AN APPROACH TO ESTIMATING THE VARIANCE COMPONENTS TO UNBALANCED CLUSTER SAMPLED SURVEY DATA AND SIMULATED DATA

by

SHAUN RAMROOP

submitted in part fulfilment of the requirements
for the degree of

MASTER OF SCIENCE

in the subject

STATISTICS

at the

UNIVERSITY OF SOUTH AFRICA

SUPERVISOR: DR R NIEUWOUDT

JOINT SUPERVISOR: PROF F E STEFFENS

NOVEMBER 2002

TABLE OF CONTENTS

Chapter 1	**An Introduction**	pg	1
Chapter 2	**Literature Survey**	pg	3
	2.1 Historical background		3
	2.2 The Mixed Model		6
	2.2.1 Introduction		6
	2.2.2 The general linear mixed model		8
	2.3 Henderson's methods		12
	2.3.1 Method I		12
	2.3.2 Method II		12
	2.3.3 Method III		12
	2.4 Henderson's mixed model equations		17
	2.5 Maximum Likelihood Estimation (ML)		19
	2.5.1 The maximum likelihood equations		19
	2.5.2 The maximum likelihood solutions using BLUP		21
	2.5.3 Notes		22
	2.6 Restricted maximum likelihood (REML)		26
	2.6.1 The basis		26
	2.6.2 REML equations		27
	2.6.3 Solutions to REML equations using BLUP		29
	2.6.4 Notes		31
	2.6.5 Weighted REML		32
	2.7 Minimum norm quadratic unbiased estimation (MINQUE)		34
	2.7.1 Notes		37
	2.8 The generalized linear mixed model		39
	2.8.1 GLMM's- The methods of Breslow and Clayton and of Schall		40
	2.8.2 The method of Schall		41
	2.8.3 Notes		43
	2.9 The future		46
Chapter 3	**Results**	pg	48
	3.1 Background and description of the data		48
	3.2 Box-and-whisker plots		50
	3.3 Cross tabulation of the data		56
	3.4 Model selection		60
	3.5 The model		63
	3.6 Henderson's method III		65
	3.7 Maximum likelihood estimation		67
	3.8 MINQUE results		70
	3.9 Weighted and unweighted REML analysis		70
	3.10 Generalised linear mixed model analysis (GLMM)		72
	3.11 Overall summary		78

Chapter 4 **Simulated data** pg 86

4.1 Introduction 86
4.2 Method and model of simulated data 86
4.3 Results of estimates for σ^2_δ 89
4.4 Results of estimates for σ^2_e 91

Chapter 5 **Conclusion** pg 93

Bibliography 95

Appendices

Appendix 1 Weighted REML output from Genstat 100
Appendix 2 SPSS output for MINQUE and ML 105

Student number: **3267-263-2**

I declare that "An approach to estimating the variance components to unbalanced cluster sampled survey data and simulated data" is my own work and that all the sources that I have used or quoted have been indicated and acknowledged by means of complete references.

.....
SIGNATURE

20/03/03
.....
DATE

Acknowledgements

This thesis has been possible through the help of Dr. R Nieuwoudt with her thoroughness, tireless guidance, time and constructive insight. I am also grateful to my wife Shanitha for all her support and encouragement.

Abstract:

The topic of variance components within the general theory of the linear model is both vast and imperative to the subject of statistics. Both balanced or unbalanced data have attractive variance component estimating procedures such as ANOVA, Maximum Likelihood (ML), Restricted Maximum Likelihood (REML) and Minimum norm quadratic unbiased estimation (MINQUE). The ANOVA procedure has attractive properties of unbiasedness but nothing about optimum properties. Both ML and REML both require normality assumptions and have nonlinear equations that must be solved numerically. MINQUE involves solving a set of linear equations and has a few disadvantages when compared to REML and ML. The aim of this thesis is to apply the ANOVA, ML, REML, MINQUE and Generalised Linear Mixed Model (GLMM) estimating procedures to unbalanced cluster sample survey data and simulated data set and compare the procedures to assess their advantages and disadvantages.

CHAPTER ONE: An Introduction

The topic of variance components within the general theory of the linear model is both vast and imperative to the subject of statistics. Variability will always exist in any set of data. It is like a gene that is inherited from a source. The variance component model applied to a set of data attempts to expose and evaluate the source of variation whilst also maintaining its aim in explaining the differences between the means and dependencies between responses and explanatory variables.

The motivation behind my interest to apply the techniques of variance component estimation to my data lies in the fact that I love the analysis of variance theory and am intrigued to infer proper conclusions and to assess the validity of the different variance component models. This has been kindled by reading and understanding much of the work done by Statisticians such as Searle S.R., prior to this thesis.

Another factor that has kindled an interest in the study of variance components is that it offers an opportunity to invoke techniques such as Maximum Likelihood (ML) and Restricted Maximum Likelihood (REML) that have been developed to estimate the variance components and draw inferences from complex data sets with their merits and demerits, very often competing as to which technique should be chosen above the other to be applied to the data set. Much of the inspiration for this thesis is drawn from the book by Searle, Casella and McCulloch (1992) called "Variance Components" and the article by Searle (1995, 1996).

The aim of this dissertation will be to explain, investigate and apply various techniques to estimating the variance components of a specific data set whilst comparing these techniques to assess their advantages and disadvantages. The dissertation is structured as follows:

In Chapter 2, I will look at the historical background, the present and future status of variance components. All relevant theory and information about different approaches of different variance component models will be discussed.

In Chapter 3, I will include the discussion of the data set and computational applications of the various techniques that I have applied to the data set with my results and conclusions. The statistical packages that have been used include GENSTAT, SPSS, SAS and HLM. Approaches such as Maximum Likelihood (ML), Restricted Maximum Likelihood (REML), Analysis of variance (ANOVA), Generalised Linear Mixed Model (GLMM) and Minimum Norm Quadratic unbiased estimation (MINQUE) will be implemented using these packages and statistical inferences will be made from the output.

In Chapter 4, I will look at the analysis of simulated data sets to uncover differences in variance component estimation between balanced and unbalanced data.

Chapter 5 will culminate in the conclusions that I will make about various variance component models relevant to the data set as well as the conclusions about the various statistical packages used.

Finally it should be noted that the scientific journal article which has prompted this dissertation is by Shayle R Searle (1995). An overview of Variance Component estimation (Metrika: Volume 42, No 3-4, Pages 215-230) followed up again in 1996 (Tatra Mountains. Math Publications. Probstat '94 (Smolenice Castle 1994) Volume 7, Pages 134-142).

CHAPTER 2: Literature survey

2.1 Historical background

Searle (1995, p.1) states that variance component estimation originated with estimating the error variance in the analysis of variance by equating the error mean square to its expected value. This procedure was then extended to random effects models for balanced data and then for unbalanced data. The relevance of estimating the variance components is that it affects the estimation of other parameters in the model fitted. Khuri and Sahai (1985) have given an excellent exposition of the history of variance components.

To understand the concept of variance components, one needs to distinguish between fixed effects models and random effects models. The beginnings of variance components has revolved around the l-way random effects classification model as

$y_{ij} = \mu + \alpha_i + e_{ij}$ where $i = 1, 2,.....,a$ where the α_i's and e_{ij}'s are assumed to be random variables, in other words the model focuses on a random selection from a population of all possible factor levels in the ANOVA model.

μ is the mean , $var(\alpha_i) = \sigma^2_\alpha$, $var\ (e_{ij}) = \sigma^2_e$ with all covariances equal to zero,

$j = 1,....,n_i$ for unbalanced data. Thus the variances associated with the random effects are called variance components.

Scheffe` (1956) refers to the astronomer Airy as being an early user of variance components in a random effects model in 1861. Scheffe` (1956) also makes reference to Legendre and Gauss in 1806 and 1809 as using variance components in fixed effects models in an application in astronomy.

Searle et al (1992) point out that R.A. Fisher's book clearly makes a significant contribution to variance components (1925, Sec 40). The name of the book was "Statistical Methods for Research Workers" and covers an analysis of variance method of estimation by taking expectations of sums of squares. The pivotal idea of Fisher's method was that in his field of quantitive genetics, he needed a more exact analysis of human variability.

Searle et al (1992) have made reference to N.S. Urquhart et al who in 1973 have given a great deal of consideration to L. C. Tippett (1931, Sec 6.1. 6.2, 10.3) as making significant contributions to the role of the analysis of variance method.

Neyman et al (1935) have used linear mixed models in experimental design and is said to have used the term "variance component" in their paper. However it was Daniels (1939) who was first to use the term "variance component" as well as the facts that he derived from sampling variances for a three-way crossed classification random effects model with interactions and expected mean squares implying that the population effects for a random factor could be of finite size. Most of the historical cases covered above illustrated the application of linear mixed models to balanced data.

For unbalanced data, Cochran (1939) applied the 1-way classification model using the analysis of variance method to estimate the variance components. Interestingly Ganguli (1941) and Crump (1946) further showed that the analysis of variance (ANOVA) method of estimating variance components can produce negative estimates for variances.

By far the greatest advances in analysing unbalanced data were made by Henderson (1953) in his work on the analysis of dairy cow records. Henderson presented three different ways of applying the ANOVA method to unbalanced data. These three ways are known as Henderson's Methods I, II and III. Methods II and III will be discussed in detail in the following chapters.

However, Henderson et al (1974) went on to reveal that to apply Methods II and III was computationally difficult to the extent of being impractical. Henderson et al (1974) promoted the searching for other techniques to compensate for the weaknesses of the ANOVA method which could produce negative estimates for the variance components and had a lack of distributional properties of the estimates.

The other techniques that were developed after Henderson et al (1974) will be discussed but we will firstly focus our attention to Henderson's methods II and III by looking at the general formulation of the mixed model and then on methods II and III. Thereafter our link would be to look at the other techniques that have been formulated.

2.2 The Mixed Model

2.2.1 Introduction

Searle (1988, p. 940) stated that it was Jackson who used a description of a mixed model in 1939, even though it was not termed a "mixed model" as such. Searle concluded that in describing a 2-factor-no-interaction situation Jackson (1939) writes that one factor is a "measure of the trial effect" and the other is a "measure of the individual effect". This seems to be the first occurrence of the word "effect" in what is now its customary usage in linear models. Jackson also described his model as having one factor random and one non-random which seems like a clear specification of a mixed model, although not called such at the time.

Associated with the term mixed model is whether or not it has fixed or random effects and whether it can be applied to balanced data or unbalanced data. So we first need to describe and define these concepts.

A. Balanced and unbalanced data

i. Data can be described as balanced when each cell in the data set contains the same number of observations and as unbalanced when this is not the case.

ii. Unbalanced data occur when the design of the experiment force the data to be so, in other words, when there is "planned"unbalancedness.

Unbalanced data can also result from unfortunate circumstances or experimental carelessness, for example if the experimenter loses some of the data points. As a result the cell containing these missing observations vary in number of observations with respect to other cells.

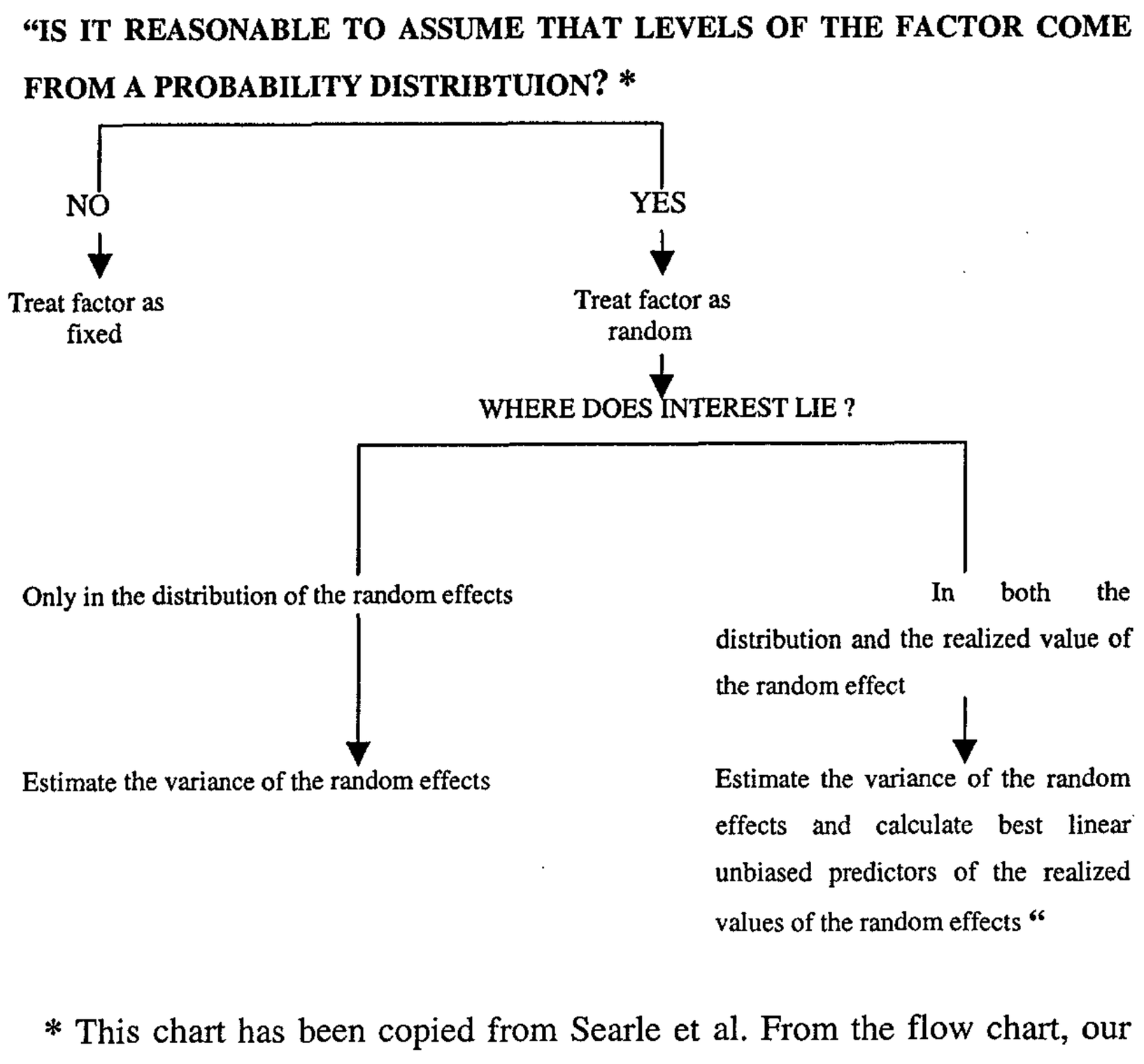
Please note that with respect to my own data set, *survey data* is always considered to be unbalanced and hence the focus of my study has to be in the analysis of unbalanced data.

B. Fixed and random effects models

i. A factor in a model is *random* if its levels consist of a random sample from a population of all possible levels. A model is termed a random effects model if all the factors in the treatment structure are random.

ii. A factor in a model *is fixed* if its levels are selected by a non-random process or consists of the entire population of all possible levels. A model is termed a fixed effects model if all the factors in the treatment structure are fixed effects.

Searle et al (1992, p.18) has given the following diagram/flow chart to help decide on the appropriate model to be fitted (as an adequate representation of the data):



* This chart has been copied from Searle et al. From the flow chart, our interest lie in the distribution of the random effects and estimating the variance of the random effects. At this point it is important to state that the

model that is an adequate representation of our smoking survey data is a mixed model i.e. a model with fixed and random effects. The variance of the random effects, σ^2(random) is called the variance component of the random effects because it refers to the components of variance of an observation.

2.2.2 The general linear mixed model

We will use the matrix formulation of the model in Searle et al (1992, pp 138-140, 170-173) and its derivation. We have already pointed out that our data is represented by a mixed model and we will now consider the general linear mixed model and its properties relevant to this thesis.

The equation of the general linear mixed model is

$$\mathbf{y}=\mathbf{X}\beta+\mathbf{Z}\mathbf{u}+\mathbf{e} \quad \ldots \quad 2.2.2.1$$

y is the vector that contains the responses;

β is the vector that contains the overall mean and all the fixed effects (main effects and possibly interactions) parameters;

X is a known matrix (design matrix, incidence matrix);

Z is a known matrix (the design or incidence matrix for the random effects);

u is the vector that contains all the random effects variables (**u** can be partitioned as $\mathbf{u}' = [\,\mathbf{u}_1' \; \mathbf{u}_2' \; \ldots \; \mathbf{u}_r' \,]$. The definition of **e**, the vector of residual errors, can be obtained from first defining:

$$E(\mathbf{y}) = \mathbf{X}\beta \text{ and } E(\mathbf{y}|\mathbf{u}) = \mathbf{X}\beta + \mathbf{Z}\mathbf{u}$$

Thus $\mathbf{e} = \mathbf{y} - E(\mathbf{y}|\mathbf{u})$ ... 2.2.2.2

From this it follows that $E(\mathbf{e}) = \mathbf{0}$ and the variance–covariance structure for $\mathbf{e}$ is

$$\text{var}(\mathbf{e}) = \sigma^2_e \mathbf{I}_N \qquad \ldots \qquad 2.2.2.3$$

If q_i represents the number of elements in $\mathbf{u_i}$, i.e. the number of random levels of the factor corresponding to $\mathbf{u_i}$ (which is 40 in the survey data set and 10 in the simulated data set) we also assume that

$$\text{Var}\ (\mathbf{u_i}) = \sigma^2_i \mathbf{I}_{q_i} \ \forall\ i \qquad \ldots \qquad 2.2.2.4$$

Also assuming that cov $(\mathbf{u_i}, \mathbf{u_j}') = \mathbf{0}$ $\quad \forall\ i \neq j$ and cov $(\mathbf{u}, \mathbf{e}') = \mathbf{0}$, the variance structure of $\mathbf{u}$ is:

$$\text{Var}(\mathbf{u}) = \mathbf{D} = \begin{bmatrix} \sigma^2_1 \mathbf{I}_{q1} & . & . & . & . \\ . & \sigma^2_2 \mathbf{I}_{q2} & . & . & . \\ . & . & . & . & . \\ . & . & . & . & . \\ . & . & . & . & \sigma^2_r \mathbf{I}_{qr} \end{bmatrix} = \{_d \sigma_i^2 \mathbf{I}_{q_i}\} \qquad \ldots \qquad 2.2.2.5$$

where $\{_d \sigma_i^2 \mathbf{I}_{q_i}\}$ is a diagonal matrix.

If we partition $\mathbf{Z}$ conformably with $\mathbf{u}$ as $\mathbf{Z} = [\mathbf{Z_1},\ldots,\mathbf{Z_r}]$ we can write our model as

$$\mathbf{y} = \mathbf{X\beta + Zu + e} = \mathbf{X\beta} + \sum_{i=1}^{r} \mathbf{Z_i u_i} + \mathbf{e}$$

So that $\mathbf{V} = \text{Var}(\mathbf{y}) = \mathbf{ZDZ}' + \sigma^2_e \mathbf{I} = \sum_{i=1}^{r} \sigma^2_i \mathbf{Z_i Z_i}' + \sigma^2_e \mathbf{I}_N \qquad \ldots \qquad 2.2.2.6$

Alternatively I could use the form of the model given by Hartley and Rao (1967) i.e. to put $\mathbf{V} = \sigma^2\ \mathbf{H}$

where $\mathbf{H} = \sum_{i=1}^{r} \gamma_i \mathbf{Z_i Z_i}' + \mathbf{I_n}$ and $\gamma_i = \dfrac{\sigma_i^2}{\sigma_o^2}$ $\qquad i = 1,\ldots,r.$

Hartley and Rao (1967) then give us a different formulation of the mixed model i.e.

$\mathbf{u_0} \equiv \mathbf{e}$, $z_o = \mathbf{I_N}$ and $\sigma^2_0 = \sigma^2_e$

Then $\mathbf{y} = \mathbf{X\beta} + \sum_{i=1}^{r} \mathbf{Z_i}\, u_i' \qquad \ldots \qquad 2.2.2.7$

So that **V** from equation 2.2.2.6 will read

$$\mathbf{V} = \sum_{i=1}^{r} \mathbf{Z}_i \mathbf{Z}_i' \sigma^2_i \qquad \ldots \qquad 2.2.2.8$$

There are several advantages to the application of the mixed model to survey data and Duchateau et al (1998, p.18)) gives the specific features of the mixed model as advantages that include:

- Complex data structures can be described in a natural way by the mixed model;
- The analysis of unbalanced data is a natural extension of the analysis of balanced data in the mixed model framework;
- The mixed model framework allows a flexible choice of the appropriate inference space;
- A mixed model also allows the prediction of random effects of interest by best linear unbiased prediction (BLUP).

Applying the general mixed model to my data set would firstly require me to give a brief description of the data set. The reason for its brief description is because the data set will be described in greater detail in Chapter 3. The data was taken as part of the October Omnibus Smoking Survey carried out by the Human Sciences Research Council of South Africa. The survey was carried out as a multistage cluster survey. The 40 clusters sampled from Gauteng have come from a random sample of clusters from different provinces in South Africa. I have all the clusters from Gauteng. In my data set there are 343 observations of which we have unequal numbers of repeated measurements of each cluster i.e we have 7 observations representing cluster 1 but 9 observations representing cluster 28 etc. Clearly now it is evident that I have unbalanced data. The variables that were surveyed include smoking status, race, sex, age, socio-economic status, marital status and education level. These variables are all fixed effects because their levels were selected by a nonrandom process (for example race has 4 levels and it was decided beforehand that race will have the levels : black, coloured, indian and

white) and are represented by β: 45x1 in equation 2.2.2.1. The clusters represent our random effect i.e we have only one random effect in this model and $\mathbf{u} = [\mathbf{u}_1]$ represents it in equation 2.2.2.1 and the rest of the derivation. The clusters are random effects because they were clearly selected by a random sample from all levels of the population. Our incident matrices from equation 2.2.2.1 are **X:** 343 x 45 and

Z : 343 x 40 whilst our error terms, **e** :343 x 1 is a vector of 343 rows. It is important to note here that I only have one random variable for both the data sets i.e. the smoking survey data set and the simulated data set. Therefore we have a mixed model with one random effect and seven fixed effects to analyse. The application of the theory of the general mixed model to my data set will lead to the following:

$\text{Var}\ (u_1) = \sigma^2_{cl}$, $\text{Var}(\mathbf{u}) = \sigma^2_1 I_{40}$, $\mathbf{Z} = Z_1$,

$$\mathbf{y} = \mathbf{X\beta+Zu+e} = \mathbf{X\beta}+\sum_{i=1}^{r} \mathbf{Z}_i \mathbf{u}_i + \mathbf{e} = \mathbf{X\beta}+Z_1 u_1 + \mathbf{e}$$

$$\mathbf{V}= \text{Var}(\mathbf{y}) = \mathbf{ZDZ'} + \sigma^2_e \mathbf{I} = \sum_{i=1}^{r} \sigma^2_i \mathbf{Z}_i \mathbf{Z}_i' + \sigma^2_e \mathbf{I}_N = \sigma^2_1 Z_1 Z_1' + \sigma^2_e \mathbf{I}_N \ (r = 1)$$

We will consider now the first technique for estimating variance components for mixed models i.e Henderson's method III. The reason we consider this method is because:

- Henderson produced a first landmark paper to deal with analysing unbalanced data and mixed models using three methods viz. I, II, III for estimating variance components from dairy cow records;
- method III is used by a SAS procedure Varcomp which calculates ANOVA(via method III) estimators from the SAS type I sums of squares .

2.3 Henderson's methods

2.3.1 Method I

It should be noted that method I can not be applied to mixed models and hence it has been left out.

2.3.2 Method II

Method II can be applied to mixed models with no interactions between fixed and random effects and is beyond the scope of this thesis. Method II is not very practical in estimating variance components and some of its disadvantages include:

- negative estimates can arise ;
- sampling variances of estimators in closed forms are not obtainable except under certain conditions.

2.3.3 Method III

Henderson's Method III entails the borrowing of the sum of squares from the fixed model which is utilised as the reduction in sums of squares due to the fitting of one model and thereafter fitting various submodels of it. We firstly need to look at some of the theory of method III from Searle et al (1992; Chapter 5; pp 202-208) to give us some indication about what method III incorporates and how it can be applied to my data set.

A. Theory

The following theory has been taken from Searle et al (1992, pp.202-205). The general mixed model equation is given as $\mathbf{y = X\beta + Zu + e}$ as in equation 2.2.2.1. If we say that there is no distinction between β and **u** i.e. between the fixed and random effects then we can rewrite our model as

$$\mathbf{y = Wb + e} \qquad \ldots \qquad 2.3.3.1$$

The model in 2.3.3.1 can be considered to be a fixed effects model forgetting all about the differences between fixed and random effects. Hence our best linear unbiased estimator of **Wb** is given as

$$\text{BLUE } (\mathbf{Wb}) = \mathbf{Wb^0 = W(W'W)^- W'Y = WW^+Y}$$

where $\mathbf{W^+ = (W'W)^- W'}$

Our residual sum of squares is now (after fitting the model)

$$\mathbf{SSE = (y-Wb^0)' (y-Wb^0)}$$

$$\mathbf{= y'y - y'Wb^0 - (Wb^0)'y + (Wb^0)' (Wb^0)}$$

$$\mathbf{= y'y - y'W\,W^+ y}$$

so that $\mathbf{y'y = SSE + y'WW^+y}$

$$\mathbf{SST = SSE + SSR}$$

To pay attention to estimating the variance components of the mixed model we look closely at the reduction in sum of squares due to fitting the model $\mathbf{y = Wb+e}$ i.e. the

SSR = R(**b**) [give SSR the R(b) notation]

$$\mathbf{= y'W\,(W'W)^- W'y = y'WW^+y}$$

To estimate the variance components of the random effects, partition **Wb** in the following way

$E(\mathbf{y}) = \mathbf{W_1 b_1} + \mathbf{W_2 b_2}$ with the following results that

$R(\mathbf{b}) = R(\mathbf{b_1}, \mathbf{b_2})$

$R(\mathbf{b_1},\mathbf{b_2}) = \mathbf{y}' [\mathbf{W_1}\ \mathbf{W_2}][\mathbf{W_1}\ \mathbf{W_2}]^{+}\mathbf{y}$

So that the difference between the two reductions in sum of squares SSR is given as

$R(\mathbf{b_2}|\mathbf{b_1}) = R(\mathbf{b_1},\mathbf{b_2}) - R(\mathbf{b_1})$

$= \mathbf{y}' [\mathbf{W_1 W_2}][\mathbf{W_1 W_2}]^{+} - \mathbf{y}'\mathbf{W_1}\mathbf{W_1}^{+}\mathbf{y}$

(note that $\mathbf{WW}^{+} = \mathbf{W}(\mathbf{W}'\mathbf{W})^{-}\mathbf{W}$)

$\therefore R(\mathbf{b_2}|\mathbf{b_1}) = \mathbf{y}'\mathbf{M_1 W_2}\ (\mathbf{W_2}'\mathbf{M_1 W_2})^{-}\mathbf{W_2}'\mathbf{M_1 y}$

Where $\mathbf{M_1} = \mathbf{I} - \mathbf{W_1}\ (\mathbf{W_1}'\mathbf{W_1})^{-}\mathbf{W_1}'$ ... 2.3.3.2

$= \mathbf{M_1}'$

$= \mathbf{M_1}^2$ with $\mathbf{M_1 W_1} = \mathbf{O}$

For our mixed model $\mathbf{y}=\mathbf{X}\beta + \mathbf{Zu} + \mathbf{e}$ in equation 2.2.2.1 it follows that

$\mathbf{M} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}' = \mathbf{M}' = \mathbf{M}^2$ with $\mathbf{MX} = \mathbf{O}$

Keeping in mind that $R(\mathbf{b_1}|\mathbf{b_2}) = \mathbf{y}'\mathbf{M_1 W_2}\ (\mathbf{W_2}'\mathbf{M_1 W_2})^{-}\mathbf{W_2}^{-}\mathbf{M_1 y}$ and $\mathbf{M_1}$ is given in equation 2.3.3.2 we see that

$$E\,R(\mathbf{b_1}|\mathbf{b_2}) = \mathrm{tr}\,\{ \begin{bmatrix} \mathbf{W_1}' \\ \mathbf{W_2}' \end{bmatrix} [\mathbf{M_1 W_2}\ (\mathbf{W_2}'\mathbf{M_1 W_2})^{-}\mathbf{W_2}'\mathbf{M_1}] \cdot$$

$$[\mathbf{W_1 W_2}]\ E\ (\mathbf{bb}')\} + \sigma^2_e\, \mathrm{tr}\ [\mathbf{M_1 W_2}(\mathbf{W_2}'\mathbf{M_1 W_2})^{-}\mathbf{W_2}'\mathbf{M_1}]$$

Analogously for our mixed model we have:

$E\,R(\mathbf{u}|\beta) = \mathrm{tr}\ [\mathbf{Z}'\mathbf{MZ}E(\mathbf{uu}')] + \sigma^2_e\,(r[\mathbf{X}\ \mathbf{Z}] - r\ [\mathbf{X}])$

$E\,R(\mathbf{u}|\beta) = \mathrm{tr}\ [\mathbf{Z}'\mathbf{MZ}\ \{_d\ \sigma^2_i\ \mathbf{I_{qi}}\}] + \sigma^2_e\,(r[\mathbf{X}\ \mathbf{Z}] - r\ [\mathbf{X}])$

$= \Sigma_i\ \mathrm{tr}\ (\mathbf{Z_i}'\mathbf{MZ_i})\ \sigma^2_i + \sigma^2_e\,(r[\mathbf{X}\ \mathbf{Z}] - r\ [\mathbf{X}])$ ... 2.3.3.3

B. Advantages and disadvantages

- method III can be applied to ALL models;
- all estimates are unbiased;
- methods I, II, III are the same for balanced data.
- method III can be expressed in a variety of ways when there are two or more crossed factors i.e. there is indecision in which way of Method III to use ;
- closed form expressions of sampling variance are NOT available.

Henderson's method III is also known as the " fitting of constants" method. Applying method III to my data set i.e. applying method III to a mixed model with just one random factor to compute the variance components of the random term (clusters) and the error term yields the following results given by Searle (1989 , pp. 465-466). Our mixed model is given as in equation 2.2.2.1 i.e

$\mathbf{y=X\beta +Zu+e}$ with rank of $\mathbf{X}$= r($\mathbf{X}$)= r and $\mathbf{u}$ representing the random levels as stated earlier, containing t levels for just one random factor with variance σ_u^2 . As a result $\mathbf{Z}$ has full column rank, t, (i.e. t = 40) with its columns summing to **1**. Hence we have the following relationship of the columns of $\mathbf{Z}$ to $\mathbf{X}$:

$$r[\,\mathbf{X}\;\mathbf{Z}\,] = r(\mathbf{X}) + t - 1 = r + t - 1$$

Searle (1997, p. 466) then gives the following estimation using Henderson' s method III:

SSE = sums of squares of the error terms = $\mathbf{y'y}$ - R($\mathbf{b}$, $\mathbf{u}$)

R($\mathbf{u}$ | $\mathbf{b}$)= R($\mathbf{b}$, $\mathbf{u}$) - R($\mathbf{b}$)

With E(SSE) = [N- (r + t – 1)] σ_e^2 and

$$E[R(\mathbf{u} \mid \mathbf{b})] = \sigma_u^2 \, tr[\mathbf{Z'Z} - \mathbf{Z'X(X'X)^{-}X'Z}] + \sigma_e^2 [r(\mathbf{X}) + t - 1 - r(\mathbf{X})]$$

Hence our estimators are

$$\hat{\sigma}_e^2 = \frac{\mathbf{y'y} - R(\mathbf{b},\mathbf{u})}{N - r(\mathbf{X}) - t + 1}$$

$$\hat{\sigma}_u^2 = \frac{R(\mathbf{u} \mid \mathbf{b}) - \hat{\sigma}_e^2 (t - 1)}{tr[\mathbf{Z'Z} - \mathbf{Z'X(X'X)^{-}X'Z}]}$$

Our link to the next section would be to look at Henderson's Mixed Model equations with their (BLUP) solutions because Searle et al (1992, Chapter 7, pp 277) states that an interesting aspect of the mixed model equation is that the elements in them can be used to set up iterative procedures for calculating solutions to the Maximum Likelihood (ML) and the Restricted Maximum Likelihood equations for estimating variance components. Furthermore the same elements can also be used for setting up and calculating the information matrices of ML and REML. Therefore we consider Henderson's Mixed Model equations from Searle et al (1992, pp. 275 – 276).

2.4 Henderson's mixed model equations (MME)

The reason for considering Henderson's mixed model equations and their BLUP solutions is because the algorithms and methods of variance component estimation (such as REML to be used later on) used by statistical packages such as GENSTAT (which will be used extensively for this thesis) is based on the mixed model equations attributed to Henderson (1950). The algorithm uses BLUP estimation of the random effects (clusters). Robinson (1991) gives an excellent exposition of BLUP. A good general summary of the derivation of the mixed model equations is taken from Searle et al (1992, pp 275-276) as follows:
They are derived by maximizing the joint probability density function of **y** and **u**, for which the var (**e**) = **R** and var (**u**) = **D** (as in equation 2.2.2.5) is of order q.

$$f(\mathbf{y}, \mathbf{u}) = f(\mathbf{y}|\mathbf{u}).f(\mathbf{u})$$

$$= \frac{\exp\{-1/2\,[\mathbf{y}-\mathbf{X}\beta-\mathbf{Z}\mathbf{u}]'\mathbf{R}^{-1}[\mathbf{y}-\mathbf{X}\beta-\mathbf{Z}\mathbf{u}]+\mathbf{u}'\mathbf{D}^{-1}\mathbf{u}\}}{(2\pi)^{1/2(N+q)}\,|\mathbf{R}|^{1/2}|\mathbf{D}|^{1/2}}$$

If we take the partial derivatives of $f(\mathbf{y}, \mathbf{u})$ and equate them to zero with respect to elements first of β and then of **u** gives, using $\tilde{\beta}$ and $\tilde{\mathbf{u}}$ to denote the solutions

$$\mathbf{X}'\mathbf{R}^{-1}\mathbf{X}\tilde{\beta} + \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z}\tilde{\mathbf{u}} = \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \qquad \text{and}$$

$$\mathbf{Z}'\mathbf{R}^{-1}\mathbf{X}\tilde{\beta} + (\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{D}^{-1})\,\tilde{\mathbf{u}} = \mathbf{Z}'\mathbf{R}^{-1}\mathbf{y}$$

We can also write these equations as

$$\begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z}+\mathbf{D}^{-1} \end{bmatrix} \begin{bmatrix} \tilde{\beta} \\ \tilde{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{y} \end{bmatrix} \quad \ldots \quad 2.4.1$$

where $\mathbf{R} = \sigma_e^2\mathbf{I}$

Searle et al (1992, pp.276-277) gives the BLUP solutions as

$\mathbf{X'BX}\tilde{\boldsymbol{\beta}} = \mathbf{X'By}$ where $\mathbf{B} = \mathbf{V}^{-1}$ so that $\mathbf{X'}\,\mathbf{V}^{-1}\mathbf{X}\tilde{\boldsymbol{\beta}} = \mathbf{X'}\,\mathbf{V}^{-1}\,\mathbf{y}$.

And $\tilde{\mathbf{u}} = (\mathbf{Z'R}^{-1}\mathbf{Z} + \mathbf{D}^{-1})^{-1}\,\mathbf{Z'R}^{-1}(\mathbf{y}-\mathbf{X}\tilde{\boldsymbol{\beta}})$ reducing to $\tilde{\mathbf{u}} = \mathbf{DZ'V}^{-1}\,(\mathbf{y}-\mathbf{X}\tilde{\boldsymbol{\beta}})$.

... 2.4.2

A good example of BLUP that is very relevant to survey analysis and hence relevant to this thesis is that of small area estimation in Robinson (1991, section 6.8). This involves utilizing direct survey information from areas of individual interest together with information on similar or related areas. It has been found that more precise estimates can be made on the quantities of interest if information on the other areas have been ignored. One approach is to regard the quantities of interest in the individual areas as being the random effects which are to be estimated. Simpson et al (1996) gives a more recent application.

Searle et al (1992, pp 276) gives a worthy comment i.e. that equation 2.4.1 will give us the ML equations for $\mathbf{y}=\mathbf{X}\boldsymbol{\beta} + \mathbf{Zu} + \mathbf{e}$ if we treated $\mathbf{u}$ as representing fixed effects rather than random effects, only if we removed $\mathbf{D}^{-1}$ from the lower right hand sub matrix of the matrix on the left. With this in mind we move on to Maximum Likelihood estimation of variance components.

2.5 Maximum likelihood estimation (ML)

Searle et al (1992, pp. 40- 41) state that the method of maximum likelihood estimation was developed by Fisher in 1925 and then applied by Crump in 1947 and 1951. Searle et al also go on to state that by far the outstanding work in ML estimation is by Hartley and Rao in 1967 where their specification of a matrix of a mixed model led them to deriving the ML equations which can be solved iteratively. Searle (1970) has also contributed to ML estimation by deriving ML estimators for the 2-Way classification model for unbalanced data. Maximum likelihood estimation is an old, yet clear and efficient way of estimating variance components of data. The crucial requirement in estimating the variance components of a set of data using the maximum likelihood technique is the assumption of an underlying probability distribution for the data. Inclusive of unbalanced data, the assumption of a normal distribution is chosen as an underlying distribution. The following ML derivation, estimation and solutions with properties is done using Fisher's Scoring method and taken from Searle et al (1992; pp232 – 242; pp278 – 281; p.285).

2.5.1 The maximum likelihood equations

In general, if a vector $\mathbf{x}$ of random variables is distributed $\sim N(\mu, \mathbf{V})$ we know that the density function of $\mathbf{x}$ is given as

$$f(x) = \frac{\exp(-(1/2)(\mathbf{x}-\mu)'\mathbf{V}^{-1}(\mathbf{x}-\mu))}{(2\pi)^{(1/2)N}\,|\mathbf{V}|^{1/2}}$$

So for our $\mathbf{y}$ we have $\mathbf{y} \sim N_N(\mathbf{X}\beta, \mathbf{V})$ to lead us to the likelihood function with parameters in β and $\mathbf{V}$ to be

$$\mathbf{L=L(\beta\,,\,V|\,y)=}\frac{\exp(-(1/2)(\mathbf{y-X\beta})'\mathbf{V^{-1}}(\mathbf{y-X\beta}))}{(2\pi)^{(1/2)N}\,|\mathbf{V}|^{1/2}} \quad \ldots \quad 2.5.1$$

We now maximize **L** in equation 2.5.1 by taking the log of both sides of the equation and maximise log **L** with respect to β and σ_i^2 in equation 2.5.1 Therefore we denote the log L in equation 2.5.1 as *l* to give us:

$$l = \log L = -\tfrac{1}{2}N \log 2\pi - \tfrac{1}{2}\log|\mathbf{V}| - \tfrac{1}{2}(\mathbf{y}-\mathbf{X\beta})'\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X\beta}) \quad \ldots \quad 2.5.2$$

Differentiating *l* with respect to β will be denoted as l_β so we get

$$l_\beta = \frac{\partial l}{\partial \beta} = \mathbf{X'V^{-1}y} - \mathbf{X'V^{-1}X\beta} \quad \ldots \quad 2.5.3$$

Now differentiating equation 2.5.2 with respect to σ^2_i and making use of

$$\frac{\partial \mathbf{V}}{\partial \sigma_i^{\,2}} = \mathbf{Z_i\,Z_i'} \quad \text{for i} = 0,\ldots..\text{r} \quad \text{leads to}$$

$$l_{\sigma^2_i} = \frac{\partial l}{\partial \sigma^2_i} = -\tfrac{1}{2}\,\mathrm{tr}\,(\mathbf{V^{-1}Z_i\,Z_i'}) + \tfrac{1}{2}\,(\mathbf{y}-\mathbf{X\beta})'\mathbf{V^{-1}}\,\mathbf{Z_i\,Z_i'}\,\mathbf{V^{-1}}\,(\mathbf{y}-\mathbf{X\beta})$$

$$\ldots \quad 2.5.4$$

Searle et al (1992, p. 236) then gives the ML equations as

$$\mathbf{X'\dot{V}^{-1}X\dot{\beta}} = \mathbf{X'\dot{V}^{-1}\,y} \quad \ldots \quad 2.5.5$$

and

$$\{_c\,\mathrm{tr}(\dot{\mathbf{V}}^{-1}\mathbf{Z_iZ_i'})\}^r_{i=0} = \{_c\,\mathbf{y'\dot{P}Z_iZ_i'\dot{P}y}\}^{\,r}_{i=0} \quad \ldots \quad 2.6.6$$

where $\{_c\, tr(\dot{\mathbf{V}}^{-1}\mathbf{Z}_i\mathbf{Z}_i')\}^r_{i=0}$ is a column vector and $\dot{\mathbf{V}}$ and $\tilde{\mathbf{V}}$ are estimates of $\mathbf{V}$ where $\dot{\sigma}^2$ and $\tilde{\sigma}^2$ are used, respectively in place of σ^2.

2.5.2 The maximum likelihood solutions using BLUP (Best Linear Unbiased Predictor)

Searle et al (1992, pp 278-279) gives the following solutions to the ML equations, that with the superscript (m) denoting computed values after m rounds of iteration we have

$$\sigma^2_e{}^{(m+1)} = [\mathbf{y}'(\mathbf{y} - \mathbf{X}\beta^{o\,(m)} - \mathbf{Z}\tilde{\mathbf{u}}^{(m)})] / N$$

$$\sigma_i^{2\,(m+1)} = [\tilde{\mathbf{u}}_i'^{(m)}\,\tilde{\mathbf{u}}_i^{(m)} + \sigma_i^{2\,(m)}\,tr\,(\mathbf{W}_{ii}^{(m)})] / q_i \qquad \ldots \qquad 2.5.2.1$$

$$\text{or } \sigma_i^{2\,(m+1)} = \tilde{\mathbf{u}}_i'^{(m)}\,\tilde{\mathbf{u}}_i^{(m)} / [q_i - tr(\mathbf{W}_{ii}^{(m)})]$$

where $\mathbf{W}=(\mathbf{I}+\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z}\mathbf{D})^{-1}=\{\mathbf{W}_{ij}\}^r_{i,j=1}$ and $\mathbf{D}$ has q_i diagonal elements of σ_i^2.

Searle et al (1992, pp. 284- 285) give the working of ML estimation via iterating as follows:

Consider the set of equations as above:

$$\sigma_e^{2\,(m+1)} = [\mathbf{y}'(\mathbf{y} - \mathbf{X}\beta^{(m)} - \mathbf{Z}\mathbf{u}^{(m)})] / N \qquad \text{(equation a)}$$

$$\sigma_i^{2\,(m+1)} = [\tilde{\mathbf{u}}_i'^{(m)}\,\tilde{\mathbf{u}}_i^{(m)} + \sigma_i^{2\,(m)}\,tr\,(\mathbf{W}_{ii}^{(m)})] / q_i \qquad \text{or (equation b)}$$

$$\sigma_i^{2\,(m+1)} = \tilde{\mathbf{u}}_i'^{(m)}\,\tilde{\mathbf{u}}_i^{(m)} / [q_i - tr\,(\mathbf{W}_{ii}^{(m)})]$$

1. Decide on starting values $\sigma_e^{2\,(0)}$ and $\sigma_i^{2\,(0)}$, and set m = 0.

2. Calculate $\mathbf{W}^{(m)} = (\sigma_e^{2\,(m)}\mathbf{I} + \mathbf{Z'ZD}^{(m)})^{-1}\,\sigma_e^{2\,(m)}$ where $\mathbf{v}$ is defined as $\mathbf{Dv} = \mathbf{u}$.

(Remember var $(\mathbf{u}) = \mathbf{D}$)

Now solve for $\boldsymbol{\beta}^{(m)}$ and $\mathbf{v}^{(m)}$ and then calculate

$$\mathbf{u}^{(m)} = \mathbf{D}^{(m)}\,\mathbf{v}^{(m)} \quad \text{from} \quad \begin{bmatrix} \mathbf{X'X} & \mathbf{X'ZD}^{(m)} \\ \mathbf{Z'X} & \mathbf{W}^{(m)} \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}^{(m)} \\ \mathbf{v}^{(m)} \end{bmatrix} = \begin{bmatrix} \mathbf{X'y} \\ \mathbf{Z'y} \end{bmatrix}$$

3. Now calculate $\sigma_e^{2\,(m+1)}$ and either of the expressions for $\sigma_i^{2\,(m+1)}$.

4. If convergence is reached for the $\sigma^{2'}$s, set $\tilde{\boldsymbol{\sigma}}^2 = \boldsymbol{\sigma}^{2\,(m+1)}$. Denote the resulting terms as $\tilde{\mathbf{W}} = \mathbf{W}^{(m+1)}$, $\tilde{\boldsymbol{\beta}} = \boldsymbol{\beta}^{(m+1)}$, $\tilde{\mathbf{v}} = \mathbf{v}^{(m+1)}$ and $\tilde{\mathbf{u}} = \mathbf{u}^{(m+1)}$. Use $\tilde{\boldsymbol{\sigma}}^2$ and $\tilde{\mathbf{W}}$ to calculate the information matrix, $\mathbf{I}\,(\tilde{\boldsymbol{\sigma}}^2{}_{ML})$.

5. If convergence is not reached, increase m by unity and return to step 2. At each repeat of step 3 use which ever of the equations a or b, that was used on the first occasion.

2.5.3 Notes

Whatever form of the equations we choose to use, Searle et al (1992, pp 242-243) raise the following questions about the iterative solution to the equations;

a. What is the best method of iteration and what form of equations is most suitable for iteration?
b. Is convergence of iteration always assured, and if convergence is assured and achieved can we be sure that we have a global maximum of the likelihood rather than a local maximum?

c. Do the initial starting values of σ^2 affect the number of iterations required for convergence and if so, is there any set of starting values that will always yield convergence of the global maximum of the likelihood.

d. The matrix **V** is by definition always non-negative definite and usually positive definite. What happens then if at some step in the iteration, the calculated σ^2 yields that the calculated **V** is NOT positive definite?

e. The ML estimators σ_e^2 and σ_j^2 must satisfy the constraints that $\sigma_e^2 > 0$ and $\sigma_j^2 \geq 0$ for i = 1,.....r.

f. Computer programs must be able to satisfy these constraints and if any σ_j^2 be set to zero or can we replace the negative solution by a small positive number?

These problems/points are better left to computational experts as said by Searle et al (1992, pp 242-243).

There is another form of the maximum likelihood equations due to Hartley and Rao (1967) but we will not consider them for the derivation and properties in this thesis

- The dispersion matrix of the parameter estimates is estimated using the inverse of the information matrix, by substituting the maximum likelihood equations for the parameters in the information matrix. Longford (1993) points out that this becomes problematic for small data sets since;
    a. The asymptotic properties may not apply to them.
    b. There is a lot of uncertainty about the parameters involved in the information function and the information function may vary substantially with the parameters.
- There is an alternative method for evaluating the parameters in our model developed by Goldstein (1986, 1995) called Iterative Generalized Least Squares (IGLS). The maximum likelihood iterations involve the generalized least squares (GLS) estimator

$$\hat{\beta}_{GLS} = (X' V^{-1} X)^{-1} X V^{-1} y.$$

When V is known this gives an unbiased estimator of β, even if the outcomes are not normally distributed.

Thus if we knew V we could estimate β using the GLS estimator but seeing that in practice V is unknown, we need to estimate the parameters.

- There are other methods of estimation, including methods based on the EM Algorithm and Gibbs sampling. These methods will not be addressed in this thesis. Longford (1993, Section 2.5) covers methods based on the EM Algorithm as well as Goldstein (1995, Appendix 2.3) also covers these methods. Methods based on Gibbs sampling are discussed by Goldstein (1995; Appendix 2.4);
- Another approach to estimation is that of:

  Generalized Estimating Equations by Liang and Zeger (1986).

  This method uses the Generalized Least Squares to estimate β, as in ML and IGLS. Generalized Estimating Equations (GEE) gets estimates of V using simple regression procedures based upon functions of the actual calculated raw residuals. It is concerned with estimating and modeling the fixed coefficients rather than exploiting the random component of a model.

  The coefficient estimates (fixed) are consistent but not fully efficient.
- The algorithm as given also requires the inversion of an N x N matrix where N is the total number of units (N). In our smoking survey data N is equal to 343. Longford (1993) points out that **V** is block diagonal with blocks given by $\mathbf{V_i}$. In the case of my data, with only one level of clustering i.e with only one random component, it follows from Longford that

$$\mathbf{V_i}^{-1} = (1/\sigma_o^2)\,\mathbf{I}_{ni} - (\sigma_1^2/\sigma_o^4 g_i)\,\mathbf{J_n}$$

Where $\det(\mathbf{V_i}) = \sigma_o^{2ni}\, g_i$ and $\mathbf{J_{ni}}$ is an n_i x n_i matrix with every entry equal to 1 and $g_i = 1 + n_i\, \sigma_1^2/\sigma_o^2$.

Applying the ML equations to my data set we would have the following

$$\sigma_e^{2\,(m+1)} = \mathbf{y}'\,(\mathbf{y} - \mathbf{X}\beta^{(m)} - \mathbf{Z}u^{(m)}) / 343$$

$$\sigma_{cl}^{2\,(m+1)} = [\,\tilde{\mathbf{u}}_1'^{(m)}\,\tilde{\mathbf{u}}_1^{(m)} + \sigma_{cl}^{2\,(m)}\,\mathrm{tr}\,(\mathbf{W}^{(m)})] / 40$$

where I only have one random factor with one vector $\mathbf{u_1}' = [u_1 \ldots u_{40}]'$ and σ_{cl}^2 would be the variance of the random sample of 40 clusters whilst σ_e^2 will be our variance of our errors .

Finally to lead on to the next section in this chapter W.A. Thompson (1962) also considered the method of ML estimation except he then introduced the idea of maximizing part of the likelihood which is invariant to the model's fixed effects. This will now be considered in our next section in this chapter. Another link that leads us on is that, one of the criticisms leveled at ML, is that it does NOT take into account the loss of degrees of freedom that results from estimating the fixed effects β.

2.6 Restricted maximum likelihood (REML)

W.A. Thompson (1962) called this method restricted maximum likelihood or residual (marginal) maximum likelihood. H.D. Patterson and R. Thompson (1971) then showed how REML could be applied to unbalanced data. In maximum likelihood estimation one does not take into account the relevant degrees of freedom used to estimate the fixed effects. One of the attractive features of REML is that it takes into account the degrees of freedom in the variance component estimation and it is based on the idea of estimating the variance components via the residuals calculated after fitting by ordinary least squares of just the fixed effects part of the model. The relevance in considering REML in this detail is because we will use GENSTAT to estimate the variance components using REML. The following summary of REML using Fisher's Scoring Method with its solution and properties is taken directly out of Searle et al (1992; pp. 249 – 253; 282 – 287).

2.6.1 The basis

REML is often interpreted as a technique that is based on linear combinations of **y**, not forgetting that these linear combinations do not contain any fixed effects.

Not surprisingly these linear combinations of values not containing any fixed effects turn out to be equivalent to the residuals obtained after we fit the model the fixed effects. Consider the set of values $\mathbf{k'y}$ where vectors of the form $\mathbf{k'}$ can be chosen to satisfy

$\mathbf{k'y} = \mathbf{k'X}\,\beta + \mathbf{k'Zu}$ such that no term in β is contained. i.e. $\mathbf{k'X}\,\beta = \mathbf{0} \quad \forall \beta.$ ... 2.6.1.1

$\Rightarrow \mathbf{k'X} = \mathbf{0}$ ... 2.6.1.2

Now $\mathbf{k}'$ has to have the form of $\mathbf{k}' = \mathbf{c}' (\mathbf{I} - \mathbf{X}\mathbf{X}^-)$

or alternatively $\mathbf{k}' = \mathbf{c}' [\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^-\mathbf{X}'] = \mathbf{c}' (\mathbf{I} - \mathbf{X}\mathbf{X}^+)$

$= \mathbf{c}'\mathbf{M}$

where $\mathbf{M}$ is defined in as $\mathbf{M} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^- \mathbf{X}' = \mathbf{I} - \mathbf{X}\mathbf{X}^+$.

Now we have $\mathbf{X}$ to be of order N x p and rank r so that 2.6.1.1 is satisfied by N-r vectors $\mathbf{k}'$ linearly independent vectors $\mathbf{k}'$ so that in utilising a set of linearly independent $\mathbf{k}'$ vectors as rows of $\mathbf{K}'$ we confine attention to $\mathbf{K}'\mathbf{y}$ for $\mathbf{K}'=\mathbf{TM}$ with $\mathbf{K}'$ and $\mathbf{T}$ of rank N-r.

2.6.2 REML equations

Searle et al (1992, pp. 251-252) gives the REML equations which we summarize as follows:

We have that $\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\beta}, \mathbf{V})$ for $\mathbf{K}'\mathbf{X} = 0$ so $\mathbf{K}'\mathbf{y} \sim N(\mathbf{0}, \mathbf{K}'\mathbf{V}\mathbf{K})$.

Using the Maximum Likelihood estimation equation we have that

$$\{_c \operatorname{tr} (\dot{\mathbf{V}}^{-1} \mathbf{Z}_i \mathbf{Z}_i')\}^r_{i=0} = \{_c \mathbf{y}' \dot{\mathbf{P}} \mathbf{Z}_i \mathbf{Z}_i' \dot{\mathbf{P}} \mathbf{y}\}^r_{i=0} \quad \ldots \quad 2.6.2.1$$

so if we make the following replacements:

Replace $\mathbf{y}$ by $\mathbf{K}'\mathbf{y}$ and $\mathbf{X}$ by $\mathbf{K}'\mathbf{X} = \mathbf{0}$

Replace $\mathbf{Z}$ by $\mathbf{K}'\mathbf{Z}$ and $\mathbf{V}$ by $\mathbf{K}'\mathbf{V}\mathbf{K}$ will now gives us the following direction where equation 2.6.2.1 will become

$$\{_c \operatorname{tr} [(\mathbf{K}' \dot{\mathbf{V}} \mathbf{K})^{-1} \mathbf{K}'\mathbf{Z}_i \mathbf{Z}_i'\mathbf{K}]\}^r_{i=0} = \{_c \mathbf{y}'\mathbf{K} (\mathbf{K}' \dot{\mathbf{V}} \mathbf{K})^{-1} \mathbf{K}'\mathbf{Z}_i \mathbf{Z}_i'\mathbf{K} (\mathbf{K}' \dot{\mathbf{V}} \mathbf{K})^{-1} \mathbf{K}'\mathbf{y}\}^r_{i=0}$$

$$\ldots \quad 2.6.2.2$$

and $\mathbf{P} = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^-\mathbf{X}'\mathbf{V}^- = \mathbf{K}(\mathbf{K}'\mathbf{V}\mathbf{K})^{-1} \mathbf{K}'$ so that equation 2.6.2.2 now becomes

$$\{_c \text{ tr } (\dot{\mathbf{P}}\ \mathbf{Z}_i\ \mathbf{Z}_i')\}^r_{i=0} = \{_c\ \mathbf{y}'\dot{\mathbf{P}}\mathbf{Z}_i\ \mathbf{Z}_i'\dot{\mathbf{P}}\mathbf{y}\}^r_{i=0} \qquad \ldots \qquad 2.6.2.3$$

It is clear that **PVP = P** so we can use the following identity to give us an alternate form of the REML equations (as given by Searle et al)

$$\{_m \text{ tr } (\mathbf{Z}_i'\ \dot{\mathbf{P}}\mathbf{Z}_j\ \mathbf{Z}_j'\ \dot{\mathbf{P}}\mathbf{Z}_i)\}_{i,}{}^r_{j=0}\ \dot{\sigma}^2 = \{_c\ \mathbf{y}'\ \dot{\mathbf{P}}\mathbf{Z}_i\ \mathbf{Z}_i'\dot{\mathbf{P}}\mathbf{y}\ \}^r_{i=0} \quad \ldots \qquad 2.6.2.4$$

The information matrix is extremely important to REML (and to Genstat) in that it plays a role in estimating the variance components. Searle et al (1992; pp252-253) also gives the information matrix as follows: Now the log – likelihood of $\mathbf{K}'\mathbf{y}$ is $\mathbf{l}_R$ i.e.

$\mathbf{l}_R$ = log (likelihood of $\mathbf{K}'\mathbf{y}$)

$$\mathbf{l}_R = -\frac{1}{2}\ (N\text{-}r) \log 2\pi - \frac{1}{2} \log |\mathbf{K}'\mathbf{V}\mathbf{K}| - \frac{1}{2}\mathbf{y}'\mathbf{K}(\mathbf{K}'\mathbf{V}\mathbf{K})\ \mathbf{K}'\mathbf{y}$$

Remember that $\mathbf{P} = \mathbf{K}\ (\mathbf{K}'\mathbf{V}\mathbf{K})^{-1}\mathbf{K}'$ so that

$$\frac{\partial \mathbf{P}}{\partial \sigma^2_i} = \frac{\partial\ \mathbf{K}(\mathbf{K}'\mathbf{V}\mathbf{K})^{-1}\ \mathbf{K}'}{\partial \sigma^2_i}$$

$$= -\mathbf{P}\ \mathbf{Z}_i\ \mathbf{Z}_i'\ \mathbf{P}$$

$$\frac{\partial^2 \mathbf{l}_R}{\partial \sigma^2_i} = -\frac{1}{2}\ \text{tr}\ [(\mathbf{P}\ \mathbf{Z}_i\ \mathbf{Z}_i') + \frac{1}{2}\mathbf{y}'\ \mathbf{P}\mathbf{Z}_i\ \mathbf{Z}_i'\ \mathbf{P}\mathbf{y}$$

With second derivatives of $\mathbf{I}_R$ we have

$$\frac{\partial^2 \mathbf{l}_R}{\partial \sigma^2_i \partial \sigma^2_j} = \frac{1}{2}\ \text{tr}\ (\mathbf{P}\ \mathbf{Z}_j\ \mathbf{Z}_j'\ \mathbf{P}\mathbf{Z}_i\mathbf{Z}_i') - \mathbf{y}'\ \mathbf{P}\ \mathbf{Z}_j\ \mathbf{Z}_j'\ \mathbf{P}\mathbf{Z}_i\ \mathbf{Z}_i'\ \mathbf{P}\mathbf{y}$$

2.6.3 Solutions to the REML equations using BLUP (Best linear unbiased predictor)

Searle et al (1992, pp 282-285) gives the following BLUP solutions

$$\sigma_i^{2\,(m+1)} = \frac{\tilde{\mathbf{u}}_1^{\prime(m)}\,\tilde{\mathbf{u}}_1^{(m)} + \sigma_i^{2\,(m)}\,\mathrm{tr}\,(\mathbf{T}_{ii}^{(m)})}{q_i}$$

or

$$\sigma_i^{2\,(m+1)} = \frac{\tilde{\mathbf{u}}_1^{\prime(m)}\,\tilde{\mathbf{u}}_1^{(m)}}{q_i - \mathrm{tr}\,(\mathbf{T}_{ii}^{(m)})}$$

$$\sigma_e^{2\,(m+1)} = \frac{\mathbf{y}'(\mathbf{y} - \mathbf{X}\beta^{o(m)} - \mathbf{Z}\tilde{\underline{\mathbf{u}}}^{(m)})}{N - r_x}$$

where and **D** has q_i diagonal elements of σ_i^2, $\mathbf{T} = (\mathbf{I} + \mathbf{ZSZD})^{-1}$ and $\mathbf{S} = \mathbf{R}^{-1} - \mathbf{R}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{R}^{-1}\mathbf{X})^{-}\,\mathbf{X}'\mathbf{R}^{-1}$ if we replace **T** for **W** in equation 2.5.2.1 we have the information matrix i.e. $I(\sigma^2_{REML})$

$$I(\sigma^2_{REML}) = \frac{1}{2}\begin{bmatrix} \dfrac{N\text{-}q. + \mathrm{tr}\,(\mathbf{T}^2)}{\sigma^4_e} & \left\{ \dfrac{\mathrm{tr}\,(\mathbf{T}_{ii}) - \sum_{k=1}^{r} \mathrm{tr}\,(\mathbf{T}_{ik}\,\mathbf{T}_{ki})}{\sigma^2_e\,\sigma^2_i} \right\}_{i=1}^{r} \\ \text{symmetry} & \left\{ \delta ij\,\dfrac{(q_i - 2\mathrm{tr}\,(\mathbf{T}_{ii})) + \mathrm{tr}\,(\mathbf{T}_{ij}\,\mathbf{T}_{ji})}{\sigma^2_i\,\sigma^2_j} \right\}_{i,j=1}^{r} \end{bmatrix}$$

where δij is the Kronecker delta with $\delta ii = 1$ and $\delta ij = 0$ for $i \neq j$ and $q. = \Sigma q_i$

The procedure for iterating for REML is as follows:

Consider the above set of equations and $T^{(m)}$ where

$$T^{(m)} = \left\{ \frac{I - Z'[I - X(X'X)^{-}X']ZD}{\sigma_e^{2(m)}} \right\}^{-1}$$

$$\sigma^2_e{}^{(m+1)} = \frac{y'(y - X\beta^{(m)} - Zu^{(m)})}{N - r_x}$$

$$\sigma^2_i{}^{(m+1)} = \frac{u_i'^{(m)} u_i^{(m)} + \sigma^2_i{}^{(m)} tr\,(T_{ii}{}^{(m)})}{q_i}$$

$$\sigma^2_i{}^{(m+1)} = \frac{u_i'^{(m)} u_i^{(m)}}{[q_i - tr\,(T_{ii}{}^{(m)})]}$$

1. Decide on starting values $\sigma^2_e{}^{(o)}$ and $\sigma^2_i{}^{(o)}$ and set m=0.
2. As for ML estimation, calculate:

$$W^{(m)} = (\sigma^2_e{}^{(m)} I + Z'ZD^{(m)})^{-1}\sigma^2_e{}^{(m)}$$

Solve $\begin{bmatrix} X'X & X'ZD^{(m)} \\ Z'X & W^{(m)} \end{bmatrix} \begin{bmatrix} \beta^{(m)} \\ v^{(m)} \end{bmatrix} = \begin{bmatrix} X'y \\ Z'y \end{bmatrix}$ for $\beta^{(m)}$, $v^{(m)}$ (MME's)

Calculate $u^{(m)} = D^{(m)} v^{(m)}$.

3. Calculate $T^{(m)}$.
4. Calculate $\sigma^2_e{}^{(m+1)}$ and either of the calculation formulae for $\sigma^2_i{}^{(m+1)}$.
5. If convergence is reached for the σ^2's, set $\hat{\sigma}^2 = \sigma^{2(m+1)}$ and do steps 2 & 3 using $\sigma^2_i{}^{(m+1)}$.

   Denote the resulting terms as

   $\tilde{W} = W^{(m+1)}$, $\tilde{\beta} = \beta^{(m+1)}$, $\tilde{v} = v^{(m+1)}$, $\tilde{u} = u^{(m+1)}$ and $\tilde{T} = T^{(m+1)}$. Use $\tilde{\sigma}^2$ and $\tilde{T}$ to calculate $I(\hat{\sigma}^2)$.
6. If convergence is not reached, increase m by one and return to step 2. At each repeat of step 4 use the $\sigma_i^{2(m+1)}$ that was used on the first occasion.

2.6.4 Notes

- Searle et al (1992, p.252) state that the REML equations don't contain **K** except only through its relationship to **P**. Thus we can express **P** as $\mathbf{P} = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}\,(\mathbf{X}'\,\mathbf{V}^{-1}\mathbf{X})\text{-}\mathbf{X}'\,\mathbf{V}^{-1}$ which does not involve **K**. This shows that the REML equations are invariant to a particular set of error contrasts that are chosen i.e. $\mathbf{K}'\mathbf{y}$, provided that $\mathbf{K}'$ is of full row rank N-p;
- REML estimation includes no procedure for estimating fixed effects;
  As with ML we can use alternative parameterisation by putting $\gamma_i = \sigma_i^2/\sigma_o^2$ for $i = 1,\ldots,r$ and then estimating σ_o^2 and $\gamma_1,\ldots,\gamma_r$;
- The algorithm requires **V** (N X N) to be inverted. However methods which take into account the structure of **V** can be used for this purpose. Engel (1990) gives computational details of the REML Algorithm;
- Gilmour, Thompson and Cullis (1995) have suggested that in areas such as animal breeding, Fisher scoring for REML presents a host of computational difficulties due to the large sizes of matrices to be inverted and multiplied. They have suggested then the use of an alternative "average information" or AI Algorithm which avoids the traces of large matrices which appear in the information matrices in the Fisher Scoring Algorithm for REML;
- Robinson (1987) did a study on the comparison between ML and REML and found out that for unbalanced data, REML estimators overcame the problem of bias which plagued the ML estimators. Robinson (1987) has suggested the use of REML:
  a. if a large number of degrees of freedom are required for the fixed effects.
  b. if we wish to use variance components to obtain more efficient and better estimates of other effects;
- Goldstein (1989) shows that if the response variable is normally distributed then the REML estimates are in fact equivalent to (the Iterative Generalized

Least Squares which can be modified to produce) the Restricted Iterative Generalized Least Squares;

- A question of weighting also arises in REML especially when the data set comes with sampling weights as in our case. The weights were firstly calculated according to the selection probabilities of the enumeration areas, then to take into account the number of eligible individuals in a household (for example, in the same enumeration area, a person in a household with 2 eligible respondents has twice the chance of being selected as a person with 4 eligible people). Finally the weights were adjusted by post stratification for race, gender and age group. There has been a lot of uncertainty and controversy whether to use weights or not, in sampling survey data. Korn and Graubard (1995a, 1995b) and Pfefferman (1993) have suggested the use of sample weights and its benefits. Pfefferman, Skinner, Holmes, Goldstein and Rasbash (1998) have all suggested the use of sampling weights with relevant reasons and substantiation. Hence we will use our sampling weights in our analysis leading us to a Weighted REML which we will now consider.

2.6.5 Weighted REML

We now apply the REML procedure to our data set. This derivation is an extension to what I have used from Searle et al. Consider our previously defined mixed model i.e.

$\mathbf{y= X\beta +Zu +e}$ from equation 2.2.2.1 ... 2.6.5.1

Then $\mathbf{V}$ (N x N) = Var ($\mathbf{y}$) = $\sigma_u^2 \mathbf{ZZ'} + \sigma^2 \mathbf{W^{-1}}$

where $\mathbf{W}$ = diag ($W_1,\ldots,W_N$) and $\mathbf{W^{-1}}$ = diag ($W_1^{-1},\ldots,W_N^{-1}$)

if $\mathbf{W^{-1} = K'K = KK}$, then $\mathbf{K}$ = diag ($(W_1)^{-1/2}$, $(W_2)^{-1/2}$, …, $(W_n)^{-1/2}$)

Now define

$\mathbf{z = K^{-1} y}$

$\mathbf{B = K^{-1} X}$

$\mathbf{A = K^{-1} Z}$

$\mathbf{f = K^{-1} e}$

Pre-multiply equation 2.6.5.1 (2.2.2.1) with $\mathbf{K^{-1}}$ so that our model becomes:

$$\mathbf{K^{-1} y = K^{-1} X\beta + K^{-1} Zu + K^{-1} e}$$

$$\mathbf{z = B\beta + Au + f} \qquad \ldots \qquad 2.6.5.2$$

Now since var $(\mathbf{f}) = \text{var}(\mathbf{K^{-1} e}) = \mathbf{K^{-1}}$ var $(\mathbf{e})$ $(\mathbf{K^{-1}})' = \mathbf{K^{-1}}\sigma^2 \mathbf{K' K K^{-1\prime}} = \sigma^2 I_N$

This is analogous to equation 2.4.1 on page 17 (where var $(\mathbf{y}) = \mathbf{D}$ but for my data set var$(\mathbf{u}) = \sigma^2_{cl} I_{40}$

$$\begin{bmatrix} \mathbf{X'R^{-1}X} & \mathbf{X'R^{-1}Z} \\ \mathbf{Z'R^{-1}X} & \mathbf{Z'R^{-1}Z+D^{-1}} \end{bmatrix} \equiv \begin{bmatrix} \mathbf{B'(\sigma^2 I_N)^{-1}B} & \mathbf{B'(\sigma^2 I_N)^{-1}A} \\ \mathbf{A'(\sigma^2 I_N)^{-1}B} & \mathbf{A'(\sigma^2 I_N)^{-1}A+(\sigma^2 I_{40})^{-1}} \end{bmatrix}$$

Theoretically var$(\mathbf{f}) = \sigma^2 I_N$ but as an estimation procedure var$(\mathbf{f}) = \mathbf{R}$(say) and var$(\mathbf{u}) = \mathbf{D}$(say)

So,

$$\begin{bmatrix} \mathbf{B'R^{-1}B} & \mathbf{B'R^{-1}A} \\ \mathbf{A'R^{-1}B} & \mathbf{A'R^{-1}A+D^{-1}} \end{bmatrix} \begin{bmatrix} \tilde{\beta} \\ \tilde{u} \end{bmatrix} = \begin{bmatrix} \mathbf{B'R^{-1}y} \\ \mathbf{A'R^{-1}y} \end{bmatrix}$$

are the mixed model equations for the weighted transformations that can be solved. The estimation procedure can now be applied to compute our REML estimates.

There are now two key points that will lead us on to the next section. Firstly ML and REML both require a normality assumption. Secondly the ML and REML estimators get their properties such as consistency, efficiency etc. through being the outcome of the maximum likelihood method and not by specifying those properties as at the outset and then developing estimators to satisfy those criteria. Comparatively the MINQUE procedure begins by not specifying the criteria of an underlying distribution of any of the factors (fixed or random) and properties of estimators at the outset but simply develops estimators of the variance components and then inference can be made about these estimators.

2.7 Minimum norm quadratic unbiased estimation(MINQUE)

As a result of the two points mentioned at the end of the previous section we now consider the procedure called MINQUE.

Searle et al (1992, pp. 41-42) states that the attempt of finding minimum variance quadratic unbiased estimators of variance components began with Townsend in 1968, Harville in 1969, Townsend and Searle in 1971 followed by La Motte in 1970, 1971, 1973a and 1976.

As mentioned before, at the end of section 2.6, MINQUE demands no distributional properties of the residuals or random effects. This method also does not involve any iterations, just solutions to linear equations obtained by using a *priori* values in place of the variance components in the estimation procedure. This is a serious deficiency in that the minimality property applies only at those *a priori values*.

The methodology that underlies the procedure of MINQUE is to specify criteria at the outset and then develop an estimator that satisfies these criteria. All three criteria hinges around $\mathbf{y'Ay}$ for which Searle et al (1992, p. 392) point out:

" Since a variance is a second moment, it seems natural to estimate it by a quadratic form of data. We therefore consider estimating $\mathbf{p'}\sigma^2$ by $\mathbf{y'Ay}$ for

imperative as an underlying requirement for MINQUE as compared to other variance component estimation techniques. MINQUE involves estimating $\mathbf{p'\sigma^2}$ by $\mathbf{y'Ay}$ subject to the following:

1. $\mathbf{A = A'}$ (symmetry)
2. $\mathbf{AX = 0}$
3. $p_i = \text{tr}\,(\mathbf{A\,Z_i\,Z_i'})$ and $\mathbf{X'AX = 0}$

By deriving **A** from minimizing a known Euclidean norm, Searle et al (1992, pp. 397-398) summarizes the MINQUE technique as follows :

If we consider the equation of the general mixed model (as previously defined) as

$$\mathbf{y = X\beta + \sum_{i=0}^{r} Z_i\,u_i}$$

the "natural" estimator (Rao's word) of σ_i^2 is $\tilde{\sigma}_i^{\,2} = \mathbf{u_i'u_i}/q_i$

and hence $\mathbf{p'}\tilde{\sigma}_i^{\,2} = \sum_i p_i \dfrac{\mathbf{u_i'u_i}}{q_i} = \mathbf{u'}\left\{ {}_d\dfrac{p_i}{q_i}\mathbf{I}_{q_i} \right\}\mathbf{u} = \mathbf{u'\Delta u}$

where $\Delta = \left\{ {}_d\dfrac{p_i}{q_i}\mathbf{I}_{q_i} \right\}$. If on the other hand we consider estimating $\mathbf{p'\sigma^2}$ by $\mathbf{y'Ay}$,

this will imply $\mathbf{p'}\hat{\sigma}^2 - \mathbf{p'}\tilde{\sigma}_i^{\,2} = \mathbf{u'[\,Z'AZ - \Delta\,]u}$.

MINQUE is based on minimizing a weighted Euclidean norm based on this difference. Using pre-assigned values $\sigma^2_{o,i}$ as values of σ_i^2 the matrix **D** is defined as $\mathbf{D} = \left\{ {}_d\sigma_{0,i}^{\,2}\mathbf{I}_{q_i} \right\}$.

The actual norm to be minimized is

$$\text{tr}\,(\mathbf{FF'}) = \text{tr}\,[(\mathbf{AV_o})^2] - \sum_i \frac{p_i^{\,2}\ \sigma^4_{o,i}}{q_i}$$

symmetric **A**, but with **A** to be determined by whatever criteria we wish to impose on $\mathbf{y'Ay}$ as an estimator of $\mathbf{p'\sigma^2}$ ".

These three criteria for our usual mixed model $\mathbf{y} = \mathbf{X\beta} + \sum_{i=0}^{r} \mathbf{Z_i u_i}$

with $\mathbf{V} = \text{Var}(\mathbf{y}) = \sum_{i=0}^{r} \mathbf{Z_i Z_i'}\sigma_i^2$ are given in Searle et al (1992, pp. 392-393):

A. Unbiasedness

Since $E(\mathbf{y'Ay}) = \text{tr}(\mathbf{AV}) + \beta'\mathbf{X'AX}\beta$ unbiasedness demands

$$\mathbf{p'\sigma^2} = \sum_{i} \text{tr}(\mathbf{A Z_i Z_i'})\sigma_i^2 + \beta'\mathbf{X'AX}\beta$$

so for this to be true we require $\forall \sigma_i^2, \beta$ leads us to

$p_i = \text{tr}(\mathbf{A Z_i Z_i'})$ and $\mathbf{X'AX = 0}$.

B. Translation Invariance

A quadratic form $\mathbf{y'Ay}$ is said to be translation–invariant for our above model if its value is unaltered by location changes in β i.e. if β becomes $\beta + \delta$ the $\mathbf{y'Ay}$ is translation invariant if $\mathbf{y'Ay} = (\mathbf{y} - \mathbf{X}\delta)'\mathbf{A}(\mathbf{y}-\mathbf{X}\delta)$.

C. Minimum Variance

The variance of $\mathbf{y'Ay}$, confining our attention to be on the fact that $\mathbf{y}$ is normally distributed (Using Searle et al (1992), Theorem S4 of Appendix S.5) is $\text{Var}(\mathbf{y'Ay}) = 2\,\text{tr}[(\mathbf{AV})^2] + 4\beta'\mathbf{X'AVAX}\beta$.

And the mean squared error (MSE) of $\mathbf{y'Ay}$ as an estimator of $\mathbf{p'\sigma^2}$ for given $\mathbf{p}$ is

$$\text{MSE}(\mathbf{y'Ay}) = \text{var}(\mathbf{y'Ay}) + [E(\mathbf{y'Ay}) - \mathbf{p'\sigma^2}]^2$$

$$\text{MSE}(\mathbf{y'Ay}) = 2\,\text{tr}[(\mathbf{AV})^2] + 4\beta'\mathbf{X'AVAX}\beta + [\text{tr}(\mathbf{AV}) + \beta'\mathbf{X'AX}\beta - \mathbf{p'\sigma^2}]^2$$

We will also now use the derivation of Minque with its estimators from Searle et al (1992; pp.397 – 399). The assumption of normality for our data is not

Where $\mathbf{F} = \mathbf{D}^{1/2}\ (\mathbf{Z'AZ} - \Delta)\ \mathbf{D}^{1/2}$, $p_i = \text{tr}\ (\mathbf{A}\ \mathbf{Z_i}\ \mathbf{Z_i'})$

We want to minimize the tr $[(\mathbf{AV_o})^2]$ subject to $\mathbf{A = A'}$, $\mathbf{AX=0}$ and $p_i = \mathbf{tr}\ (\mathbf{A}\ \mathbf{Z_i}\ \mathbf{Z_i'})$. This leads to the following MINQUE equation

$$\{_m \text{tr}\ (\mathbf{P_o}\ \mathbf{Z_i}\ \mathbf{Z_i'}\mathbf{P_o}\ \mathbf{Z_j}\ \mathbf{Z_j'})\}\ \sigma^2 = \{_c\ \mathbf{y'}\ \mathbf{P_o}\ \mathbf{Z_i}\ \mathbf{Z_i'}\ \mathbf{P_o}\ \mathbf{y}\} \qquad \ldots \qquad 2.7.1$$

This equation is similar to the REML equation in section 2.6 and the equation can be solved without iteration. Searle (1995, pp 228) gives the following comments about the MINQUE equation 2.7.1:

He states that $\mathbf{P_0}$ is $\mathbf{P}$ from $\mathbf{P = V^{-1} - V^{-1}X(X'V^{-1}X)\text{-}X'V^{-1}}$ which involves $\mathbf{V}$, but with a set of pre-assigned values $\sigma^2_{0,i}$ used in place of σ^2_i in $\mathbf{V}$. These values are denoted by σ^2_0 . Then replacing σ^2 by σ^2_0 in $\mathbf{V}$ gives $\mathbf{V_0}$ replacing $\mathbf{V}$ in $\mathbf{P}$ gives $\mathbf{P_0}$. Thus the solution to equation 2.7.1 depends upon whatever values are chosen as elements of σ^2_0. People with different choices will, from the same set of data get different estimates. Searle also states that the MINQUE equations only have to be solved as they stand and that is easy to work out.

2.7.1 Notes

- MINQUE estimators are biased. Searle et al (1992, p.42) suggest that the MINQUE procedure has a deficiency in that the minimality property applies only at those *a priori* values;
- In the presence of a priori values iteration results using successive solutions as a priori values leading to Iterative MINQUE i.e. IMINQUE which are the same as REML estimates;
- Searle et al (1992, p.398) suggests the impracticality of the MINQUE method, primarily because a MINQUE = 1st iteration of REML;
- MINQUE can produce negative estimates;
- Seifert (1985) shows how MINQUE can be efficiently used to handle various topics such as negative estimates, estimation of fixed effects;

- Prediction of random effects as well as the estimation of variance and co-variances;
- The topic of MINQUE is dealt in a very indepth and exclusive way (mathematically) by Rao and Kleffe (1988) who give complements and problems of Minque;
- Kleffe and Seifert (1986) give a full computation of the variance components by using the MINQUE method.

Applying MINQUE to our data set we would simply have the weighted REML equations as discussed in previous chapters to be solved linearly.

Up to this point we have considered the following techniques:
Henderson's methods II and III, ML, REML and MINQUE. We will move now to looking at the (GLMM) Generalized Linear Mixed Models because this can also be used to estimate variance components.

2.8 The generalized linear mixed model (GLMM)

The GLMM is an extension to the general linear model ($\mathbf{y} = \mathbf{X}\beta + \mathbf{e}$) since we can allow for a non-normal underlying distribution and a non-identity link.

A good exposition of the general linear model is given by Firth (1991) and McCullagh and Nelder (1989). The following description is from Littell et al (1996; pp 434-435).

In our "normal errors" mixed model (from section 2.2.2)

$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\mathbf{u} + \mathbf{e}$ where the conditional mean of the observations, given the random model effects is

$E(\mathbf{y}|\mathbf{u}) = \mathbf{X}\beta + \mathbf{Z}\mathbf{u}$ and the conditional variance is var $[\mathbf{y}|\mathbf{u}] = \mathbf{R} = \mathrm{Var}(\mathbf{e})$. The observations can then be described as $\mathbf{y} = \mu + \mathbf{e}$ where μ denotes the conditional mean $E[\mathbf{y}|\mathbf{u}]$.

The Generalized Linear Mixed Model (GLMM) can also be described in terms of conditional means. It takes the form $\eta = \mathbf{X}\beta + \mathbf{Z}\mathbf{u}$ where η is the link function $g(\mathbf{u})$ and $\mathbf{u}$ is the conditional mean $E(\mathbf{y}|\mathbf{u})$. Then we can write the GLMM as

$g\{E[\mathbf{y}|\mathbf{u}]\} = \mathbf{X}\beta + \mathbf{Z}\mathbf{u}$.

As in the conventional mixed model, the random model effects $\mathbf{u}$ are assumed to have a multivariate normal distribution with mean zero and covariance matrix $\mathbf{G}$, while as in the conventional generalized linear model, the underlying distribution of y is assumed to be a member of the exponential family.

Seeing that we will be using Genstat 5 extensively to analyse our data, we will now consider two methods for obtaining parameter estimates for

GLMM that have been implemented by Genstat 5 procedure library (Welham, 1993).

2.8.1 GLMMs – The methods of Breslow and Clayton and of Schall

Genstat uses an algorithm that is (for the GLMM) analogous to generalized linear models. We consider the following derivation where the description is from Waddington et al [(1994); pp.13-24].

A. Derivation

For data **y** with mean μ, the mean is related to the linear predictor η by the link function $g(\mu) = \eta = \mathbf{X}\beta + \mathbf{Zu}$ where β represents the fixed effects with var $(\mathbf{u}) = \mathbf{G}$, a function of the unknown variance components σ_1^2 , σ_2^2
A "working dependent variate" **y*** is created by linearizing the link function applied to the data about their mean values as follows:

$$\mathbf{y}^* = \mathbf{X}\beta + \mathbf{Z}u + \mathbf{D}(\mu)(\mathbf{y}-\mu) \quad \ldots \quad 2.8.1.1$$

Where $\mathbf{D}(\mu) = \frac{dg}{d\mu} = \text{diag}\{g'(u_1), g'(u_2),.....,\}.$

Thus the working variate has 3 components: fixed effects represented by $\mathbf{X}\beta$, random effects represented by **Zu** and an error term represented by $\mathbf{D}(\mu)(\mathbf{y}-\mu)$ which depends on the distribution of **y**, and on the link function g through $\mathbf{D}(\mu)$. The second and third terms of equation 2.8.1.1 gives us:

$$\text{var}(\mathbf{y}^*) = \mathbf{ZGZ}' + \mathbf{DVD} \quad \ldots \quad 2.8.1.2$$

where **V** is the variance of **y** conditional on the random effects **u** and **DVD** is a diagonal matrix. Thus the working variate **y*** is described by a linear mixed model with fixed effects β, random effects **u** and weights

$\mathbf{W} = (\mathbf{DVD})^{-1}$. Given an estimate of μ, the standard mixed model equation can be used to estimate β and **u.**

$$\begin{bmatrix} \mathbf{X'WX} & \mathbf{X'WZ} \\ \mathbf{Z'WX} & \mathbf{Z'WZ+G^{-1}} \end{bmatrix} \begin{bmatrix} \tilde{\beta} \\ \tilde{u} \end{bmatrix} = \begin{bmatrix} \mathbf{X'Wy^*} \\ \mathbf{Z'Wy^*} \end{bmatrix} \quad \ldots \quad 2.8.1.3$$

These equations can be solved using REML which also estimates the variance components in G. Welham (1993) has showed that REML should be restricted to two iterations to provide a solution rather than allowing it to converge also thereby saving computing time. Given new estimates of β and **u**, further estimates of μ are formed and used to update the working variate **y*** and the weights **W**, and the process is repeated until convergence.

2.8.2 The method of Schall

Schall's method (1991) uses an estimate of the conditional mean of **y** given **u** viz.

$\mu = g^{-1} (\mathbf{X}\beta + \mathbf{Zu})$ ... 2.8.2.1

Breslow and Clayton (1993) refer to this method as penalized quasi-likelihood or PQL. The following method is now derived using a quasi-likelihood argument by following Littell et al (1996, pp. 443-435).

For generalized linear models the estimating equations are determined by

$$Q(u_i, y_i) = \frac{y_i \theta_i - b(\theta_i)}{a(\phi_i)}$$

where μ_i is involved in the right hand side of the equation through the mean function θ_i.

In the generalized linear mixed model, the conditional distribution of **y** given **u** plays the same role as the distribution of **y** in the fixed effects generalized linear model i.e. the conditional quasi-likelihood of an observation y_i given u_i is

$$Q(u_i, y_i \mid u_i) = \frac{y_i \theta_i - b(\theta_i)}{a(\phi_i)}$$

The joint quasi-likelihood of the observations is the sum of the quasi-likelihood of y given u and the quasi-likelihood of u. In matrix terms the joint quasi-likelihood is

$$Q(\mu, u; y) = y' A^{-1}\theta - (b_\theta^{0.5} A^{-1} b_\theta^{0.5}) + 0.5\, u'G^{-1}u \quad \ldots \quad 2.8.2.2$$

where **A** is the matrix of $a(\phi_i)$'s

θ is the vector of θ (μ_i)'s

$\mathbf{b_\theta}$ is the vector of $b(\theta_i)$'s and **G**

G is the covariance matrix

Schall's method assumes that the scale parameter $a(\phi) \equiv 1$. Breslow and Clayton (1993) have shown that solutions for β can be obtained from $Q(\mu,u,y)$ by iteratively solving the mixed model in equation 2.8.1.3.

An alternative to Schall's model is what Genstat refers to as the marginal model of Breslow and Clayton. The algorithm is identical to Schall's method/model except for the estimation of the mean as $\mu=g^{-1}(\mathbf{X}\beta)$.
Breslow and Clayton (1993) derive the model leading to the marginal quasi-likelihood estimation by specifying the generalized linear model in terms of the marginal mean as

$$E(y_i) = \mu_i = f(\mathbf{x}_i'\beta) \quad \ldots \quad 2.8.2.3$$

Where f is the inverse link function. If the link function is not the identity, then the marginal mean so defined does not confirm with the marginal mean calculated from the conditional formulation

$$E[\mathbf{y}|\mathbf{u}] = f(\mathbf{X}\beta + \mathbf{Zu}) \quad \ldots \quad 2.8.2.4$$

2.8.3 Notes

- Genstat allows the GLMM procedure to be implemented as well as specifying which model (Schall or Breslow & Clayton) is to be used. However there are a number of improvements to algorithms for penalized (PQL) quasi-likelihood and marginal quasi-likelihood (MQL);
- Breslow and Clayton (1993) compare PQL and MQL showing that regression coefficients of PQL depend strongly on the estimated variance components when we do not have an identity link whereas this is NOT true for MQL, they also suggested that there is a clear downward bias which is exhibited by the fixed effects;
- Goldstein and Rasbash (1996) suggested an improved approximation which eliminates downward biases from GLMM's i.e. an iterated form of bootstrapping by Kuk (1995);
- Goldstein (1995, Chapter 5, p. 78) provides an alternative approach to the GLMM's by rewriting the model and fitting it as an approximate linear model using IGLS;
- Engel and his colleagues have worked extensively in researching the problem of bias in estimates from GLMM particularly in the estimates of the variance components. Engel and Keen (1994) point out that the estimates of the iterative weights used in GLMM are

$$\mathbf{W}^{-1} = \mathbf{V}(\mu)\,[g'(\mathbf{u})]^2$$

where the estimate μ is obtained from the first step of the estimation procedure.

They note that while $\mathbf{V(\mu)\ [g'(u)]^2}$ may often be an accurate prediction of $\mathbf{V(\mu)\ [g'(u)]^2}$, particularly if (in the case of a single random effect as in our data set) the variance component σ^2_u is small, it is not necessarly a consistent estimator. They recommend using alternative weights which depend on both the link and the assumed variance function;

- Lin and Breslow (1996) derived a four step algorithm to correct the bias of PQL estimates of regression coefficients and variance components. We will not consider the algorithm for this thesis. They also considered approximations based on Laplace approximations of the integrated likelihood used by Liu and Pierce (1993);
- Engel and Buist (1998) looked at bias in GLMM's in animal breeding and concluded that using the alternative weights of Engel and Keen (1994) could remove the bias and decrease the Mean Square Error but not for a large number of clusters with many observations (say more than 40) per cluster. In my data set we have about 40 clusters so perhaps the methods and Lin and Breslow and Engel and Keen are worth considering;
- Engel and Keen (1994) suggested the use of alternative weights depending on the link and the assumed variance function of our data to eliminate bias;
- Engel, Buist and Visscher (1995) carried out simulation studies by fitting models in animal breeding and found that both the magnitude and direction of the bias in the estimate of the variance component depend on the number of fixed effects and also on the overall underlying response probality, with over estimation of the variance component σ_u^2 when there are a large number of fixed effects and the overall incidence is above 0.9. This has proved to be the case with the data set considered in this thesis;

- Engel (1998) uses a simple example to illustrate the asymptotic bias in the GLMM estimate of the variance component σ_u^2 and deduces that this can be underestimated by almost half. He then says that this could be overcome by use of a Markov Chain Monte Carlo Method such as the Gibbs Sampler.

Applying the GLMM to our data set the link function η would be the identity whilst our distribution would be normal. The rest of the parameters will remain the same i.e the fixed and random effects.

This brings me to a conclusion of my literature survey of the overview of variance components (models). In the last section I will consider possible future developments in the area of variance components.

2.9 The future

The future of variance components is clearly very bright and hopeful with statisticians/scientists making all kinds of wonderful breakthroughs and contributions in this area. There has been alot of work done where the design of an experiment or survey is taken into consideration to improve the estimation of variance components, especially with certain colleagues of mine here at the University of Natal.

Aitkin M (1999) has done work on describing an EM algorithm for nonparametric ML estimation in generalized linear models with variance component structure. The algorithm can be applied to mixed models as well as providing an alternative analysis to approximate MQL and PQL analyses. (Goldstein 1995, Breslow and Clayton 1993). The algorithm also saves computational time. (He used GLIM 4).

R.R. Hocking (1990) suggested concern over estimating variance components to derive a diagnostic analysis in balanced linear mixed models. He derives two sets of functions, one is an interpretable set by sums of the original components and the other set is for simple expressions of estimators.

Chow and Shao (1988) have given time to research new estimators for variance components which avoid non-positive estimates in random effects models. M. Ghosh (1995) uses a decision theoretic and Bayesian method for estimating variance components. The estimators that are constructed are non-negative and some of them have smaller mean squared errors than the classical estimators of ANOVA.

C. Lavergne (1995) has proposed a Jacknife method for estimation of variance components in an observed regression model. He shows that by using the least squares approach for his estimators that MINQUE and Jacknife are very close. Perhaps the Jacknife method to estimate variance components could be extended to other models (mixed models perhaps).

J. Ofversten (1993) has given a method of deriving exact tests for variance components in unbalanced mixed models. In particular he looks at a

hypothesis of variance components of a model with one random factor (P46 – 47) which is a representation of our model for our data. He goes on to state that his proposed tests are unbiased and under reasonable conditions consistent. The test for our model (i.e. a mixed model with one random component) is identical to the traditional exact F–statistics for fixed effects analysis. Perhaps traditional analyses with our approximate F-tests could be replaced by more exact F-tests for varaince components.

R. Christensen (1996) discusses generalizations of Ofverstens tests extending them generally.

Callanan and Harville (1991) proposed four new algorithms viz. Linearized Newton Algorithm (LN1), Linearized Newton Algorithm (LN2), Linearized Method of of Successive Approximations (LMSA1) and Linearized Method OF Successive Approximations (LMSA2) which are more efficient and less sensitive to their choice of starting points. They compare these 4 algorithms to the EM and method of scoring algorithms.

Finally Sun et al (1996) has explored the area of Bayesian methods for estimating variance components. This area is by far the "hot" area of research i.e. MCMC methods and is turning out to be a promising area as well as being fruitful in results associated with variance components. Sun et al(1996) introduce the "collapsing phenomenon" which I will not elaborate on further seeing that it is beyond the scope of this thesis.

In conclusion it is appropriate for me to say that in variance component estimation we always keep the essential estimation procedures! MCMC methods are thriving in the statistical community but more work has still to be done in areas of MCMC model checking and validity. The comparison of the different approaches will be discussed in my conclusion.

CHAPTER 3: Results

3.1 Background and description of the data

South African adults' smoking status, perceptions of the health effects of nicotine and cigarettes, attitudes toward smoking control and awareness of Government health warnings on the harmful effects of smoking were investigated by means of a series of interviewer- administered questionnaires conducted by fieldworkers of the Human Sciences Research Council through the Omnibus surveys in February 1995, February 1996 and October 1996. The first survey preceded the implementation of the Tobacco Control Act whilst the second and third surveys followed the act. The data which were part of the October 1996 omnibus smoking survey in South Africa were graciously given to me by Dr Jonathan Levin of the Medical Research Council (MRC). We have the data of the survey that was carried out in Gauteng. There are 343 values of the dependent variable y (which are the "logtots" in our case - this response variable will be described later on). The surveyed variables are race, sex, marital status, socio-economic status, smoking status, age, awareness of the link of smoking (to cancer, heart disease and lung disease) and education level. The sampling methodology for the data is as follows:

The study population consisted of South African residents of 18 years and older, in all 9 provinces. The population was stratified by province and type of area. The sample allocation to the resulting strata was done proportionally to the 1991 census figures except that disproportion was introduced to give 120 respondents per province and the minimum number of Indians in the sample was fixed at 120. Multistage cluster sampling with probability proportional to size was used to draw respondents, with the adjusted 1991 population census figures as a measure of size. Census enumeration areas and similar areas were used as the clusters. There were 40 clusters in total in Gauteng. The clusters for which I have data are considered to be a random selection of clusters from all the clusters in South Africa, even though all the clusters are from Gauteng. A random selection of respondents was then drawn from the clusters, for example

we have 4 respondents from cluster 1. All clusters were drawn with probability proportional to size, whilst households were drawn from the final clusters with equal probability. One respondent aged 18 years or older was selected from each household by applying a grid. For each selected respondent a sampling weight was calculated, using the stratification variables of province and type of area and by post-stratification for age, gender, education and race. Respondents in the survey were asked to recall a number of anti-smoking messages which appeared as warning messages on cigarette advertisements. There were 9 different warning messages viz.:

1. Danger: Smoking can kill you
2. Danger: Smoking causes Cancer
3. Smoking Damages Your Lungs
4. Warning: Don't Smoke near Children
5. Pregnant? Breast Feeding? Your Smoking Can Harm Your Baby
6. Danger: Smoking Causes Heart Disease
7. Tobacco is addictive
8. Your smoking can harm those around you
9. Tobacco causes cancer

For each respondent the total number of messages spontaneously recalled was noted. The response variable used was log tot = $\log_e$ (total number of messages recalled + 0.5). An amount of 0.5 has to be added in because the log of zero is not defined and some respondents recalled zero messages, hence the need to add on a half when calculating the logtot which is our dependent variable. The main aim of the survey was to investigate the prevalence of smoking and the factors associated with it. Also with respect to the media, the aims were to investigate firstly the distribution of the number of messages recalled and how this differed with respect to the explanatory variables and secondly, how the proportions receiving the message from each of the following media (cigarette packets, cinema, radios, television, newspaper, magazines, billboards, posters in shops, pamphlets, family/friends) depended on smoking status and other explanatory

variables specified above were part of the objectives and finally how the attitude of the respondents towards media messages differed between smokers and non smokers. We are interested in applying the various techniques to estimating the variance components by fitting variance component models to the data and comparing them to assess their advantages and disadvantages. As a result to determine the effects of various explanatory variables on the number of messages recalled, this will lead us to ANOVA and the mixed model applications.

3.2 Box-and-whisker-plots

We now consider the boxplots as a graphical representation of the data to check for outliers, trends and differences between groups of the same variable, for example differences in the education levels with respect to the logtot variable can give us insight and summary between the groups of individuals with different educational levels. The dependent variable is plotted on the y-axis whilst the other variables like race, sex, age, marital status, socio-economic status, smoking status and educational status will be on the x-axis. Let us now look at these plots done by SPSS:

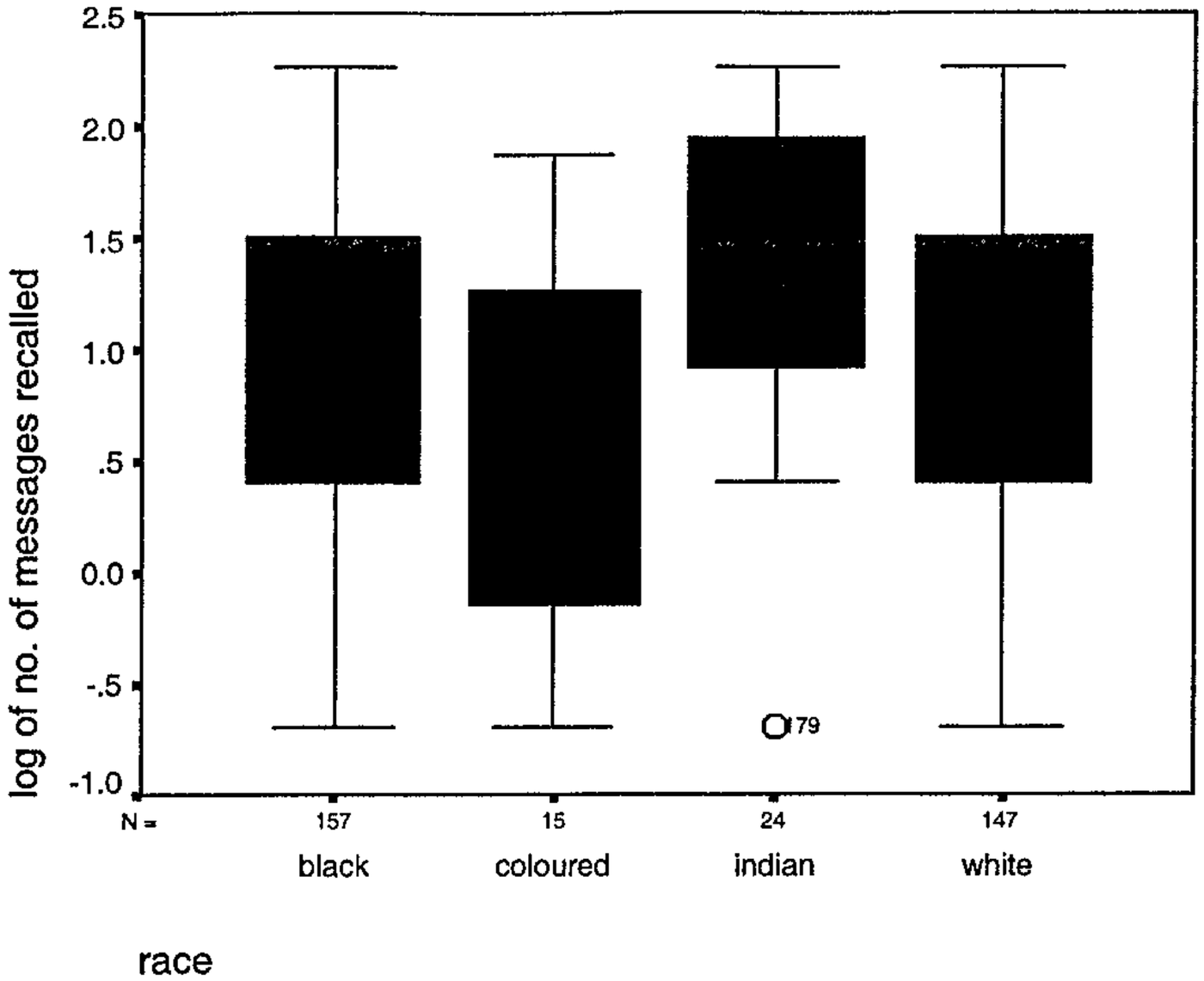


Figure 3.2.1

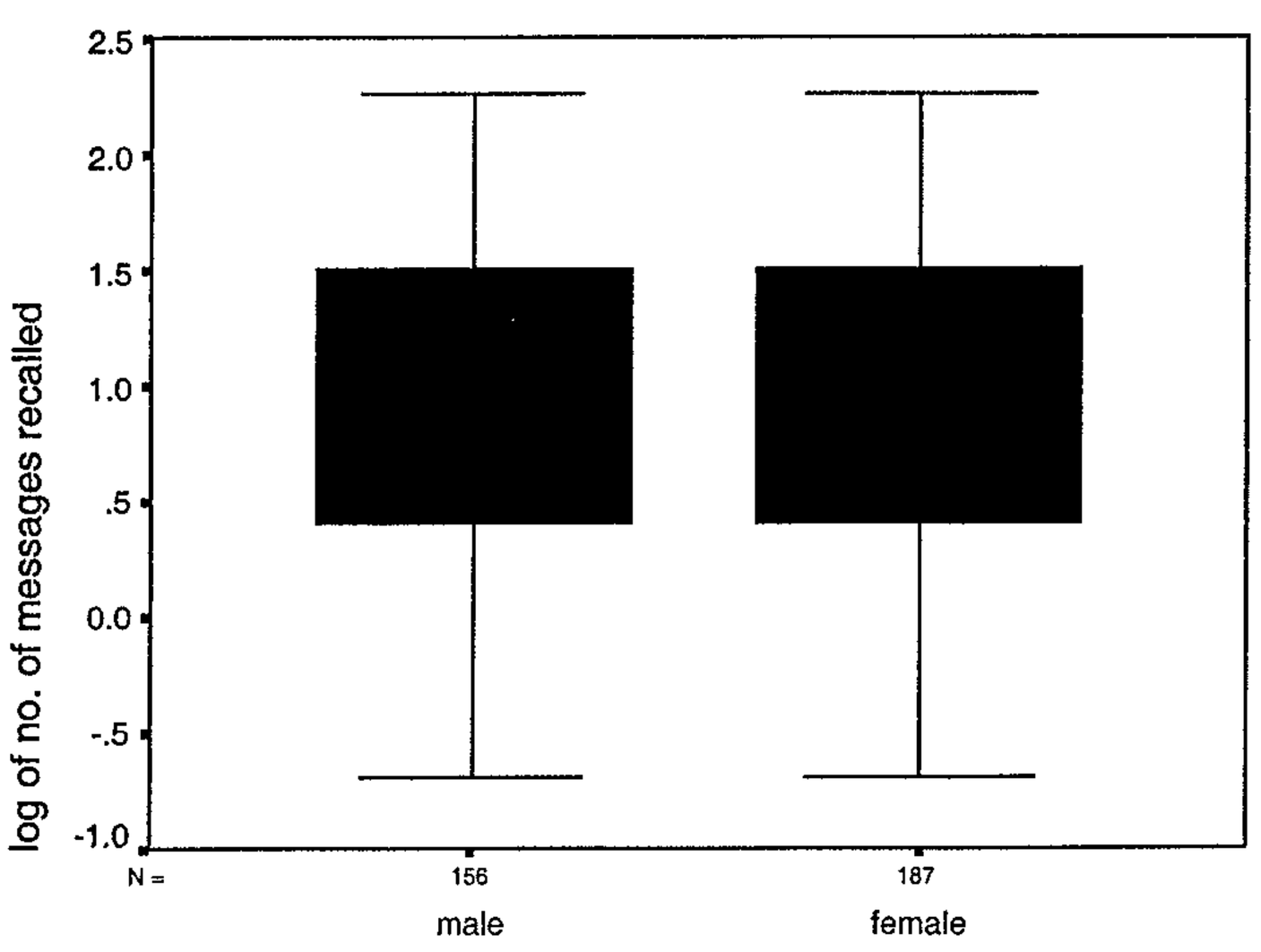


Figure 3.2.2

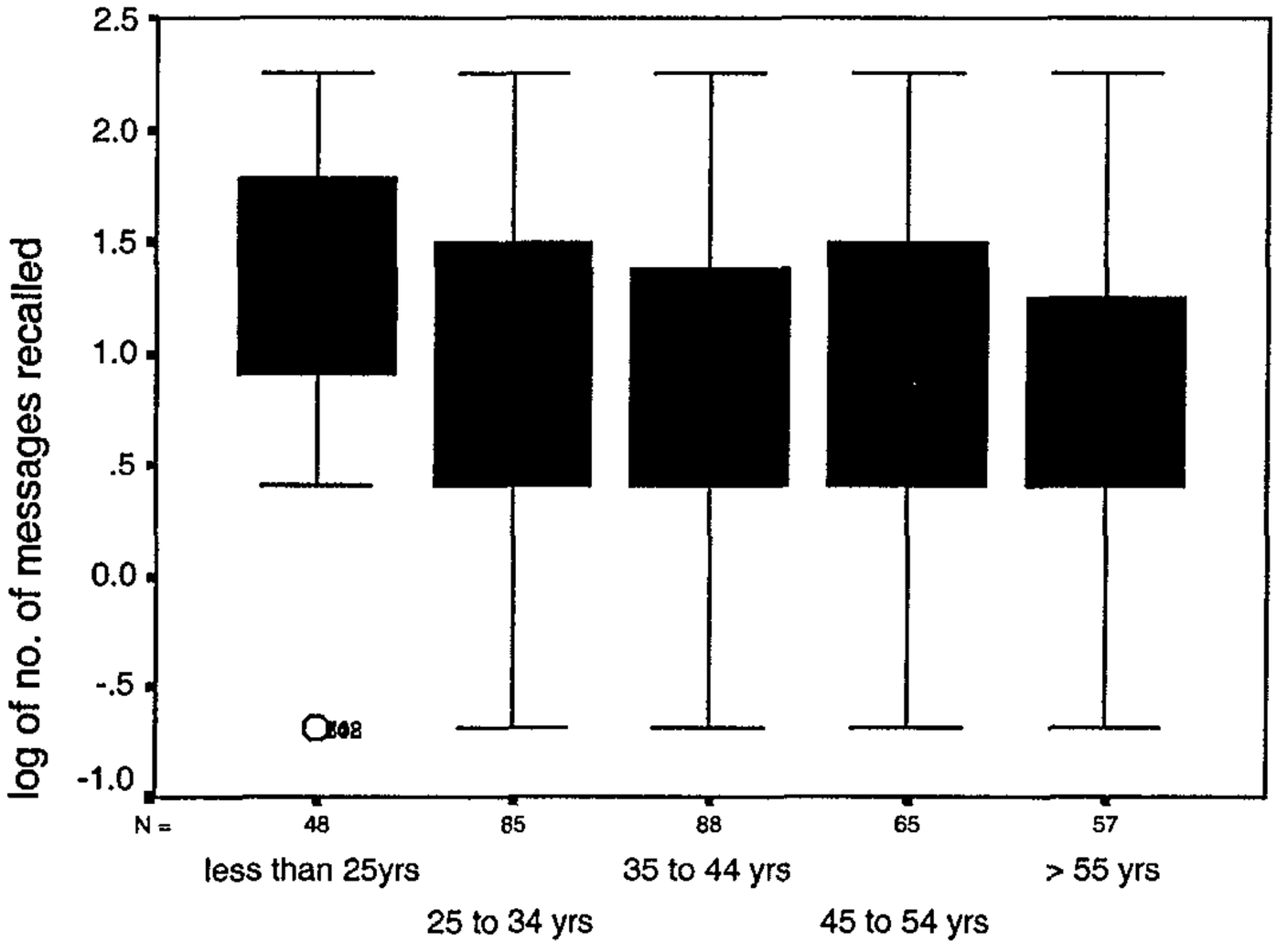


Figure 3.2.3

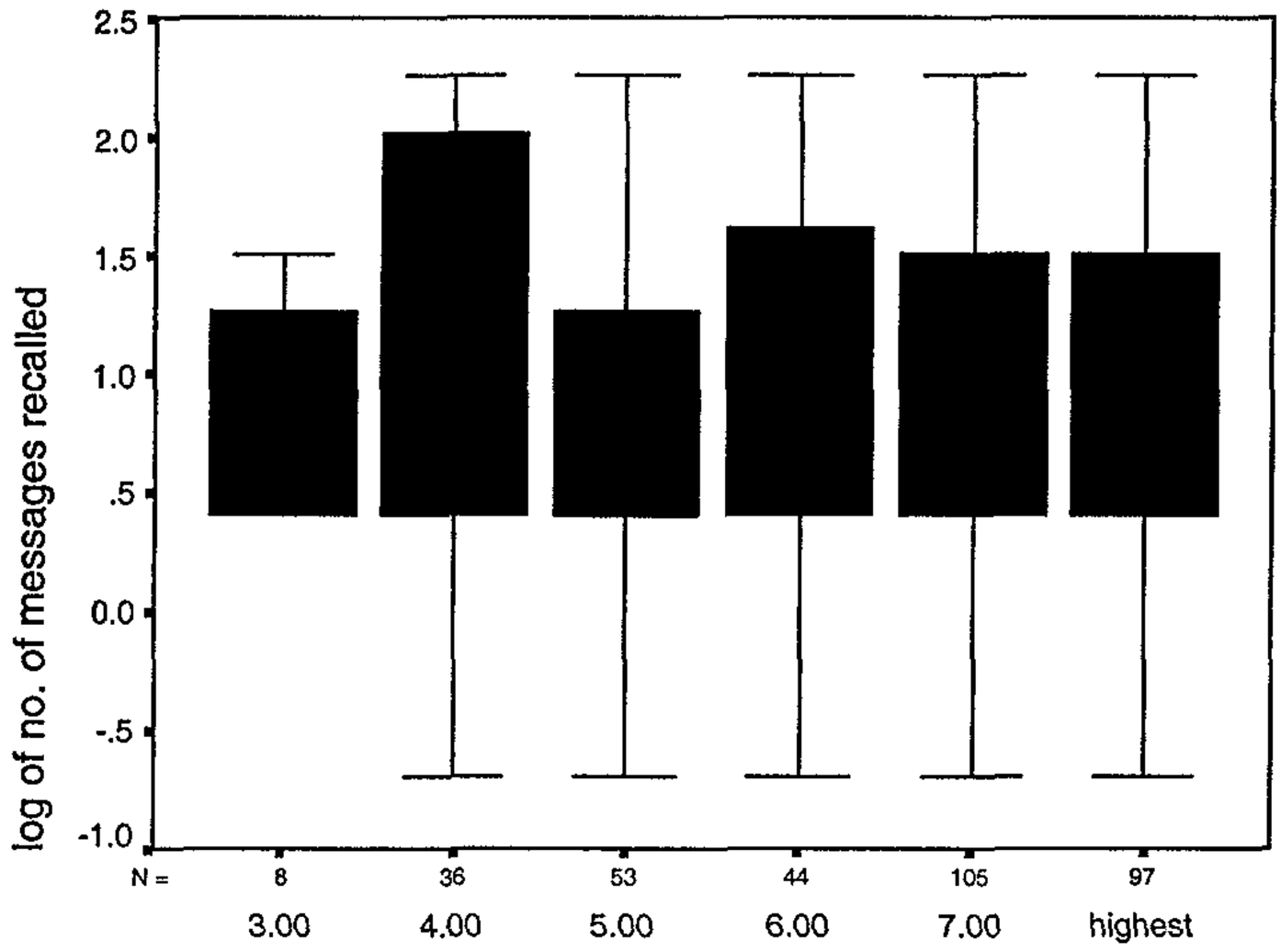


Figure 3.2.4

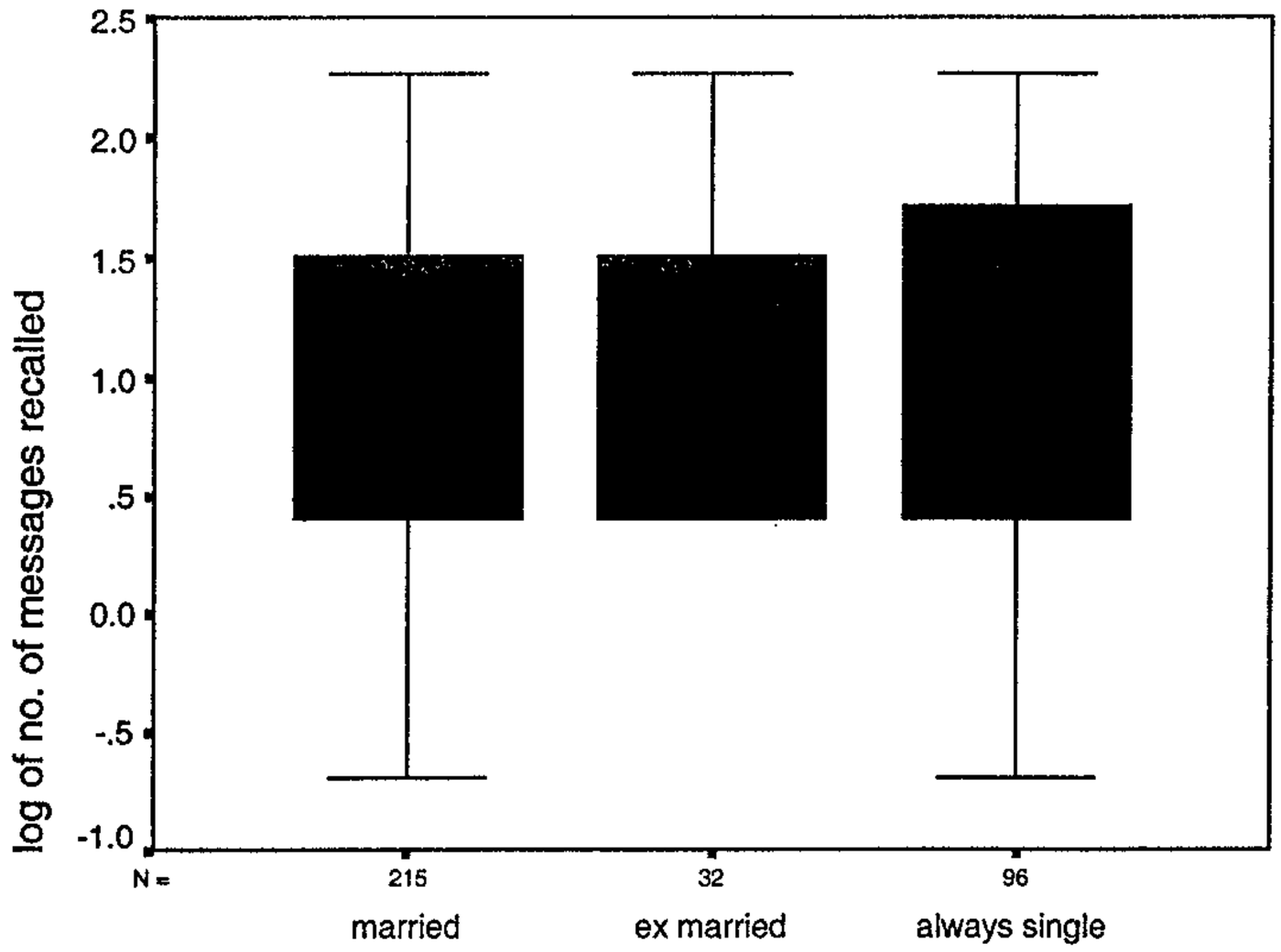


Figure 3.2.5

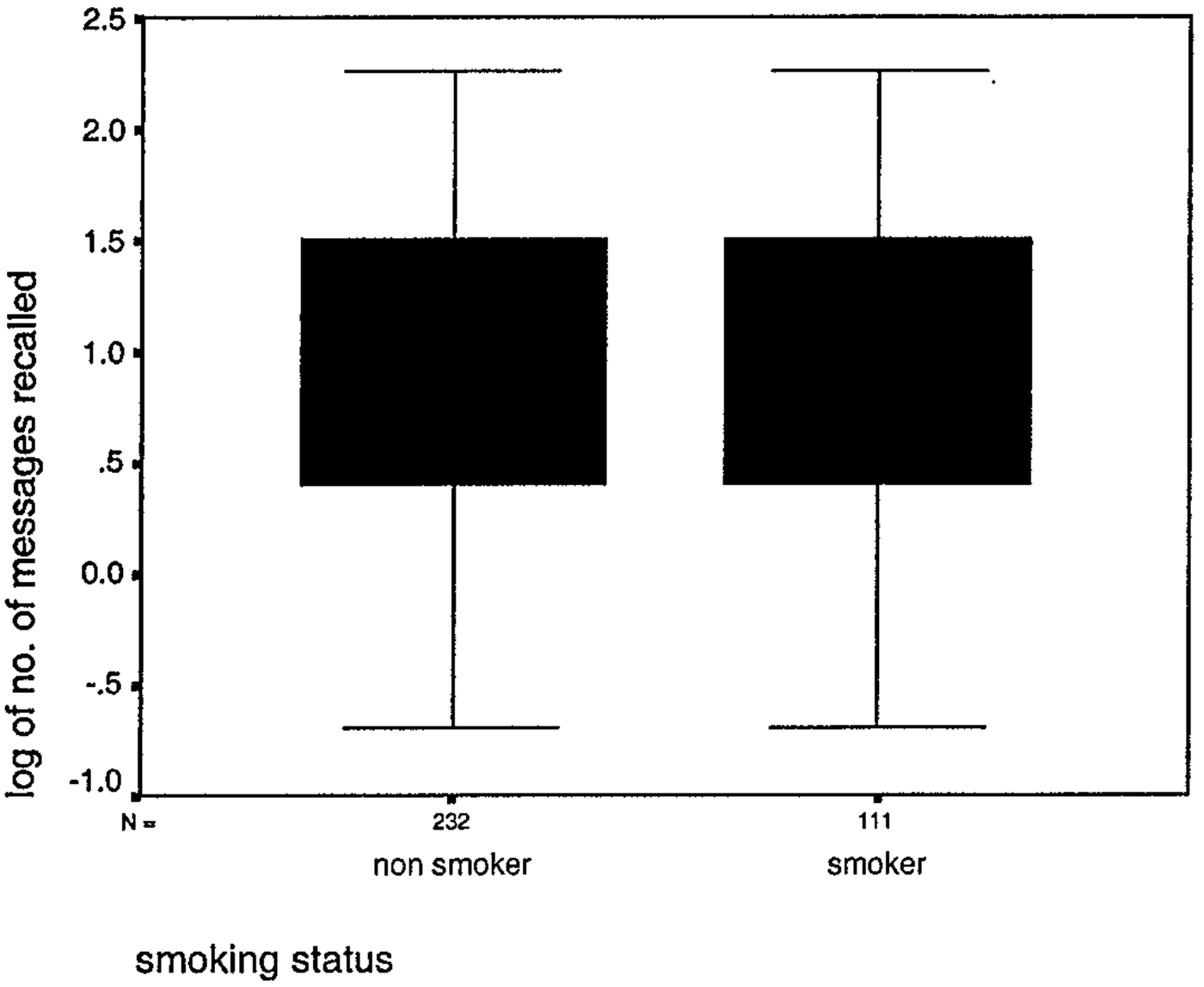


Figure 3.2.6

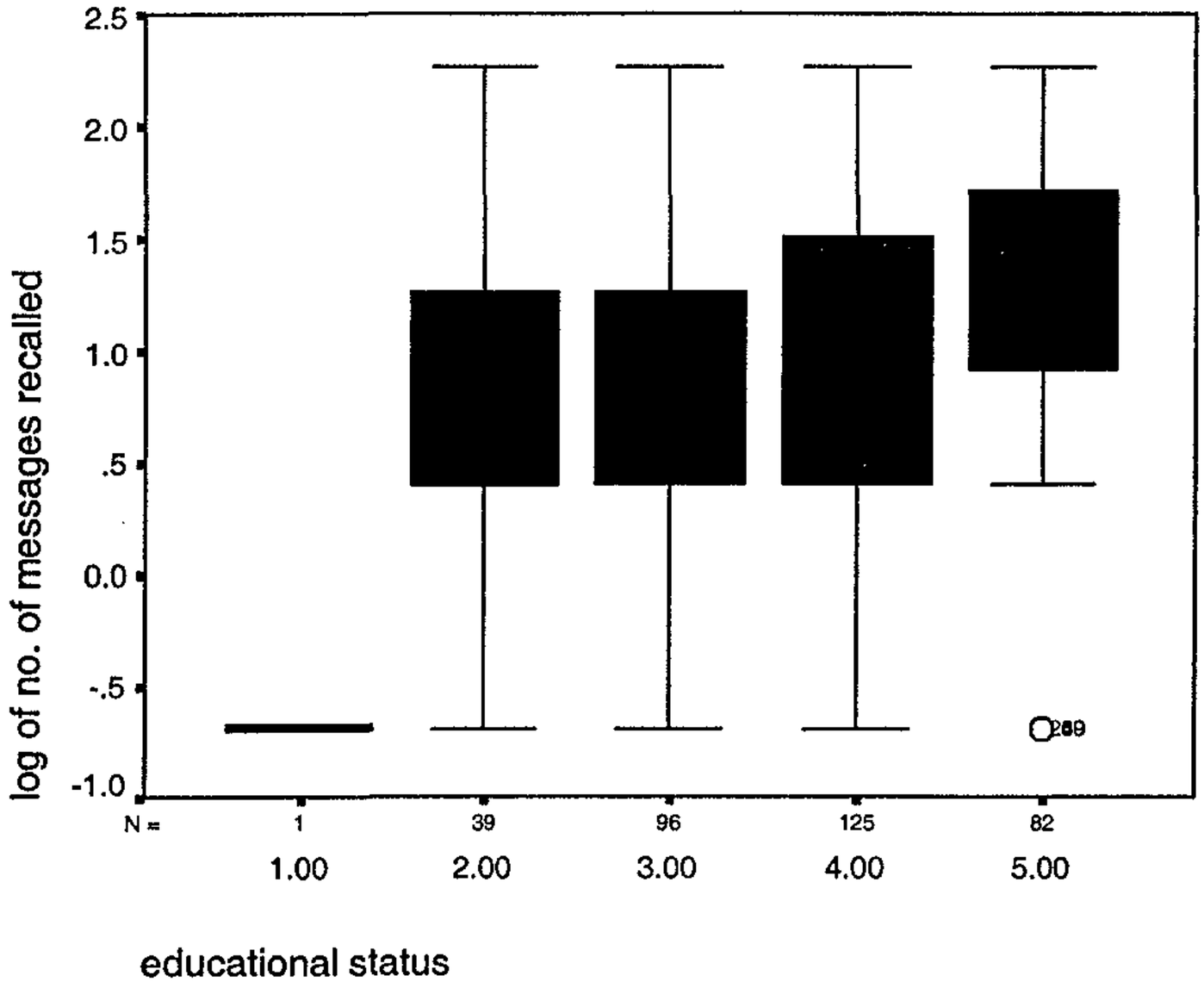


Figure 3.2.7

Figure 3.2.1 shows us that the responses of the black race (level 1) and white race (level 4) are negatively skewed i.e. skewed toward lower values whilst the response of the coloured race (level 2) and the Indian race (level 3) are skewed toward higher values in their responses. The median logtot value is the lowest for the coloured race group meaning that they recalled the lowest number of messages of the four race groups. In general the distribution of the logtot is lowest for coloureds and highest for Indians.

Figure 3.2.2 clearly shows us that the distribution of the males and females are very similar whilst only the medians differ slightly. The median number of messages recalled was higher for males than for females. The distribution of the females seem to be symmetrical in the number of messages that they recalled.

Figure 3.2.3 clearly shows us that the two age groups from 25-34(level 2) years and 45-54(level 4) years are skewed to lower number of messages recalled

whilst the < 25 (level 1) years group are skewed toward a higher number of messages recalled. The 35-44 (level 3) years group are about symmetrical in their distribution. The 55+ years group (level 5) are skewed slightly to lower number of messages recalled. The 45-54 (level 4) age group recalled the highest number of messages while the 35-44 (level 3) age group recalled the lowest number of messages.

Figure 3.2.4 shows that the number of messages recalled clearly increased after the 5^{th} group with the 6^{th}, 7^{th} and 8^{th} group having distributions that are skewed toward a lower number of messages recalled. Group 3 and 4 have distributions that are skewed to a higher number of messages recalled while group 5 have the lowest number of messages recalled. Group 6 recalled the highest number of messages. These groups are defined by socio-economic status.

Figure 3.2.5 showed the three marital status groups viz the married, ex-married and single to have the number of messages recalled to be skewed to lower number of messages recalled and their medians are almost equal.

Figure 3.2.6 showed smokers and non-smokers recalled a low number of anti-smoking messages having their distributions skewed to a low logtot value.

Figure 3.2.7 showed that the highly educated respondents in Gauteng recalled more anti-smoking messages than the lesser educated people. The 3^{rd} and 4^{th} groups have distributions that are skewed toward a lower number of messages recalled whilst for level 1 SPSS did not construct a plot due to there being only one respondent in group 1. Also the median number of messages recalled increased as the level of education increased.

3.3 Cross tabulation of the data

As part of the exploratory data analysis, I needed to look at the cross tabulations of the data in the two way interactions. The reason for doing this was to look for possible empty cells in that may exist between certain interactions. The presence of empty cells would have an effect when it comes to estimating the degrees of freedom for the interaction terms and furthermore, the software packages such as Genstat take into account any degrees of freedom lost due to empty cells in it's overall estimation of the degrees of freedom for the interaction terms. Fienburg (1980, p. 140) state that empty cells or "zero entries" as he terms it, are of two types viz. fixed and sampling zeros. Fixed zeros occur when it is impossible to observe values for certain combinations of the variables, for example, in my data set there may not be Whites with education in the rc x edu interaction and this could be regarded as an empty cell. Sampling zeros are due to sampling variation and the relatively small size of the sample when compared with a large number of cells; they disappear when we increase the sample size sufficiently.

The easiest way to think about the empty cells or fixed zeros and it's effect on estimating the degrees of freedom for the interaction term is to consider the following example in terms of the number of observations and the number of parameters estimated: If I am modeling an A x B table of means with 5 x 2 levels, then there are 10 cells with 9 degrees of freedom (9 df). The main effects A and B will have 4 and 1 degrees of freedom respectively i.e. (5-1) and (2-1) df. There are now 4 df left for the interaction term A.B i.e. (5-1) x (2-1) = 4 df. Let us say that there are 2 empty cells, then the degrees of freedom for the interaction term A.B will be (5-1) x (2-1) – 2 = 2 degrees of freedom. The output from SPSS using the cross tabulations function for the interaction terms yielded the following tables of cross tabulations with missing cells :

Table 3.3.1: Interaction of socio-economic status(lsm)* race(rc) cross tabulation

Count

		lsm						Total
		3.00	4.00	5.00	6.00	7.00	8.00	
rc	black	7	34	45	32	29	10	157
	coloured	1	2	7	2	2	1	15
	indian				1	12	11	24
	white			1	9	62	75	147
Total		8	36	53	44	105	97	343

Table 3.3.2: Interaction of race(rc) * marital status(mar) cross tabulation

Count

		mar			Total
		married	ex married	always single	
rc	black	79	18	60	157
	coloured	11		4	15
	indian	17	2	5	24
	white	108	12	27	147
Total		215	32	96	343

Table 3.3.3: Interaction of race(rc) * educational status(edu) cross tabulation

Count

		edu					Total
		1.00	2.00	3.00	4.00	5.00	
rc	black	1	35	63	44	14	157
	coloured		4	6	4	1	15
	indian			8	9	7	24
	white			19	68	60	147
Total		1	39	96	125	82	343

Table 3.3.4 Interaction of sex(sx) * educational status(edu) cross tabulation

Count

		edu					Total
		1.00	2.00	3.00	4.00	5.00	
sx	male	1	20	39	52	44	156
	female		19	57	73	38	187
Total		1	39	96	125	82	343

Table 3.3.5 Interaction of age * marital status(mar) cross tabulation

Count

		mar			Total
		married	ex married	always single	
age	less than 25yrs	5	1	42	48
	25 to 34 yrs	45	2	38	85
	35 to 44 yrs	66	8	14	88
	45 to 54 yrs	54	9	2	65
	> 55 yrs	45	12		57
Total		215	32	96	343

Table 3.3.6: Interaction of age * educational status(edu) cross tabulation

Count

		edu					Total
		1.00	2.00	3.00	4.00	5.00	
age	less than 25yrs			12	30	6	48
	25 to 34 yrs		7	15	36	27	85
	35 to 44 yrs	1	9	31	25	22	88
	45 to 54 yrs		10	24	15	16	65
	> 55 yrs		13	14	19	11	57
Total		1	39	96	125	82	343

Table 3.3.7: Interaction of socio-economic status(lsm) * educational status(edu) cross tabulation

Count

		edu					Total
		1.00	2.00	3.00	4.00	5.00	
lsm	3.00		1	6	1		8
	4.00		14	13	7	2	36
	5.00		16	24	11	2	53
	6.00		6	18	18	2	44
	7.00	1	2	19	48	35	105
	highest			16	40	41	97
Total		1	39	96	125	82	343

Table 3.3.8: Interaction of marital status(mar) * educational status(edu) cross tabulation

Count

		edu					Total
		1.00	2.00	3.00	4.00	5.00	
mar	married	1	26	53	74	61	215
	ex married		8	14	6	4	32
	always single		5	29	45	17	96
Total		1	39	96	125	82	343

Table 3.3.9: Interaction of smoking status(ys) * educational status(edu) cross tabulation

Count

		edu					Total
		1.00	2.00	3.00	4.00	5.00	
ys	non smoker	1	26	65	81	59	232
	smoker		13	31	44	23	111
Total		1	39	96	125	82	343

Table 3.3.10: Interaction of sex(sx) * socio-economic status(lsm) cross tabulation

Count

	lsm									Total
		1.00	2.00	3.00	4.00	5.00	6.00	7.00	8.00	
sx	male			4	17	26	17	50	42	156
	female			4	19	27	27	55	55	187
Total				8	36	53	44	105	97	343

Table 3.3.1 reveals that there are 2 levels missing from the lsm (socio economic status) leaving lsm with 6 levels and there are 4 levels for race. There are 5 empty cells which means there are (6-1) x (4-1) –5 = 10 df for the race xlsm interaction.

Table 3.3.2 reveals that there are 3 levels for the marital status main effect and there are 4 levels for race main effect. There is 1 empty cell which means there are (3-1) x (4-1) –1 = 5 df for the race x marital interaction.

Table 3.3.3 reveals that there are 5 levels for the educational main effect and there are 4 levels for race main effect. There are 5 empty cells which means there are (5-1) x (4-1) –5 = 7 df for the race x educational interaction.

Table 3.3.4 reveals that there are 5 levels for the educational main effect and there are 2 levels for sex main effect. There is 1 empty cell which means there are (5-1) x (2-1) –1 = 3 df for the sex x educational interaction.

Table 3.3.5 reveals that there are 5 levels for the age main effect and there are 3 levels for marital status main effect. There is 1 empty cell which means there are (5-1) x (3-1) –1 = 7 df for the age x marital interaction.

Table 3.3.6 reveals that there are 5 levels for the educational main effect and there are 5 levels for age main effect. There is 5 empty cells which means there are (5-1) x (5-1) –5 = 11 df for the educational x age interaction.

Table 3.3.7 reveals that there are 2 levels missing from the lsm (socio economic status) leaving lsm with 6 levels and there are 5 levels for educational status. There are 7 empty cells which means there are (6-1) x (5-1) –7 = 13 df for the socio economic x educational interaction.

Table 3.3.8 reveals that there are 5 levels for the educational main effect and there are 3 levels for marital status main effect. There are 2 empty cells which means there are (5-1) x (3-1) –2 = 6 df for the educational x marital status interaction.

Table 3.3.9 reveals that there are 5 levels for the educational main effect and there are 2 levels for smoking status main effect. There is 1 empty cell which means there are (5-1) x (2-1) –1 = 3 df for the educational x smoking status interaction.

Table 3.3.10 reveals that there are 2 levels for the sex main effect and there are 8 levels for socio-economic main effect. There are 4 empty cells which means there are (2-1) x (8-1) – 4 = 3 df for the sex x socio-economic interaction.

3.4 Model selection

In order to estimate the variance components of the data we firstly need to have a model that is an adequate representation of the data. GENSTAT was used to do this. The "REML" command in GENSTAT has a "submodel" and "full model" procedure built into it to check for interaction between fixed factors. Genstat has implemented a test statistic due to Welham and Thompson (1997). Unlike in the general linear model, the difference in log-likelihoods between two nested fixed models on REML does not give a sensible test statistic. This is because it is the residual likelihood ℓ_R (i.e. the likelihood of the data after projection into the residual space) that is maximized, rather than the likelihood of the original data. For the residual likelihood two different fixed models correspond to two different projections and hence effectively to two different data sets on which the same random terms are estimated. The statistic proposed by Welham and Thompson (1997) can be used to test a fixed model against a nested submodel. The method calculates the likelihood for the fixed

model as usual. The same projection is then used for the submodel and fixed effects to be dropped in the submodel are constrained to be zero. This gives the likelihood calculated from the same projected data set using the random model, but with some fixed effects constrained to zero for the submodel. The difference in the loglikelihoods therefore gives a likelihood ratio test in the usual way, where

$-2\ (\ell_R - \ell_{OR})$ is the test statistic which has an asymptotic chi-squared distribution with degrees of freedom equal to the degrees of freedom of the fixed model terms constrained to be zero in the submodel. Welham and Thomson state that for small samples this statistic tends to be slightly conservative. I have used this procedure to check for all possible two-way interactions to the full model. Now the maximal model fitted had as its explanatory variables race, sex, age, socio-economic status, marital status, smoking status, educational status, every possible two way interaction between the fixed terms and the random term i.e our clusters. Statistically the model that I fitted using Genstat looked like this :

logtot = μ + rc + sx + age + lsm + mar + ys + edu + rc.sx+…+ys.edu + cl+ error
Interaction terms were then dropped from the model to see if they made any significant differences to the model so that the interaction terms could be included in the model.

The results are summarized in the following table:

Table 3.4.1 : Changes in deviance for significant interaction terms

Interaction Term	Change in deviance $-2(\ell_R - \ell_{OR})$	Degrees of freedom
ys.ed	2.460	3
Lsm.edu	5.857	13
Mar.edu	8.352	6
Age.edu	8.336	11

Table 3.4.1 : Changes in deviance for significant interaction terms		
Sx.edu	2.385	3
Rc.edu	7.448	7
Lsm.ys	9.119	4
Mar.ys	1.957	2
Age.ys	6.323	4
Sx.ys	-0.02565	1
Rc.ys	0.9773	3
Mar.lsm	13.11	10
Age.lsm	21.55	20
Sx.lsm	12.82	5
Rc.lsm	12.30	10
Age.mar	4.721	7
Sx.mar	5.648	2
Rc.mar	2.345	5
Sx.age	5.093	4
Rc.age	7.106	12
Rc.sx	4.702	3

From the above table the only interaction term that is significant at the $\alpha = 0.05$ level when compared to the appropriate chi-square table value is the sx.lsm interaction (since $\chi^2_{0.05;5} = 11.070$) and will be added into the model where the fixed terms are race (rc), sex (sx), age (age), socio-economic status (lsm), marital status (mar), smoking status (ys), educational status (edu), interaction between sex and socio-economic status (sx.lsm) and our random term "cluster". The "strange" degrees of freedom that are present for certain interaction terms are the result of empty cells. Theoretically there are 16 (2 x 8) degrees of freedom for the interaction between sex and socio-economic status (sx.lsm) but from table 3.4.1 there are only 5 degrees of freedom due to empty cells. Higher order interactions, for example 3 way interactions were looked at and all the interaction terms were found to be not significant.

3.5 The model

Now the mixed model from equation 2.2.2.1 i.e.

y=Xβ+ Zu +e can be fitted to our data as follows :

y : 343 x1 is a vector representing our logtot values which is the logarithm of the score of the total number of messages recalled as defined above.

X : 343 x 45 is our incidence matrix for the fixed effects (all two-way interactions included).

β : 45 x1 is a vector that represents the main effects parameters of our fixed effects as well as some of the interactions i.e

$\beta = (rc_2, rc_3, rc_4, sx_2, age_2, age_3, age_4, age_5, lsm_2, lsm_3, lsm_4, lsm_5 , lsm_6, lsm_7, lsm_8, mar_2, mar_3, ys_2, edu_2, edu_3, edu_4, edu_5, (sx.lsm)_{12}, (sx.lsm)_{13}, (sx.lsm)_{14}, (sx.lsm)_{15}, (sx.lsm)_{16}, (sx.lsm)_{17}, (sx.lsm)_{18}, (sx.lsm)_{21}, (sx.lsm)_{22}, (sx.lsm)_{23}, (sx.lsm)_{24}, (sx.lsm)_{25}, (sx.lsm)_{26}, (sx.lsm)_{27}, (sx.lsm)_{28})$.

The fixed effects can be summarized into the following table :

FIXED EFFECT	SURVEY CODE	ABBREVIATION
Race (4 levels)	1=black 2=coloured 3=indian 4=white	rc
Sex (2 levels)	1=male 2=female	sx
Age (5 levels)	1:<25yrs 2:25-34 3:35-44 4:45-54 5:55+	age
socio economic status (8 levels)	1=lowest …. 8=highest	lsm
marital status (3 levels)	1=married 2=ex-married 3=always single	mar
smoking status (2 levels)	0=nonsmoker 1=smoker	ys
education on a 5 point scale (5 levels)	1=lowest … 5=highest	edu

I also had sampling weights (wt) which play a role in the estimation of the variance components by using them to fit different models to the data. The weights were firstly calculated according to the selection probabilities of the enumeration areas, then to take into account the number of eligible individuals in a household (for example, in the same enumeration area, a person in a household with 2 eligible respondents has twice the chance of being selected as a person with 4 eligible people). Finally the weights were adjusted by post - stratification for race, gender and age group. To assess the importance of the weights, the analyses will be carried out with and without the weights and the different estimates will then be compared.

The variance components model is as follows:

$y_{ijklmnopq} = \mu + \alpha_i + \beta_j + \gamma_k + \delta_l + \xi_m + \eta_n + \theta_o + (\beta\delta)_{jl} + \omega_p + \varepsilon_{ijklmnopq}$ where α represents our race effect for i = 1; …; 4, β represents our sex effect for j =1;2, γ represents our age effect for k =1; …;5, δ represents out socio-economic status for l = 1; …;8, ξ represents our marital status for m = 1;…;3, η represents our smoking status for n= 1;2, θ represents our education effect for o = 1; …;5, ω represents our random effects of the random factor "clusters" for p = 1; …;40 and lastly ε represents our residuals for the specified levels.

Z : 343x40 represents the incidence matrix for the one random effect whilst **u** represents our random effects which in our case is the 40 randomly selected clusters from the variable clusters(cl).

In matrix notation our data can be represented as follows:

$$\underline{y} = \mathbf{X}\ \underline{\beta} + \mathbf{Z}\ \underline{u} + \underline{e}$$

$$y_{ijklmnop} = \begin{bmatrix} 1 & 1 & 1 & . & . & 1 & . & . & . & 1 \\ . & . & . & . & . & 1 & . & . & . & . \\ . & 0 & . & . & . & 1 & . & 0 & . & . \\ . & . & 0 & . & . & 1 & . & . & . & . \\ 1 & . & . & 0 & . & 1 & . & . & . & 1 \end{bmatrix} \begin{bmatrix} \mu \\ \mathbf{rc1} \\ . \\ . \\ \mathbf{age1} \\ . \\ . \\ \mathbf{ys1} \\ . \\ \mathbf{edu5} \\ . \\ \mathbf{sx.lsm} \end{bmatrix} + \begin{bmatrix} 1 & . & . & 0 & . & 0 & . & . & . & 0 \\ 0 & 0 & 0 & . & . & . & . & . & . & 0 \\ 1 & 0 & 0 & . & . & . & . & . & . & . \\ 0 & 1 & . & . & . & . & . & . & 0 & . \\ . & 0 & . & . & . & . & 0 & 0 & . & . \\ . & 1 & 0 & . & 0 & . & . & . & 0 & 0 \\ . & . & . & . & . & . & 0 & . & 0 & 1 \\ 0 & 0 & . & . & . & . & . & 0 & 0 & 1 \\ 0 & . & . & . & . & . & 0 & . & 0 & 1 \end{bmatrix} \begin{bmatrix} \mathbf{cl1} \\ . \\ . \\ . \\ \mathbf{cl40} \end{bmatrix} + e_{ij\ldots p}$$

3.6 Henderson's method III

Searle et al (1992, p. 312) summarizes different computer packages with their respective procedures to estimate variance components and in particular states that the SAS procedure VARCOMP calculates ANOVA estimators based on Henderson's method III from the SAS Type I sums of squares.

The model that was fitted is the model obtained from the model selection procedure and that was specified above i.e logtot as the dependent variable, the fixed effects rc, sx, age, lsm, mar, ys, edu, the interaction sx.lsm and the random effects cl. The results are summarized below:

Table 3.6.1: Output from SAS using Varcomp procedure

Variance Components Estimation Procedure

Dependent Variable: LOGTOT

Source	DF	Type I SS	Type I MS
RC	3	5.95240361	1.98413454
SX	1	0.11924355	0.11924355
AGE	4	6.56816874	1.64204218
LSM	5	5.87403490	1.17480698
MAR	2	2.71999920	1.35999960
YS	1	0.27836338	0.27836338
EDU	3	6.04424031	2.01474677
SX*LSM	5	5.33062141	1.06612428
CL	39	57.46697858	1.47351227
Error	279	130.71831225	0.46852442
Corrected Total	342	221.07236593	

Source	Expected Mean Square
RC	Var(Error)+5.807+Var(CL)+Q(RC,SX,AGE,LSM,MAR,YS,EDU,SX*LSM)
SX	Var(Error) + 1.1795 Var(CL) + Q(SX,AGE,LSM,MAR,YS,EDU,SX*LSM)
AGE	Var(Error) + 1.2291 Var(CL) + Q(AGE,LSM,MAR,YS,EDU,SX*LSM)
LSM	Var(Error) + 1.7539 Var(CL) + Q(LSM,MAR,YS,EDU,SX*LSM)
MAR	Var(Error) + 1.0814 Var(CL) + Q(MAR,YS,EDU,SX*LSM)
YS	Var(Error) + 1.0988 Var(CL) + Q(YS,EDU,SX*LSM)
EDU	Var(Error) + 1.1906 Var(CL) + Q(EDU,SX*LSM)
SX*LSM	Var(Error) + 1.0291 Var(CL) + Q(SX*LSM)
CL	Var(Error) + 7.4233 Var(CL)
Error	Var(Error)

Variance Components Estimation Procedure

Dependent Variable: LOGTOT

Variance Component	Estimate
Var(CL)	0.13538350
Var(Error)	0.46852442

The variance component that was estimated for the random part of the model i.e for the clusters, σ^2_{cl} was 0.13538350 whilst the variance component of the error term , σ^2_{error} was 0.46852442. This model was fitted without the sampling weights because the VARCOMP procedure does not have that option in it. Henderson's method III in SAS does not give us any fixed effects estimates or means, it is simply for estimating variance components. The degrees of freedom from the output is due to the fact that SAS takes into account the degrees of freedom that are lost due to empty cells.

3.7 Maximum likelihood estimation

The maximum likelihood estimation technique for estimating the variance components was done using " VARCOMP" and " PROC MIXED" using SAS and the variance components procedure using SPSS. The model fitted was the one as selected from above with rc, sx, age, lsm, mar, ys, sx.lsm and edu as the fixed effects and the clusters (cl) as the random effect. The models using " PROC MIXED" in SAS were fitted with and without sampling weights(as discussed on p.32) whilst SPSS had the sampling weights (wt) fitted along with it. The " VARCOMP" procedure did not allow this. The results are summarized below:

Table 3.7.1: Maximum likelihood results for estimating variance components

	PROC MIXED	PROC MIXED	VARCOMP	SPSS
	estimate (weighted)	estimate (unweighted)	estimate (unweighted)	estimate (weighted)
σ^2_{cl}	0.1932	0.1253	0.1253	0.1940
σ^2_{error}	0.3907	0.4366	0.4366	0.3870

The " PROC MIXED"(weighted) and " SPSS" (weighted) estimates are consistent whilst the " VARCOMP" and unweighted " PROC MIXED" produce smaller estimates of the variance component (σ^2_{cl} and σ^2_{error}) due to the fact that the sampling weights have not been used. The difference between the weighted and unweighted estimates using the " PROC MIXED" procedures is 0.0679 and -0.0459 respectively for σ^2_{cl} and σ^2_{error}. From these differences, with one being positive (for σ^2_{cl}) and one being negative (for σ^2_{error}) we can conclude that σ^2_{error} "benefits" from weights but σ^2_{cl} does not.

The issue of using the sampling weights to estimate the variance components is very important because if the variance component tended to zero then using a model to fit the survey data would lead to the exact same solution as to using the design base approach. Goldstein (1995a, p. 59) states the following:

"…. Weighting should be used when the modeling of a population structure is inadequate, that is the model is misspecified ….As a general procedure, therefore, to guard against model misspecification weighting is to be recommended."

The question of the impact of sample weights on the analysis of survey data has been discussed in two articles by Korn and Graubard (1995a, 1995b) and in a review article by Pfefferman (1993).

The " PROC MIXED" procedure also produces further output with:

- Akaikes's information criterion (AIC) (Akaike 1974) which is computed as AIC = $p(\theta) - q$ where $p(\theta)$ is either REML or ML

loglikelihood and q is the number of variance parameters in the model. The model with the largest AIC is considered to be the best one;

- Schwarz's Bayesian Criterion (SBC) (Schwarz 1978) is computed as SBC = $p(\theta) - (q \log n^*)/2$ where n^* is n for ML estimation and n-p for REML estimation. Models with large SBC are preferred;
- The Null model LRT chi-square statistic is (-2 loglikelihood) from the null model minus (-2 loglikelihood) from the fitted model, where the null model is the model including only the fixed effects i.e assuming a single variance component which is the error variance. Thus this statistic can be used to check whether we need additional components of variance.

The following output was obtained from the weighted " PROC MIXED" procedure with the above statistics :

Table 3.7.2: SAS output using "PROC MIXED"

Model Fitting Information for LOGTOT
Weighted by WT

Description	Value
Observations	343.0000
Log Likelihood	-403.730
Akaike's Information Criterion	-405.730
Schwarz's Bayesian Criterion	-409.568
-2 Log Likelihood	807.4607

The AIC and the SBC are extremely valuable when comparing two or more models to assess which is the best model. I have only fitted this model in SAS and cannot comment on the values of the SBC and the AIC because I have only fitted one model whereas two or more models must be fitted and then the AIC and the SBC can provide statistical insight.

3.8 MINQUE results

The Minque technique was applied to the model using SPSS and yielded the following results:

Table 3.8.1: Variance component estimation using MINQUE

MINQUE SPSS	Estimates
σ^2_{cl}	0.213
σ^2_{error}	0.418

The cluster σ^2_{cl} estimate is smaller than the σ^2_{error} estimate and this implies that the factor " cluster" could be treated as a fixed effect as well. Also SPSS does not give any other output regarding the model itself.

3.9 Weighted and unweighted REML analysis

Genstat was used to calculate the variance components of our model. The REML algorithm (using Fisher's scoring method) was used with and without sampling weights to estimate the treatment effects and variance components in a linear mixed model with both fixed and one random effect. Genstat uses two directives called VCOMPONENTS and REML. The VCOMPONENTS directive specifies the linear mixed model to be fitted into a fixed part and a random one. The REML directive specifies the variable on which the analysis is to be carried out i.e in our case it would be the logtot. This directive also allows an output of the model, variance components, model effects, means, stratum variances from a decomposition of the information matrix for the variance components, variance-covariance matrix of the estimated components, deviance of the fitted model and a submodel if fitted and Wald tests for all fixed terms in the model. Genstat firstly produces an output of descriptive statistics separating

the variables into factors and variates. The means, maximum and the minimum of the variates are calculated whilst the factors are categorized by the number of levels, missing values and a count of the observations in each factor.

The output given by Genstat is as follows:

Table 3.9.1: Output using Genstat with the REML procedure

Identifier	Minimum	Mean	Maximum	Values	Missing
wt	0.04640	0.9287	3.435	343	0
logtot	-0.6931	0.9830	2.251	343	0

Identifier	Values	Missing	Levels
cl	343	0	40
rc	343	0	4
sx	343	0	2
age	343	0	5
lsm	343	0	8
mar	343	0	3
ys	343	0	2
edu	343	0	5

It is also important to note that there is use of of an absorbing factor in the output which are the clusters (cl). The absorbing factor is used to divide the model terms into two groups. This partition is then used in calculations during the fitting process to reduce the size of the matrices that have to be inverted and stored. Therefore the use of the absorbing factor saves computation time and data space. Usually the absorbing factor is chosen as a fixed or random variable that have the largest number of levels as a factor. The actual variance components of the clusters (σ^2_{cl}) and error term (σ^2_{error}) was estimated to be the following values :

Table 3.9.2: Variance component estimates using REML

REML (Genstat)	Estimate (Weighted)	Standard error (Weighted)	Estimate (unweighted)	Standard error (unweighted)
σ^2_{cl}	0.2142	0.065	0.1466	0.0486
σ^2_{error}	0.407	0.0344	0.449	0.0380

The differences between the weighted and unweighted estimates of the clusters and the errors are 0.0676 and 0.042 respectively. The weighted estimates of the clusters and the errors differ somewhat slightly from the weighted ML and Minque estimates by about 0.02 and 0.0163 (σ^2_{cl} REML - σ^2_{cl} ML , σ^2_{error} REML - σ^2_{error} ML) and 0.0012 and 0.011(σ^2_{cl} REML - σ^2_{cl} MINQUE, σ^2_{error} REML - σ^2_{error} MINQUE) because REML takes into account the number of degrees of freedom lost in estimating parameters of the model. These differences are not extreme and large which means that each variance component estimation technique is producing an estimate that is roughly uniform with the other and no evidence is available right now to state that a technique is superior or more precise than the other.

The PROC MIXED procedure in SAS was also used to carry out a REML analysis. The variance component estimates are summarized in the following table:

Table 3.9.3: Variance component estimates using REML in SAS

REML (SAS)	Estimate (Weighted)	Estimate (unweighted)
σ^2_{cl}	0.2119	0.1400
σ^2_{error}	0.4207	0.4695

The differences between the variance component estimates are 0.0719 and 0.0488 respectively.

3.10 Generalised linear mixed model analysis (GLMM)

Genstat has a procedure that is called GLMM where it fits a generalised linear mixed model to the data using the method of either Schall or the marginal method of Breslow and Clayton. Ignoring the random effects **u**, this gives a linear predictor $\mathbf{X}\beta$ on the scale of the link function g(.). Predicted means are

then calculated on this transformed scale, in exactly the same way as they are for REML i.e ignoring the random effects. These means are referred to as the "predicted means". GLMM will also print a table headed "Back-transformed means (on the original scale)" which is simply found as the inverse-function of the link itself. A warning is then printed to emphasize that the means are probabilities not expected values. Thus inference should be carried out on the transformed scale and the back-transformed means are provided only to give an intuitive guide in interpreting results. This same method is followed for

Schall (1991) and Breslow and Clayton's (1993) methods with the only difference being in the way the parameter estimates are found. The result for the unweighted model that was fitted using the method of Schall are given below :

The general table is given with the estimated variance component with a variance covariance matrix. The general information table gives us the type of model fitted with its link (in our case the link is our identity and we have a normal distribution)

Table 3.10.1: Genstat output of the generalized linear mixed model

```
****** Generalised Linear Mixed Model Analysis ******

                Method:   cf Schall (1991) Biometrika
     Response variate:   logtot
         Distribution:   NORMAL
        Link function:   IDENTITY

          Random model:  cl
           Fixed model:  Constant + ((((((rc + sx) + age) + mar)
+ lsm) + ys) + edu) + (sx . lsm)


* Dispersion parameter estimated


**** Monitoring information ****

 Iteration      Gammas Dispersion    Max change
         1      0.3259     0.4492    0.5508E+00
         2      0.3264     0.4492    0.4660E-03
         3      0.3264     0.4492    0.1804E-04

*** Estimated Variance Components ***
```

Random term	Component	S.e.
cl	0.1466	0.0485

*** Residual variance model ***

Term	Factor	Model(order)	Parameter	Estimate	S.e.
Dispersn		Identity	Sigma2	0.449	0.0380

*** Estimated Variance matrix for Variance Components ***

cl	1	0.002354	
Dispersn	2	-0.000201	0.001444
		1	2

The variance component of the clusters σ^2_{cl} is given as 0.1466 which is the exact same estimate as an unweighted REML estimate. The reason for all the estimates of the fixed effects and predicted means (and other parameters) to be the same is the fact that the GLMM procedure in Genstat calls upon the REML procedure. Hence all the results of the unweighted GLMM's can be summarized as the unweighted REML procedure.

Finally the IRREML procedure was used to fit a weighted generalized linear mixed model which was downloaded from the Centre for Biostatistics at Wageningen at their website on the internet at www.plant.wageningen-ur.nl/product/genstat

Now IRREML fits a GLMM, applying iterative reweighted REML to the link adjusted dependent variate. The estimation procedure can be derived as an extension of the iterative reweighted least squares algorithm for GLMs (Schall 1991; Engel and Keen 1994) replacing weighted least squares by weighted REML or by approximation of the likelihood equations by iterative use of Laplace integration (Breslow and Clayton 1993; Engel 1997). The options of IRREML are very similar to REML. Firstly the model is fitted with GLM and then, for estimating the variance components, weighted REML is carried out

iteratively. Weights are the usual weights of a GLM and are functions of the estimated conditional means μ, in which the predicted random effects are included. The output for the model will be now considered :

A table is printed with the general information and the variance component information. The variance component estimate is exactly the same as the REML estimate for this model. This is followed by the table of effects with their standard errors.

Table 3.10.2: Genstat output using IRREML to fit a weighted Generalised Linear Mixed model

```
Generalized Linear Mixed Model Analysis
 ---------------------------------------------------
 Distribution:     Normal
 Link function:    Identity
 Dispersion:       *
 Fixed effects:    ((((((rc + sx) + age) + lsm) + mar) + ys) +
edu) + (sx . lsm)
 Random effects:   cl
 Absorbing factor: cl

***** REML Variance Components Analysis *****

Response Variate : Zeta  ; link-adjusted response variate logtot

Fixed model      : Constant+rc+sx+age+lsm+mar+ys+edu+sx.lsm
Random model     : cl+Dispersn


Number of units  : 343
Weights variate  : W ; iterative weights at link-scale
Absorbing factor : cl

* Dispersn used as residual term



*** Estimated Variance Components ***

Random term                    Component          S.e.

cl                                0.2142        0.0650
Dispersn                          0.4065        0.0344
```

```
*** Approximate stratum variances ***

                                                    Effective d.f.
cl                                    1.7305            37.25
Dispersn                              0.4065           279.75

* Matrix of coefficients of components for each stratum *

          cl        6.19        1.00
    Dispersn        0.00        1.00
```

The fixed effects estimates and predicted means are the same as the weighted REML estimates and will be reproduced in the Overall Summary at the end of this section. The differences between the GLMMs fitted are best summarized in the following tables:

Table 3.10.3: Comparison of variance component estimates with weighted and unweighted GLMMs

Variance component	Schall & Breslow and Clayton(unweighted)	Irreml (weighted)
	estimate	estimate
σ^2_{cl}	0.1466	0.2142
σ^2_{error}	0.449	0.407

The differences in the variance component estimates are clearly evident thereby stressing the importance of using sampling weights in fitting models to estimate the variance components and other parameters of interest. The use of sampling weights ensures that if the variance component tended to zero (i.e. the random cluster effects were negligible) then by fitting a variance component model to the data, the solution of this model will agree with the solution if I had only considered the design of the survey i.e. cluster sampling and worked with an aggregated design approach. Korn and Graubard (1995a, 1995b) give an excellent review of the use of sampling weights and its importance to estimate

the variance components and other parameters. Korn and Graubard(1995b) give four examples of weighted and unweighted analyses using data from the 1988 National Maternal Infant Health Survey and come to the following conclusions:

1. If we wish to use the unweighted estimators we should include the exogenous design variables of the survey as explanatory variables in the model.
2. In some cases we will not be able to reconcile the differences between the weighted and unweighted estimators; in this case we should use the weighted estimators.

However it is still difficult for me to say which is the better analysis, whether the weighted model is better than the unweighted model. The standard errors of the variance components from the REML analysis tell a different story:

Table 3.10.4: Standard errors of variance components from REML analysis

REML (Genstat)	Estimate (Weighted)	Standard error (Weighted)	Estimate (unweighted)	Standard error (unweighted)
σ^2_{cl}	0.2142	0.065	0.1466	0.0486

It is seen that the unweighted estimate of the variance component has a smaller error than the weighted estimate of the variance component. This difference (standard error weighted – standard error unweighted) is 0.0164. This is still not enough evidence for a conclusive decision.

3.11 Overall summary

The following tables give an overall summary of all the techniques used to estimate the variance components of the data as well as the other estimates of other parameters such as the means and effects.

Table 3.11.1: Estimates of the variance components

	Software & Procedure	σ^2_{cl}	σ^2_{error}
Henderson III (unweighted)	Varcomp (SAS)	0.1353	0.4685
ML (weighted)	Proc mixed(SAS)	0.1932	0.3907
ML (unweighted)	Proc mixed(SAS)	0.1253	0.4366
ML (unweighted)	Varcomp (SAS)	0.1253	0.4366
ML (weighted)	SPSS	0.1940	0.3870
Minque (weighted)	SPSS	0.2130	0.418
Unweighted REML	Proc mixed	0.1400	0.4695
Weighted REML	Proc mixed (SAS)	0.2119	0.4207
Weighted REML	REML (Genstat)	0.2142	0.407
REML & GLMM (unweighted)	GLMM (Genstat)	0.1466	0.449
Weighted GLMM	IRREML (Genstat)	0.2142	0.407

From table 3.11.1 it is clear that the estimates of the variance components of the clusters are smaller for all the unweighted methods when compared to the methods where weights were used. The smallest value being 0.1253 and the largest value being 0.1466. For the methods where weights were used the smallest value was 0.1932 and the largest value was 0.2142 (IRREML Genstat). The estimates for the variance of the clusters vary from 0.1253 to 0.214. The variance component estimate of the error terms do not vary widely, since each of the estimates are more or less consistent with one another. The REML and IRREML estimates of the error terms are the lowest. There was also a slight difference in the weighted REML estimates between REML in Genstat and PROC MIXED in SAS by about 0.0023. The IRREML estimate of the variance component of the clusters is the largest. Minqué is also equal to the first iteration estimate of REML.

However the question still remains unanswered as to which method of variance component estimation is the best one? Clearly from table 3.11.1 the methods lead to different answers, so this gives me a good "connecting point" and so serves as a motivation for a simulation study. With a simulation study I will already have known what the unknown parameters are and I can then investigate which methods will lead me to an answer closest to the real values.

The predicted means of the fixed effects are summarized in the following table :

Table 3.11.2: Estimates of the predicted means

Factor	REML & GLMM (unweight)	ML (weight)	ML (unweight)	IRREML (weight)	REML (weight)
Rc (level 1)	0.7283	1.079	1.012	0.7304	0.7305
Rc (level 2)	0.4596	0.720	0.837	0.3652	0.3652
Rc (level 3)	0.7145	1.246	1.129	0.8499	0.8500

Table 3.11.2: (Continued) Estimates of the predicted means

Rc (level 4)	0.4540	0.873	0.839	0.5094	0.5044
Sx(level 0)	*	1.014	1.035	*	*
sx (level 1)	*	0.945	0.918	*	*
age(level 1)	0.7871	*	*	0.7808	0.7808
age (level 2)	0.6285	1.124	1.155	0.6217	0.6218
Age (level 3)	0.5094	0.962	0.991	0.5551	0.5551
Age (level 4)	0.5230	0.913	0.885	0.4658	0.4659
Age (level 5)	0.4975	0.853	0.931	0.6390	0.6390
Lsm (level 3)	0.4877	1.045	0.922	0.6090	0.6090
Lsm (level 4)	0.5220	0.989	0.880	0.4109	0.4109
Lsm (level 5)	0.4372	0.793	0.929	0.4776	0.4776
Lsm (level 6)	0.5508	0.863	0.838	0.6589	0.6589
Lsm (level 7)	0.7889	1.038	0.956	0.8095	0.8095
Lsm (level 8)	0.7570	1.158	1.151	0.7091	0.7091
Mar (level 1)	0.4918	1.035	1.107	0.5218	0.5218
Mar (level 2)	0.7244	0.872	0.864	0.7353	0.7353
Mar (level 3)	0.5510	1.105	1.108	0.5804	0.5804
Ys (level 0)	0.6048	0.962	0.959	0.6412	0.6412
Ys (level 1)	0.5734	0.937	0.946	0.5838	0.5838
Edu (level 1)	-0.9048	1.022	1.001	-0.7985	-0.7895
Edu (level 2)	0.7165	0.655	0.716	0.6564	0.6564
Edu (level 3)	0.8543	0.886	0.850	0.8843	0.8843
Edu (level 4)	1.0365	1.043	1.057	1.0239	1.024
Edu (level 5)	1.2429	1.334	1.285	1.2964	1.296
Sx.lsm (1,3)	0.4665	1.001	0.850	0.6275	0.6276
Sx.lsm (1,4)	0.9037	1.195	1.307	0.8189	0.8190
Sx.lsm (1,5)	0.5629	0.858	0.923	0.4758	0.4758
Sx.lsm (1,6)	0.5738	1.048	0.960	0.6924	0.6924
Sx.lsm (1,7)	0.7643	0.992	1.093	0.6671	0.6671

Table 3.11.2: (Continued) Estimates of the predicted means					
Sx.lsm (1,8)	0.7230	0.991	1.080	0.6500	0.6500
Sx.lsm (2,3)	0.4909	0.979	0.911	0.5905	0.5905
Sx.lsm (2,4)	0.1403	0.390	0.551	0.0028	0.0029
Sx.lsm (2,5)	0.3476	0.869	0.723	0.4795	0.4795
Sx.lsm (2,6)	0.5278	1.028	0.952	0.6253	0.6253
Sx.lsm (2,7)	0.8135	1.324	1.209	0.9519	0.9519
Sx.lsm (2,8)	0.7910	1.079	1.134	0.7682	0.7682

The predicted means vary numerically using the methods of REML, IRREML, ML and GLMM due to the use of sampling weights. The Indian race group which is race group/level 3 has the highest number of messages recalled whilst the fifth educational group which is the highest educational level group recalled the highest number of messages. The Black population, race group/level 1 were able to recall the second highest number of anti-smoking messages. As the educational level increased on the group the number of messages recalled increased pointing logically that education is a crucial element in one's understanding of health risks and diseases in relationship to smoking. Genstat does not produce a set of p-values for the means to test whether they are significantly different from zero but SAS does. From the SAS output, the weighted REML and weighted ML both showed that all the least squares means had p-values less than 0.05 which means that they were significantly different from zero with the exception of the sx.lsm (2,4) mean which was not significant in both ML and REML. In the unweighted ML and unweighted REML all the least squares means were significantly different from zero and had p-values less than 0.05.

The table for the estimates of the fixed effects is as follows:

Table 3.11.3: Estimates of the fixed effects

Factor	REML & GLMM (unweighted)	ML (weighted)	ML (unweighted)	IRREML (weighted)	REML (weighted)
Rc (level 1)	0.000	0.2058	0.2627	0.000	0.000
Rc (level 2)	-0.2687	-0.1532	-0.003	-0.3653	-0.3653
Rc (level 3)	-0.0138	0.3274	0.2899	0.1195	0.1195
Rc (level 4)	0.2743	0.000	0.000	-0.2260	-0.2260
Sx(level 0)	0.000	-0.878	-0.053	0.000	0.000
Sx (level 1)	0.0680	0.000	0.000	0.1185	0.1185
Age(level 1)	0.000	-0.788	0.2331	0.000	0.000
Age (level 2)	-0.1586	-0.827	0.069	-0.1590	-0.1590
Age (level 3)	-0.2776	-0.1321	-0.037	-0.2257	-0.2257
Age (level 4)	-0.2641	-0.1914	0.009	-0.3150	-0.3150
Age (level 5)	-0.2896	0.000	0.000	-0.1480	-0.1480
Lsm (level 3)	-0.2656	-0.100	-0.2231	-0.0224	-0.0224
Lsm (level 4)	0.1807	-0.6889	-0.5829	0.1690	0.1690
Lsm (level 5)	-0.1961	-0.2103	-0.3810	-0.1741	-0.1741
Lsm (level 6)	-0.1492	-0.051	-0.1819	0.0424	0.0424
Lsm (level 7)	0.0413	0.2449	0.075	0.0171	0.0171
Lsm (level 8)	0.000	0.000	0.000	0.000	0.000
Mar (level 1)	0.000	-0.090	-0.095	0.000	0.000
Mar (level 2)	0.2326	0.1431	0.1498	0.21353	0.21353
Mar (level 3)	0.059	0.000	0.000	0.0586	0.0586
Ys (level 0)	0.000	-0.086	-0.061	0.000	0.000
Ys (level 1)	-0.0313	0.000	0.000	-0.0574	-0.0574
Edu (level 2)	1.621	-0.6783	-0.5693	1.455	1.455
Edu (level 3)	1.759	-0.4476	-0.4356	1.683	1.683
Edu (level 4)	1.941	-0.2907	0.2284	1.822	1.822
Edu (level 5)	2.148	0.000	0.000	2.095	2.095

Table 3.11.3:(Continued) Estimates of the fixed effects					
Sx.lsm (1,3)	0.000	0.1097	-0.008	0.000	0.000
Sx.lsm (1,4)	0.000	0.8928	0.8091	0.000	0.000
Sx.lsm (1,5)	0.000	0.077	0.2235	0.000	0.000
Sx.lsm (1,6)	0.000	0.1073	0.061	0.000	0.000
Sx.lsm (1,7)	0.000	-0.2441	-0.062	0.000	0.000
Sx.lsm (1,8)	0.000	0.000	0.000	0.000	0.000
Sx.lsm (2,3)	-0.0436	0.000	0.000	-0.1553	-0.1553
Sx.lsm (2,4)	-0.8314	0.000	0.000	-0.9343	-0.9343
Sx.lsm (2,5)	-0.2474	0.000	0.000	-0.1146	-0.1146
Sx.lsm (2,6)	-0.114	0.000	0.000	-0.1854	-0.1854
Sx.lsm (2,7)	-0.0188	0.000	0.000	0.1660	0.1660
Sx.lsm (2,8)	0.000	0.000	0.000	0.000	0.000

There are considerable differences between the estimates of the Indian race group (race group /level 3) between the unweighted GLMM and the weighted GLMM and REML. This also is true for the education levels 3 and 4. The issue of whether to use the sampling weights clearly has an effect on the estimates of the fixed effects. The differences between the PROC MIXED weighted and PROC MIXED unweighted estimates are clearly evident with a large difference being between the first race group (blacks). Genstat does not produce a set of p-values to test which fixed effects estimates are significantly different from zero but SAS does. In both the weighted ML and weighted REML procedures the following fixed effects estimates had p-values less than zero and were significantly different from zero, they were lsm 4, edu 2, edu 3, edu 4 and sx.lsm (1,4). The rest of the fixed effects estimates were not significantly different from zero. In both the unweighted ML and unweighted REML the very same fixed effects estimates as in the weighted ML and weighted REML cases were also significantly different from zero. The rest of the fixed effects estimates were not significantly different from zero. Now by looking at the weighted ML

and unweighted ML fixed effects estimates, the estimates of the fixed effects indicate that as compared to race group 4 the race group 3 exhibits a higher effect on the number of messages recalled for the weighted and unweighted analyses. When compared to age group 5, age group 3 contributes a lower effect to the number of messages recalled in the weighted analysis whilst in the unweighted analysis age group 1 contributes a higher effect. In both the weighted and unweighted cases, the lsm level 7 contributes a higher effect to our dependent variable when compared to lsm level 8. Again in the weighted and unweighted cases, marital group 2(ex-married) exhibits a higher effect when compared to marital level 3(single) in recalling the number of anti-smoking messages. In the weighted case education level 5 has the highest effect towards the number of messages recalled whilst in the unweighted case education level 4 has the highest effect towards the number of messages recalled.

From the above table one can see the following for the REML weighted case:
The effect of race group 3 (Indians) is higher than that of race group 1(Blacks). The effects of the females contributing to the number of messages recalled is higher than that of the males. Age group 2 (24-34yrs) has a lesser effect on the number of messages recalled than that of age group 1(<25 yrs). The lsm level 4 has a higher effect when compared to lsm level 2 in its contribution to the number of messages recalled. The ex-married contribute a higher effect than the single to the dependent variable whilst the non smokers contribute a higher effect than the smokers in recalling messages. The more educated respondents (edu level 5) have a higher effect as compared to the least educated respondents in recalling messages.

From this output one can see that for the REML unweighted case:
The effect of race group 3 (Indians) is lower than that of race group 1(Blacks). The effects of the females contributing to the number of messages recalled is higher than that of the males. All the other Age groups have a lesser effect on the number of messages recalled than that of age group 1(<25 yrs). The lsm level 4 has a higher effect when compared to lsm level 2 in its contribution to the number of messages recalled. The ex-married contribute a higher effect than the single to the dependent variable whilst the non smokers contribute a higher

effect than the smokers in recalling messages. The more educated respondents (edu level 5) have a higher effect as compared to the least educated respondents in recalling messages.

It is interesting to note that for the race group 4 i.e. the Whites, the estimate is higher than that of race group 1(Blacks) in the unweighted case but lower in the weighted case. The lsm level 6 estimate is also higher than the lsm level 2 in the unweighted case but smaller than lsm level 2 in the weighted case.

The survey itself had a small variance component associated with the clusters and maybe an extension to the survey would be to use the variance components to estimate the intra-cluster correlation coefficient, to assess the extent to which individuals in the same cluster are similar and then use the intra-cluster correlation coefficient to look at the design effect i.e Kish's design effect to assess the validity on the sample size with respect to the design of the survey itself. The intra-cluster correlation coefficient is given as $\sigma^2_{cl} / (\sigma^2_{cl} + \sigma^2_{error})$.

For the future, maybe society should be constantly reminded by Health and Medical Institutions through media and advertising of the health risks of smoking in order to increase the ability to recall warning slogans about the dangers of smoking and thereby decrease the fatality of diseases associated with smoking. Hence I can state that an educated society is a healthy society. My recommendations to the relevant people at the MRC (even though this survey was carried out a few years ago) is perhaps to embark on a full, drawn out campaign to create a health awareness using every form of media (as well as the internet) on the health hazards associated with smoking.

Chapter 4: Simulated data

4.1 Introduction

The key underlying factors that have motivated this chapter is that real life examples of the mixed model are difficult to come by and the fact that the dependent variable i.e. the "logtot" of my smoking data set is not such an excellent example of a continuous dependent variable. As a result I have had to simulate a mixed model data set using SPSS and apply the various variance component estimation techniques to this balanced data set as well as an unbalanced data set derived from the balanced data set to look at the differences of the techniques as to how they are able to "handle" balanced and unbalanced data in their parameter estimates.

4.2 Method and model of the simulated data

The data set was simulated in SPSS. Theoretically the model looked like this:

$$y_{ijklm} = \mu + \alpha_i + \beta_j + \gamma_k + \delta_l + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + \varepsilon_{ijklm} \quad \ldots \quad 4.2.1$$

where α, β and γ were the fixed effects where α_i has two levels for i = 1,2 , β_j has four levels for j = 1,2,3,4 and γ_k has three levels for k = 1,2,3. δ_l has ten levels for l = 1,...,10 and will be considered to be the random effect. The interactions $(\alpha\beta)_{ij}$ and $(\alpha\gamma)_{ik}$ have eight (2x4) and six (2x3) levels respectively.

I carried out many simulations to arrive at the "best possible model". By fixing the variance of the random component at different values (from 100, 50, 20, 11 and other values) and comparing the estimate of the variance of the random component against the fixed or chosen value of the variance of the random component before simulation, I was able to see the difference/deviation numerically and this led me to an appropriate chosen value of the variance of the random component (as well as looking at the standard error of the estimates of the variance of the random component). Hence I settled for a variance value of 4

for the random component. I then simulated 4 replicates within each fixed effect combination. The random component, δ_l was simulated from a normal distribution with a mean of 0 and a variance of 4. The error component, ε_{ijklm} was simulated from a normal distribution with a mean of 0 and a variance of 16. The three fixed effects were created with chosen values, so such that $\Sigma\alpha_i = 0$, $\Sigma\beta_j = 0$ and $\Sigma\gamma_k = 0$ with interactions between two of them $\Sigma(\alpha\beta)_{ij} = 0$ and $\Sigma(\alpha\gamma)_{ik} = 0$. The values of the fixed effects were chosen as follows:

Table 4.2.1: Values of fixed effects that were set before simulation

Fixed effect	Value
μ	10
α_1	-6
α_2	6
β_1	-12
β_2	-4
β_3	4
β_4	12
γ_1	-9
γ_2	0
γ_3	9
$(\alpha\beta)_{11}$	3
$(\alpha\beta)_{12}$	-5
$(\alpha\beta)_{13}$	0
$(\alpha\beta)_{14}$	2
$(\alpha\beta)_{21}$	-3
$(\alpha\beta)_{22}$	5
$(\alpha\beta)_{23}$	0
$(\alpha\beta)_{24}$	-2
$(\alpha\gamma)_{11}$	1
$(\alpha\gamma)_{12}$	-4
$(\alpha\gamma)_{13}$	3
$(\alpha\gamma)_{21}$	-1
$(\alpha\gamma)_{22}$	4
$(\alpha\gamma)_{23}$	-3

The y_{ijklm}, dependent values were then simulated as the sum of the fixed and random components. In the matrix notation the mixed model is as follows:

The mixed model from equation 2.2.2.1 i.e.

$\mathbf{y=X\beta+Zu+e}$ can be fitted to the data as follows:

$\mathbf{y}$: 960 x1 is a vector representing the y values.

$\mathbf{X}$: 960 x 24 is the incidence matrix for the fixed effects.

β : 24 x1 is a vector that represents the main effects parameters of the fixed effects as well as interactions i.e.

$\beta = (\mu, \alpha_1, \alpha_2, \beta_1, \beta_2, \beta_3, \beta_4, \gamma_1, \gamma_2, \gamma_3, (\alpha\beta)_{11}, \ldots, (\alpha\gamma)_{23})$

$\mathbf{Z}$: 960x10 represents the incidence matrix for the one random effect δ_l.

In matrix notation the data can be represented as follows:

$$\underline{\mathbf{y}} = \mathbf{X}\ \underline{\beta} + \mathbf{Z}\ \underline{\mathbf{u}} + \underline{\mathbf{e}}$$

$$\mathbf{y} = \begin{bmatrix} 1 & . & . & . & 1 & 1 \\ . & . & . & 1 & . & . \\ 1 & . & 1 & . & . & . \\ . & . & 1 & . & . & . \\ . & . & . & 1 & . & . \\ 1 & 1 & . & . & . & 1 \end{bmatrix} \begin{bmatrix} \mu \\ \alpha_1 \\ . \\ . \\ \beta_4 \\ . \\ . \\ \gamma_3 \\ . \\ \alpha\beta_{24} \\ . \\ \alpha\gamma_{23} \end{bmatrix} + \begin{bmatrix} 1 & . & . & . & . & 0 \\ . & . & . & 1 & . & 0 \\ 0 & . & 1 & . & . & . \\ . & . & . & 1 & 0 & . \\ 0 & . & . & . & . & 1 \\ 0 & 0 & . & . & . & 1 \end{bmatrix} \begin{bmatrix} \delta_1 \\ . \\ . \\ . \\ \delta_{10} \end{bmatrix} + \mathbf{e}$$

The unbalanced data set was derived from the above balanced data set by generating random numbers from 1 to 960 and then taking the first 100 numbers without replacement (ensuring a different set of 100 numbers at each time) and

deleting the y values at those number positions, giving me a new data set of size 860 with one hundred missing y values. Twenty independent repetitions (with different random seeds) of the same model of the data simulated were carried out i.e. there were twenty data sets with balanced data and twenty data sets with unbalanced data.

4.3 Results of estimates for σ^2_δ

Using the MINQUE, ANOVA (Henderson's method III), Maximum likelihood (ML) and Restricted Maximum Likelihood (REML) procedures in SPSS, the following variance component estimates resulted for the balanced and unbalanced simulated data:

Table 4.3.1: Estimates of variance components using the different procedures for σ^2_δ

Case No.	MINQUE $\hat{\sigma}_\delta^2$		ANOVA $\hat{\sigma}_\delta^2$		ML $\hat{\sigma}_\delta^2$		REML $\hat{\sigma}_\delta^2$	
	Balanced	Unbalanced	Balanced	Unbalanced	Balanced	Unbalanced	Balanced	Unbalanced
1	3.859	3.916	3.859	3.810	3.458	3.500	3.859	3.908
2	3.862	8.371	3.862	8.703	3.462	7.841	3.862	8.730
3	5.998	6.305	5.998	6.540	5.384	5.988	5.998	5.398
4	4.195	7.135	4.195	7.209	3.760	7.003	4.195	6.892
5	3.359	5.344	3.359	5.369	3.009	4.092	3.359	3.569
6	2.887	3.802	2.887	3.744	2.575	3.403	2.887	3.798
7	3.896	4.497	3.896	4.461	3.492	4.029	3.896	4.495
8	5.469	3.618	5.469	3.573	4.907	3.236	5.469	3.614
9	3.610	2.436	3.610	2.421	3.235	2.174	3.610	2.434
10	4.736	4.909	4.736	4.807	4.248	4.398	4.736	4.904
11	3.549	5.355	3.549	5.351	3.481	5.302	3.549	5.353
12	3.500	4.059	3.501	4.057	3.635	3.637	3.501	4.059
13	3.550	6.019	3.505	5.906	3.002	5.397	3.505	6.015
14	3.720	4.241	3.720	4.324	3.682	3.806	3.720	4.246
15	4.535	4.281	4.535	4.330	4.067	3.839	4.535	4.284
16	3.839	2.801	3.836	2.746	3.441	2.500	3.836	2.796
17	5.054	4.121	5.054	4.048	4.534	3.689	5.054	4.117
18	3.702	5.757	3.702	5.760	3.317	5.165	3.702	5.757
19	2.234	5.493	2.237	5.508	1.996	5.129	2.237	5.494
20	2.656	3.583	3.656	3.626	2.381	3.211	3.656	3.587
Mean	**3.881**	**4.802**	**3.958**	**4.815**	**3.553**	**4.376**	**3.958**	**4.673**
Std. dev	**0.958**	**1.450**	**0.8602**	**1.516**	**0.8114**	**1.436**	**0.8602**	**1.146**

From the above table it is seen that the REML and the ANOVA balanced estimates are the closest to 4 and the standard deviation is low as well when compared to the estimate(s) of the MINQUE standard deviation(s). The ML balanced estimate of the variance component is not as close to 4 as the REML and ANOVA counterparts are, but it has a lower standard deviation than the REML balanced standard deviation. The ML balanced standard deviation is also smaller than those of the MINQUE and ANOVA estimates. Searle (1995,1996) has pointed out that one of the attractive features of REML, for balanced data, is that it produces exactly the same estimates as ANOVA does and this is true in my case from the above table. The difference between the standard deviation of the REML and ML estimates is 0.0386. This difference is small and the attractive feature about REML, in this case is that the estimate is much closer to 4 (3.963) than that of the ML estimate (3.553). From this simulation exercise I can conclude that REML and ANOVA have been the procedures that have yielded more accurate estimates than the other procedures. Furthermore, the unbalanced data estimates were more biased than the balanced data estimates i.e. they yielded higher values (> 4) for their estimates and were less accurate than the balanced data estimates because of the higher standard deviations than those of the balanced data. The reasons that ML did not yield an estimate very close to 4 could be the fact that the ML procedure does not take into account the number of degrees of freedom lost when estimating parameters of the model. I will now look at the estimates of the error term σ^2_e.

4.4 Results of estimates for σ^2_e

Table 4.4.1: Estimates of variance components using the different procedures for σ^2_e

Case no.	MINQUE $\hat{\sigma}_e^2$		ANOVA $\hat{\sigma}_e^2$		ML $\hat{\sigma}_e^2$		REML $\hat{\sigma}_e^2$	
	Balanced	Unbalanced	Balanced	Unbalanced	Balanced	Unbalanced	Balanced	Unbalanced
1	16.113	15.947	16.113	15.947	15.926	15.741	16.113	15.948
2	15.189	15.299	15.189	15.299	15.013	15.101	15.299	15.299
3	15.135	15.431	15.135	15.431	14.960	15.231	15.431	15.439
4	16.204	15.896	16.204	15.899	16.017	15.690	15.899	15.896
5	15.007	14.847	15.007	14.876	14.833	14.655	14.876	14.879
6	15.199	15.504	15.199	15.504	15.023	15.304	15.504	15.506
7	15.044	14.681	15.044	14.569	14.877	14.491	14.569	14.562
8	16.156	16.624	16.153	16.624	15.969	16.053	16.624	16.624
9	15.961	16.009	15.961	16.009	15.777	15.802	16.009	16.009
10	15.007	14.864	15.007	16.052	14.833	14.647	16.052	16.052
11	14.798	15.623	14.798	16.921	14.627	16.003	16.921	15.924
12	16.236	16.161	16.236	16.106	16.048	15.952	16.106	16.106
13	15.623	16.023	15.623	14.982	14.995	15.962	14.982	16.026
14	15.988	16.029	15.988	14.369	15.803	16.033	14.369	16.923
15	16.162	16.525	16.162	16.525	15.974	16.311	16.525	16.524
16	15.031	15.090	15.031	14.592	14.857	14.895	14.592	15.050
17	16.185	16.147	16.185	15.938	15.998	15.938	15.938	16.009
18	16.052	15.717	16.052	15.717	15.866	15.514	15.717	16.002
19	15.971	15.818	15.971	15.818	15.786	15.613	15.818	15.818
20	15.786	16.975	15.786	16.958	15.971	16.755	16.958	16.953
Mean	**15.642**	**15.761**	**15.642**	**15.710**	**15.548**	**15.585**	**15.715**	**15.877**
Std. dev	**0.521**	**0.609**	**0.520**	**0.758**	**0.537**	**0.600**	**0.761**	**0.627**

From the above table it is seen that all the estimates of the error component of the model is below 16 (which was the value that I had chosen as the variance of the error terms). The balanced MINQUE and ANOVA estimates are the same in their magnitude and have the smallest standard deviations making them accurate estimates as compared to the other estimates. The REML unbalanced estimate is closest to 16 but with a higher standard deviation than that of the MINQUE and ANOVA standard deviations. The ML balanced estimate has the smallest estimate with a standard deviation of 0.537. There are not huge differences in the magnitudes of the estimates of the error component between the balanced and unbalanced simulated data.

In conclusion, I would like to state that the simulated data was much easier to work with and interpret than the smoking survey data. The simulated data gave me a chance to assess the advantages and disadvantages of each technique, for example, how important it is that REML does take into account the degrees of freedom lost in the model when estimating parameters. The ML estimates were clearly smaller than the other estimates of the other techniques. MINQUE estimates are equal to the first iterative estimate of REML. The differences between the balanced and unbalanced data estimates of different parameters was different but not by much.

Chapter 5: Conclusion

Firstly there are huge research avenues in the application of the mixed model to cluster sampled survey data to estimate the variance components. The use of sampling weights especially for estimating the variance components for survey data is imperative and crucial. I have found the topic of estimating the variance components both stimulating and interesting.

The data itself was not the easiest data set to work with because the dependent variable not being a good example of a continuous variable. The original y-total values made analyses invalid due to heteroskedascity of the data hence the use of the "logtot" as a dependent variable but still this was one of the criticisms of the data set itself.

The simulated data was easier to work with, presenting me with an opportunity to look at the different variance component models and assess them. The simulated data also presented me with an opportunity to experience proper academic research and that was truly rewarding. There are always differences in analyses and parameter estimates between balanced and unbalanced data, but because most "real life" data sets are unbalanced, more research seems to favour this area.

The techniques used to estimate the variance components were fascinating to work with leaving me to conclude that the REML and IRREML techniques were the most versatile and viable having many features and characteristics built into them in theory and practice. Every method seems to have an advantage and a disadvantage. ML estimation does not take into account the number of degrees of freedom with the fixed effects of the model. REML does take into account the number of degrees of freedom for the fixed effects of the model. In practice the best approach to estimating the variance components is to work out both the REML and ML estimates and compare them. If the difference is not too great then either method can be chosen. If the difference is too great then one must possibly look at the standard errors of the variance component estimates.

GLMMs have been an excellent tool for estimating variance components but their estimates have a downward bias.

Currently there are many new Bayesian approaches available to estimate the variance components of data but the Bayesian approach has still many questions to answer and the software is not as friendly to work with.

The software involved showed that all three packages i.e. Genstat, SAS and SPSS are statistically viable for estimating variance components of different models. SPSS was very user friendly but limited in its output power. Genstat and SAS were excellent as a programming tool and to handle various problems with a wide variety of output and numerical control. It seems to me that the approach to estimating the variance components might be to use REML, IRREML (weighted GLMMs) because they suit the design of the experiment in the sense that they take into account the actual foundation plan of the survey i.e. the details of the variables are taken into account.

Diagnostics for the mixed model applied to survey data for estimating variance components is also perhaps an area that needs exploration. I have found this area of estimating variance components both stimulating and exhilarating.

BIBLIOGRAPHY

1. **Aitkin, M.** (1999). A general maximum likelihood analysis of variance components in generalized linear models. *Biometrics*, 55, 117-128.

2. **Akaike, H.** (1974). A new look at the statistical model identification. *IEEE Transaction on automatic control*. AC-19, 716-723.

3. **Breslow, N.E.** and **Clayton, D.G.** (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88, 9-25.

4. **Callan, T.P.** and **Harville, D.A.** (1991). Some new algorithms for computing restricted maximum likelihood estimates of variance components. *J. Statist. Comput. Simulation,* 38(1-4).239-259.

5. **Chow, S.C.** and **Shao, J.** (1988). A New Procedure for the estimation of variance components. *Statist. Prof. Letter* 6(5) : 349 – 355.

6. **Cochran, W.G.** (1939). The use of analysis of variance in enumeration by sampling. *J. Amer. Stat. Assoc.*, 34, 492 – 510.

7. **Crump, S.L.** (1946). The estimation of variance components in analysis of variances. *Biometrics Bull.*, 2, 7 – 11.

8. **Daniels, H.E.** (1939). The estimation of components of variance. *J. R. Stat. Soc. Suppl.*, 6, 186 – 197.

9. **Duchateau, L.** and **Janssen, P.** and **Rowlands, J.G.** (1998). *"Linear Mixed Models: An Introduction with Applications in Veterinary research'*, ILRI (International Livestock Research Institute) Nairobi, Kenya.

10. **Engel, B.** (1988). A simple illustration of the failure of PQL, IRREML and APHL as approximate ML methods for mixed models for binary data. *Biometrical Journal,* 40, 141-154.

11. **Engel, B.** and **Keen, A.** (1994). A simple approach for the analysis of generalized linear mixed models. *Statistica Neerlandica,* 48, 1-22.

12. **Engel, B.** (1990). The analysis of unbalanced linear models with variance components. *Statistica Neerlandica,* 44, 195 – 219.

13. **Engel, B.** and **Buist, W.** (1988). Bias reduction of approximate maximum likelihood estimates for heritability in threshold models. *Biometrics,* 54, 1155 – 1164.

14. **Engel, B.** and **Busit, W.** and **Visscher, A.** (1995). Inference for threshold models with variance components from the generalized linear mixed model perspective. *Genetics Selection Evolution,* 27, 15 – 32.

15. **Fienburg, S.E** (1980). *The analysis of cross classified data* (2nd edition). Cambridge, M.A: MIT PRESS.

16. **Firth, D**. (1991). Generalised Linear Models. In: *Statistical Theory and Modelling: In Honour of Sir David Cox*.Eds.D.V.Hinkley, N.Reid and E.J.Snell. London: Chapman and Hall, 55-82.

17. **Ganguli, M.** (1941). A note on nested sampling. *Sankhya 5*, 449 – 452.

18. **Ghosh, M.** (1995). New directions for estimation of variance components. *J. Statist. Res 29*(2): 1-15.

19. **Gilmour, A.R.** and **Thompson, R.** and **Cullis, B.R.** (1995). Average Information REML: An efficient algorithm for variance parameter estimation in linear mixed models. *Biometrics*, 51, 1440 – 1450.

20. **Goldstein, H.** (1986). Multilevel mixed linear model analysis using iterative generalised least squares. *Biometrika* 73, 43 – 56.

21. **Goldstein, H.** (1989). Restricted unbiased iterative generalised least squares estimation. *Biometrika* 76, 622 – 623.

22. **Goldstein, H.** (1995). *Multilevel Statistical Models* (2nd edition). London: Arnold.

23. **Goldstein, H.** and **Rasbash, J.** (1996). Improved approximations for multilevel models with binary responses. *Journal of the Royal Statistical Society* A 159, 505 – 513.

24. **Hartley, H.O**. and **Rao, J.N.K.** (1967). Maximum likelihood estimation for the mixed analysis of variance model. *Biometrika* 54, 93 – 108.

25. **Henderson, C.R.** (1950). Estimation of genetic parameters. (Abstract). *Annals of Mathematical Statistics*, 21, 309 – 310.

26. **Henderson, C.R.** (1953). Estimation of variance and covariance components. *Biometrics* 9, 226 – 252.

27. **Henderson, C.R., Searle ,S.R.** and **Schaeffer, L.R.** (1974). The invariance and calculation method 2 for estimating variance components. *Biometrics*, 30, 583-588.

28. **Hocking, R.R.** (1990) A new approach to variance component estimation with diagnostic application. *Comm. Statist. Theory. Methods,* 19(12),4591-4617.

29. **Jackson, R.W.B.** (1939), Reliability of mental tests. *Brit. J. Psychol.* 29, 267 – 287.

30. **Khuri, A.** and **Sahai, H.** (1985). Variance components Analysis: A Selective Literature Survey. *Inter. Stat. Rev* 53(3): 279 – 300.

31. **Kish, L.** (1965). *Survey Sampling*. New York: Wiley.

32. **Kleffe, J.** and **Seifert, B.** (1986). Computation of variance components by using the MINQUE Method. *Journal of the Multivariate Analysis* 18, 107 – 116.

33. **Kleffe, J.** and **Rao, J.N.K.** (1986). The existence of asymptotically unbiased non-negative quadratic estimates of variance components in ANOVA models. *J. Amer. Stat. Assoc.* 81, 692 – 698.

34. **Kleffe, J.** and **Rao, J.N.K.** (1988). *Estimation of variance components and applications.*(1st edition) Netherlands: Elsevier Science Publishers.

35. **Korn, E.L.** and **Graubard, B.I.** (1995a). Analysis of large health surveys: According for the sampling design. *Journal of the Royal Statistical Society* A 158, 263 – 295.

36. **Korn, E.L.** and **Graubard, B.I.** (1995b). Examples of differing weighted and unweighted estimates from a sample survey. *The American Statistican* 49, 291 – 295.

37. **Kuk, A.Y.C.** (1995). Asymptotically unbiased estimation in generalized linear models with random effects. *Journal of the Royal Statistical Society* B 57, 395 – 407.

38. **Lee, Y.** and **Nelder, J.A.** (1996). Hierarchical generalized linear models (with discussion). *Journal of the Royal Statistical Society* B58, 619 – 678.

39. **Liang, K.Y.** and **Zeger, S.L.** (1986). Longitudinal data analysis using generalised linear models. *Biometrika* 73, 45 – 51.

40. **Lin, X.** and **Breslow, N.E.** (1996). Bias correction in generalized linear mixed models with multiple components of dispersion. *Journal of the American Statistical Association* 91, 1007 – 1016.

41. **Littell, R.C.**and **Milliken, G.A.**, **Stroup, W.W.** and **Wolfinger, R.D.** (1996). *SAS System for Mixed Models*. Cary, NC : SAS Institute Inc.

42. **Liu, Q.** and **Pieree, D.A.** (1993). Heterogeneity in Mantel-Haenszel type models. *Biometrika* 80, 543 – 556.

43. **Longford, N.T.** (1993). *Random Coefficient Models*, 2nd ed. New York: Oxford University Press.

44. **McCullagh, P** and **Nelder, J.A** (1989).*Generalised Linear Models*(2nd edition). London: Chapman and Hall.

45. **Neyman, J., Iwaszkiewicz, K.** and **Kolodziejczyk, S.T.** (1935). Statistical problems in agricultural experimentation. *J.R. Stat. Soc. Suppl.* 2, 107 – 154.

46. **Ofversten, J.** (1993). Exact tests for Variance Components in Unbalanced Mixed Linear Models. *Biometrics* 49(1) : 45-57.

47. **Patterson, H.D.** and **Thompson, R.** (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika* 58, 545 – 554.

48. **Pfefferman, D.** (1993). The role of sampling weights when modeling survey data. *International Statistical Review* 51, 317 – 337.

49. **Pfefferman, D., Skinner, C.J.** and **Holmes, D.J., Goldstein, H.** and **Rasbash, J.** (1998). Weighting for unequal selection probabilities in multilevel models (with discussion). *Journal of the Royal Statistical Society* B 60, 23 – 56.

50. **Robinson, D.L.** (1987). Estimation and use of variance components. *The Statistician* 36, 3 – 14.

51. **Robinson, G.K.** (1991). That BLUP is a good thing: the estimation of random effects. *Statistical Science* 6, 15 – 51.

52. **Rodriguez, G.** and **Goldman, N.** (1995). An assessment of estimation procedures for multilevel models with binary responses. *Journal of Royal Statistical Society* A 158, 73-89.

53. **Schall, R.** (1991). Estimation in generalized linear models with random effects. *Biometrika* 78, 719 – 727.

54. **Schwartrz, G.** (1978). Estimating the dimension of a model. *Annals of Statistics*, 6, 461-464.

55. **Scheffe, H.** (1956). Alternative models for the analysis of variance. *Ann. Math. Stat.* 27, 251 – 271.

56. **Searle, S.R.** (1970). Large sample variances of maximum likelihood estimators of variance components using unbalanced data. *Biometrika* 26, 505 – 524.

57. **Searle, S.R.** (1987). *Linear Models for unbalanced data*. 2nd ed. New York: John Wiley & Sons, Inc.

58. **Searle, S.R.** (1988). Mixed Models and Unbalanced Data: Wherefrom, whereat, whereto?-*Comm. Statist. Theory Methods* 17(4) : 935 – 968

59. **Searle, S.R.** (1988). Mixed Models and unbalanced data: wherefrom, whereat and Whereto. *Commun. Stat. A: Theory & Methods (Special Issue on Analysis of the Unbalanced Mixed Model)* 17, 935 – 968.

60. **Searle, S.R., Casella, G.** and **McCulloch, C.E.** (1992). *Variance Components*.New York: John Wiley & Sons, Inc.

61. **Searle, S.R.** (1995). An overview of variance component estimation. *Metrika* 42 (3-4) : 215 – 230.

62. **Searle, S.R.** (1996). An overview of variance component estimation. *Tatra Mt. Math. Public.* 7: 139 – 142.

63. **Searle, S.R.** (1997). *Linear Models*, New York: John Wiley and Sons, Inc

64. **Seifert, B.** (1985). Estimation and test of variance components using MINQUE Method. *Statistics* 16 (4), 621 – 635.

65. **Simpson, S., Diamond, I., Tonkin, P.** and **Tye, R.** (1996). Updating small area population estimates in England and Wales. *Journal of the Royal Statistical Society* A 159, 235-247.

66. **Sun, L., John, S.J., Guttman, I.**, and **Leonard, T.** (1996). Bayesian methods for variance components models. *J. AM. Statist. Assoc.* 91 (434) : 743 – 752

67. **Thompson, W.A., Jr.** (1962). The problem of negative estimates of variance components. *Ann. Math. Stat.* 33, 273 – 289.

68. **Tippett, L.H.C.** (1931). *The Methods of Statistics*, 1st edn. Williams and Norgate, London.

69. **Waddington, D., Welham, S.J., Gilmour, A.R.** and **Thompson, R.** (1994). Comparisons of some GLMM estimators for a simple binomial model. *Genstat Newsletter* 30, 13 – 24.

70. **Welham, S.J.** (1993). Procedure GLMM. In: *Genstat 5 procedure Library Manual, Release 2[3]* eds. R.W. Payne and G.M. Arnold. Oxford: Numerical Algorithms Group.

71. Welham, S.J. and Thompson, R. (1997). Likelihood ratio tests for fixed model terms using residual maximum likelihood. *Journal of the Royal Statistical Society* B 59, 701-714.

APPENDIX 1 Weighted REML output from Genstat

```
GenStat Release 4.21  (PC/Windows 98)                 04 December 2001
09:39:54
Copyright 2001, Lawes Agricultural Trust (Rothamsted Experimental
Station)

 GenStat Fifth Edition (Service Pack 1)
GenStat Procedure Library Release PL12.1

 GenStat Fifth Edition (Service Pack 1)
GenStat Procedure Library Release PL12.1

 17  %CD 'C:/GEN5ED/BIN'
 18  job ' a reml analysis with weights'
 19  open 'a:smoket.txt';channel=2;width=500
 20  factor[lev=40]cl
 21  factor[lev=4]rc
 22  factor[lev=2]sx
 23  factor[lev=5]age
 24  factor[lev=3]mar
 25  factor[lev=5]edu
 26  factor[lev=8]lsm
 27  factor[lev=!(1,0);lab=!t(smoker,nonsmoker)]ys
 28  factor[lev=!(1,0);lab=!t(yes,no)]d1
 29  factor[lev=!(1,0);lab=!t(yes,no)]d2
 30  factor[lev=!(1,0);lab=!t(yes,no)]d3
 31  read[channel=2]cl,rc,sx,age,lsm,mar,ys,wt,d1,d2,d3,edu,logtot

 Identifier     Minimum        Mean     Maximum      Values     Missing
         wt     0.04640      0.9287       3.435         343           0
     logtot     -0.6931      0.9830       2.251         343           0

 Identifier      Values     Missing      Levels
         cl         343           0          40
         rc         343           0           4
         sx         343           0           2
        age         343           0           5
        lsm         343           0           8
        mar         343           0           3
         ys         343           0           2
         d1         343           0           2
         d2         343           0           2
         d3         343           0           2
        edu         343           0           5

 32  model[weight=wt] logtot
 33  vcomponents[fixed=rc+sx+age+mar+lsm+ys+edu+sx.lsm;absorb=cl]cl
 34  reml[print=model ,comp, stra
;means,deviance;weights=wt;method=fisher]logtot
```

34.....
.....

***** REML Variance Components Analysis *****

Response Variate : logtot

Fixed model : Constant+rc+sx+age+mar+lsm+ys+edu+sx.lsm
Random model : cl

Number of units : 343
Weights variate : wt
Absorbing factor : cl

* Residual term has been added to model

* Non-sparse algorithm with Fisher scoring

*** Estimated Variance Components ***

Random term	Component	S.e.
cl	0.2142	0.0650

*** Residual variance model ***

Term	Factor	Model(order)	Parameter
Estimate	S.e.		
Residual		Identity	Sigma2
0.407	0.0344		

*** Approximate stratum variances ***

		Effective d.f.
cl	1.7299	37.25
units	0.4065	279.75

* Matrix of coefficients of components for each stratum *

cl	6.19	1.00
units	0.00	1.00

*** Deviance: -2*Log-Likelihood ***

Deviance	d.f.
262.67	315

Note: deviance omits constants which depend on fixed model fitted.

*** Table of predicted means for Constant ***

0.6125 Standard error: 0.17659

*** Table of predicted means for rc ***

rc	1	2	3	4
	0.7305	0.3652	0.8500	0.5044

Standard error of differences: Average 0.3153
Maximum 0.4412
Minimum 0.1808

Average variance of differences: 0.1080

*** Table of predicted means for sx ***

sx	1	2
	*	*

* All values in table missing: due to missing factor combinations in higher
order interactions

*** Table of predicted means for age ***

age	1	2	3	4	5
	0.7808	0.6218	0.5551	0.4659	0.6390

Standard error of differences: Average 0.1487
Maximum 0.1844
Minimum 0.1197

Average variance of differences: 0.02247

*** Table of predicted means for mar ***

mar	1	2	3
	0.5218	0.7353	0.5804

Standard error of differences: Average 0.1489
Maximum 0.1753
Minimum 0.1201

Average variance of differences: 0.02268

*** Table of predicted means for lsm ***

lsm	1	2	3	4	5	6

7 8

```
                  *         *    0.6090    0.4109    0.4776    0.6589
0.8095    0.7091

Standard error of differences:      Average        0.2104
                                    Maximum        0.3142
                                    Minimum        0.1163

Average variance of differences:                   0.04843


*** Table of predicted means for ys ***

          ys         smoker     nonsmoker
                     0.6412        0.5838

Standard error of differences:      0.08960


*** Table of predicted means for edu ***

          edu              1              2              3              4
5
                      -0.7985         0.6564         0.8843         1.0239
1.2964

Standard error of differences:      Average        0.3078
                                    Maximum        0.5624
                                    Minimum        0.1121

Average variance of differences:                   0.1338


*** Table of predicted means for sx.lsm ***

          lsm         1         2         3         4         5         6
7
           sx
            1         *         *    0.6276    0.8190    0.4758    0.6924
0.6671
            2         *         *    0.5905    0.0029    0.4795    0.6253
0.9519


          lsm         8
           sx
            1    0.6500
            2    0.7682

Standard error of differences:      Average        0.2757
                                    Maximum        0.5130
                                    Minimum        0.1475

Average variance of differences:                   0.08524

Standard error of differences for same level of factor:

                        sx         lsm
Average             0.2773      0.2452
Maximum             0.4377      0.5130
Minimum             0.1573      0.1475
```

```
Average variance of differences:
                 0.08529     0.07556


  35   stop
```

Appendix 2 SPSS output for MINQUE and ML

Factor Level Information

		N	Value Label
CL	1.00	4	
	2.00	4	
	3.00	9	
	4.00	9	
	5.00	8	
	6.00	9	
	7.00	8	
	8.00	11	
	9.00	10	
	10.00	8	
	11.00	10	
	12.00	11	
	13.00	8	
	14.00	9	
	15.00	9	
	16.00	10	
	17.00	8	
	18.00	9	
	19.00	10	
	20.00	9	
	21.00	10	
	22.00	8	
	23.00	4	
	24.00	8	
	25.00	8	
	26.00	8	
	27.00	8	
	28.00	8	
	29.00	8	
	30.00	8	
	31.00	4	
	32.00	8	
	33.00	16	
	34.00	8	
	35.00	8	
	36.00	8	
	37.00	8	
	38.00	8	
	39.00	16	
	40.00	8	
rc	1.00	157	black
	2.00	15	coloured
	3.00	24	indian
	4.00	147	white
sx	1.00	156	male
	2.00	187	female
age	1.00	48	less than 25yrs
	2.00	85	25 to 34 yrs
	3.00	88	35 to 44 yrs
	4.00	65	45 to 54

			yrs
	5.00	57	> 55 yrs
lsm	3.00	8	
	4.00	36	
	5.00	53	
	6.00	44	
	7.00	105	
	8.00	97	highest
mar	1.00	215	married
	2.00	32	ex married
	3.00	96	always single
ys	.00	232	non smoker
	1.00	111	smoker
edu	1.00	1	
	2.00	39	
	3.00	96	
	4.00	125	
	5.00	82	

Dependent Variable: LOGTOT

Variance Estimates

Component	Estimate
Var(CL)	.213
Var(Error)	.418

Dependent Variable: LOGTOT Method: Minimum Norm Quadratic Unbiased Estimation
(Weight = 1 for Random Effects and Residual)
a Weighted Analysis - Weighted by WT

Variance Estimates

Component	Estimate
Var(CL)	.194
Var(Error)	.387

Dependent Variable: LOGTOT Method: Maximum Likelihood Estimation
a Weighted Analysis - Weighted by WT