture complicated data structures. In mixed modeling, complications that arise include unbalanced structures, nesting vs. crossed structure, size of data and negativity of variance, residual analysis and diagnostics due to assumptions on the residuals and random effects(normality assumption), among others. Restricted/Residual Maximum Likelihood Estimation (REML) is well suited to handle the negativity of variance estimates, unlike ANOVA or Maximum Likelihood Estimation (MLE).

The data used in the study are simulated using information from data originally collected from a survey of cow Milk Off-takes from a group of herds in Uganda, Africa. For a comprehensive literature review of some of the key contributions to the area of mixed modeling, some suitable literature include Searle, Casella and McCulloch (1992), and Khuri and Sahai (1985). Computers have played an even a bigger role in mixed-model estimation, enabling easy handling of the large sample dispersion matrices involved. A key advance in mixed-model analysis in the R statistical software is the work of Pinheiro and Bates (2000). Thanks to their work, we use the *lmer* function in the *lme4* package here in detail. The purpose of the survey was to identify quality breeds of African indigenous cattle for the purposes of genetic breeding following the difficulties involved in the implantation of the foreign breeds of cattle in Africa.

The data are collected from three main regions, herein referred to as clusters. The clusters, just like herd groups, represent different ecological regions and vegetation types. Since the herd groups represent exhaustively the ecological differences between the regions studied, they are considered as fixed effects in the models at the analysis stage.

A fixed effect factor is a factor whose levels are the only possible levels in the population being studied. This is opposed to a random effect factor whose levels in the study are just a sample of all the other possible choices.

For example, in each herd group, a few herds were randomly chosen, out of many other herds that were present in the herd group. In the analysis, we therefore model herd as a random effect factor. Any cow studied was either in lactation number 1 or 2. Thus, Lactation Number is considered a fixed effect at the modeling stage. Then, due to repeated measurements, we have Subject and Time factors that come into play. They are considered as random effect factors, as their levels in the data are also samples of the whole population. Thus Cluster, herd group and Lactation Number are considered fixed effects while herd, Subject and Time as random effects.

The multilevel structure in which some factors are considered fixed and others random defines the mixed model scenario. The key steps of mixed model analysis involves estimating variance component parameters using Restricted Maximum Likelihood (REML), then estimating fixed effects parameters using Generalized Least Squares(GLS).

Best Linear Unbiased Predictors (BLUPs) of random effects are obtained using the obtained REML and GLS estimates. For animal breed selection, BLUPs play a very significant role. We will also refer to the GLS estimates of fixed effects as Best Linear Unbiased Estimators (BLUEs). Note that BLUP and BLUE are sufficient initials but we add small "s" to make it plural.

The R statistical software is gaining popularity among many data analysts (students and researchers). It is similar in many features to S-Plus and any experience with S-Plus is more than sufficient for using R. One can download R from the site (http://www.r-project.org). It is mainly a command language software with option for pull down menus in R-Commander, a separate package that could be downloaded alongside.

The paper is organized as follows: section 2 considers the study design and elicits the main multi-level data structures. In section three, an exploratory data analysis including a discussion on model selection is done to justify our model selection. This is followed by model specification. Section four addresses some of the theoretical technicalities involved in parameter estimation for the unbalanced multi-level nested data structure, before we fit the data to the selected model in section five to obtain the results. Most of these complications are already handled in software algorithms. We end with a discussion, highlighting the *lmer* function and its potential in comparison to *lme* function.

2. STUDY DESIGN

The primary survey dataset was collected with the aim of identifying quality breeds (high milk producers) of African indigenous cattle, for purposes of genetic breeding. Such cattle from Europe and other temperate regions have been introduced in Africa without much success due to relatively harsh climatic conditions.

The survey was conducted in Mbarara district, Uganda for a period of 12 months, among 40 cattle keepers. The eight regions studied represent different production systems and vegetation types (see Table 1). Milk off take data (amount of milk obtained at a milking moment), as opposed to Milk intake by Calves (MC) was collected from 467 Subjects (cows). Two stage cluster sampling was used to collect the data, with purposive sampling (where subjects are selected because of some characteristic) conducted at each stage. The first stage selected the eight herd groups, each representing a certain ecological environment and the second stage selected the herds within the herd groups.

The herds represented different herd management activities, for instance how the cows are fed, treated for illnesses or even milked. The Lactation Number was also recorded for each cow studied, cows in lactation 1 were undergoing their first lactation (milking) in life while those in lactation 2 were in their second or later lactation in life. The clusters in Table 2 represented main vegetation types and ecological regions that were thought to be a possible source of variation in milk production. herd groups were chosen from the clusters in such a way that they represented all the specific ecological/vegetation-type differences.

A comparison of the herd groups was therefore a comparison of milk productions in these diverse set ups.

Table 1. Milk production environments and their characteristics

Herd groups	Prod. system	Veg. type	No. Cows	Herds
Kanyanya	Pastoral	Cymbo' Afronadus	89	6
Kashongi	Pastoral, Agro-Pastoral	Cleared thickets	75	7
Kikaatsi	Pastoral	Cymbo' Afronadus	55	5
Ruhengere	Pastoral	Acacia thickets	52	5
Rushere	Pastoral	Shrub anthills, Acacia thickets	26	2
Masha	Pastoral	Acacia thickets	13	2
Mutonto	Agro-Pastoral	Cleared thickets	62	5
Kanyaryeru	Pastoral	Cleared thickets	95	5

Table 2. The Study Design

Clusters	Herd groups	Herds	no. of Cows
Nyabus-hozi	Kanyanya	Bek, Bwe, Kab, Kaf, Kir, Rub	16, 10, 12, 18, 14, 19
	Kashongi	Bih, Bir, BukC, Gan, Kah, Mug, Tum	10, 10, 13, 18, 4, 13, 7
	Kikaatsi	Aga, Bar, Kav, Mor, Uka	10, 10, 7, 16, 12
	Ruhengere	KAC, Kam, Mpo, Rug, Tin	13, 10, 14, 9, 6
	Rushere	Rute, Tume	13, 13
Isingiro North	Masha	Kak, Bat	4, 9
Kahsari	Mutonto	Bahw, Bak, Kan, Mas, Muh	12, 8, 4, 16, 12
	Kanyaryeru	Bah, BukY, Kat, Man, Nab	21, 13, 11, 20, 30

The dataset of this case study is hierarchical with nesting since the random effects (herds) are nested in their herd group. The *lmer* function in *lme4* package can easily handle both nested and crossed cases without model modification (Bates 2005, Quiné and Berg 2008). A simple illustration of crossed versus nested data is given in Table 3. In the crossed case, all levels of one factor (Fertilizer) appear in each level of the other factor (Farm). In the nested case however, we see that levels of one factor (Teacher) occur in only one of the levels of the other factor (School), e.g., John is employed and teaches only in school *A* but not in school *B*.

Similarly, in the present case study, one notices that herds are nested in herd group where each herd is studied in only one particular herd group, as illustrated in Table(2). For example, herd *Bek* is only studied in herd group *Kanyana*. This is in contrast with the crossed data case where one or more levels of herd would be studied in more than one herd group.

Table 3. Nested versus Crossed Datasets

Crossed case		Nested case	
Farm	Fertiliser	School	Teacher
A	S1	A	John
A	S2	A	Jack
A	S3	A	Tom
B	S1	B	Mary
B	S2	B	Mat
B	S3	B	Rose

The *xtabs* command in R helps to observe the data structure.

```
>xtabs(~herdgroup + herd, AnkoleRepeated)
Herd
Herd group Aga Bah BahW Bak Bar Bat Bek Bih Bir BukC BukY
Kanyanya   0   0   0    0   0   0   16  0   0   0    0   ...
Kashongi   0   0   0    0   0   0   0   10  10  13   0   ...
Kikaatsi   10  0   0    0   10  0   0   0   0   0    0   ...
```

3. EXPLORATORY DATA ANALYSIS AND MODEL SELECTION

The dataset is imported from an excel spreadsheet using the *read.table* command and the first 5-rows of the *AnkoleRepeated* dataset are displayed below. The *AnkoleRepeated* dataset has four weeks of milk yield recorded per cow (longitudinal/repeated measurements).

```
>AnkoleRepeated[1:5, ]
  id cluster      herdgroup  herd lacno yield.1wk y.2wk y.3wk y.4wk
1  1 Nyabushozi Kanyanya   Bek    1     379.82  377   376   377
2  2 Nyabushozi Kanyanya   Bek    2     394.90  391   391   390
3  3 Nyabushozi Kanyanya   Bek    2     385.22  386   382   378
4  4 Nyabushozi Kanyanya   Bek    2     381.68  381   379   376
5  5 Nyabushozi Kanyanya   Bek    2     390.53  389   386   382
```

The data occurs in its "wide form", having 467 rows, each row representing a Subject (cow) observed. Each cow has 4 milk offtake values recorded in 4-different columns. For analysis, we ought to transform the data to its "long form".

The long form of the data has 1868 rows (467 by 4), where each unit of observation (cow) has information in four different rows. All the 467 week one(yield.1wk) yield observations are lined up first in the data *AnkoleRepeatedLong* depicted here below, all yield values fall in only one column (see the R-code in the appendix

for achieving this).

```
>AnkoleRepeatedLong[1:10,]
      id cluster       herdgroup herd lacno_f subject time yield
1.1wk 1 Nyabushozi Kanyanya   Bek  1       1       1    379.8200
2.1wk 2 Nyabushozi Kanyanya   Bek  2       2       1    394.9000
3.1wk 3 Nyabushozi Kanyanya   Bek  2       3       1    385.2200
4.1wk 4 Nyabushozi Kanyanya   Bek  2       4       1    381.6800
5.1wk 5 Nyabushozi Kanyanya   Bek  2       5       1    390.5300
```

We now observe that in the long form, Subject and Time factors come into play.

It is customary to obtain summary details such as means and counts as in Table 4. For instance, for herd groups and herds, we have

```
> attach(AnkoleRepeated)
>meanHG<-tapply(yield, herdgroup, mean)
> sigmaHG<-tapply(yield, herdgroup, sd)
> summary(herdgroup)
> plot(yield~herdgroup, data=AnkoleRepeatedLong) # Box plot
> meansH<-tapply(yield, herd, mean) #means for Herd
> sigmaH<-tapply(yield, herd, sd) #standard errors, Herds
> detach(AnkoleRepeated)
```

Table 4. Summary, herd groups

Clusters	Herd groups	mean yield(kg)	Units (no. cows)	Observ-ations
Nyabus-hozi	Kanyanya	383.4068	89	356
	Kashongi	317.439	75	300
	Kikaatsi	242.6298	55	220
	Ruhengere	215.9975	52	208
	Rushere	266.6886	26	104
Isingiro-North	Masha	275.1479	13	52
Kahsari	Mutonto	260.8632	62	248
	Kanyaryeru	323.4036	95	380

We consider a box plot for the herd groups. Except for two herd groups, all the others seem to have a median yield in the range of 250-350 kg. We also notice that the Kanyanya herd group has a higher median milk production, followed by Ryeru (last herd group on the right on Figure 1). This position is later confirmed by the mixed model analysis estimates of fixed effects. Some values appear as outliers and could be easily identified and removed. The analysis however is performed with these outliers as they may represent possible quality milk producers and we cannot afford to do away with them in this case (see Figures 1 and 2).

```
>plot(yield.1wk~herdgroup, data=AnkoleRepeated)
> plot(yield~herdgroup, data= AnkoleRepeated)
```

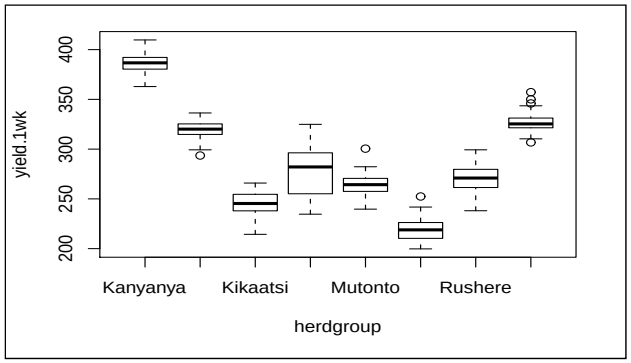

Figure 1. Box plots for herd groups, week 1 yield

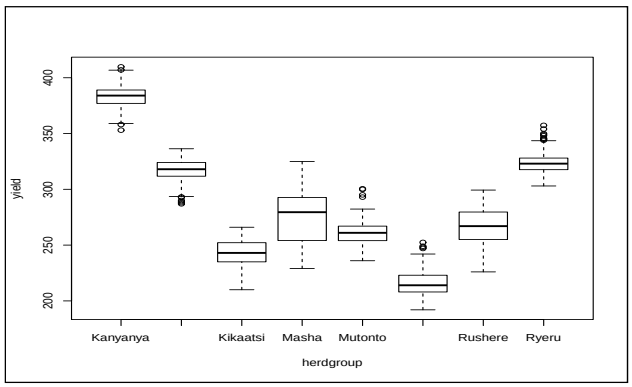

Figure 2. Box plots for herd groups, whole data

To ascertain the independence of the residuals and the homoskedasticity in a typical model to be considered later in the paper (model sm2 defined later in the paper), we use a plot of the observed yield values versus residuals, noting that one could also use a plot of residuals versus predicted yield values. The points should be randomly scattered with constant spread if independence and homoskedasticity hold.

The normality assumption does not quite hold for the residuals. A qq-plot of the residuals versus a normal distribution shows some deviation from the normality assumption. This is due to values in the qq-plot that deviate from the qq-line at the extreme ends of the graph, see Figure 4. Neither the logarithm transformation on the data nor the square root transformation rectifies this situation. The removal of the outliers identified in the box plot does not help either (R-Code for transformations and qq-plot construction is included in the appendix).

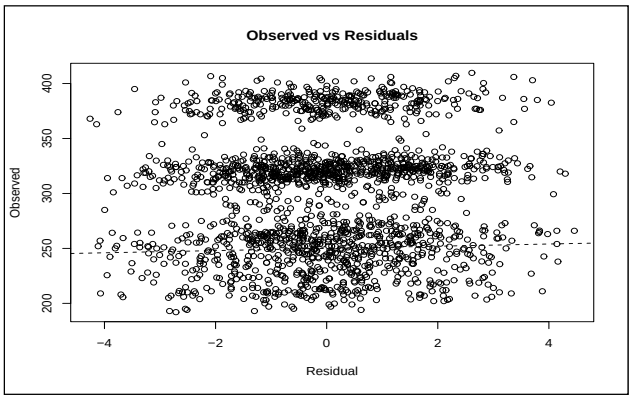

Figure 3. Observed versus residuals in model *sm2*.

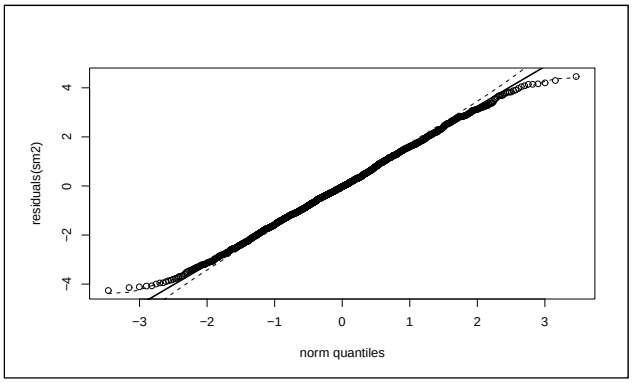

Figure 4. A qq-plot of the residuals versus a normal distribution.
Note: Although observations do not lie outside the 95% confidence interval, the normality assumption is not met in the strict sense. We assume robustness to this violation in the mixed modeling context.

Nobre and Singer (2007, pg 867] undertake an elaborate discussion about residual analysis in mixed models and how it helps to verify homoscedasticity, linearity of effects, presence of outliers, normality and independence of the errors. They note that the estimates of the parameters of the model $Y = X\beta + Z\alpha + \varepsilon$ obtained under normality assumptions are asymptotically consistent even when the distribution of α is not normal but has third finite absolute moment, and only requires a correction in the covariance matrix of the fixed effects estimators. The assumption about normality of the errors (random effects and residual error) is violated (since observations on the extreme end of the qq-plot deviate from the diagonal qq-line). The BLUEs of fixed effects are robust to non-normality in the random error distribution but may influence BLUPs of random effects and tests of hypothesis on parameters (Nobre and Singer 2007). The normality assumptions on errors is addressed already in the *lme4* package of R (Pinheiro and Bates 2000, Chapter 4.

The model selection is informed by the multilevel mixed data structure and the aim which is to conduct a breed selection, which requires BLUPs of random effect factors. The linear relationship in the data is also confirmed and hence fitting a linear model is in order.

3.1 Model Selection

We entertain a series of models. We begin with the complete model $sm1$, but eliminate it since the matrix $X'X$ is not positive definite, and poses an analysis problem(X is the design matrix discussed ahead). This problem is removed when we fit $sm2$ which ignores fixed effect due to Cluster.

```
>sm1<-lmer(yield~cluster+herdgroup+lacno
+(1|herd)+(1|subject)+(1|time), AnkoleRepeatedLong)
Error in mer_finalize(ans):Downdated X'X not positive definite.

>sm2<-lmer(yield ~ herdgroup + lacno +
(1|herd)+(1|subject)
+(1|time), AnkoleRepeatedLong)

>(sm3<-lmer(yield ~ herdgroup + lacno +(1|subject)
+(1|time), AnkoleRepeatedLong))
```

Since herd is a random effect nested in herd group, we consider other model formulations with that fact in mind, before we do an ANOVA comparison of the models to be able to chose the optimal one.

```
>(sm4<-lmer(yield ~ herdgroup + lacno
+(herdgroup|herd), AnkoleRepeatedLong))
>(sm5<-lmer(yield~herdgroup+lacno+(herdgroup|herd)
+(1|subject)+(1|time), AnkoleRepeatedLong))
>(sm6<-lmer(yield~herdgroup+lacno+(herdgroup|herd)
+(1|subject), data=AnkoleRepeatedLong))
```

The Akaike Information Criterion(AIC) and Bayesian Information Criterion(BIC) information for model *sm2* are AIC=9907, BIC=9979.6 and the Log Likelihood equals -4940.8. This is a fairly good model comparing its criterion values to those of models *sma4, sm5, sm6*. We cannot compare *sm2* directly with the other models since BIC for instance, requires that we compare nested models. Model *sm5* is significantly different and has a lower AIC value compared to model *sm4* and *sm6*. See for example Burnham and Anderson (2002) for a useful reference in model selection. We note that smaller values for AIC, BIC define the better model.

>anova(sm4, sm5, sm6)

	sm4	sm6	sm5
Df	46	47	48
AIC	14,403.1	11,573.3	9,971.7
BIC	14,658	11,833	10,237
logLik	-7,155.5	-5,739.7	-4,937.9
Chisq		2,831.8	1,603.6
Chi Df		1	1
Pr(>Chisq)		<2.2e-16**	-2.2e-16***

Signif. Codes: 0.001 ***.

We recall here that the AIC is given by $AIC = -2logL + 2k$, where *logL* is the maximum log-likelihood and *k* is the number of parameters in the model. The BIC is given by $BIC = -2log(L) + k\,log(n)$ and for normally distributed errors, we have $BIC = log(\sigma_\varepsilon^2) + \frac{k}{n} log(n)$, where *logL* is the maximum log-likelihood, *k* is the number of parameters in the model, *n* is the number of observations and σ_ε^2 is the error variance. BIC is thus an increasing function of error variance and the number of parameters.

By the "principle of parsimony" we choose the model

>(sm5<lmer(yield~herdgroup+lacno+(herdgroup|herd)+ (1|subject)+(1|time), AnkoleRepeatedLong))

as the preferred model as it has a lower AIC value (9971) and contains most information required. The model considers herd group and Lactation numbers as fixed effects while herd is considered as a random effect in herd group. Subject and Time are considered random effects.

A linear mixed effects model is fitted of the form

$$Y_{ijklm} = \mu + p_i + q_{ij} + r_k + s_l + t_m + \varepsilon_{ijklm} \quad (1)$$

where Y_{ijklm} is the m^{th} milk off take of animal l in the j^{th} herd (j=1,2,...,36), i^{th} herd group (1=1,2,...8) and in the k^{th} Lactation number (k=1,2). We have μ as the overall mean, p_i is the fixed effect of the i^{th} herd group, q_{ij} is the random effect of the j^{th} herd in the *ith* herd group, r_k is the fixed effect of the k^{th} Lactation number s_l is the random effect due to subject (l=1,2,...,467) t_m is the random effect due to Time (m=1,2,3,4) and ε_{ijklm} is the random error.

A model with crossed effects only would have been stated as $Y_{ijklm} = \mu + p_i + q_j + r_k + s_l + t_m + \varepsilon_{ijklm}$.

The matrix form for the mixed model is

$$Y = X\beta + Z\alpha + \varepsilon \quad (2)$$

where the vector β represents the fixed effect parameters, usually estimated by Generalized Least Squares (GLS) approach, the vector α represents the random effects and are estimated as BLUPs. The dimensions of the vectors and matrices in (2) are, Y_{nX1}, for n=1868 observations, X_{nXp}, for p=11 fixed effects parameters (8 levels of herd group, 2 for Lactation Number, plus the overall mean μ), β_{pX1}, Z_{nXh}, for *h* levels of random effects. In the case of model *sm5* we have h=10 (ee Section 5). Finally we have ε_{nX1}.

Assume ε and α are normally distributed with

$$E\begin{bmatrix}\alpha\\ \varepsilon\end{bmatrix} = \begin{bmatrix}0\\ 0\end{bmatrix}$$

and

$$Var\begin{bmatrix}\alpha\\ \varepsilon\end{bmatrix} = \begin{bmatrix}\sigma_\alpha^2 I_\alpha & 0\\ 0 & \sigma_\varepsilon^2 I_\varepsilon\end{bmatrix}$$

The assumption is that $Y \sim N(X\hat{\beta}, V)$, where $V = Z'\sigma_\alpha^2 Z + \sigma_\varepsilon^2 I_\varepsilon$. The vector of fixed effects in this case reads $\beta = \{\mu, p_1, p_2, ..., p_8, r_1, r_2\}$ where μ is the grand mean, p_i is the $\{i-th: \ i = 1,2,...,8\}$ herd group effect and $\{r_k: \ k = 1,2\}$ is the *kth* lactation number effect.

The vector for random effects reads $\alpha = \{q_1, q_2, ..., q_a\}$ where α represents the number of random factors defined in the model of choice.

The design matrices X and Z are of the form given below.

$$X = \begin{bmatrix} 1 & 1 & 0 & \cdots & 1 \\ 1 & 1 & 0 & \cdots & 0 \\ 1 & 1 & 0 & \cdots & 1 \\ 1 & 1 & 0 & \cdots & 0 \\ 1 & 0 & 0 & \cdots & 1 \\ 1 & 0 & 0 & \cdots & 0 \\ 1 & 0 & 0 & \cdots & 1 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & 0 & \cdots & 1 \\ 1 & 0 & 0 & \cdots & 0 \\ 1 & 0 & 0 & \cdots & 1 \\ 1 & 0 & 0 & \cdots & 0 \end{bmatrix} \quad Z = \begin{bmatrix} 1 & 1 & 0 & \cdots & 0 \\ 1 & 0 & 1 & \cdots & 0 \\ 1 & 0 & 0 & \cdots & 0 \\ 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ 0 & 0 & 0 & \cdots & 1 \\ 0 & 0 & 0 & \cdots & 1 \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix}$$

4. CONCEPTUAL RE-EXAMINATION OF PARAMETER ESTIMATION

The mathematical theory of Mixed Model Analysis based on model (2) and well illustrated in literature such as Searle, Casella and McColloch (1992) requires that we estimate the parameters β and α. To estimate β, the fixed effects parameters, we use the method of GLS that maximizes the log-likelihood function with respect to β. As detailed in the appendix, we obtain BLUEs β via,

$$X'V^{-1}Y = X'V^{-1}X\beta \tag{3}$$

The GLS function (3) depends on the variance components via the matrix V and one has to obtain an estimate of matrix V as a first step. This is done by REML (Restricted or Residual Maximum Likelihood) estimation. Then one needs the estimates of random effects. The estimates are referred to as "Predictors" to distinguish them from fixed effects for which the word "Estimates" has been used. The BLUPs of α are obtained from the equation,

$$BLUP(\alpha) = (Z'Z + \hat{\Gamma}^{-1})^{-1} Z'(Y - X\hat{\beta}). \tag{4}$$

In (4), the parameters to be estimated include Γ and β. The BLUE of β and the REML estimate of the variance components contained now in Γ, are substituted in the equation to finally obtain the BLUPs.

4.1 Restricted Maximum Likelihood Estimation

The variance components are two main parameters σ_α^2 and σ_ε^2 contained in the matrix V. Note that σ_α^2 may have sub-variances, for each of the levels of random factors included in the model. The variance components can be estimated by a number of methods, including, ANOVA, Maximum Likelihood, Bayesian Estimation and Method of Moments. But REML, developed by Patterson and Thompson (1971) is more attractive since it offers unbiased and non-negative estimates of variance components. Maximum likelihood estimates (MLE) of variance components may turn out to be negative (see for example Duchateaux et al. 1998). Searle et al. (1992)mentions that such variance components can be set to zero.

The REML procedure maximizes the part of the likelihood function that is location invariant. Location parameters are fixed effect parameters, and one has to split the likelihood function into a part that depends on fixed effects and another that is independent. This is possible for balanced datasets, and is not straightforward for unbalanced cases.

Verbyla (1990) gives a clear interpretation of the REML method as follows: partition the likelihood into two independent parts, splitting the vector of observations Y into $(Y_1 = K_1'Y)$ relating to the fixed effects and $(Y_2 = K_2'Y)$ relating to the residual contrasts (zero expectation) with,

- K_1 an n by p matrix of full column rank
- K_2 an n by (n-p) matrix of full column rank
- $K_1'X = I, \quad K_2'X = 0$, where X is the design matrix for fixed effects.

The residual contrast K_2 is used in the estimation of the variance components for general cases, by maximizing a linear combination $K_2'Y$ of the vector of observed values Y with the properties:

$$E(K_2'Y) = 0, \qquad K_2'Y \sim N(0, K_2'VK_2).$$

Take $K_2 = K$ for simplicity. Replacing $Y\ by\ K'Y$ translates to replacing $Z\ by\ K'Z$, $X\ by\ K'X = 0$ and $V\ by\ K'VK$.

Note that if $E(K'Y) = 0$ then $K'E(Y) = K'X\beta = 0$ so $K'X = 0.$

Evaluating (14) of the appendix with these replacements, one obtains

$$[tr(PZ_iZ_i')]_{i=\alpha,\varepsilon} = [Y'PZ_iZ_i'PY]_{i=\alpha,\varepsilon}, \quad (5)$$

where

$$P = V^{-1} - V^{-1}X(X'V^{-1}X)^{-}X'V^{-1} = K(K'VK)^{-1}K'.$$

This is similar to the Maximum Likelihood equations but with V^{-1} replaced by $P.$ These equations are usually a complex non-linear function of the variance components through P and cannot be solved (by setting equal to zero) directly. Note that balanced designs are a special case. Iterative computer algorithms are thus used to solve them. These are already written as programs in software toolboxes such as *lmer* in this case.

4.2 Hypothesis Testing on the Parameters

In the context of a fixed effects model, there is only one source of random variation, and the test of hypothesis relies on the ratio

$$\frac{ms(factor)}{ms(residual)}. \quad (6)$$

The ratio follows an F-distribution with degrees of freedom due to fixed effect factors in the numerator and the degrees of freedom due to residual in the denominator. If the null hypothesis is true, then this ratio simplifies to one.

In the case of a mixed effects model, the denominator is often (not always) a linear combination of the different sources of random variation. The degrees of freedom due to the denominator is therefore derived using Welch-Satterthwaite equation which is used to calculate an approximation to the effective degrees of freedom of a linear combination of sample variances (Satterthwaite 1946). Then the challenge of computing the mean square errors for the factors in the model arises due to imbalance in data. Suppose we had the same number of herds in each herd group and some number of observations (cows) in each herd, we could use the following computations:

Herd group: $ms_{HG} = \sigma_\varepsilon^2 + a\sigma_h^2 + bg\sigma_h^2,$

Herds: $ms_H = \sigma_\varepsilon^2 + c\sigma_h^2$ and the residual mean square by $ms_R = \sigma_\varepsilon^2.$ Here σ_ε^2 is the residual mean square error at unit (cow) level, σ_h^2 is at herd level and $g\sigma_h^2$ at herd group level. A hypothesis test on effect of herd groups could be conducted using the ratio

$$\frac{ms(Herdgroup)}{ms(Herds\ within\ Herdgroup)}.$$

The test statistic does not follow an F distribution under the null hypothesis. Moreover, we have an unbalanced data set which does not allow for this sort of computation. It makes no sense to obtain the F test for levels of factors in this study, and correctly so, the *lmer* function does not offer F-test results in its output. The imbalance in the dataset makes it inappropriate to use the F-test, and often, the Wald test is used instead.

5. DATA ANALYSIS

5.1 Results of Fitting the Linear Mixed Model Using *lmer*

We consider the output for model *sm5*. The output shows that REML was the tool that produced the variance estimates. REML is the default tool under the *lmer* function, and if specified to be false, them the MLE is then used instead. The other information includes criterion for model choice, including AIC, BIC and the other criteria.

Linear mixed model fit by REML Formula:

yield ~ herdgroup + lacno + (herdgroup | herd) + (1 | subject) + (1 | time)
Data: AnkoleRepeatedLong

AIC	BIC	logLik	deviance	REMLdev
9941	10207	-4923	9876	9845

Random effects:

Groups Name	Variance	Std.Dev.
subject (Intercept)	118.7943	10.8993
herd (Intercept)	3.5525	1.8848
herdgroupKashongi	2.3110	1.5202
herdgroupKikaatsi	13.5369	3.6792
herdgroupMasha	3.5486	1.8838
herdgroupMutonto	2.5166	1.5864
herdgroupRuhengere	6.8931	2.6255
herdgroupRushere	28.7822	5.3649
herdgroupRyeru	3.5526	1.8848
time (Intercept)	7.1735	2.6783
Residual	3.2725	1.8090

Number of obs: 1868, groups: subject, 467; herd, 37; time, 4

The output includes estimates of the variance components, including σ_a^2 which has ten sub-

components while $\sigma_\varepsilon^2 = 3.2725$. Most variance components have a standard deviation above that of the residual error, hence justifying their inclusion as random effects in the model. Subject contributes the most variation in the data, with a variance estimate of 118.7943. These variance estimates which are elements of matrix V are then used to compute the fixed effect parameters via the formula $X'V^{-1}Y = X'V^{-1}X\beta$.

Fixed effects:

	Estimate	SE	t-val
(Intercept)	384.335	2.328	165.11
herdgroupKashongi	-66.003	2.016	-32.74
herdgroupKikaatsi	-140.103	3.118	-44.93
herdgroupMasha	-108.322	3.342	-32.41
herdgroupMutonto	-122.787	2.046	-60.02
herdgroupRuhengere	-168.036	2.787	-60.29
herdgroupRushere	-116.781	5.421	-21.54
herdgroupRyeru	-60.038	1.795	-33.45
lacno2	-1.022	1.387	-0.74

Note that Kanyanya is the baseline herd group. All the others are compared to Kanyanya, and 384.335kg is the average performance in Kanyanya. The estimate -66.003 represents the difference between performance in the Kanyanya and Kashongi herd groups. The average performance in Kashongi is thus (384.335-66.003) kg. The p-values for the t-test are not given. We could however infer that since the absolute values of the t-values for herd groups are very large (greater that approximately 2 is a sign of significant difference), all the fixed effect factors are significant in the model. Lactation number two has a poor performance compared to Lactation number one, its performance being one unit below that of Lactation one. However, the effect of Lactation number is probably not significant given its low t-value. The best performing herd groups are Kanyanya, followed by Ryeru and the Kashongi. Milk offtake in Lactation one is also averagely higher than that of Lactation 2.

Correlation coefficients for the estimates are provided in the *lmer* output. Low values of correlation are preferred. High correlation is an indication of some high relationship among the groups which is often not true.

One notices that the degrees of freedom and the p-value for the t-tests and the F-tests done using the *lmer* function are not provided, because the "F-statistics" used in the nested mixed effects model with unbalanced groups (such as in this case study) do not exactly follow an F distribution. The degrees of freedom used are not statistical reliable. The Markov Chain Monte Carlo (MCMC) approach, and Parametric bootstrap techniques provide alternatives in dealing with this problems (Faraway 2009).

5.2 Residual Analysis

In addition to the information on the exploratory data analysis, it is as useful to consider the residual plots and analysis for purposes of diagnostics. Normal QQ-plots, factor versus residual plots, residual versus fitted value plots are some of the common tools in understanding the residual structure in the data. Using the model *sm5*, we obtain the residual plots (Figure 5). A more liberal test of normality done on the residuals of model $sm5$ is given by the Shapiro Wilk test,

>shapiro.test(residuals(sm5))
Shapiro-Wilk normality test data:
residuals(sm5) W = 0.9982, p-value = 0.0404

The normality assumption is violated by the results of Shapiro-Wilk test, i.e., the null hypothesis of normally distributed data is rejected as the p-value is significant at 5% level, (p-value = 0.04). The stem and leaf diagram obtained by "*> stem(residuals(sm5))*" has a skew to the left, also confirming the non-normality assumption.

5.3 BLUPs and Selection

The *ranef()* commands returns the BLUP for both Subject and herd random factors. The command is given by *BLUPS<-ranef(sm5)* which could also be modified to obtain BLUP for either herd using *BLUPS<-ranef(sm5)[["herd"]]* and *BLUPS<-ranef(sm5)[["subject"]]* for Subjects.

Once the BLUP for Subjects are obtained, we sorted the data to obtain the ranking of the first 40 cattle, starting with largest BLUP values to the lowest. The first 40 cows were chosen as the quality breeds, out of the 467 cows studied. Examining the BLUP for selected cows, we notice that the average milk off take from the selected herd groups was higher than from the non-selected herd groups in all four weeks.

Among the production systems, the Pastoral system was more suited for quality milk production while the Cleared Thickets provided the most suitable Vegetation type. The results of the exploratory analysis, fixed effects estimates and the selected group obtained from BLUPs all concur in terms of the best herd groups.

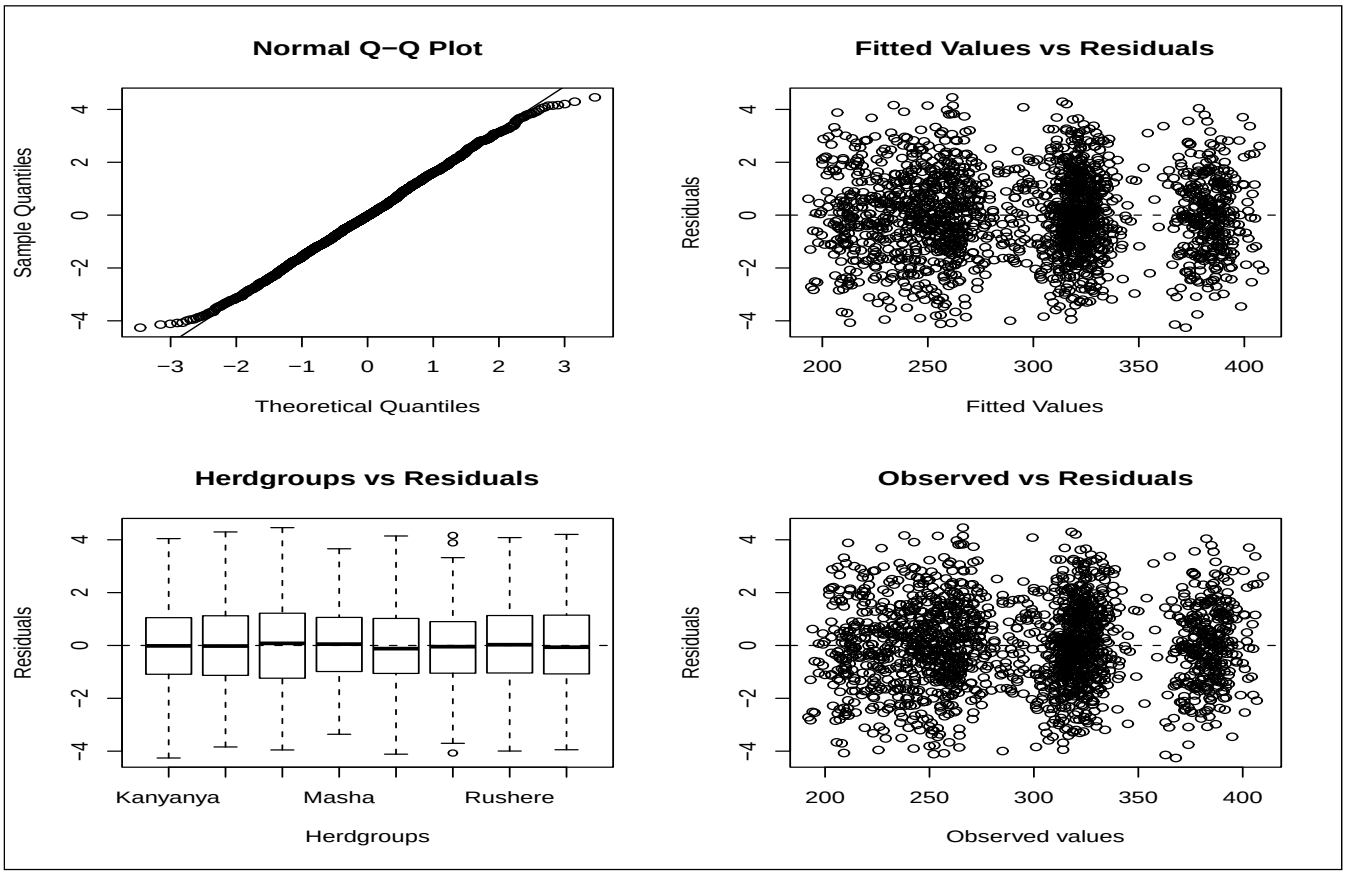

Figure 5. Residual plots of AnkoleRepeated dataset

Notes: The QQ-plot has plotted points lying along the qqline, except at extreme regions. This difficulty indicates non-normality hence other tests such as the Shapiro Wilk test need to be run.

Table 5: Performance among the 40 selected versus the 427 non-selected groups

In kg.	yield.1wk	yield.1wk	yield.1wk	yield.1wk
Selected	313.712	311.775	309.575	307.75
Non-selected	301.635	299.611	297.482	295.48

Table 6: Percentages of selected cows across herd groups

herd group	No. selected	% out of 40.
Kanyanya	6	15
Kashongi	4	10
Kikaatsi	5	12.5
Masha	4	10
Mutonto	4	10
Ruhengere	4	10
Rushere	7	17.5
Ryeru	6	15
Total	**40**	**100**

To demonstrate that BLUP results are fairly accurate, we compare the best to the worst cow, by BLUP ranking. Cow ID number 220 from Isingiro North, a pastoral region with Acacia-thickets, belonged to the Masha herd group, Bat herd, and was in Lactation 1. It recorded a BLUP value of 42.17 and weekly milk offtakes of 324.90kg, 320kg, 318kg and 317kg. Incidentally, the cow(ID 223) with the least BLUP value comes from the same herd group and herd, but in its second Lactation. It had a BLUP value of -41.797 and weekly milk yield of 234.62kg, 231kg, 232kg and 232kg.

6. DISCUSSION

This study aimed at identifying and characterizing quality breeds of cattle for milk production. Mixed modeling played a very important role in the analysis and specifically BLUP is a key tool in a selection study of this kind. Though we used BLUPs for the selection of quality cows in this case, there are however other ways of selection based on Yield values(Y) or even Residuals(ε).

We notice that the assumption of normality of residuals in our model is violated, as evidenced from qq-plots of yields and the outcome of the residual analysis. However we assume robustness in the parameter estimation and proceed with the modeling based on normality assumption.

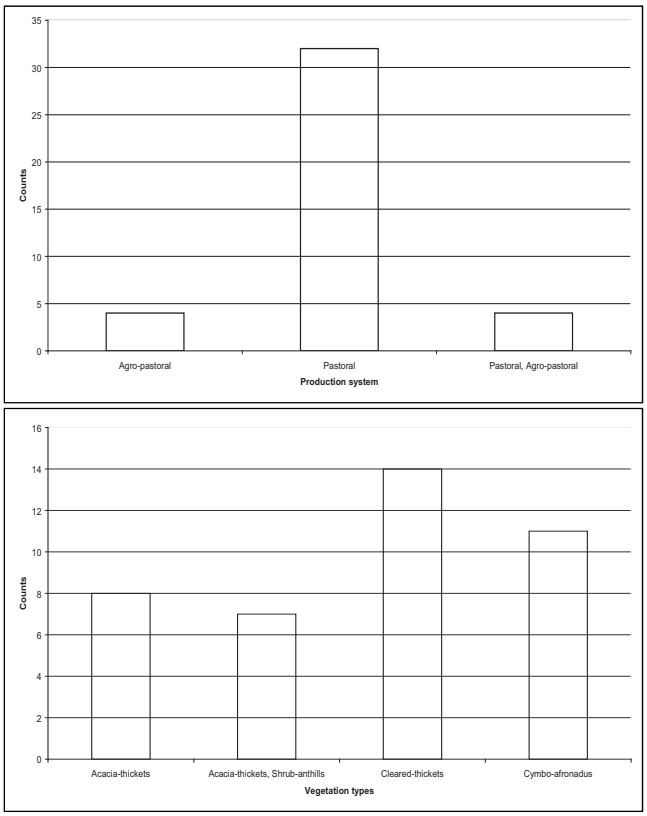

Figure 6: Number of selected cows by production systems and vegetation types

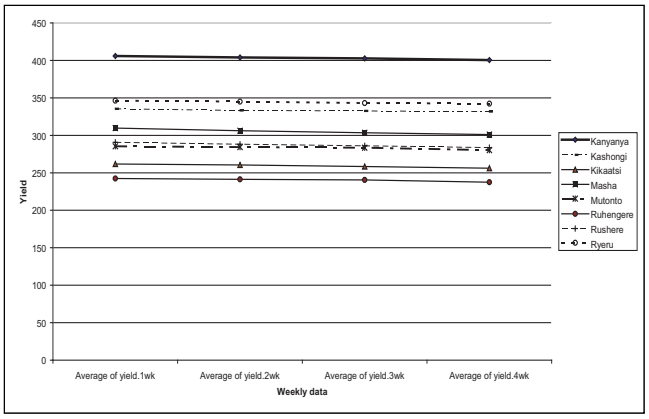

Figure 7: Performance among herd groups across weeks for selected cows.

We note that the *lme* and *lmer* functions both have their advantages and disadvantages. The function *lmer* provides an improvement over *lme* and is obviously stronger in some aspects; it is faster even for large datasets, can handle both crossed and nested data using the same model specification and is obviously more user friendly in terms of specifying models for fixed and random effects. However, it provides no p-values for t-tests and F-tests within its output. This is because the p-values even if given are not so useful or appropriate due to the problem of computation of degrees of freedom involved especially for nested cases. It is hard to observe which effects are significant directly from the output, though the absolute values of t-statistics can be a pointer in this direction. The *lme* model for repeated cases can handle covariance structures as opposed to *lmer* function explicitly. Depending on situations, one can switch from *lmer* to *lme* to achieve certain targets.

It is expected that data from an individual cow taken repeatedly over time has some correlation. Since it is expected that there is a gradual decrease of milk yield from the first week of lactation to the successive weeks, the first order autoregressive covariance structure would be more appropriate for analyzing the *AnkoleRepeated* dataset. First order covariance structure takes the form:

$$\sigma\begin{bmatrix} 1 & \Phi & \Phi^2 & \Phi^3 & \cdots & \Phi^h \\ \Phi & 1 & \Phi & \Phi^2 & \cdots & \Phi^h \\ \Phi^2 & \Phi & 1 & \Phi & \cdots & \Phi^h \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \Phi^h & \Phi^{h-1} & \Phi^{h-2} & \Phi^{h-3} & \cdots & 1 \end{bmatrix}$$

where $h = 4$ is the number of repeated observations per cow and Φ is the correlation between any two adjacent observations in time. Hence observations four weeks apart have a lower correlation of Φ^4. Taking this correlation into account makes the variance estimates more reliable as well as the parameter estimates and tests of hypotheses that follow.

The function *lme* has an option of including a given covariance structure, whether simple, unstructured, compound or first order autoregressive, among others. The command line in *lme* for incorporating the first order autoregressive structure would be
> AnkoleRepeated_lmer4<- update(AnkoleRepeated_lmer1, + correlation = corCAR1(form = ~ time |subject)). This is however not so straightforward with the *lmer* function.

We finally note that since the data of this case study are simulated, we do not stress on the values reported in this paper, but rather, on the methods and stages of mixed modeling analysis and animal breed selection.

REFERENCES

Bates, D. 2005. *Fitting linear mixed models in R*. R News, 5 no. 1: 27-30.

Burnham. K.P. and Anderson, D.R. 2002. *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach. Second ed.* New York: Springer-Verlag.

Duchateau, L. P. Janssen and Rowlands, J. 1998. *Linear Mixed Models. An Introduction with Applications in Veterinary Research*,Kenya: ILRI.

Faraway, J. 2009. *Linear Models with R*. Chapman and Hall. London.

Khuri, A.I. and Sahai, H., 1985. *Variance components analysis: A selective literature survey*, International Statistical Review, 53 no. 3: 279-300.

Ndumu,D. B. 2000. *Identification and Characterization of Elite Performing Ankole Longhorn Cattle for Milk Production*, M. Sc. Thesis, Uganda: Makerere University.

Nobre, J.S. and Singer, J.M. 2007. Residual Analysis for Linear Mixed Models. *Biometrical Journal*, 49 no. 6: 863-875

Patterson, H. D. and Thompson, R. 1971. Recovery of interblock information when block sizes are unequal. *Biometrika*, 58 no. 3: 545-554.

Pinheiro, J. and Bates, D. M. 2000. *Mixed Effects Models in S and S-Plus*, New York: Springer-Verlag.

Quené, H. and Berg, H. Examples of mixed-effects modeling with crossed random effects and with binomial data. *Journal of Memory and Language* 59: 413-425.

Satterthwaite, F.E. 1946. *An* approximate distribution of estimates of variance components. *Biometrics Bulletin* 2: 110-114.

Searle, S.R.,Casella, G. and McCulloch, C.E. 1992. *Variance Components*, New York: J.W. Wiley.

Verbyla, A.P. 1990. *A conditional derivation of residual maximum likelihood*, Australian Journal of Statistics, 32: 227-230.

Acknowledgements

The original dataset which we refer to in this case study was collected by Ndumu D.B.(Makerere University,Uganda, year 2000 as part of research towards his Masters thesis). I analyzed a subset of the original dataset using GENSTAT for my M.Sc thesis, during an internship position at the International Livestock Research Institute (ILRI-Kenya), 2003 under supervision of Dr. John Rowlands and Dr. Thomas Achia at the institute of Biometry-ILRI-Kenya. We use the names in Ndumu's dataset, however simulated the current dataset of this case study using averages and standard deviations from the original dataset. I therefore acknowledge ILRI and the support from staff of the Biometry institute during my research internship (2003).

Correspondence: nelsonowuor@gmail.com

Appendix

A: Fixed effect parameters

The mathematical theory of Mixed Model Analysis is well illustrated in e.g.,[12]. We note key steps that are useful for the context of data analysis. The model (2) requires that we estimate the parameters β and $\propto$. To estimate β, the fixed effects parameters, we solve the equation set,

$$\frac{\partial}{\partial\beta}(l_Y[\beta,V]) = 0 \tag{7}$$

where in this case, $V = Z'\sigma_\alpha^2 + \sigma_\varepsilon^2 I_\varepsilon$ with two variance components (σ_α^2 and σ_ε^2) and l_Y is the likelihood function for the model, given by

$$l_Y = constant\, w.r.t\, \beta - \frac{1}{2}(Y - X\beta)'V^{-1}(Y - X\beta).$$

From (7) one obtains,

$$X'V^{-1}Y = X'V^{-1}X\beta \tag{8}$$

The GLS function (8) depends on the variance components, and one has to obtain an estimate of matrix $\mathbf{V}$ as a first step.

B: Random Effects-BLUP

In the linear model (2), $E(Y) = X\hat{\beta}$ and $\mathrm{Var}(Y) = V$, where,

$$V = ZCov(\alpha)Z' + var(\varepsilon)I_n \tag{9}$$

$$= \sum_{i=1}^{a}(\sigma_i^2 Z_i Z_i') + \sigma_\varepsilon^2 I_n, \tag{10}$$

where a=36 denotes the number of parameters in α. It is possible(this is often done for simplification) to re-parameterize $\mathbf{V}$ as follows.

$$V = \sigma_\varepsilon^2\left(\sum_{i=1}^{s}[\Phi_i Z_i Z_i'] + I_n\right) = \sigma_\varepsilon^2 H, \tag{11}$$

where $\Phi_i = \sigma_i^2/\sigma_\varepsilon^2$, $H = ZIZ + I_n$ and Γ is a variance-covariance matrix having entries $\sigma_i^2/\sigma_\varepsilon^2$ along its main diagonals and $\sigma_{ij} = 0\ \forall\ i \neq j$.

The BLUP of α is a solution to the equation,

$$BLUP(\alpha) = (Z'Z + \hat{\Gamma}^{-1})^{-1}Z'(Y - X\hat{\beta}). \tag{12}$$

In (12), the parameters to be estimated include Γ and β. The BLUE of β and the REML estimate of the variance components contained now in Γ are substituted in the equation to finally obtain the BLUPs.

C: Maximum Likelihood Estimation

We illustrate the maximum likelihood estimation theory as a preamble to REML, due to similarity in approach. Consider the data $Y \sim N(X\beta, V)$, such that the log likelihood function of Y is,

$$l_Y = constant\, w.r.t\, \beta - \frac{1}{2}(Y - X\beta)'V^{-1}(Y - X\beta).$$

Differentiating the log likelihood with respect to the variance components σ_α^2 and σ_ε^2 we get a summarized expression of the maximum likelihood equation as,

$$\frac{\partial}{\partial\sigma_i^2}l_Y(\beta,V) = \frac{-1}{2}(V^{-1}Z_i Z_i') + \frac{1}{2}(Y - X\beta)'V^{-1}Z_i Z_i'V^{-1}(Y - X\beta), \tag{13}$$

where we have only two components, $i = \alpha, \varepsilon$. We note that

$$\frac{\partial V}{\partial\sigma_i^2} = Z_i Z_i'.$$

To obtain σ_i^2 that maximizes the likelihood function, we equate (13) to zero and solve for each σ_i^2. We consider the following useful results on matrix differentiation:

For a general matrix $A(\theta)$,

$$\frac{\partial ln|A|}{\partial \theta} = tr\left(A^{-1} \frac{\partial A}{\partial \theta} \right).$$

Also,

$$\frac{\partial A^{-1}}{\partial \theta} = -A^{-1} \frac{\partial A}{\partial \theta} A^{-1}.$$

For a matrix

$$P = H^{-1} - H^{-1}X(X'H^{-1}X)^{-}X'H^{-1},$$

with matrix $H = H(\theta)$,

$$\frac{\partial P}{\partial \theta} = -P \frac{\partial H}{\partial \theta} P.$$

The resultant maximum likelihood equations can be represented in the form below(see e.g.,Searle [12]), using the matrix differentiation results.

$$\left[tr(V^{-1}Z_i Z_i') \right]_{i=\alpha,\varepsilon} = \left[Y'PZ_i Z_i' PY \right]_{i=\alpha,\varepsilon}. \quad (14)$$

Closed form solutions for (14) cannot be obtained as the equations are usually a complex non-linear function of σ_i^2 (through V and P) and cannot be solved directly(just by setting (13) equal to zero), unless we are dealing with balanced designs(e.g., a case where each Herdgroup has equal number of Herds, and each Herd, equal number of cattle selected, also each Heardgroup/Herd has equal number of cows in each Lactation group).

Numerical solutions of (13) are therefore used. However, the Maximum Likelihood estimates of the variance components are often negative, which is not in the required parameter space for variance parameters. When this occurs, the Maximum Likelihood(ML) estimate is taken to be zero, and the residual variance component is re-estimated from data, dropping the corresponding random factor from the data. Note that assuming that a given variance component is zero is tantamount to dropping the corresponding random factor(in this case, the corresponding Herd) from the data.

D: Welch-Satterthwaite procedure

According to the Welch-Satterthwaite procedure, we have a situation in which we are creating a composite variable $G = \sum_{i=1}^{n} k_i V_i$, where k_i are arbitrary constants and each V_i is a sample variance that is proportional to a chi-square variable with known degrees of freedom ϑ_i, i.e., $V_i \sim \chi^2_{\vartheta_i}$. Then **G** is assumed to be approximately proportional to a chi-square variable with ρ degrees of freedom, where

$$\rho = (\sum_{i=1}^{n} k_i V_i)^2 / \sum_{i=1}^{n} \{ \frac{(k_i V_i)^2}{\vartheta_i} \}.$$

In practice, ρ is being estimated by substituting the observed V_i by its expected value.It can be shown that

$$min\vartheta_i \le \rho \le \sum \vartheta_i.$$

ρ will attain its upper bound when each V_i is proportional to its degrees of freedom. In fact, if each V_i is proportional to ϑ_i, then **G** is exactly, not approximately, proportional to a chi-square variable with $\rho = \sum \vartheta_i$. The other extreme occur when any one V_i, say V_j, is so much larger than all the others that the sample **G** effectively is V_j, regardless of the other V_i values. Then ρ approaches ϑ_j, its minimum.

E: R-codes

E.1: qq-plots and data transformation

```
#.....................................................................................
#a simple qq-plot
>qq.plot(AnkoleRepeated$yield.1wk, + dist= "norm",
+ labels=FALSE, col="BLACK")

#.....................................................................................
#log(base-e) transformation
>logyield.1wk<- log(AnkoleRepeated$yield.1wk) ;
>AnkoleRepeatedLog<-cbind(AnkoleRepeated, logyield.1wk);
>qq.plot(AnkoleRepeatedLog$logyield.1wk,
+ dist= "norm",
```

```
+ labels=FALSE, col="BLACK")

#.....................................................................................
 #removal of outliers, indicated rows removed.
>AnkoleRepeatedLogSample<-
+AnkoleRepeatedLog[-c(123,236,313,323,332,355,462),]
>qq.plot(AnkoleRepeatedLogSample$logyield.1wk,
+ dist= "norm", labels=FALSE, col="BLACK")

#.....................................................................................
#squareroot transformation.
>sqrtyield.1wk<- sqrt(AnkoleRepeated$yield.1wk)
>AnkoleRepeatedSqrt<-cbind(AnkoleRepeated,
+sqrtyield.1wk);
>qq.plot(AnkoleRepeatedSqrt$sqrtyield.1wk,
+ dist= "norm", labels=FALSE, col="BLACK")
```

E.2: Residual plots

```
> (sm2<-lmer(yield ~ herdgroup + lacno +
+ (1| herd)+(1|subject)+(1|time), AnkoleRepeatedLong))
> op<-par(mfrow=c(2,2))
> qqnorm(residuals(sm2)); qqline(residuals(sm2))
> plot(fitted(sm2), residuals(sm2), xlab="Fitted Values",
+ ylab="Residuals", + main="Fitted Values vs Residuals");
+ abline(a=0, b=0, lty=2)
> plot(AnkoleRepeatedLong$herdgroup,
+ residuals(sm2), xlab="Herdgroups",
+ ylab="Residuals",  main="Herdgroups vs Residuals");
+ abline(a=0, b=0, lty=2)
> plot(AnkoleRepeatedLong$yield, residuals(sm2),
+ xlab="Observed values", ylab="Residuals",
+ main="Observed vs Residuals"); abline(a=0, b=0, lty=2)
> par(op) > shapiro.test(residuals(sm2))
```

E.3: Transforming repeated measures data into "long form" from "wide form"

```
>AnkoleRepeated<-read.table("clipboard",
+ header=TRUE, sep="\t")
>AnkoleRepeated$subject<-factor(rownames(AnkoleRepeated))
>nobs<-nrow(AnkoleRepeated)
>AnkoleRepeatedLong<-reshape(AnkoleRepeated, idvar="subject",
+ varying=c("yield.1wk", "yield.2wk", "yield.3wk", "yield.4wk"),
+ direction="long")
>AnkoleRepeatedLong$time<-rep(c(1,2,3,4), rep(nobs,4))
> AnkoleRepeatedLong[1:10,]
```