ANALYSIS OF LINEAR MIXED MODELS WITH AN EXTENSION TO THREE OR MORE FACTORS EACH HAVING BOTH FIXED AND RANDOM LEVELS

by

LYSON CHAKA

46287582

Submitted in accordance with the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

STATISTICS

at the

University of South Africa

(UNISA)

Supervisor: PROF. PETER NJUHO

08 January 2023

DECLARATION

Name: Lyson Chaka

Student Number: 46287582

Degree: PhD

ANALYSIS OF LINEAR MIXED MODELS WITH AN EXTENSION TO THREE OR MORE FACTORS EACH HAVING BOTH FIXED AND RANDOM LEVELS

I declare that the above thesis 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.

I further declare that I submitted the thesis to originality checking software and that it falls within the accepted requirements for originality.

I further declare that I have not previously submitted this work, or part of it, for examination at UNISA for another qualification or at any other higher education institution.

....   08 January 2023 ....

Signature   Date

FULL DISCLOSURE INFORMATION

(i) **Plagiarism Report**:

Turnitin was used for checking plagiarism on the final thesis, and the digital receipt is available (see Appendix E). The percentage of similarity was 28%, of which the first two greatest sources (similarity score of 9%) point to some of my published articles derived from this study. The rest of the sources had a percentage of 1% or less each.

(ii) **Ethical Clearance**:

The research study reported in this thesis received ethical clearance (ERC Reference No.: 2021/CSET/SOS/012) from UNISA's College of Science, Engineering, and Technology Research and Ethics Committee (see certificate in Appendix F). Therefore, the research was conducted following the methods and procedures outlined in the ethical clearance.

(iii) **Professional Language Editing**:

This thesis was professionally edited by Ms Lianne Hugo. The editing certificate is attached (see Appendix G).

(iv) **Literature Referencing**:

The Harvard referencing style for in-text citation and references was used throughout the thesis. The compilation of the reference list in publications and the final thesis was managed by the Zotero citation manager.

(v) **Peer-reviewed Journal Publications**:

The following publications are part of the work reported in this thesis.

- Chaka, L. and Njuho, P. (2022). Repeated-Measures Analysis in the Context of Heteroscedastic Error Terms with Factors Having Both Fixed and Random Levels. *Stats*, **5**(2), 458-476. https://doi.org/ 10.3390/stats5020027
- Chaka, L. and Njuho, P. (2021). Construction of a linear mixed model with each factor having both fixed and random levels: A case of split-split-plot structure in a RCBD. *International Journal of Agricultural and Statistical Sciences*, **17**(2): 501-518. https://connectjournals.com/03899.2021.17.501

(vi) **Conference Presentations**:

The following abstract presentations were part of the work reported in this thesis.

- Chaka, L. (2022). A general approach to linear mixed models with factors having both fixed and random levels in the presence of heteroscedastic error terms. *The 31st International Biometric Conference (IBC2022)*, Riga, Latvia (Accepted Abstract).
- Chaka, L. (2021). Construction and analysis of experimental designs with three factors each having both fixed and random levels. SASA2021, *The 62nd Annual Conference of the South African Statistical Association Conference (SASA2021)*, Stellenbosch University, South Africa. https://app.glueup.com/event/sasa2021-33624/

Acknowledgments

Firstly, I would like to thank God Almighty, for giving me the wisdom and strength to start and finish this long journey. This journey would not have been possible without the support I received from some noble individuals who stood by me. Secondly, my sincere appreciation goes to my Supervisor, Professor Peter Njuho, who has been the pillar of my success. His guidance, support and mentoring inspired me. The most valuable lesson he imparted on me is to accept disappointments and stay focused. As my supervisor, he is a true role model and reliable source of wisdom, encouragement and determination. I am greatly indebted to him for his patience with me, lovely corrections, wonderful experience and above all, the network of researchers in my field that I could access through his guidance. May you live long Professor Njuho!

I would like to thank both the Sol Plaatje University and the University of South Africa for financial, professional, and academic support that I received. I would not have been able to accomplish this academic journey without the valuable financial support for my studies and the numerous academic trainings and workshops in scholarly research.

My special thanks to the SAS Global family for their commitment to their interaction platforms for knowledge and skills sharing as well as data analysis tools. Many thanks to the unknown reviewers of my publications, to the South African Statistical Association (SASA2021) and International Biometric Conference (IBC2022) organisers for memorable events and platforms for knowledge sharing and collaborations.

Last but not least, I wish to thank my family for their unconditional love and prayers. Without these, I could have never traversed this far. I owe my wife, Beauty, my two daughters, Delight and Bliss, and my son, Hiswish, a great amount of gratitude for their unconditional love, support, and care during the time I could neither reciprocate the same, nor avail myself for the expected family duties during the study period. Your hard work and love have never gone unnoticed. I thank you!

List of Abbreviations

AIC	:	Akaike's Information Criteria
ANOVA	:	Analysis of variance
BLUE	:	Best linear unbiased estimator
BLUP	:	Best linear unbiased predictor
CRD	:	Completely randomised design
EBLUE	:	Empirical best linear unbiased estimator
EBLUP	:	Empirical best linear unbiased predictor
FELM	:	Fixed-effects linear model
GLMs	:	Generalised linear models
LME	:	Linear mixed effects
LMM	:	Linear mixed model
LRT	:	Likelihood ratio test
ML	:	Maximum likelihood
MLE	:	Maximum likelihood estimator
MINQUE	:	Minimum norm quadratic unbiased estimator
MIVQUE	:	Minimum variance quadratic unbiased estimator
MOM	:	Methods of moments
RCBD	:	Randomised complete block design
RELM	:	Random-effects linear model
REML	:	Restricted maximum likelihood estimator
RMD	:	Repeated measures design
SPD	:	Split-plot design
SSPD	:	Split-split plot design

Abstract

Studies or experiments involving three or more factors, each having both fixed and random levels, usually require the use of linear mixed models on treatments arranged in completely randomised design (CRD), randomised complete block design (RCBD), or any other design. These scenarios require analysts to be more accurate when measuring some of the effects of factors. When independent factors have a dichotomous composition in factor levels (either fixed or random effects) two issues that need careful attention immediately emerge: (i) the assumed linear mixed model under the partitioning approach will involve a design matrix that is either a full-rank or less-than-full-rank form, (ii) the approach leaves the partitioned data subsets vulnerable to outlier contamination, which might subsequently compromise the level of accuracy and precision of the selected partitioned models. Traditional statistical approaches have to be reoriented in order to extract all the variations in the data sufficiently. This study builds upon the partitioning approach by Njuho and Milliken (2005, 2009), and extends the concept to the case of contaminated linear mixed model estimation (Koller and Stahel, 2011), and the issue of characterising treatment effect variation (Dixon, 2016; Ding *et al.*, 2019) in various experimental designs that involve three or more factors, each having both fixed and random levels. The *robustlmm* package, available in the Comprehensive R Archive (CRAN), was used to robustly fit linear mixed models when considerably little outlier contamination exists in the data set. To circumvent the tedious process of creating partitions of experimental data based on targeted factor levels (data scrapping), a SAS code was proposed for generating partitioned and combined analyses. The partitioning approach effectively offered alternative ways of getting more accurate estimation and analysis of fixed effects and variance components in the case of a non-full-rank design matrix scenario by considering the partitioned and combined analysis of experimental data based on the targeted factor level combinations and the desired inference scope. The study confirmed that the partitioning approach is compatible with the use of robust estimation methods, which resulted in improved precision in the model estimates. In addition, the partitioning approach was considered for multi-stratum experimental designs where randomisation at different levels is necessary to achieve better model precision at different levels of the experiment such as the split-split-plot treatment structure, where all the three factors, each with both fixed and random levels, are laid in an RCBD. The essence of the approach was in manipulating the appropriate factor combinations in order to allow for narrow, intermediate and broad inferential space on the levels of each of the factors as well as their associated interactions. Furthermore, the approach proved to be useful beyond the fundamental consideration of homogeneous and uncorrelated error variance in the estimation process of linear mixed models. In essence, the study provides solutions for regaining the information that could be lost in various experimental designs if traditional analysis approaches are not improved.

Key Terms:

Linear mixed models; inference space; contaminated linear models; split-split-plot design; completely randomised design; randomised complete block design; repeated measures design; covariance structure; Kronecker product; robust estimation

CONTENTS

List of Tables

List of Figures

CHAPTER 1

INTRODUCTION

1.1 Motivation

Many agronomic and industrial experimentation involve two types of factors, i.e., treatment factors which are intentionally selected to address research questions, and other factors that do not directly relate to the main objectives of the experiment but have extraneous variations which impact the effects of the treatments (Jayalath and Ng, 2018; Stroup *et al.*, 2018). These experiments often involve factorial designs where two kinds of factor levels combine to influence the experimental results (Oliveira *et al.*, 2019). In any experimental design in which analysis of variance (ANOVA) is applied, fixed factors and random factors are technically defined in a different statistical manner (Janssen, 2012). According to Milliken and Johnson (2009), a factor is classified as fixed when all of its levels constitute the entire population of possible levels. On the other hand, when the levels of a factor are a random sample from a large population of possible levels, the factor is considered a random factor (Jayalath and Ng, 2018). In planned experiments involving fixed and/or random factors, the analysis of variance procedure plays a crucial role in the testing of significance of treatment effects. If the two types of factors are simultaneously used in an experiment, care must be taken to decide when to consider a factor effect as fixed or random (Yang, 2010).

When constructing a statistical model, the most crucial step begins by clearly defining effects as either fixed or random. Harrison *et al.* (2018) supported the idea that classification of independent factors as fixed or random effects is critical in model construction. The concept is also essential for studying the estimators' statistical properties and deciding which inference to take. However, the decision to designate an effect as either fixed or random is not an easy task (Yang, 2010). This decision is informed by how the factor levels are to be selected for the experiment. One criterion for distinguishing random and fixed effects used in modern linear model theory is to classify an effect as random if the effect levels are a subset of a population of effects; otherwise, the effect is fixed (Littell, *et al.*, 2010). Fixed effects are generally defined as fields whose factor levels represent a set of all possible levels of primary interest such that the inference is applied only to the entire set of levels in the study. Subsequently, a statistical model whose factors of primary interest in the treatment structure are assumed to be fixed is known as a fixed-effects model. A fixed-effects model exists when the entire set of treatment levels in the study are treated as known constants (Littell *et al.*, 2006), and are the only ones to which inference is to be made. In this case, the main objective of fitting a fixed-effects model is to estimate treatment means and establish if treatment differences exist (Stroup *et al.*, 2018).

A random-effects model exists when the treatment levels in the study are a small subset of a more extensive set of all possible treatment levels. The inferences about treatment means from a random-effects model are generalisable to all possible treatment levels. However, considering a factor with too few levels as random suffers the risk of obtaining a highly unreliable estimate of its variance and covariance parameters due to insufficient information to produce the precise estimates (Yang, 2010). Stroup and Mulitze (1991) further argued that if a factor has more than ten levels, it should be considered random, otherwise the most common approach is to consider the factor as fixed. The primary focus of a random-effects model, in addition to estimating variance components among random factor levels, is to estimate and to deduce some inference not only about specific levels in the model, but also the whole set of possible levels from which the sample was drawn. Random effects are useful for explaining excess variability in the target. Therefore, effects are defined as fixed if the main interest is in the effects themselves, or random if the interest extends to the entire population.

Subsequently, a mixed model combines both random and fixed effects in a single model. The technique of using both fixed and random effects simultaneously in a single analysis is what defines mixed modelling (Winter, 2013; Pusponegoro *et al.*, 2017; Seltman, 2018). Linear mixed models have the essential property of accommodating both fixed and random effects in a single model in order to estimate the treatment means and covariances of the data. To that effect, the extent to which a linear mixed model accurately predicts and fits the data depends on the appropriate classification of fixed and random effects in the model (Pan and Shang, 2018). Experimental research frequently deals with linear models in which the researcher is concerned with making comparisons among specific factors, or in some cases, extending the conclusions beyond the set of specific factors included in the study, thereby introducing the concept of random factors. One of the significant strengths of linear mixed models in statistical analysis is that they reduce statistical inference bias, which generally prevails when inappropriate models are used. According to Runcie and Crawford (2019), the inclusion of random effect terms in mixed models accounts for specific correlations among observations. However, fitting this mixed model requires estimating the importance of each random effect. There is a need to determine the sources of variation in a system rather than simply making specific comparisons in most cases. The flexibility of a mixed effects model is heplful for analysing data with more than one source of random variability as well as defining the covariance structure.

The development of Linear Mixed Models (LMMs) or Mixed Linear Models (MLMs) has its origin in animal breeding (Henderson, 1984), where hybrid factor analysis cases were solved in nested and crossed data structures. The extension of the concept to other disciplines involving industrial experiments (McLean *et al.*, 1991), longitudinal data (Molenberghs and Verbeke, 2000), plants, machinery and human subjects (Tan and Nott, 2014; Searle and Gruber, 2017) followed later with time. In the past decades, researchers have been struggling with how to analyse experimental data that contain mixed factors when the classical analysis of variance does not apply. Currently, significant contributions to how experimental designs are built and analysed when the factors involved are either fixed, random or mixed exist in literature (Harville, 1978; Ferreira *et al.*, 2017). The strength of the linear mixed modelling approach in such designs lies in their flexibility and ability to cater for a combined analysis of both fixed and random effects. Some improvements in statistical software such as the R and SAS has made data analyses using linear mixed models much easier and more flexible (Littell *et al.*, 2006) especially for non-independent, hierarchical, longitudinal, or correlated data. These types of data are crucial for decision-making purposes, among other things.

Matuschek *et al.* (2017) noted that linear mixed models (LMMs) have been, for the past decade, widely used in place of mixed model analyses of variance (ANOVAs) for statistical inferences in factorial designs. The application of LMMs has brought a number of advantages over ANOVAs. Some advantages, for example, include better preservation of statistical power (Baayen *et al.*, 2008) and replacing separate ANOVAs, which eliminates ambiguities of interpretation when effects are significant in only one of the default ANOVAs. However, setting up a linear mixed model is different from running an ANOVA model. Harrison *et al.* (2018) argued that certain types of data, such as in ecology, often have complex structures that require sophisticated linear mixed models and interpretation. Selection of a random structure that supports the data is needed to minimise type I error rate. Matuschek *et al.* (2017) strongly agreed that model selection is a highly active research field, which has considerably significant implications on type I error rate and significance tests in statistics.

Although linear mixed models are a powerful tool, their complexity cannot be underestimated (Harrison *et al.*, 2018). The application of linear mixed models (LMMs) in various disciplines which involve experimental or treatment designs, has received considerable attention in literature. However, it has been discovered that, in some experiments, factors under consideration might be sharing both fixed and random levels (Njuho and Milliken, 2005). As a result, analysis of variance in such linear mixed models requires consideration of both fixed and random intercepts as well as their slopes to guard against anti-conservative conclusions. When appropriately considered, random-effect structure is important because it provides a basis for an appropriate test of fixed effects, and a valuable source of information on the processes underlying the effect. While some researchers have made remarkable contributions to this kind of modelling, for example, one-factor and two-factor linear mixed models with factors that consist of a mixture of fixed and random levels (Njuho and Milliken, 2005, 2009), the need to explore beyond these less complicated models has become a desirable aspect.

The current Fourth Industrial Revolution, characterised by complex technological advancement, has propelled industries to transform their production and management systems and adopt new technological innovations. As new technologies are introduced where indigenous technologies have been in existence, or intervention strategies engaged to boost efficiency, experimental designs involving mixed treatment effects emerge. For example, an on-farm trial in agricultural experimentation might involve the comparison of effects of irrigation methods, types of fertiliser and paste control methods. With technological advancement and climatic changes constantly affecting the farming industry, each of these farming strategies is characterised by new inventions or improvements that are proposed for improved efficiency and production. Farmers are often found in a situation where they need to compare the class of new technologies against the several traditional ones that have been in existence before. Such cases are common in precision farming. This enables the farmer to decide whether they need to replace, upgrade, or maintain the old farming methods or combine some of the strategies with the newly invented ones. The approach requires one to conceptualise the factors involved as sharing both fixed (new strategies) and random (traditional strategies). In the production industry, where different indigenous equipment is in place, the adoption of new machinery might be necessary to improve production. In this case, the newly introduced technology or machinery is considered a fixed treatment level, while the existing indigenous equipment constitutes the random treatment levels. This study focuses on mixed modelling, a technique for combining fixed and random factors into one analysis, with a particular focus on modelling three or more factors, each consisting of both fixed and random levels.

1.2 Background

Analysis of variance is a widely used statistical technique built to conduct tests of significance of treatment effects for planned experiments. The concept of analysis of variance was introduced by Sir Ronald A. Fisher in the 1920's, when he was working on the estimation and analysis of correlations and error variances among and between relatives in agricultural experimental designs. The concept was developed further by Crump (1946) with an aim to incorporate random effects, and their variance component estimates in different varietal experiments. These variances can be presented as linear combinations of mean squares, contributing to the total variance estimate involved (Satterthwaite, 1946). Building from Arnold Fisher's work, an American statistician and a pioneer in animal breeding, Henderson (1953), introduced the best linear unbiased estimator (BLUE) of fixed effects and the best linear unbiased predictions (BLUP) for random effects (Robinson, 1991).

The study by Kackar and Harville (1984) portrays that standard errors associated with the estimation of fixed and random effects, in conjunction with the use of analysis of variance, are the basis of other approximations in mixed models. From the first industrial era (1951 to late 1970's) to the modern era, characterised by competitiveness and globalisation, ANOVA has been a key component in quality improvement procedures, research projects and other different types of study ranging from designed experiments, sample surveys, or observational studies. It has been a vital analysis tool for partitioning variability in experimental data into systematic fixed factors and/or random factors in statistical modelling.

An effort to express the response variable as the sum of the population parameters of the predictor variables is the basis of the formation of a linear model. Littell *et al.* (2006) define a statistical model as a function that connects the response variable to the predictor variables under an assumed probability distribution that characterises the random variation affecting the response variable. A statistical model is characterised by either fixed effects, random effects or both. Fitting linear mixed models to experimental data requires improved modelling techniques to specify factor effects (Smith and Edwards, 2017), which is important for estimation of treatment means, variance components, and drawing of inference.

There are a number of precautions that need to be considered when using mixed models to analyse experimental data. One such important aspect of mixed model analysis that, one has to be familiar with is the choice of inference space (McLean *et al.*, 1991; Yang, 2010). The scope of inference depends on how random effects are selected into a predictable function, which leads to three types of inference space: narrow, intermediate and broad inference scope. (Yang, 2010). Inference space is classified as "narrow" if inference is applied to certain specific random effects levels. In contrast, inference applied to the entire set of random effects, including random interaction effects, is classified as "broad" (McLean *et al.*, 1991). A class of inference scope is called "intermediate" when inference is narrow to specific random effects and broad to others. The accuracy of prediction and inference is determined by the selection of fixed and random effects in linear mixed models. Through manipulating the random effects involved in a mixed model, the researcher is flexible in deciding on the space to draw an inference. Inappropriate specification of the intended inference space results in biased parameter estimates, standard errors and predictions.

The technique of combining random and fixed effects into one analysis as a means to bypass some of the problems suffered by the classical analysis of variance (ANOVA) can be traced back to the 1980s. Literature shows that mixed models are becoming a prevalent feature in studies involving longitudinal data in medicine, public health, psychology, biology and other fields (Gad and EL-Zayat, 2018). Boisgontier and Cheval (2016) share the same sentiment that mixed models provide a better framework for using analysis of variances in current scientific studies. In order to apply mixed modelling in experimental research, there is a need to establish whether one is dealing with a design that contains factors that are entirely defined as either random, fixed, or mixed effects.

Recently, Njuho and Milliken (2005) presented a one-way treatment model structure where a factor is conceptualised as having both fixed and random effect levels. An example was given in a farming setup, where new farming strategies, such as newly developed technologies, are to be considered in determining the production efficiency. The concept was considered for a balanced one-way treatment structure (factor A, say), with treatments arranged in a completely randomised design. The proposed linear mixed model for this scenario is given by

$$y_{ij} = \mu + \tau_i + \epsilon_{ij}, \qquad i=1,\ldots,a \qquad j=1,\ldots,n, \tag{1.1}$$

where τ_1, τ_2, ..., τ_f ($f<a$) represent the fixed-effect levels and $\tau_{f+1}, \tau_{f+2}, ..., \tau_a$ ($a=f+r$) denote the random-effect levels of factor A, μ is the overall mean, and y_{ij} is the j^{th} observation receiving the i^{th} treatment, assuming that $\tau_{f+i} \sim$ i.i.d $N(0, \sigma_a^2)$, $\epsilon_{ij} \sim$ i.i.d $N(0, \sigma_\epsilon^2)$ and τ_{f+i} and ϵ_{ij} are pairwise independent, $i=1,2,\ldots,a$, $j=1,2,\ldots,n$.

In their study, Njuho and Milliken (2009) demontrated the application of mixed model analyses procedures used in on-farm trials and experiments which usually involve factors that have hybrid (fixed and random) effects. The concept was extended to a two-way treatment structure scenario where each factor (say A and B) is made up of fixed and random levels. Analogous to (1.1), the two-way linear mixed model involving factors A and B is expressed in its general form:

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}, \qquad i=1,2,...,a \qquad j=1,2,...,b \qquad k = 1,2,..., n. \tag{1.2}$$

where α_1, α_2, ..., α_f $(f < a)$ denote the fixed-effect levels of factor A and α_{f+1}, α_{f+2}, ..., α_a $(a = f_a + r_a)$ are the random-effect levels of factor A. Also β_1, β_2, ..., β_f $(f < b)$ represent the fixed-effect levels of factor B and β_{f+1}, β_{f+2}, ..., β_b $(b = f_b + r_b)$ are the random-effect levels of factor B. μ is the overall mean, and y_{ijk} is the kth observation receiving the ith treatment of factor A and the jth treatment of factor B. We assume that $\alpha_{f+i} \sim i.i.d\ N(0, \sigma_a^2)$; $\beta_{f+j} \sim i.i.d\ N(0, \sigma_b^2)$; $\epsilon_{ijk} \sim i.i.d\ N(0, \sigma_\epsilon^2)$ and α_{f+i}, β_{f+j} and ϵ_{ijk} are pairwise independent, $i = 1,2,...,f_a + 1,\ f_a + 2,...,a, \quad j = 1,2,...,f_b + 1,\ f_b + 2,...,b, \quad k=1,2,...,n.$

According to Njuho and Milliken (2009), on-farm trials and experiments often involve factors consisting of both fixed and random levels. Therefore, modelling these scenarios requires several approaches for estimating the fixed effects and variance components.

1.3 Justification

Experimental research common in literature involves factors that are regarded as either fully fixed or fully random with respect to their levels. According to Harrison *et al.* (2018), the ability to correctly specify model predictors as fixed or random effects is key to the modelling process. However, distinguishing between a fixed and random effect may be a challenging task. The way fixed and random effects are conceptualised in contemporary factorial designs has recently shifted (Njuho and Milliken, 2009). In the agricultural and production industries, the need to develop new strategies, methods and technologies to upgrade or substitute the indigenous ones in order to increase productivity and efficiency has inspired researchers to revisit how treatment factors are structured. The extension of the concept of hybrid factors (factors made up of a combination of fixed and random levels) is a phenomenon that has to be studied in depth. This kind of modelling allows for other factor interactions that could have been ignored in the linear model, yet they contribute to the most desired variation in treatment effects. The construction of this model results in the estimation of various effects, testing of homogeneity of variances and hypotheses of interest (Njuho and Milliken, 2005, 2009).

Based on this conceptual framework, this study aims to:

- expand the existing knowledge and techniques in the analysis of linear mixed models (LMMs) as applied in agriculture, industry and other research fields where new strategies, methods and technologies are tested and compared against the existing ones.
- promote the use of linear mixed models in complex scenarios involving factors that are made up of both fixed and random levels.
- serve as a stepping stone and a bridging gap to further research and similar applications of the analysis of variance techniques in linear models.

1.4 Research Problem

The structure and analysis of linear mixed models with factors that have both fixed and random levels are not straightforward concepts and, as such, can never be underestimated, especially in the realm of linear mixed model analysis. The complexity of mixed effects analysis of this nature has triggered the motivation to bridge the gap by further investigating the ways of accommodating and analysing linear mixed models with three or more factors, each consisting of both fixed and random levels.

1.5 Objectives

This research aims at unearthing the conceptualisation and procedural aspects involved when analysing linear mixed models which involve three or more factors, each having both fixed and random factor levels. This will be achieved by addressing the following objectives:

(i) construct models for various experimental designs involving three or more factors, each having both fixed and random levels,

(ii) estimate various fixed effects and provide an estimation of variance components from different experimental designs,

(iii) estimate approximate standard errors from the variance components of the estimable functions,

(iv) construct linear mixed models under the influence of considerably low outlier contamination, and assess its effects on model parameter estimates,

(v) construct linear mixed models, and test appropriate hypotheses in cases where the constant covariance structure of homoscedastic error terms is not appropriate,

(vi) generate simulation samples to establish consistency of the constructed models.

1.6 Layout of the Thesis

Chapter 1 Introduction: This chapter introduced the background, motivation and justification of the research problem. A brief overview of the important concepts and aspects that underpin the development of the new analytical concept was discussed. The bridging gaps in relation to the existing body of knowledge were identified and briefly discussed.

Chapter 2 Literature review: The chapter will present a review of well-known documented concepts about linear mixed models. A brief explanation of the theory of linear mixed models, their assumptions, estimation and approximation of degrees of freedom for standard errors, testing of hypotheses and their application in research and analysis of experimental designs is presented.

Chapter 3 Partitioning of factors in linear mixed models: An outline of the methodology and techniques used to construct a general linear mixed model is given. The main focus will be on developing linear mixed models with factors that consist of both fixed and random factor levels. The work done by Njuho and Milliken (2005, 2009) will be extended to linear mixed models with three or more factors in different treatment designs such as the completely randomised design (CRD), randomised complete block design (RCBD) and split-split-plot design (SPD). This includes defining the research designs, methodologies and testing of model assumptions, hypotheses testing, estimation of model parameters and variance components.

Chapter 4 Linear mixed models for contaminated data: Alternative approaches to assessing and detecting outlier contamination in linear mixed models are compared to the classical approach when relatively little contamination is permitted in the data. The application will be considered in experimental designs with treatment structures arranged in CRD and RCBD. Simulation samples are generated to validate the use of the proposed methods.

Chapter 5 Linear mixed models in split-split-plot design: We present the results of the article (Chaka and Njuho, 2021) published in the International Journal of Agricultural and Statistical Sciences. Statistical approaches to model construction, testing model assumptions and statistical analysis procedures for factorial experiments arranged in split-split-plot design under the condition that factors have fixed and random levels are discussed.

Chapter 6 The partitioning approach in repeated-measures design: We present the results of the article (Chaka and Njuho, 2022) published in the Stats Journal. The new approach is extended to linear mixed models in factorial experiments where the elements of the error vector are unequal and correlated. We consider a repeated measures design with multiple between-subjects factors where each of these factors has both fixed and random levels. We present the theories and methodologies that relate to the construction of linear mixed models, variance components and testing hypotheses, when the default covariance structure of homoscedastic error terms is inappropriate.

Chapter 7 Extention to multiple factors: In this chapter, the partitioning approach is extended to a general linear mixed model constructed from balanced or unbalanced data. Simple algorithms for constructing covariance matrices and sums of squares are suggested. The associated concepts of estimable functions, estimability, and variance-covariance structures are discussed.

Chapter 8 Conclusion, recommendations and future works: The chapter will present the conclusions derived from the research, specifying the major findings; the limitations of the study; and the suggested research gaps for future study that the current research could not fully explore.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

A review based on published articles about the theory and application of linear models, construction of linear mixed models (LMMs), estimation of fixed and random effects, inference space, estimation of standard errors, and hypothesis testing is presented. In addition, a brief review of the application of LMMs in research is included.

2.2 Developments in Linear Mixed Model (LMMs)

The concept of linear models develops from a basic linear regression model, whose regression coefficients (also called factors) are either fixed (fixed-effects model) or random (random-effects model) or a mixture of the two (mixed-effects model). The nature of factors involved in the model determines the editions of the basic linear regression model. These editions include, among many others, analysis of variance (ANOVA), analysis of covariance (ANCOVA), and linear mixed models (LMMs). Scientific research usually involves the construction of an appropriate statistical model that adequately characterises particular relationships or phenomena (Smith and Edwards, 2017).

Linear mixed models are perhaps the most popular class of models for statistical analysis, which includes analysis of variance (ANOVA) models of a broad spectrum of areas, such as multilevel, clustered data, repeated measures, and longitudinal data (Agostinelli and Yohai, 2016; Kuran and Özkale, 2020). These models are applicable to data that satisfy the normality assumption, making it possible to use of maximum likelihood (ML) principle in parameter estimation

(Agostinelli and Yohai, 2016). In other words, the distribution of both random effects and the within-subject measurement error is commonly assumed to be normal (Aghamohammadi and Meshkani, 2017). Due to its restrictive nature, the normality assumption usually succumbs to a lack of robustness against departures from the normal distribution. Confirming the normality assumption is not an easy task for random effects; hence its suitability becomes questionable (Ghidey *et al.*, 2004). Thus, more flexible and robust approaches which replace the normality assumption in linear mixed models have been proposed (Pinheiro *et al.*, 2001; Lachos *et al.*, 2012; Aghamohammadi and Meshkani, 2017). Furthermore, linear mixed models have a traditional assumption that fixed-effects variables are observed with neglible error; otherwise, the ordinary maximum likelihood estimators would become inconsistent (Yavarizadeh *et al.*, 2020). The use of the method of moments (Cui *et al.*, 2004), among other alternative approaches for estimation of fixed and random effects parameters, has been suggested (Zhong *et al.*, 2002; Zare *et al.*, 2012) for cases where the assumption is not satisfied.

Linear models are commonly used tools in the analysis of various experimental and scientific research in agriculture (Yang, 2010). With the advantage of increased computing power, more advanced versions of the linear models, such as generalised linear models (GLMs), mixed models and the Bayesian linear models, have taken centre stage in statistical research. Linear models are categorised into three classes: fixed, random and mixed effects models. Workable definitions of these types of linear models were given by Milliken and Johnson (2009).

2.2.1 Fixed-effects Linear Model (FELM)

Searle and Gruber (2017) define fixed effects as the treatments of an experiment upon which the primary focus is, and no others. In the single-factor model, a factor is characterised by either fixed or random effects. A general fixed-effect linear model (FELM) in an experiment without replications is given by:

$$y_{ij} = \mu + \alpha_i + \epsilon_{ij}, \tag{2.1}$$

where y_{ij} is the j^{th} response observation in the i^{th} level of the fixed explanatory factor, α_i is the average effects of the fixed factor involved, μ is the overall mean, and ϵ_{ij} is the random error term. The fixed-effects linear model (2.1) is a standard linear regression models if the factor levels α_i's are regarded as fixed observations of the predictor variable. Factor effects are called fixed-effects when the α_i's remain the same in each replication of the experiment. The general assumptions of model (2.1) are that the $\sum_{all\ i} \alpha_i = 0$ since the mean is μ over all the treatments, and the residuals (ϵ_{ij}) are a random sample from a normally distributed population of errors with mean 0 and variance σ^2. Under the fixed-effects analysis model, the main aim of the experiment is to estimate and compare the treatment means differences if they exist.

2.2.2 Random-effects Linear Model (RELM)

On the other hand, a random-effects model is analogous to (2.1), except that the treatment levels (α_i) are a random sample drawn from a larger population of treatments. In this case, the treatment effect is a random variable where for any given random sample of treatment levels, $\sum \alpha_i = \sum(\bar{Y_i} - \mu) \neq 0$, while the population of these treatment effects (α_i) has mean 0 and variance σ^2_α. The main objective of this model is to determine and test the presence of the additional variance component ($r\sigma^2_\tau$ for r replicates), and to estimate its magnitude. The conclusions can be extended to any scope of interpretation of random effects incorporated into the predictable function (i.e., the inference space), predetermined by the researcher (McLean *et al.*, 1991).

2.3 Linear Mixed Model Framework

The mixed model methodology can be traced back to the 1940s. A comprehensive illustration of the mixed model methodology was developed by Henderson (1984), which included the mixed model equations, their properties and procedures to estimate the fixed effects as well as the random effects. Basically, a linear mixed model or simply linear mixed model is an extension of a linear regression model generated from a continuous response variable being influenced by one or more factors considered as the predictor variables. A linear mixed model is capable of accounting for both the between-and within-subject variabilities (Aghamohammadi and Meshkani, 2017). The inclusion of random effects in linear models is a means to consider the within-subject correlation as well as the variability of the response among the different subjects (Martinez and Holian, 2014). Mixed model methodology was first introduced in animal breeding (Henderson, 1984), and then extended to other disciplines (Searle and Gruber, 2017). Several authors, notably Harville (1977, 1978), Robinson (1991), and McLean *et al.* (1991) contributed to building up and explaining the mixed model procedure further. This approach can be applied in a variety of experimental designs, such as the split-plot designs, location experiments, and many more (Stroup and Kachman, 1994).

As proposed by Anderson and McLean (2019), the linear mixed model is written in general form as

$$y_{ijkh} = \mu + F_i + F_j + R_k + (FF)_{ij} + (FR)_{ik} + (FR)_{jk} + (FFR)_{ijk} + \epsilon_{ijkh}, \tag{2.2}$$

where $i = 1, 2, ..., a$ and $j = 1, 2, ..., b$ are fixed factor levels of factor A and B, respectively; $k = 1, 2, ..., c$ are random factor levels of factor C; $k = 1, 2, ..., n_{ijk}$ replicates (all $n_{ijk} = n$ for balanced data). The assumptions that have dominated in literature on the model parameters are that: $\sum F_i = \sum FR_{ik} = 0$ over all i; $\sum F_j = \sum FR_{jk} = 0$ over all j; $R_k \sim N(0, \sigma^2_R)$; $FR_{ik} \sim N(0, \sigma^2_{FR})$; $FR_{jk} \sim N(0, \sigma^2_{FR})$; $FFR_{ijk} \sim N(0, \sigma^2_{FFR})$; and $\epsilon_{ijkh} \sim N(0, \sigma^2_\epsilon)$ with all random effects being pairwise independent. There have been varying opinions concerning the model assumptions employed in the analysis of mixed models. Wilk and Kempthorne (1955) strongly emphasised that whenever the analysis of variance is to be utilised in the interpretation of experimental data, its meaning and justification should transcend the set of arbitrary assumptions employed.

According to Smith and Edwards (2017), the general linear mixed model for a single response data is thus expressed in matrix form as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\epsilon}, \tag{2.3}$$

where $\mathbf{y}$: $n \times 1$ is the response vector (data) of observations; $\mathbf{X}$:$n{\times}p$ is a known fixed-effects design matrix which links $\boldsymbol{\beta}$ to $\mathbf{y}$; $\boldsymbol{\beta} : p \times 1$ is a vector of unknown fixed-effects parameters to be estimated; $\mathbf{Z}$: $n{\times}q$ is a known incidence matrix of random-effects parameters; $\mathbf{u} : q \times 1$ is a vector of unobservable random-effects parameters; $\boldsymbol{\epsilon}$ is a vector of independent and identically distributed Gaussian random errors. The term $\mathbf{X}\boldsymbol{\beta}$ is the fixed component of the mixed model, while $\mathbf{Zu}$ is the random component of the model.

We assume that $E(\mathbf{u}) = \mathbf{0}$, $E(\boldsymbol{\epsilon}) = \mathbf{0}$, $\mathbf{u}$ and $\boldsymbol{\epsilon}$ are uncorrelated, i.e $Cov(\mathbf{u}, \boldsymbol{\epsilon}) = 0$. It is also further assumed that $Var(\mathbf{u}) = \sigma^2\mathbf{G}$ and $Var(\boldsymbol{\epsilon}) = \sigma^2\mathbf{R}$, where σ^2 is an unknown positive scalar while $\mathbf{G}$ (the variance-covariance matrix of the random effects in $\mathbf{u}$) and $\mathbf{R}$ (the variance-covariance matrix of the random error terms in $\boldsymbol{\epsilon}$) are known nonsingular matrices. We have,

$$E\begin{bmatrix}\mathbf{u}\\ \boldsymbol{\epsilon}\end{bmatrix} = \begin{bmatrix}\mathbf{0}\\ \mathbf{0}\end{bmatrix}$$

and

$$Var\begin{bmatrix}\mathbf{u}\\ \boldsymbol{\epsilon}\end{bmatrix} = \begin{bmatrix}\mathbf{G} & \mathbf{0}\\ \mathbf{0} & \mathbf{R}\end{bmatrix}\sigma^2,$$

where $\mathbf{G}$ and $\mathbf{R}$ are known positive definite matrices which depend on some vector of dispersion parameters $\boldsymbol{\theta}$, and σ^2 is a positive constant (Robinson, 1991). The elements of matrix $\mathbf{X}$ are usually known, but the elements of $\mathbf{G}$, $\mathbf{R}$ may be functions of an $m \times 1$ vector $\boldsymbol{\theta} = (\theta_1, ..., \theta_m)'$ of unknown parameters. The parameter space for $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ is taken to be $\{\beta, \theta : \beta \in E^p, \theta \in \Omega\}$, where Ω is some given subset of Euclidean space E^m (Harville, 1977). With, the rank($\mathbf{X}$) denoted by p^*, where $\mathbf{X} : n \times p^*$ is a matrix whose any p^* columns are linearly independent, an unbiased estimator of σ^2 as proposed by Harville (1976) is given by:

$$\begin{aligned}\hat{\sigma}^2 &= (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{R} + \mathbf{ZGZ}')^{-1}(\mathbf{y} - \mathbf{X}\hat{\beta})/(n - p^*)\\ &= \mathbf{y}'\mathbf{R}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{Z}\hat{\mathbf{u}})/(n - p^*).\end{aligned}$$

For the given mixed model whose structure of $\mathbf{Z}$, $\mathbf{G}$ and $\mathbf{R}$ is specified, the variance of $\mathbf{y}$ is given by $\mathbf{V} = Var(\mathbf{y}) = \mathbf{ZGZ}' + \mathbf{R}$. Depending on the specified structure of $\mathbf{Z}$, $\mathbf{G}$ and $\mathbf{R}$, different models of variance-covariance of the data emerge. The structure of the covariance matrices $\mathbf{G}$ and $\mathbf{R}$ depends on the assumptions made on them, which define the subsequent model for the variance-covariance of the data. For instance, one can specify $\mathbf{Z}$ as a matrix of dummy variables, whereas $\mathbf{G}$ is a diagonal matrix of variance components. As indicated by Laird and Ware (1982), a simple case is that the model (2.3) can be simplified to a "conditional-independence model" when $\mathbf{R}= \sigma^2 I$ since the responses, $\mathbf{y}$, are conditional on $\boldsymbol{\beta}$ and $\mathbf{u}$.

2.3.1 Heterogeneity Linear Mixed Models

LMMs are developed on the fundamental normality and homogeneous error variance assumptions of the random effects among all the subjects (Martinez and Holian, 2014). The presence of heterogeneity causes variance components to have a bias on the estimation of the whole linear mixed model, including the fixed effects (Du and Wang, 2020). Several approaches to relax the strong normality and/or variance homogeneity assumption documented in literature include the extension of the classical LMM by allowing sampling of random effects from a finite mixture of normal distributions with a common covariance matrix (Martinez and Holian, 2014), parametric bootstrap approach (Xu *et al.*, 2015), computational approach test (CAT) for one-way analysis of variance (Mutlu *et al.* 2017), replacing the traditional least squares (LS) estimators with the modified maximum likelihood (MML) estimators in the test statistics (Güven *et al.*, 2019), and nonparametric analysis of variance methods (Luepsen, 2018), among others.

2.3.2 Estimation of Fixed and Random Effects Parameters

A powerful technique for estimating random effects (also known as predictors) in mixed models, known by Harville as the Best Linear Unbiased Prediction (BLUP), exists. According to Robinson (1991), the BLUP estimates are known as the "best" from the fact that they have minimum mean squared error within the class of linear unbiased estimators; "linear" in the sense that the estimates of the realised random variables, $\mathbf{u}$, are linear functions of the response data, $\mathbf{y}$; "unbiased" because the average value of the estimate is equal to the average value of the quantity being estimated; and the estimates are called "predictors" to differentiate them from the fixed effects estimators.

The model (2.3) has expected value of $\mathbf{y}$ estimated as

$$\begin{aligned} E(\mathbf{y}) &= E(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\epsilon}) \\ &= \mathbf{X}\hat{\boldsymbol{\beta}} \end{aligned}$$

and variance of y estimated by

$$\begin{aligned} Var(\mathbf{y}) &= Var(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\epsilon}) \\ &= Var(\mathbf{Z}\mathbf{u} + \boldsymbol{\epsilon}) \\ &= \mathbf{Z}\mathbf{G}\mathbf{Z}^T + \mathbf{R} \\ &= \mathbf{V} \end{aligned}$$

Assuming that $(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})$ is nonsingular, $(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}$ exists, and that $\hat{\boldsymbol{\beta}}$ is estimable, the maximum likelihood estimate of $\boldsymbol{\beta}$, $(\hat{\boldsymbol{\beta}})$, is a solution to the equation

$$\begin{aligned} \mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}\hat{\boldsymbol{\beta}} &= \mathbf{X}^T\mathbf{V}^{-1}\mathbf{y} \\ \mathbf{X}^T(\mathbf{Z}\mathbf{G}\mathbf{Z}^T + \mathbf{R})^{-1}\mathbf{X}\hat{\boldsymbol{\beta}} &= \mathbf{X}^T(\mathbf{Z}\mathbf{G}\mathbf{Z}^T + \mathbf{R})^{-1}\mathbf{y} \\ \hat{\boldsymbol{\beta}} &= [\mathbf{X}^T(\mathbf{Z}\mathbf{G}\mathbf{Z}^T + \mathbf{R})^{-1}\mathbf{X}]^{-1}\mathbf{X}^T(\mathbf{Z}\mathbf{G}\mathbf{Z}^T + \mathbf{R})^{-1}\mathbf{y} \end{aligned}$$

Therefore, the MLE of $\boldsymbol{\beta}$ is given by

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y}. \tag{2.4}$$

One of the main objectives of mixed model methodology is to estimate the fixed unknown parameters $\hat{\boldsymbol{\beta}}$ and predict the random variables $\mathbf{u}$ in (2.3). When matrices $\mathbf{G}$ and $\mathbf{R}$ are known, then the best linear unbiased estimator (BLUE) $\hat{\boldsymbol{\beta}}$ and the best linear unbiased predictor (BLUP) are obtained from (2.3). The estimates of random effects (BLUP) in this model are derived from the fixed effects estimator (BLUE) by expressing $\hat{\mathbf{u}}$ in the form (Henderson, 1953):

$$\hat{\mathbf{u}} = \mathbf{C}^T(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}), \tag{2.5}$$

where $\hat{\mathbf{u}} : q \times 1$ is a vector corresponding to unobservable random effects, and $\mathbf{C}$ is an $N \times q$ matrix. Using the fact that $\mathbf{V} = \mathbf{Z}\mathbf{G}\mathbf{Z}^T + \mathbf{R}$, and covariance between $Cov(\mathbf{y}, \mathbf{u}) = \mathbf{Z}\mathbf{G}$, then

$$\begin{aligned} \mathbf{V}\mathbf{C} &= \mathbf{Z}\mathbf{G}; \\ \mathbf{C} &= \mathbf{V}^{-1}\mathbf{Z}\mathbf{G}. \end{aligned}$$

Thus, the estimator $\hat{\mathbf{u}}$ in (2.5) becomes

$$\hat{\mathbf{u}} = \mathbf{G}\mathbf{Z}^T\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}). \tag{2.6}$$

Assuming $\mathbf{u}$ as fixed in (2.3), an easier alternative approach for obtaining the estimates $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{u}}$, exists (see Appendix A).

In practice, the matrix $\mathbf{V}$ is usually very large and non-diagonal, making it difficult to obtain its inverse $\mathbf{V}^{-1}$. Henderson's (1953) approach is computationally simpler due to the fact that neither matrix $\mathbf{V}$ nor its inverse $\mathbf{V}^{-1}$ is needed. The only required pieces of information are matrix R (usually identity) and matrix $\mathbf{G}$ (often diagonal). Subsequently, matrix $\mathbf{V}$ is either diagonal or has a large diagonal sub-matrix.

The estimates $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{u}}$ are known as the mixed model solutions for the fixed and random effects, respectively. The BLUE of fixed effects ($\hat{\boldsymbol{\beta}}$) is unbiased since its expectation is $\boldsymbol{\beta}$. The BLUP of random effects ($\hat{\mathbf{u}}$), also known as shrinkage estimator (Harville, 1977), has the same property, with an additional tendency of shrinking the fixed-effects estimates of $\mathbf{u}$ towards zero (McLean *et al.*, 1991). Through matrix operations, (2.4) and (2.6) have been proven to be the same as (see 8.1 in Appendix A); hence we express the BLUE and BLUP as,

$$\begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} (\mathbf{X}'V^{-1}\mathbf{X})^{-1}\mathbf{X}'V^{-1}\mathbf{y} \\ \mathbf{G}\mathbf{Z}'V^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \end{bmatrix}. \tag{2.7}$$

As shown by Henderson (1975), provided matrix $\mathbf{X}$ is of full rank, p, the BLUP estimates in equation (2.7) have an error variance-covariance matrix given by:

$$Var\begin{bmatrix} \hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \\ \hat{\mathbf{u}} - \mathbf{u} \end{bmatrix} = \begin{bmatrix} \mathbf{X}^T\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}^T\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}^T\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}^T\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1} \end{bmatrix}^{-1} \sigma^2. \tag{2.8}$$

As the $\mathbf{G}^{-1}$ tends to a zero matrix, $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{u}}$ from the mixed model equations give the same result as the one from the generalised least squares solutions when the random components $\mathbf{u}$ are considered as fixed effects (Robinson, 1991). Provided matrices $\mathbf{G}$ and $\mathbf{R}$ are nonsingular, the BLUE and BLUP solutions of equation (2.7) can be easily estimated for problems with both small and large numbers of observations using computer software. There are cases where $\mathbf{G}$ and $\mathbf{R}$ are known diagonal matrices, or a combination of $\mathbf{G}$ and $\mathbf{R}$ being block-diagonals which are easily invertible (McLean *et al.*, 1991).

In practice, $\mathbf{G}$ and $\mathbf{R}$ are not always known. Harville (1976) and Dempster *et al.* (1981) proposed the empirical Bayes formulation of the mixed model as an alternative approach to derive the fixed effects $\boldsymbol{\beta}$ and random effects $\mathbf{u}$. The approach involves obtaining the estimates, $\hat{\mathbf{G}}$ and $\hat{\mathbf{R}}$, and use them in equation (2.9) to obtain empirical BLUE (EBLUE) of $\boldsymbol{\beta}$ and empirical BLUP (EBLUP) of $\mathbf{u}$. Thus, the estimation of covariance parameters is usually done before the BLUE and BLUP estimates, $\boldsymbol{\beta}$ and $\mathbf{u}$, respectively. Various methods of estimating the covariance parameters exist (Keele *et al.*, 1991), but the restricted maximum likelihood (REML) is the most preferred. Henderson (1984) suggested some procedures which can be used when these variance-covariance matrices are unknown. Most statistical software have provision for estimating the matrices $\mathbf{G}$ and $\mathbf{R}$ using the method of moments or, more precisely, the restricted maximum likelihood (REML) (Stroup and Kachman, 1994). In the case where $\mathbf{G}$ and $\mathbf{R}$ matrices are singular, and with the assumption of $\mathbf{u}$ and $\boldsymbol{\epsilon}$ described earlier on, a generalised inverse (g-inverse) matrix denoted by $\mathbf{C}$ below can be used in (8.1) (Harville, 1976; Henderson, 1984).

$$\mathbf{C} = \begin{bmatrix} \mathbf{X}^T\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}^T\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}^T\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}^T\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1} \end{bmatrix}^{-} \tag{2.9}$$

McLean *et al.* (1991) developed an approach to handle mixed linear models based on Goldberger's (1962), Henderson's (1953, 1984) and Harville's (1976) approaches in various experimental designs, error structures, balanced and unbalanced data. One of the limitations of the least squares method when estimating the BLUE is that the inverse of $Var(\mathbf{y})$ is usually non-diagonal. The problem can be averted by computing the g-inverse (2.9) and applying the mixed model procedures (McLean *et al.*, 1991). Laird and Ware (1982) concurred that, from the marginal likelihood of $\mathbf{y}$ after integrating out $\boldsymbol{\beta}$ and $\mathbf{u}$, the introduction of a flat prior to the location parameters, $\boldsymbol{\beta}$, and the estimate $\boldsymbol{\theta}$ yields the restricted maximum likelihood (REML) estimates of $\boldsymbol{\theta}$. Hence, the estimated means of the posterior distribution are known as the empirical Bayes estimates of $\boldsymbol{\beta}$ and $\mathbf{u}$. McLean *et al.* (1991) demonstrated that, with appropriate estimates of variance components, the model (2.3) is a useful tool for either balanced or unbalanced data when expressed in the form

$$\mathbf{y} = (\mathbf{X}\ \mathbf{Z})(\mathbf{B}\ \mathbf{U})' + \boldsymbol{\epsilon}, \tag{2.10}$$

where the order of $\mathbf{X}$:$n{\times}p$ and $\mathbf{Z}$:$n{\times}q$ are determined by the number of effects in the ($\mathbf{B}\ \mathbf{U}$) vector, $\mathbf{B}$: $p \times 1$ is a vector of unknown parameters for the fixed effects to be estimated, $\mathbf{U}$:$q \times 1$ is a vector of unobservable random effects, and $\boldsymbol{\epsilon} : n \times 1$ is a vector of errors.

2.3.3 Estimation of Variance Components

In experimental designs, the idea of understanding variability and experimental error is of paramount importance. Identifying and understanding the different sources of experimental errors is equally important. These errors can be peculiar to the environment, personal circumstances, or as a result of measurement tool applied (Robinson, 1987). The idea of estimating error variation from different sources has been known in different expressions for a long time. For example, Fisher (1925) termed it "components of variation", some authors call it "error components", while other authors refer to it as "component of variance". The estimation of the variance components for random effects has been widely documented (Harville, 1977). With known variances and covariances of random effects, or at least their estimates, mixed model analysis of fixed effects is achievable (Yang, 2010). The total variance-covariance matrix, $\mathbf{V} = \mathbf{ZGZ}' + \mathbf{R}$, consists of a variance-covariance matrix component for the random effects $\mathbf{u}$ where $var(\mathbf{u}) = \mathbf{G}$, and the variance-covariance matrix for the random residuals $\boldsymbol{\epsilon}$ with $var(\boldsymbol{\epsilon}) = \mathbf{R}$.

Numerous methods of obtaining estimates of variance components exist. According to Laird and Ware (1982), most of the approaches to the estimation of variance components in literature are in the context of analysis of variance (ANOVA). In the 1920s, Sir R.A. Fisher developed a basic procedure for estimating the error variance by equating the mean square for error (MSE) to its expected value, E(MSE). Thus we have,

$$E(MSE) = \sigma_e^2, \tag{2.11}$$

which subsequently gives, $\hat{\sigma}_e^2 = MSE$. Fisher's traditional analysis of variance (ANOVA) procedure was initially designed for fixed effects models, for which F-statistics (named by Snedecor in honour of Fisher) are used to test the factor effects (Searle and Gruber, 2017). The approach was fashioned to identify a single source of variation, the error variance, which amounts to the total error variation (Robinson, 1987). Some important features about the ANOVA procedure are that:

- the estimators for the variance components are unbiased regardless of whether the data are normally distributed;
- estimates of variance components require that the data set be balanced, i.e. having equal cell sizes, or unbalanced, provided the data is classified by only one factor.

The foundational approach for estimating variance components is usually by equating observed mean squares statistics to the expressions which describe their expected values (Satterthwaite, 1946). In the analysis of variance, estimates of variance components parameters are found by solving a linear function of the random effects connecting the mean squares in the ANOVA table to their expected values (Harville, 1977). The average value of any other mean square is a linear function of the mean squares within subclasses (σ_ϵ^2) and the other variances (Crump, 1946). The ANOVA-based estimates of variance components require that cell sizes be well balanced, which is often not the case in field studies. ANOVA method provides an integrative approach to variance effect parameter estimation. However, the ordinary ANOVA estimates of variance components are generally biased and can be negative, even though, by their definition, variances must be greater or equal to zero.

Henderson (1953) extended the technique to cater for unbalanced data. These approaches are popularly known as Henderson's methods I, II and III. The methods are basically alternative ways of using the traditional ANOVA approach. The difference in the methods is found on the quadratics that are analogous to the sum of squares used in the linearly independent quadratic forms of observations. Other modifications to the traditional ANOVA have been suggested to cater for unbalanced data (Henderson, 1953; Searle and Gruber, 2017) and find unbiased estimates of variance components. Robust approaches, such as the maximum likelihood (ML) and restricted maximum likelihood (REML) estimators (Hartley and Rao, 1967; Dempster *et al.*, 1977), which do not demand balancedness in the data were introduced. When data are balanced, these estimators produce the same estimates of variance components as those obtained from solving appropriate linear functions of mean squares in ANOVA.

The traditional analysis of variance procedure, sometimes known as the conventional least squares approach, was extended to cater for variance components models (random effects and/or mixed effects models), which involve random variables and more than one variance. It is the total sum of these variances, resulting in the variance of the response variable, that is called variance components. The purpose of finding estimates of variance components in mixed models is to estimate the contribution of each random effect to the variance of the dependent variable and determine where to concentrate in order to reduce the variance. Robinson (1987) highlighted that variance component techniques are useful for three reasons:

- they provide information about the experimental material and variances used in optimisation, programme selection and system evaluation,
- analysis of individual experiments for effectiveness,
- combined information can be extracted from several different experiments or trials.

Alternatives to the traditional ANOVA estimation of variance components include the minimum norm quadratic unbiased estimation (MINQUE; La Motte, 1971; Rao, 1971) which depends on some pre-assigned values through extensive algebra to estimate variance components, and the Minimum Variance Quadratic Unbiased Estimators (MIVQUE; La Motte, 1971; Rao, 1971), which assumes no assumption about the form of distribution of the response variable, are relatively suitable for unbalanced data when normality is not assumed. If normality is assumed, MINQUE solution bears the same properties as the MIVQUE solution. The maximum likelihood (ML) method (Hartley and Rao, 1967), which is based on maximising the likelihood function to yield ML estimators of fixed effects and variance components, and the restricted maximum likelihood (REML) method (Patterson and Thompson, 1971), which provides statistics for the variance components based on maximising only the part of the likelihood which is invariant to the location (fixed effects) parameter are preferred alternatives over the ANOVA approach. The other most widely used estimates of variance components listed by Keele *et al.* (1991) include the symmetric differences squared (SDS; Grimes and Harvey, 1980); weighted SDS (Christian, 1980; Keele and Harvey, 1989); pseudo expectation approach (PE; Schaeffer, 1986); tilde-hat approach (TH; van Raden and Jung, 1988).

Maximum likelihood (ML) approach for estimating variance components was first suggested by Crump (1946), and further refined by Hartley and Rao (1967). Although the ML approach is technically tedious, it is conceptually simple since it does not require assumptions concerning the structure or balance of data. The beauty of the ML approach is that the estimates of variance components are easily obtained together with the approximate standard errors. One of the major drawbacks of the ML approach is that all fixed effects are assumed to be known without error, resulting in biased estimates of variance components. According to Patterson and Thompson (1971, 1975) and Harville (1977), ML estimators are generally biased downwards, a problem that can be solved by making use of residual maximum likelihood (REML), which Patterson and Thompson (1971) formally described as a Maximum-likelihood (ML) method that accounts for the loss of degrees of freedom due to fitting fixed effects. Hence, RELM is referred to as a restricted version of ML due to its ability to eliminate bias as it maximises only the portion of likelihood that does not depend on the fixed effects. This makes REML the preferred method for analysing large data sets with complex structures. Both the ML and REML methods aim to find the set of parameters which maximises the likelihood of the data for a given model. In the case of a completely balanced design, the traditional ANOVA and REML yield the same estimates of variance components.

Most computer statistical packages provide these estimates of variance with ease. For example, the *PROC MIXED* statement in SAS provides six options of estimators of **G** and **R** through the METHOD option. The first three methods of moments (MOM) estimators, which are calculated based on Type I, II and III sums of squares, are designated in *PROC GLM* as TYPE 1, TYPE 2 and TYPE 3, respectively. In addition, the fourth methods-of-moments estimator is known as the minimum variance quadratic unbiased (MIVQUE) estimator (Rao, 1971; Swallow and Searle, 1978), which is used when data is unbalanced. Provided the random effects and errors are uncorrelated, MOM estimators are unbiased, and can be easily derived by equating the observed mean squares to their expected values. The other two are maximum-likelihood estimators, the restricted maximum likelihood (REML) (Patterson and Thompson, 1971) and maximum likelihood (ML) estimators (Hartley and Rao, 1967), which are obtained by maximising the likelihood function over the parameter space. Due to the availability of computing facilities, maximum-likelihood (ML) methods are widely used, despite the fact that methodology is mathematically complex (Ferreira *et al.*, 2017).

The major advantages of the MOM estimators over the likelihood-based estimators are that the former are computationally less demanding and can be obtained with no distributional assumptions, whereas the normality assumption is required for the latter. However, research shows that the REML procedure is generally preferred, particularly with unbalanced data (Searle and Gruber, 2017). Provided the solution to the MOM estimates are positive and pairwise correlations range from -1 to +1, all the likelihood-based estimators and the MOM estimators are identical for balanced data sets (Yang, 2010).

2.4 Kronecker Products in Linear Mixed Models

Kronecker products of identity matrices and matrices with all elements equal to 1 have a wide application in most linear models where sums of squares and covariance matrices are needed for the analysis of variance. Design matrices in linear models and mixed linear models are usually expressed using several submatrices consisting of Kronecker products, which are subsequently used in expressing sums of squares as matrix quadratic form. Some algebraic and computational algorithms for constructing these Kronecker products of matrices exist in literature. These include cases such as the one-way and two-way balanced analysis of variance (ANOVA) model (Saw, 1992), the balanced and unbalanced two-way (ANOVA) model (Rogers, 1984), the k-factor classification model (Takemura, 1983; Sunwoo, 1996), and the mixed model (Moser and Sawyer, 1998). We provide the formal definition of Kronecker product and its application in some of the experimental designs.

The right Kronecker (direct or tensor) product of matrix $\mathbf{A} : m \times n$ and matrix $\mathbf{B} : p \times q$, denoted by $\mathbf{A} \otimes \mathbf{B} : mp \times qn$, where $\otimes$ denotes the Kronecker matrix product, is defined by an $mp \times qn$ block matrix

$$\begin{aligned}\mathbf{A} \otimes \mathbf{B} &= [a_{ij}\mathbf{B}] \\ &= \begin{bmatrix} a_{11}\mathbf{B} & \cdots & a_{1n}\mathbf{B} \\ a_{21}\mathbf{B} & \cdots & a_{2n}\mathbf{B} \\ \vdots & \ddots & \vdots \\ a_{m1}\mathbf{B} & \cdots & a_{mn}\mathbf{B} \end{bmatrix},\end{aligned}$$

where each submatrix is a result of scalar multiplication of an element of A with the matrix B.

The Kronecker product, named after the German mathematician Leopold Kronecker (1827-1891), has several properties related to vector operators, matrix products, determinants, trace, rank, and polynomial matrix products (Graybill, 1983; Rogers, 1984; Zhang and Ding, 2013). Important theorems and proofs of Kronecker products are provided in these and many other sources.

Linear mixed models conform with the linear matrix equation theory where the Kronecker product plays an essential role (Zhang and Ding, 2013). The linear model (2.10) compressed in matrix form becomes

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{2.12}$$

where $\mathbf{y} = (\mathbf{y}'_1 ... \mathbf{y}'_N)'$, the coefficient matrix $\boldsymbol{\beta} = (\mu, \boldsymbol{\alpha}'_i, \boldsymbol{\beta}'_j, \boldsymbol{\gamma}'_k, ..., (\boldsymbol{\alpha\beta\gamma}'_{ijk}))'$ and the error matrix $\boldsymbol{\epsilon} = (\epsilon'_1 ... \epsilon'_N)'$. The incidence matrix $\mathbf{X} = (\mathbf{X}'_1 ... \mathbf{X}'_N)'$, where the incidence matrix, $\mathbf{X}_i$, corresponding to either the main effects (A, B or C) or interaction effects (AB, AC, BC, or ABC) in the model, can be partitioned into several submatrices, which consist of Kronecker products of matrices of 1's and identity matrices (Sunwoo, 1996).

For example, Sunwoo (1996) considered a general k-factor classification balanced model with r replications per cell, having n_i levels for each factor i. Expanding model (2.12) and expressing it in matrix form:

$$\mathbf{y} = \mathbf{1}_N \mu + \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + ... + \mathbf{X}_m\boldsymbol{\beta}_m + \boldsymbol{\epsilon}, \tag{2.13}$$

where μ is the overall mean, N is the total number of observations, $\boldsymbol{\beta}_j$ $(j = 1, 2, ..., m)$ is either a main effect or interaction effect. The incidence matrix $\mathbf{X}_j$ in a k-factor classification model is made up of Kronecker products of $(k+1)$ matrices $\mathbf{Q}_i$ $(i = 1, 2, ..., k)$ which are either $\mathbf{I}_{n_i}$ $(n_i \times n_i$ identity matrix) or $\mathbf{1}_{n_i}$ (a vector with all n_i components equal to 1), provided the model (2.13) is balanced. That is, $\mathbf{X}_j = \mathbf{Q}_1 \otimes ... \otimes \mathbf{Q}_k \otimes \mathbf{1}_r$, where $\mathbf{Q}_i = \mathbf{I}_{n_i}$ if the index i corresponds to the i^{th} factor in the model, or $\mathbf{Q}_i = \mathbf{1}_{n_i}$ if not. The last submatrix, $\mathbf{1}_r$ (a vector with all r components equal to 1), represents the replications per cell in the case of a balanced model.

Assume the three-factor linear model (2.13) has a levels of factor A, b levels of factor B, and c levels of factor C, with full interaction. The linear model can be expressed in matrix form as (Sunwoo, 1996)

$$\mathbf{y} = \mathbf{1}_N\mu + \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \mathbf{X}_3\boldsymbol{\beta}_3 + \mathbf{X}_4\boldsymbol{\beta}_4 + \mathbf{X}_5\boldsymbol{\beta}_5 + \mathbf{X}_6\boldsymbol{\beta}_6 + \mathbf{X}_7\boldsymbol{\beta}_7 + \boldsymbol{\epsilon}, \tag{2.14}$$

where μ is the overall mean, N is the total number of observations, $\boldsymbol{\beta}_m$ $(m = 1, 2, ..., 7)$ is either a main effect or interaction effect. The matrix $\mathbf{X}_m$ $(m = 1, 2, 3)$ is an incidence matrix corresponding to the main effects (A, B or C), respectively, and matrix $\mathbf{X}_m$ $(i = 4, ..., 7)$ is an incidence matrix corresponding to the interaction effects AB, AC, BC, or ABC, respectively. Each incidence matrix $\mathbf{X}_m$ is made up of Kronecker products of $(3+1)$ matrices $\mathbf{Q}_i$ $(i = 1, 2, 3)$ which are either $\mathbf{I}_{n_i}$ or $\mathbf{1}_{n_i}$. If we assume the model is balanced, with r replications per cell, then, the Kronecker products of the intercept, main and interaction effect incidence submatrices will be given by

$$\begin{aligned}
\mathbf{1}_N &= \mathbf{1}_a \otimes \mathbf{1}_b \otimes \mathbf{1}_c \otimes \mathbf{1}_r = \mathbf{J}_N,\\
\mathbf{X}_1 &= \mathbf{I}_a \otimes \mathbf{1}_b \otimes \mathbf{1}_c \otimes \mathbf{1}_r = \mathbf{D}_a(\mathbf{J}_r),\\
\mathbf{X}_2 &= \mathbf{1}_a \otimes \mathbf{I}_b \otimes \mathbf{1}_c \otimes \mathbf{1}_r = \mathbf{D}_b(\mathbf{J}_r),\\
\mathbf{X}_3 &= \mathbf{1}_a \otimes \mathbf{1}_b \otimes \mathbf{I}_c \otimes \mathbf{1}_r = \mathbf{D}_c(\mathbf{J}_r),\\
\mathbf{X}_4 &= \mathbf{I}_a \otimes \mathbf{I}_b \otimes \mathbf{1}_c \otimes \mathbf{1}_r = \mathbf{D}_{ab}(\mathbf{J}_r),\\
\mathbf{X}_5 &= \mathbf{I}_a \otimes \mathbf{1}_b \otimes \mathbf{I}_c \otimes \mathbf{1}_r = \mathbf{D}_{ac}(\mathbf{J}_r),\\
\mathbf{X}_6 &= \mathbf{1}_a \otimes \mathbf{I}_b \otimes \mathbf{I}_c \otimes \mathbf{1}_r = \mathbf{D}_{bc}(\mathbf{J}_r),\\
\mathbf{X}_7 &= \mathbf{I}_a \otimes \mathbf{I}_b \otimes \mathbf{I}_c \otimes \mathbf{1}_r = \mathbf{D}_{abc}(\mathbf{J}_r),
\end{aligned} \tag{2.15}$$

where $\mathbf{D}_a(\mathbf{J}_r)$ is a diagonal matrix of a column vectors of ones, each of length r (i.e. $\mathbf{J}_r$ is a vector of r ones), and $\mathbf{X}_4$ to $\mathbf{X}_7$ the Kronecker products for the interaction effect $A \times B; A \times C; B \times C$; and $A \times B \times C$, respectively. Combining the incidence submatrices for a 3-factor linear mixed model with full interaction results in a single incidence matrix $\mathbf{X} = (\mathbf{1}_N \quad \mathbf{X}_1 \quad \mathbf{X}_2 \quad \mathbf{X}_3 \quad \mathbf{X}_4 \quad \mathbf{X}_5 \quad \mathbf{X}_6 \quad \mathbf{X}_7)$.

In the case of an unbalanced dataset, which could be either by chance or by design, the incidence submatrices ($\mathbf{X}_m$) in (2.5) will have strings of ones of unequal length (Hocking, 1985). The Kronecker product notation demonstrated in (2.15) is suitable only for balanced data, and hence cannot be applied to cases of study with unequal observations per cell or treatment combination. However, as noted by Hocking (1985), the notation for diagonal matrices can be used for unbalanced data cases. Hence, the incidence submatrices of a three-factor unbalanced model linear mixed model take the form

$$\begin{aligned}
&\mathbf{1}_N = \mathbf{J}_N, \\
&\mathbf{X}_1 = \mathbf{D}_a(\mathbf{J}_{n_i}), \\
&\mathbf{X}_2 = \mathbf{D}_b(\mathbf{J}_{n_j}), \\
&\mathbf{X}_3 = \mathbf{D}_c(\mathbf{J}_{n_h}), \\
&\mathbf{X}_4 = \mathbf{D}_{ab}(\mathbf{J}_{n_{ij}}), \\
&\mathbf{X}_5 = \mathbf{D}_{ac}(\mathbf{J}_{n_{ih}}), \\
&\mathbf{X}_6 = \mathbf{D}_{bc}(\mathbf{J}_{n_{jh}}), \\
&\mathbf{X}_7 = \mathbf{D}_{abc}(\mathbf{J}_{n_{ijh}}),
\end{aligned} \tag{2.16}$$

where $\mathbf{J}_{n_{ijk}}$ is an $n_{ijk} \times 1$ vector of ones $(i = 1, ..., a; j = 1, ..., b; h = 1, ..., c)$, $\mathbf{D}_{abc}(\mathbf{J}_{n_{ijh}})$ is diagonal matrix of order $N = abc$ with strings of ones of unequal length.

Therefore, with appropriately defined diagonal submatrices for balanced data (2.15) and unbalanced data (2.16), the incidence matrix $\mathbf{X}$ for a k-factor linear model is generally expressed as

$$\mathbf{X} = (\mathbf{J}_N \quad \mathbf{X}_1 \quad \mathbf{X}_2 \quad ... \quad \mathbf{X}_k \quad \mathbf{X}_{k+1} ... \quad \mathbf{X}_{k+s}), \tag{2.17}$$

where $\mathbf{X}_{k+1}, ..., \mathbf{X}_{k+s}$ are coefficient matrices for the interaction effects.

2.4.1 Kronecker Products for the Covariance Structure in LMM

Linear mixed model analysis includes covariance structures and quadratic forms, also known as sum of squares, which are easily constructed using Kronecker products. Moser and Sawyer (1998) presented some general algorithms for constructing covariance matrices and sums of squares in Kronecker form in complete or incomplete balanced linear mixed models under both the infinite and finite model assumptions. We build on the algorithms proposed by Moser and Sawyer (1998) to construct a covariance matrix for a three-factor balanced, infinite linear mixed model using Kronecker products

Consider a three-factor experiment with a levels of fixed factor A, b levels of fixed factor B, c levels of random factor C, and r replicates in each treatment combination of the three factors. The linear mixed model for the experiment is

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\epsilon}, \tag{2.18}$$

where $\mathbf{y}$: $N\times 1$ is the response vector with mean vector $\mathbf{X}\hat{\boldsymbol{\beta}}$; known fixed-effects incidence matrix $\mathbf{X}$:$N\times ab$; known random-effects incidence matrix $\mathbf{Z}$: $N\times q$; $\boldsymbol{\beta}$: $ab\times 1$ and $\mathbf{u}$: $q\times 1$ are unknown vectors fixed and random effects, respectively; $\boldsymbol{\epsilon}$ is a vector of random errors. The covariance structure of model (2.18) is $\mathbf{V} = \sum = \mathbf{Z}\mathbf{G}\mathbf{Z}' + \mathbf{R}$, where $\mathbf{G} = Cov(\mathbf{u})$ and $\mathbf{R} = Cov(\boldsymbol{\epsilon})$. The incidence matrix $\mathbf{X} = (\mathbf{1}_N \quad \mathbf{X}_1 \quad \mathbf{X}_2 \qquad \mathbf{X}_3)$, where

$$\begin{aligned}
\mathbf{1}_N &= \mathbf{1}_a \otimes \mathbf{1}_b \otimes \mathbf{1}_c \otimes \mathbf{1}_r = \mathbf{J}_N,\\
\mathbf{X}_1 &= \mathbf{I}_a \otimes \mathbf{1}_b \otimes \mathbf{1}_c \otimes \mathbf{1}_r = \mathbf{D}_a(\mathbf{J}_r),\\
\mathbf{X}_2 &= \mathbf{1}_a \otimes \mathbf{I}_b \otimes \mathbf{1}_c \otimes \mathbf{1}_r = \mathbf{D}_b(\mathbf{J}_r),\\
\mathbf{X}_3 &= \mathbf{I}_a \otimes \mathbf{I}_b \otimes \mathbf{1}_c \otimes \mathbf{1}_r = \mathbf{D}_{ab}(\mathbf{J}_r),
\end{aligned} \tag{2.19}$$

and $\boldsymbol{\beta} = [\theta_0, \alpha_1, ..., \alpha_a, \beta_1, ..., \beta_b, (\alpha\beta)_{11}, ..., (\alpha\beta)_{ab}]'$. The incidence matrix $\mathbf{Z} = (\mathbf{Z}_1 \quad \mathbf{Z}_2 \quad \mathbf{Z}_3 \quad \mathbf{Z}_4)$, where

$$
\begin{aligned}
\mathbf{Z}_1 &= \mathbf{1}_a \otimes \mathbf{1}_b \otimes \mathbf{I}_c \otimes \mathbf{1}_r = \mathbf{D}_c(\mathbf{J}_r),\\
\mathbf{Z}_2 &= \mathbf{I}_a \otimes \mathbf{1}_b \otimes \mathbf{I}_c \otimes \mathbf{1}_r = \mathbf{D}_{ac}(\mathbf{J}_r),\\
\mathbf{Z}_3 &= \mathbf{1}_a \otimes \mathbf{I}_b \otimes \mathbf{I}_c \otimes \mathbf{1}_r = \mathbf{D}_{bc}(\mathbf{J}_r),\\
\mathbf{Z}_4 &= \mathbf{I}_a \otimes \mathbf{I}_b \otimes \mathbf{I}_c \otimes \mathbf{1}_r = \mathbf{D}_{abc}(\mathbf{J}_r),
\end{aligned}
\tag{2.20}
$$

and $\mathbf{u} = [\gamma_1, ..., \gamma_c, (\alpha\gamma)_{11}, ..., (\alpha\gamma)_{ac}, (\beta\gamma)_{11}, ..., (\beta\gamma)_{bc}, (\alpha\beta\gamma)_{111}, ..., (\alpha\beta\gamma)_{abc}]'$.

We use the fact that $\mathbf{1}_n\mathbf{1}_n' = \mathbf{J}_n$ in conjunction with the simple structures of $\mathbf{G}$ and $\mathbf{R}$ expressed as, respectively

$$
\mathbf{G} = \begin{bmatrix} \sigma_C^2\mathbf{I}_c & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \sigma_{AC}^2\mathbf{I}_a \otimes \mathbf{I}_c & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \sigma_{BC}^2\mathbf{I}_b \otimes \mathbf{I}_c & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \sigma_{ABC}^2\mathbf{I}_a \otimes \mathbf{I}_b \otimes \mathbf{I}_c \end{bmatrix}
\tag{2.21}
$$

and

$$
\mathbf{R} = \begin{bmatrix} \sigma_{R(ABC)}^2 & \cdots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \sigma_{R(ABC)}^2 \end{bmatrix}.
\tag{2.22}
$$

Thus, combining $\mathbf{Z}$, $\mathbf{G}$ and $\mathbf{R}$ in (2.20) - (2.22), and using the definition of the covariance matrix $\mathbf{V}$, we obtain

$$
\begin{aligned}
\mathbf{V} =&\mathbf{Z}\mathbf{G}\mathbf{Z}' + \mathbf{R}\\
=&(\mathbf{1}_a \otimes \mathbf{1}_b \otimes \mathbf{I}_c \otimes \mathbf{1}_r)(\sigma_C^2\mathbf{I}_c)(\mathbf{1}_a \otimes \mathbf{1}_b \otimes \mathbf{I}_c \otimes \mathbf{1}_r)'\\
&+ (\mathbf{I}_a \otimes \mathbf{1}_b \otimes \mathbf{I}_c \otimes \mathbf{1}_r)(\sigma_{AC}^2\mathbf{I}_a \otimes \mathbf{I}_b)(\mathbf{I}_a \otimes \mathbf{1}_b \otimes \mathbf{I}_c \otimes \mathbf{1}_r)'\\
&+ (\mathbf{1}_a \otimes \mathbf{I}_b \otimes \mathbf{I}_c \otimes \mathbf{1}_r)(\sigma_{BC}^2\mathbf{I}_b \otimes \mathbf{I}_c)(\mathbf{1}_a \otimes \mathbf{I}_b \otimes \mathbf{I}_c \otimes \mathbf{1}_r)'\\
&+ (\mathbf{I}_a \otimes \mathbf{I}_b \otimes \mathbf{I}_c \otimes \mathbf{1}_r)(\sigma_{ABC}^2\mathbf{I}_a \otimes \mathbf{I}_b \otimes \mathbf{I}_c)(\mathbf{I}_a \otimes \mathbf{I}_b \otimes \mathbf{I}_c \otimes \mathbf{1}_r)'\\
&+ (\sigma_{R(ABC)}^2)(\mathbf{I}_a \otimes \mathbf{I}_b \otimes \mathbf{I}_c \otimes \mathbf{I}_r)\\
=&\sigma_C^2\mathbf{J}_a \otimes \mathbf{J}_b \otimes \mathbf{I}_c \otimes \mathbf{J}_r + \sigma_{AC}^2\mathbf{I}_a \otimes \mathbf{J}_b \otimes \mathbf{I}_c \otimes \mathbf{J}_r\\
&+ \sigma_{BC}^2\mathbf{J}_a \otimes \mathbf{I}_b \otimes \mathbf{I}_c \otimes \mathbf{J}_r + \sigma_{ABC}^2\mathbf{I}_a \otimes \mathbf{I}_b \otimes \mathbf{I}_c \otimes \mathbf{J}_r\\
&+ \sigma_{R(ABC)}^2\mathbf{I}_a \otimes \mathbf{I}_b \otimes \mathbf{I}_c \otimes \mathbf{I}_r.
\end{aligned}
\tag{2.23}
$$

An algorithm for the covariance structure, $\mathbf{V} = \sum = \mathbf{Z}\mathbf{G}\mathbf{Z}' + \mathbf{R}$, is derived without the use of matrices $\mathbf{G}$, $\mathbf{R}$ and $\mathbf{Z}$ as follows (Moser and Sawyer, 1998):

Step 1: Create rows of random main and interaction effects.

Step 2: Create column headings of factor letters and subscript letters on the variance.

Step 3: If the letter in the column heading is part of the variance subscript letter-combination, write $\mathbf{I}_d$.

Step 4: Place $\mathbf{J}_d$ elsewhere.

Table 2.1 summarises the algorithm steps 1 - 4, for constructing the covariance matrix $\sum$ of a three-factor linear mixed model.

Table 2.1: Covariance Matrix Layout

Factor	A	B	C	R	
Subscript d	a	b	c	r	
σ^2_C	$\mathbf{J}_a \otimes$	$\mathbf{J}_b \otimes$	$\mathbf{I}_c \otimes$	$\mathbf{J}_r$	$+$
σ^2_{AC}	$\mathbf{I}_a \otimes$	$\mathbf{J}_b \otimes$	$\mathbf{I}_c \otimes$	$\mathbf{J}_r$	$+$
σ^2_{BC}	$\mathbf{J}_a \otimes$	$\mathbf{I}_b \otimes$	$\mathbf{I}_c \otimes$	$\mathbf{J}_r$	$+$
σ^2_{ABC}	$\mathbf{I}_a \otimes$	$\mathbf{I}_b \otimes$	$\mathbf{I}_c \otimes$	$\mathbf{J}_r$	$+$
$\sigma^2_{R(ABC)}$	$\mathbf{I}_a \otimes$	$\mathbf{I}_b \otimes$	$\mathbf{I}_c \otimes$	$\mathbf{I}_r$	

Summing up the row elements in the covariance matrix table gives the same covariance matrix result obtained in (2.21). The algorithm is applicable to an infinite model, where all random effects are assumed to be independent (Moser and Sawyer, 1998). However, for an infinite model case, the random effects maintain the same sampling distribution, but the random effects that are a result of the interaction of fixed and random factors will have to be subjected to additional restrictions that lead to a correlated error structure. Using our 3-factor finite model example, the imposed restrictions are, $\sum_j^b = 1(BC)_{jk} = 0$ for each $k = 1, ..., c$, such that $E(\mathbf{BC}) = \mathbf{0}$ and $Var(\mathbf{BC}) = \sigma^2_{BC}[(\mathbf{I}_b - b^{-1}\mathbf{J}_b \otimes \mathbf{I}_c)]$, for the random vector of the interaction of fixed factor B and random factor C. The covariance matrix for a finite model would be constructed following the same order of steps in the infinite model but with an additional step (3b) applied before step 4. The finite model covariance matrix procedure is updated as follows:

Step 1: Create rows of random main and interaction effects.

Step 2: Create column headings of factor letters and subscript letters on the variance.

Step 3a: If the letter in the column heading is not part of the variance subscript letter-combination, write $\mathbf{J}_d$ in the Kronecker product.

Step 3b: If a non-nested fixed factor letter in the column heading is found on the variance subscript letter-combination, write $(\mathbf{I}_d - d^{-1}\mathbf{J}_d)$ in the covariance matrix.
Step 4: Place $\mathbf{I}_d$ elsewhere.

Depending on the assumed scope of inference for the interaction effects, the choice of model (finite or infinite) must be appropriately considered. An infinite mode is appropriate when broad inference space is assumed, whereas a finite model is considered for the narrow inference space.

2.4.2 Kronecker Products for the Sum of Squares in LMM

The same algorithm for the covariance structure of a balanced linear mixed model is applied when constructing sums of squares. We extend the sum of squares algorithm proposed by Moser and Sawyer (1998) using our three-factor linear mixed model described in the previous section. Let $\mathbf{Y}'M_g\mathbf{Y}$, be the sum of squares in the model where M_g $(g = 1, ..., h)$ is the matrix associated with the sum of squares of the overall mean (μ), factors (A, B, and C), full interactions, and the nested factor $R(ABC)$, respectively. The respective sums of squares are constructed through the following steps (Moser and Sawyer, 1998).

Step S1: Create the first row heading for the letters of the overall mean, factors and interactions in the model, and the second row heading of the associated matrices (M_g).
Step S2: Create two column-headings, one for the factor letters and the second for the number of levels (d) of the factor.
Step S3a: If the first row-heading letter does not match the column-heading letter, write $d^{-1}\mathbf{J}_d$ in the Kronecker product.
Step S3b: If the first row-heading letter of a non-nested factor matches the column-heading letter , write $(\mathbf{I}_d - d^{-1}\mathbf{J}_d)$ in the Kronecker product.
Step S4: Place $\mathbf{I}_d$ elsewhere.

Table 2.2 summarises the algorithm steps S1 - S4, for constructing the sums of squares matrices $\mathbf{Y}'M_g\mathbf{Y}$ of a three-factor linear mixed model.

Table 2.2: Sum of Squares Matrix Layout

	Factor	A		B		C		R
	Level $d =$	a		b		c		r
μ	$M_1 =$	$a^{-1}\mathbf{J}_a$	$\otimes$	$b^{-1}\mathbf{J}_b$	$\otimes$	$c^{-1}\mathbf{J}_c$	$\otimes$	$r^{-1}\mathbf{J}_r$
A	$M_2 =$	$(\mathbf{I}_a - a^{-1}\mathbf{J}_a)$	$\otimes$	$b^{-1}\mathbf{J}_b$	$\otimes$	$c^{-1}\mathbf{J}_c$	$\otimes$	$r^{-1}\mathbf{J}_r$
B	$M_3 =$	$a^{-1}\mathbf{J}_a$	$\otimes$	$(\mathbf{I}_b - b^{-1}\mathbf{J}_b)$	$\otimes$	$c^{-1}\mathbf{J}_c$	$\otimes$	$r^{-1}\mathbf{J}_r$
C	$M_4 =$	$a^{-1}\mathbf{J}_a$	$\otimes$	$b^{-1}\mathbf{J}_b$	$\otimes$	$(\mathbf{I}_c - c^{-1}\mathbf{J}_c)$	$\otimes$	$r^{-1}\mathbf{J}_r$
AB	$M_5 =$	$(\mathbf{I}_a - a^{-1}\mathbf{J}_a)$	$\otimes$	$(\mathbf{I}_b - b^{-1}\mathbf{J}_b)$	$\otimes$	$c^{-1}\mathbf{J}_c$	$\otimes$	$r^{-1}\mathbf{J}_r$
AC	$M_6 =$	$(\mathbf{I}_a - a^{-1}\mathbf{J}_a)$	$\otimes$	$b^{-1}\mathbf{J}_b$	$\otimes$	$(\mathbf{I}_c - c^{-1}\mathbf{J}_c)$	$\otimes$	$r^{-1}\mathbf{J}_r$
BC	$M_7 =$	$a^{-1}\mathbf{J}_a$	$\otimes$	$(\mathbf{I}_b - b^{-1}\mathbf{J}_b)$	$\otimes$	$(\mathbf{I}_c - c^{-1}\mathbf{J}_c)$	$\otimes$	$r^{-1}\mathbf{J}_r$
ABC	$M_8 =$	$(\mathbf{I}_a - a^{-1}\mathbf{J}_a)$	$\otimes$	$(\mathbf{I}_b - b^{-1}\mathbf{J}_b)$	$\otimes$	$(\mathbf{I}_c - c^{-1}\mathbf{J}_c)$	$\otimes$	$r^{-1}\mathbf{J}_r$
$R(ABC)$	$M_9 =$	$\mathbf{I}_a$	$\otimes$	$\mathbf{I}_b$	$\otimes$	$\mathbf{I}_c$	$\otimes$	$(\mathbf{I}_r - r^{-1}\mathbf{J}_r)$

In the example used, it is assumed that only the term $R(ABC)$ has nested factor letters A, B and C. The Kronecker product in each row of Table 2.2 gives the sum of squares of the corresponding factor or interaction thereof. For an example, the sum of squares matrices for the main factor B and the interaction of factors A and B are given by

$$\mathbf{Y}'M_3\mathbf{Y} = \mathbf{Y}'[a^{-1}\mathbf{J}_a \otimes (\mathbf{I}_b - b^{-1}\mathbf{J}_b) \otimes c^{-1}\mathbf{J}_c \otimes r^{-1}\mathbf{J}_r]\mathbf{Y},$$

$$\mathbf{Y}'M_5\mathbf{Y} = \mathbf{Y}'[(\mathbf{I}_a - a^{-1}\mathbf{J}_a) \otimes (\mathbf{I}_b - b^{-1}\mathbf{J}_b) \otimes c^{-1}\mathbf{J}_c \otimes r^{-1}\mathbf{J}_r]\mathbf{Y},$$

respectively. Further details and alternatives of the matrix quadratic forms for the one way and two-way balanced ANOVA models sum of squares were discussed (Hocking, 1985; Rogers, 1984; Saw, 1992; Sunwoo, 1996). As highlighted by Saw (1992), the sums of squares matrix quadratic forms for unbalanced ANOVA cases may be derived by making use of direct sums.

2.5 Significance Tests for Fixed Effects

The hypotheses of interest in LMMs are two-fold: (1) for fixed effects, $H_0 : F_i = 0$ and (2) for random effects, $H_0 : \sigma^2_R > 0$ or $\sigma^2_{FR} > 0$, where estimates and magnitudes of σ^2_R and/or σ^2_{FR} are to be established (McLean *et al.*, 1991). Given the general linear mixed model defined in (2.3), the most common classic tests of significance of fixed effects include the Wald test, and the Likelihood ratio test. Luke (2017) argued that the two methods are somewhat anti-conservative, especially for smaller sample sizes. The null and alternative hypotheses for fixed effects significance tests are: $H_0 : \beta_i = 0$, against the alternative $H_1 : \beta_i \neq 0$.

2.5.1 The Wald Test

Wald test, also called the Z-test, is a classic significance test for fixed effects in general mixed models. Luke (2017) named it the t-as-z approach since its p-values are estimated by using the z-distribution. This is because, as degrees of freedom increase, the t-distribution approximates the z-distribution, and at infinite degrees of freedom, the two are identical. Hence, the Wald t-values taken to be z-distributed are used to generate p-values. Lack of clear guidelines to justify and evaluate the t-as-z approach makes the technique unreliable (Luke, 2017). The test is developed on the assumption that the Wald statistic,

$$Z = \frac{(\hat{\beta}_i - \beta_i)}{\sqrt{Var(\hat{\beta}_i)}}, \tag{2.24}$$

and can be approximated by a standard normal distribution. Therefore, with a known matrix $\mathbf{K}$, the null and alternative hypotheses for the Wald test are: $H_0 : \mathbf{K}\boldsymbol{\beta} = 0$, against the alternative $H_1 : \mathbf{K}\boldsymbol{\beta} \neq 0$. The Wald test statistic is therefore expressed as,

$$Q = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\mathbf{K}'[\mathbf{K}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{K}']^{-1}\mathbf{K}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}), \tag{2.25}$$

can be approximated by Chi-Square (χ^2) distribution with rank($\mathbf{K}$) degrees of freedom. In order to take care of the estimate of variance components ($\mathbf{V} = Var(\mathbf{y}) = \mathbf{ZGZ}' + \mathbf{R}$), Littell *et al.* (2006) suggested an approximation of the Wald test statistic by F distribution with rank($\mathbf{K}$) numerator degrees of freedom as follows,

$$F = \frac{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\mathbf{K}'[\mathbf{K}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{K}']^{-1}\mathbf{K}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})}{rank(\mathbf{K})}, \tag{2.26}$$

and the denominator degrees of freedom can be estimated from the data using approaches, such as the Satterthwaite (1946) approximation. Other methods for approximating degrees of freedom and obtaining p-values include the Kenward-Roger (Kenward and Roger, 1997), applied to restricted maximum likelihood (REML) models, and the Satterthwaite (1941) approximations, applied to both maximum likelihood (ML) models and restricted maximum likelihood (REML) models. These two methods are used to approximate the denominator degrees of freedom for the F statistics in (2.15), or the t statistics (Luke, 2017). The other method for evaluating significance while producing acceptable type I error rates, even for smaller samples, is the parametric bootstrapping (Halekoh and H∅jsgaard, 2014).

2.5.2 The Likelihood Ratio Test (LRT)

Likelihood ratio tests (LRTs) can also be used to test fixed effects in linear mixed models. The tests are used to determine a better model fit between two competing models (Luke, 2017). Maximum likelihood estimation results in a -2log likelihood value, which summarises the fit of the observed to the expected values used to compare different models. The assumption of this test is that the two models compared (one without and the other with the fixed effect) must be nested. Specifically, LRTs are used to determine whether the inclusion of a particular parameter or random effect improves the model fit, holding all other model parameters constant. If elimination of the parameter or random effect reduces the log-likelihood, then the parameter or random effect is statistically significant. The likelihood ratio test statistic is given by,

$$2log\left(\frac{L_1}{L_0}\right) = 2[log(L_1) - log(L_0)], \tag{2.27}$$

where L_0 is the likelihood for the nested null model (without the analysed fixed effect) and L_1 is the likelihood for the general model (with the analysed fixed factor), i.e. the condition $log(L_1) > log(L_0)$ must be satisfied to keep the LRT statistic positive. The test statistic for a LRT (i.e. twice the difference in log-likelihoods) is assumed to follow a chi-square distribution, with degrees of freedom equal to the number of additional parameters in the more complex model (e.g., $df = 1$ if testing a single random effect).

In support of Pinheiro and Bates (2000), Luke (2017) agrees that LRTs can be used even when the model has a complex random effects structure that includes random slopes by comparing the log-likelihoods of models with and without the random effect component. However, LTRs are anti-conservative, meaning that the p-values obtained from the tests are normally lower than the true p-values. Hence, the LRT may not always be appropriate for evaluating the significance of fixed effects when the two competing models have different fixed-effects structures. Luke (2017) argued that when evaluating significance, especially in smaller samples, the use of other methods, such as the parametric bootstrapping, Kenward-Roger and Satterthwaite approximations for deriving the p-values and using the REML when fitting the models, produce acceptable Type 1 error rates.

2.6 Significance Tests for Random Effects Variance

The conceptual idea in a fixed effect model is to estimate the effect of each of the specific treatment levels of that variable, and test whether these treatment effects significantly contribute to the estimation of the response variable. The null hypothesis, $H_0 : \beta_1 = \beta_2 = ... = \beta_k = 0$, opines that the fixed effects have no contribution in determining the amount of response variable, which can be tested using the traditional F-test or p-values. Alternatively, we can test whether fixed effects are all equal, $H_0 : \beta_1 = \beta_2 = ... = \beta_k$, against treatment effects not all equal, $H_1 : \beta_1 \neq \beta_i$ for at least one i. If the interest is on each of the fixed effects in the experiment, then the test for parameter contribution to the response is appropriate. On the other hand, if the interest is on random effects, the test of whether the variance components are significantly greater than zero is appropriate. Although there is no need to compare random effects in the random effects model, the interest lies in investigating the variation of treatment effects and estimating the values of these variance components. If random effects vary, then they cannot be all zero. In linear mixed models, the interest is not restricted to making inferences about the fixed and random effects only but also the variance components. This is achieved by testing homogeneity across units using the null hypothesis: $H_0 : \sigma^2_R = 0$, against $H_1 : \sigma^2_R > 0$.

For example, a study is conducted aiming to understand the effect of a drug on two different breeds of dogs, but more or less chose the dogs at random. Ideally, comparing the treatment effects of the drug in the dogs treated is of no important use since the dogs are a random sample in the breed, not those specific dogs. However, it is statistically important to include the breed effect in the model, to account for the possible variations in these breeds. Therefore, such random effects are not tested; they are estimated for report purposes. Testing is rarely of interest in this case because the role of the random effects is to act as the basis to estimate the variance components of interest and align the model with its assumptions.

As indicated earlier on, hypothesis testing for random effects (or mixed effects) can be done using the likelihood ratio test (LRT) based on the null hypothesis, $H_0 : \sigma^2_R > 0$ or $\sigma^2_{FR} > 0$. Significance testing of random effects in mixed effects models involves the construction of error terms that contain all the sources of random variation of interest. Satterthwaite's (1946) method of denominator synthesis, which basically decomposes sums of squares to linear combinations of all sources of random variation that serve as appropriate error terms, can be used for testing the significance of the respective effect of interest. When ANOVA methods are used for estimation, standard ANOVA techniques for testing the significance of effects by taking ratios of mean squares can be employed.

2.7 Inference Space

The way fixed and random effects are defined is directly attached to the manner in which they were gathered and the environment from which they came (McCulloch *et al.*, 2008). This parallel definition of effects forms the basis for defining inference space in mixed models. McLean *et al.* (1991) introduced the concept of an inference space to clearly define fixed and random effects. As alluded to by McCulloch *et al.* (2008), effects are considered random if inferences are generalisable to all the possible effects from which random sample was drawn, whereas fixed effects exist when inferences are confined only to the effects in the model. This definition of fixed and random effects implies that one is at liberty to consider all effects as random but taking into consideration the appropriate inference space. However, as Gelman (2005) cautioned, there are consequences for treating all effects as random effects. Thus, proper selection and specification of the effects as fixed or random are vital for the accuracy of both prediction and inference in linear mixed models.

Three types of inferential scope are defined based on how random effects are chosen into the predictable function ($\mathbf{K}'\boldsymbol{\beta} + \mathbf{M}'\mathbf{u}$). Firstly, inference scope is referred to as "narrow" if the inference is specific to levels of random effects (McLean *et al.*, 1991) or when every effect is regarded as fixed (Yang, 2010). This implies that, if a selection is to be repeated in the future, the same levels of random effects are to be used to estimate those fixed effects, and the main and interaction random effects remain unchanged (Yang, 2010). Secondly, inference space is known as "broad" if the inference is made to the whole population of random effects (McLean *et al.*, 1991), including the random interaction effects (Yang, 2010). An illustration of broad inference scope in an agricultural experiment would typically involve cultivators as fixed, whereas operators are considered random. Since human efficiency and mental and physical status are not constant, operators' performance is bound to vary. The objective of the experiment is to assess the performance of the cultivator over operator conditions, which clearly leads to a broad inference space scenario.

The third case is "intermediate" inference space, which results when inference is narrow to some random effects but broad to others (McLean *et al.*, 1991). According to Yang (2010), intermediate inference scope in multi-factor experiments is further subdivided into two, depending on whether one wants to consider random main effects as narrow and random interactions as broad, or vice versa. The idea gets more complicated as the number of factors in the experiments increases.

An important aspect of mixed model analysis is the choice of inference space and the expected consequences thereof. As highlighted by McLean *et al.* (1991), researchers have to be familiar with the inference space implicit in the predictable function they use in order to avoid the use of inappropriate standard errors of the estimator. A study by Yang (2010) also confirmed that standard error increases as the scope of inference broadens. The dilemma becomes more complicated when unbalanced data is involved. Hence, the need to make a proper choice of inferential scope to avoid incorrect estimates of standard errors and predictions. The researcher is able to select the desired inference space by manipulating the estimable function. A study by Blouin and Riopelle (2005) demonstrated that different classes of inference spaces can be accommodated in various designs of experiments and data structures. More case studies with inference spaces applied in t-test with unequal variances, randomised complete block design experiments, split-plot experiments (Yang, 2010), completely randomised design, unbalanced data, multivariate-type experiments (McLean *et al.*, 1991) exist.

2.8 Estimation of Standard Errors

Every statistical method applied to research data is associated with some degree of uncertainty that cannot be avoided. Therefore, researchers need to brace themselves for embracing uncertainty as part of research dynamics (Gelman, 2016). For example, reporting a point estimate in research is not complete if the estimate is not supported by its standard error (Wasserstein *et al.*, 2019). The major strength of mixed linear models lies in the estimation of fixed effects (point estimates) and prediction of random effects relative to the intended inference scope. The reliability of these point estimates and predictions depends on the way measures of standard errors, associated prediction intervals and degrees of freedom are calculated (McLean and Sanders, 1988).

Generally, estimable functions in mixed linear models with known variance components produce realistic measures of standard errors. However, in practice, variance components are not always readily available or known a priori, but have to be estimated from the data before standard errors can be estimated (Goldberger, 1962; Kackar and Harville, 1984). The variance component estimate is then used as an estimate of variability to compute standard error (SE), an important indicator of precision of the point estimate (McLean *et al.*, 1991). One of the most prominent variance estimation procedures for determining approximate standard errors in linear mixed models is the restricted maximum likelihood (REML). The estimated variance components are then used as if they were known variances (McLean and Sanders, 1988; McLean *et al.*, 1991). The approach can be applied to any mixed model problem, including but not limited to balanced and unbalanced designs. Depending on the researcher's preference and experimental design, various procedures for approximating the associated degrees of freedom for these estimates include the Satterthwaite's (1946) and Kenward and Roger's (1997) approaches, among others, exist.

For example, a confidence interval of a sample mean can be determined (at a predetermine α level of significance) using its standard error. Other than checking precision or uncertainty around the estimate of the mean, the confidence interval for the estimate can also be used to form a test statistic for testing hypotheses $H_0 : \mu = 0$ against $H_1 : \mu \neq 0$, where μ is the population mean. The standard error of a treatment mean in a linear mixed model proposed by McLean *et al.* (1991) can be computed from the corresponding estimate of variance ($\hat{\sigma}^2$) given by:

$$\hat{\sigma}^2 = \left(\frac{1}{n}\right)(EMS), \tag{2.28}$$

where *EMS* is the error mean square; n is the sample size.

To estimate the variance of the difference between two treatment means in a mixed model procedure, the following variance formula can be used (McLean *et al.*, 1991):

$$\hat{\sigma^2} = \left(\frac{2}{rn}\right)(EMS_{FR}), \tag{2.29}$$

where r is the random factor levels, EMS_{FR} is the interaction mean square, and n is the sample size.

There are numerous approaches to estimate standard errors or variance components. The basic approaches for estimating standard errors of estimated variance components do require assumptions about the distributions of the score effects. Due to different assumptions that have to be satisfied for a particular statistical approach to work on any given data, there is no single best approach to use. Different data structures define the reasons for preference of one method over another. For example, when the analysis of variance approach is used in linear models, and there is enough evidence that the data violate normality or homogeneity of variance assumption, the bootstrap or jackknife re-sampling methods approaches may be preferred for estimation of standard errors. In whatever circumstance or approach, the purpose of statistical analysis must be considered, i.e., quantifying the variability in data by providing the point estimates of parameters that are supported by their precision.

There have been many differences and confusions in the procedures for calculating expected mean squares, which are directly linked to the calculation of standard errors for different types of mixed models. The differences emanate from the assumption that interactions of fixed and random effects sum to zero over the fixed effect level (McLean *et al.*, 1991) when estimating the expected mean squares. Some authors (Hocking, 1985; Searle and Gruber, 2017) advocate for the relaxation of this assumption, while many consider it valid.

In a split-split-plot treatment structure, for example, expected mean squares can be used to compute estimates of variance components and standard errors for means and comparisons of the means (Littell *et al.*, 2006). Table 2.1 displays some of these computational methods for constructing standard errors for the main and interaction effects with *a*-levels of whole plot (wp) factor having, *b*-levels of split-plot (sp) factor, *c*-levels of split-split-plot (ssp) factor and *r* replicates per cell.

Table 2.3: Standard Errors for Split-Split-Plot Treatment Structure

Means	Standard Error	t-test Statistic
Whole-plot mean	$\left(\frac{E_{(wp)}}{rbc}\right)^{\frac{1}{2}}$	t_a
Split-plot mean	$\left(\frac{E_{(sp)}}{rac}\right)^{\frac{1}{2}}$	t_b
Split-plot mean for the same whole-plot	$\left(\frac{E_{(sp)}}{rc}\right)^{\frac{1}{2}}$	t_b
Split-plot mean for different whole-plots	$\left[\frac{(b-1)E_{(sp)}+E_{(wp)}}{rbc}\right]^{\frac{1}{2}}$	t_b
Split-split-plot mean	$\left(\frac{E_{(ssp)}}{rab}\right)^{\frac{1}{2}}$	t_c
Split-split-plot mean for the same whole-plot	$\left(\frac{E_{(ssp)}}{rb}\right)^{\frac{1}{2}}$	t_c
Split-split-plot mean for the same split-plot	$\left(\frac{E_{(ssp)}}{ra}\right)^{\frac{1}{2}}$	t_c
Split-split-plot mean for the same whole-plot and split-plot	$\left(\frac{E_{(ssp)}}{r}\right)^{\frac{1}{2}}$	t_c
Split-plot mean for same or different split-split-plot	$\left[\frac{(c-1)E_{(ssp)}+E_{(sp)}}{rac}\right]^{\frac{1}{2}}$	t_{bc}
Split-plot mean for same whole-plot and same or different split-split-plot	$\left[\frac{(c-1)E_{(ssp)}+E_{(sp)}}{rc}\right]^{\frac{1}{2}}$	t_{bc}
Whole-plot mean for same or different split-split-plot	$\left[\frac{(c-1)E_{(ssp)}+E_{(wp)}}{rbc}\right]^{\frac{1}{2}}$	t_{ac}
Whole-plot mean for same or different split-plot and split-split-plot	$\left[\frac{b(c-1)E_{(ssp)}+(b-1)E_{(sp)}+E_{(wp)}}{rbc}\right]^{\frac{1}{2}}$	t_{abc}

$$t_{bc} = \frac{(c-1)E_{(ssp)}(t_c)+E_{(sp)}(t_b)}{(c-1)E_{(ssp)}+E_{(sp)}}; \qquad t_{ac} = \frac{(c-1)E_{(ssp)}(t_c)+E_{(sp)}(t_a)}{(c-1)E_{(ssp)}+E_{(wp)}};$$

$$t_{abc} = \frac{b(c-1)E_{(ssp)}(t_c)+(b-1)E_{(sp)}(t_b)+E_{(wp)}(t_a)}{b(c-1)E_{(ssp)}+(b-1)E_{(sp)}+E_{(wp)}}$$

Uncertainty is a natural component of all research data. Wasserstein *et al.* (2019) argued that when a point estimate is qualified with a measure of its uncertainty, an acceptance of uncertainty becomes more meaningful. Fitting linear mixed models has been made easier in several statistical computing software, such as SAS, R and SPSS, which provide parameters estimates, estimates of variance components and their associated standard errors and or interval estimate.

2.9 Application of LMMs in Other Research Fields

The use of linear mixed models has taken the central stage in various research fields. Research in most of these fields requires a comprehensive mastery of the linear mixed model framework as a crucial tool for both designing and analysis purposes. Linear mixed models have been routinely used to model scenarios involving both fixed and random effects. These factors are usually defined as either fully fixed or random in nature. There are circumstances where the convenience of conceptualizing factors as having both fixed and random levels is necessary. This section highlights some of these research areas where the proposed concept can be useful.

In most cases, linear mixed models of ANOVA type have been applied to estimation and testing procedures involving both a selected few or fixed number of fixed effects out of possibly many fixed effects. Usually, only those fixed factors that are believed to have a significant contribution to the response variable are considered in model construction. There are situations when the number of fixed effects is large (high-dimensional case), and the number of fixed effects diverges as the sample size goes to infinity (Chen *et al.*, 2015). Linear mixed models of high-dimensional data would require a modification to the traditional linear mixed model tests to accommodate the problem with a sparse model structure. Chen *et al.* (2015) proposed tests that are post-selection-based with an orthogonality-based selection of the smoothly clipped absolute deviation penalty (SCAD) type applied when selecting significant fixed effects into the working model. Most classical approaches to testing fixed effects linear mixed models (Kenward and Roger, 1997; Wang and Dai, 2014) are robust in small datasets, but they tend to break down when the number of covariates increases (Bradic *et al.*, 2020). These high-dimensional environments naturally create possible complex interactions and unexpected heterogeneity which procedural restrictions in common approaches may not be able to encompass. (Bradic *et al.*, 2020) developed a family of moment matching tests that use penalised estimators to deal with misspecification and/or misestimation of random effects in high-dimensional linear models.

Linear mixed models have also been a powerful tool for analysing correlated and longitudinal data. Studies involving skew-normal, longitudinal (Ye and Wang, 2015; Laird and Ware, 1982) and panel data are common in fields such as demography, biomedicine, economics and finance, etc. The normality assumption is too restrictive in most linear mixed models' practical applications (Wu *et al.*, 2017). For this reason, in studies involving longitudinal or correlated data, the application of linear mixed models requires some relaxation of the normality assumption in order to make reasonable inferences on the unknown parameters. More flexible approaches are needed to handle the analysis of data from these scenarios. Recently, Wu *et al.* (2017) proposed the ANOVA-type F-tests for testing hypothesis on fixed effects of interest as well as the significance of random effects in linear mixed effect models with skew-normal errors and distribution-free random errors.

When a study is designed so that certain data points are expected to be, on average, more similar to each other than other data points, the issue of non-independence in data emerges especially in within-subject or within-item designs (Brauer and Curtin, 2018). For example, when multiple data points are collected from each subject or unit (longitudinal research), when clustered subjects can influence one another within individual clusters, or when subjects are exposed to the same set of items. Such cases are common in psychology studies (Barr *et al.*, 2013; Westfall *et al.*, 2015), where the same subjects are exposed to the same set of treatments, stimuli or targets, etc. In such cases, the predictors vary within the subjects. Brauer and Curtin (2018) provided a guiding framework for analysing data with one or more sources of non-independence, either involving categorical within-unit predictors or continuous within-unit predictors using linear mixed models. They cautioned researchers and practitioners on the importance of including all possible sources of error (appropriate type of random effects) when estimating linear mixed effects models involving non-independent data to minimise the type I error rate.

In certain studies where linear mixed models include covariates, random effects may become confounding in fixed effects, causing unexpected substantial changes in fixed effect estimates relative to the same model without random effects (Schnell and Bose, 2019). In such cases, the fixed effects and the random effects compete with each other to capture the variation in the data. How random effects impact the fitted values of a linear mixed model is better understood by reparameterising the random effect component into a canonical form with independent and identically distributed random effects, a process (Schnell and Bose, 2019) referred to as spectral decomposition of a linear mixed model. Spectral reparameterisation of random effects provides a simpler way to formulate the model and understand and assess the model fits. According to Schnell and Bose (2019), certain trends and features inherent in the data can be visualised and weighted through spectral decomposition. It is a useful mechanism to interpret and possibly plan to avoid confounding fixed effects by random effects.

Central to linear mixed model inference is the estimation of fixed effects and variance components. The least squares method is the most popular method for estimating model parameters. Some of the most common methods used for estimation of variance components in LMMs include the maximum likelihood (ML), restricted maximum likelihood (REML) (Harville, 1977), analysis of variance (ANOVA) (Henderson, 1953), and Bayesian methods (Agresti, 2015). However, these methods work effectively well in orthogonal linear mixed models (when the normality assumption is assumed) and balanced data. When normality assumption is not considered, nonorthogonal mixed models are involved, and inference becomes a challenge (Ferreira *et al.*, 2020). To this effect, Ferreira *et al.* (2020) proposed a general least squares estimation method for estimating variance components and estimable vectors in nonorthogonal and orthogonal linear mixed models, without assuming normality or any other prior distribution.

Luke (2017) confirmed that the use of linear mixed effects models (LMMs) is becoming increasingly common in many real-world applications ranging from clustered, nested, longitudinal, genome-wide association and spatial data analysis. Thereby providing results with acceptable type I error rates. One area that is populated by LMMs is psycholinguistics studies. A study by Hohenstein *et al.* (2017) used LMMs in psycholinguistics to analyse eye-movement control. The research applied linear mixed models (LMMs) to model fixation duration that takes into account the predetermined order of their occurrence in the behavioural stream.

Another study by Liu *et al.* (2016) involves the Genome-Wide Association Studies (GWAS) exploring how to control genetic-phenotypic relationships using Mixed Linear Models (MLMs). Liu *et al.* (2016) made use of direct statistical tests in Genome-Wide Association Studies (GWAS) as a strategy to eliminate false positives by fitting both population structure and each individual's total genetic effect as covariates in a Mixed Linear Model (MLM) to make adjustments for testing markers. In support of this, Runcie and Crawford (2019) confirmed that linear mixed effect models are powerful tools used to account for population structure in genome-wide association studies (GWAS) and estimate the genetic architecture of complex traits.

Vargas *et al.* (2015) provided experiments involving three-way linear mixed effects models with interaction in agronomy and breeding research. There are many types of LMMs that are appropriate for different types of data. However, Bates *et al.* (2015), argued that, for models with several fixed factors (such as experimental manipulations) and several random factors (like subjects and items), the question of how to choose the appropriate random-effects structure becomes substantially more complex. Schielzeth and Forstmeier (2009) cautioned that both random intercepts and random slopes need to be considered in LMMs to guard against anti-conservative conclusions, like accepting an experimental effect more frequently as significant than warranted by the data.

2.10 Conclusion

The mixed model methodology can be applied in various of experimental designs, including completely randomised design (CRD), randomised complete block design (RCBD) and split-plot design. In each design, various hypotheses can be tested by making use of appropriate linear combinations of mean squares from both fixed and random effects. However, researchers should pay particular attention to the inference space they intend to use. A review of the theory of linear models, mixed model framework, estimation of standard errors, hypothesis testing procedures, estimating degrees of freedom, and determining intended inference scope have been reviewed. Recent literature on the application of linear mixed models in scientific research has been highlighted. Chapter 3 focuses on the principles around the partitioning approach to linear mixed models, emphasising on the procedure of partitioning factors based on their levels before model construction.

CHAPTER 3

PARTITIONING OF FACTORS IN LINEAR MIXED MODELS

This chapter introduces the concept of partitioning of factors of a linear mixed model based on targeted factor levels, intended inference space and the researcher's objectives. Consideration is given to the design of experiment, appropriate tretment structure, formulation of a three-way linear mixed model, hypothesis testing, analysis of variance and associated degrees of freedom.

3.1 Introduction

An outline of the process of partitioning factors, construction of appropriate linear mixed models and the analysis experimental data with two or more factors is the main focus of this chapter. A general theoretical framework covering fundamental principles of linear mixed models, testing of model assumptions underlying the linear mixed models (LMMs), testing of hypotheses, and testing of fixed and random effects is presented. Based on the work done by Njuho and Milliken (2005, 2009), the main focus is on developing linear mixed models with three or more factors, each consisting of both fixed and random factor levels, and installing these models in different treatment designs such as completely randomised design (CRD), randomised complete block design (RCBD), Latin Square design (LSD) and Split-plot design (SPD).

3.2 Choice of an Appropriate Linear Mixed Model

The previous sections provided some background information and the developments in linear models. The methodology is applicable in various scenarios depending on the type of study and research objectives to be achieved. The scenario where three or more factors share both fixed and random levels is one of the complex setting that needs caution when linear mixed models are applied.

Design of statistical experiments was first introduced by Sir Ronald A. Fisher in the 1920's, as a correctional tool to some flaws in the way the experimental data was analysed (Montgomery, 2013). Statistical design of experiments is a planning process to obtain valid and authentic conclusions from the experiments. Hence, experimental design entails laying out a detailed plan that defines the objectives of the experiment, choice of design, variables involved, analysis procedure, control for extraneous noise around the experimental information and statistical analysis approach. Bate and Chatfield (2016) cautioned that failing to identify the appropriate experimental design and its structure correctly can lead to incorrect model selection and misleading inferences. When a complex experimental design has been chosen, the construction of an appropriate statistical model for analysis is not a straightforward exercise (Bate and Chatfield, 2016).

Littell *et al.* (2006) defined a statistical model for a given data as a mathematical description of how the data can be conceivably produced. In this description of a statistical model, at least two features emerge: (1) a function describing the relationship between the response and all the explanatory variables, and (2) the assumed distribution of the error terms to characterise the random variation in the observed response. These tow features represent the simplest way of describing how the experimental data can be produced using a linear statistical model. Depending on the assumed probability distribution of the error terms or, in particular, the response variable, the form of the general function turns out to be linear or nonlinear. When the distribution of the response variable is normal, ANOVA and linear regression models are examples. An extension beyond linear statistical models arises when the distribution of the response variable is non-normal, leading to a class of models known as the generalised linear models.

LMMs have application in a broad spectrum of areas, including agriculture, biology and medicine. However, as Pan and Shang (2018) noted, selecting an appropriate structure of a linear mixed model is not easy especially in when modelling multi-factor experimental data. On the other hand, the choosing appropriate fixed and random effects is essential for accurate inference and prediction of both the means and covariances in linear mixed models. The development of computing facilities provides various methods for yielding parameter estimates for a statistical model of interest. Maximum-likelihood (ML) and Restricted Maximum-likelihood (REML) methods are standard examples among the various techniques widely used.

The choice of a model is informed by the type of experiment, the objective(s) to be achieved, and the type of data collected. The way experimental data is collected, with proper preparation for the purpose of meeting the specified objective(s), is termed experimental design. Experimental design is therefore defined by the experimental units used, the types of variables involved and the structure of treatment(s) applied to experimental units. The analysis of variance (ANOVA) approach is one of the most popular techniques used to test the treatment effects through comparing the treatment means (fixed effects) or estimating the variance components (random effects).

Montgomery (2013) highlighted that design of experiments is an important tool that scientists often use during the product design, development and improvement processes. With the current Fourth Industrial Revolution taking space in industries, the need for experimental designs in developing products that match environmental factors and other sources of variability is increasing. This much-needed development and progression in sciences are achieved through performing experiments. Experimental designs often involve linear mixed models problems or other complex model structures, which can be analysed easily by means of investigating the difference among treatment effects or predicting the variability to get informed conclusions.

3.3 Linear Models (LMs)

The basic linear regression model forms the basis for the partitioning approach (to be discussed later) in linear mixed models and builds up into various scenarios and experimental designs. For convenience purposes, we revisit the linear regression model before extending the concept to complex scenarios involving factors with hybrid levels. LMs encompass regression analysis, analysis of variance (ANOVA) and analysis of covariance (ANCOVA). A general linear model (LM) approach makes use of a random sample of treatment observations drawn a given population, after a treatment has been administered to each individual in the sample. The data obtained is then analysed using the analysis of variance (ANOVA) approach that produces an F-test or a p-value at a predetermined level of significance (α).

3.3.1 Formulation of a Linear Model

A general linear model is expressed as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{3.1}$$

where $\mathbf{Y}_{n\times 1}$: is an n-dimensional vector of the dependent random variable corresponding to the response variables $\mathbf{X}_{n\times p}$: where $\mathbf{X}_{n\times p}$= $(\mathbf{1} \;\; \mathbf{X}_1 \;\; \mathbf{X}_2 \; ... \; \mathbf{X}_{p-1})$ is a known matrix of explanatory variables with $p \leq n$, $\boldsymbol{\beta}_{p\times 1} = (\beta_0, \beta_1 ..., \beta_k)'$ is a parameter vector whose elements correspond to each X_i variable, $i = 1, ..., p-1$, and $\boldsymbol{\epsilon}_{n\times 1} = (\epsilon_1, ..., \epsilon_n)'$ is vector of error terms. The matrix $\mathbf{X}$ is also known as the design or an incidence matrix, depending on the type of linear model in question, whereas the corresponding elements β_k's are the regression coefficients.

Case I (Regression Model):
The variables forming the columns of the design matrix $\mathbf{X} = (\mathbf{1} \;\; \mathbf{X}_1 \;\; \mathbf{X}_2 \; ... \; \mathbf{X}_{p-1})$ are continuous predictor variables in a regression model.

Case II (ANOVA Model):
The variables forming the columns of the design or an incidence matrix $\mathbf{X} = (\mathbf{1} \;\; \mathbf{X}_1 \;\; \mathbf{X}_2 \; ... \; \mathbf{X}_{p-1})$ are discrete and often coded as 0 and 1. These codes correspond to the levels of factors in the analysis.

This study will focus on Case II (ANOVA models), where both fixed and random factors are involved in a single model. We build on a two-way mixed model that was reviewed in the previous chapter, and extend the LMM with three or more factors.

3.3.2 Three-way Fixed-effects Model

A three-way fixed-effects model is considered, with a single response variable, Y, predicted by three categorical factors: A and B and C, whose levels are fixed and exhaustive. The factor levels are determined specifically by the researcher. One of the primary objectives of the experiment is to determine if these specific factor levels differ in terms of their contribution to the variation in the response variable. Considering the case of unbalanced data, the three-way fixed-effects model with interactions is expressed as

$$y_{ijkh} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk} + \epsilon_{ijkh}, \tag{3.2}$$

where y_{ijkh} is the response of the h^{th} replicate on the i^{th} level of factor A, j^{th} level of factor B, and k^{th} level of factor C, for $i = 1, 2, ..., a$; and $j = 1, 2, ..., b$; $k = 1, 2, ..., c$; and $h = 1, 2, ..., n_{ijk}$ (all $n_{ijk} = n$ when data is balanced). The term μ is the overall mean; α_i, β_j, and γ_k are the main effects of the i^{th}, j^{th} and k^{th} levels of factors A, B and C, respectively. The terms $(\alpha\beta)_{ij}$, $(\alpha\gamma)_{ik}$, $(\beta\gamma)_{jk}$ and $(\alpha\beta\gamma)_{ijk}$, are the interaction effects in the model. We assumed that $\epsilon_{ijkh} \sim i.i.d.$ $N(0; \sigma_e^2)$, which implies that $\mathrm{E}(\mathrm{y}_{ijkh}) = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk}$, whereas $\mathrm{Var}(\mathrm{y}_{ijkh}) = \sigma_e^2$.

The null hypotheses associated with a three-way fixed-effects model include:

$$\begin{aligned}
H_{0(A)} :&\alpha_i = \alpha_{i^*} = 0, \qquad for \;\; i \neq i^*, \\
H_{0(B)} :&\beta_j = \beta_{j^*} = 0, \qquad for \;\; j \neq j^*, \\
H_{0(C)} :&\gamma_k = \gamma_{k^*} = 0, \qquad for \;\; k \neq k^*, \\
H_{0(AB)} :&(\alpha\beta)_{ij} = (\alpha\beta)_{i^*j} = (\alpha\beta)_{ij^*} = (\alpha\beta)_{i^*j^*} = 0, \\
H_{0(AC)} :&(\alpha\gamma)_{ik} = (\alpha\gamma)_{i^*k} = (\alpha\gamma)_{ik^*} = (\alpha\gamma)_{i^*k^*} = 0, \\
H_{0(BC)} :&(\beta\gamma)_{jk} = (\beta\gamma)_{j^*k} = (\beta\gamma)_{jk^*} = (\beta\gamma)_{j^*k^*} = 0, \\
H_{0(ABC)} :&(\alpha\beta\gamma)_{ijk} = (\alpha\beta\gamma)_{i^*jk} = (\alpha\beta\gamma)_{ij^*k} = (\alpha\beta\gamma)_{ij^*k} = (\alpha\beta\gamma)_{i^*j^*k^*} = 0.
\end{aligned} \tag{3.3}$$

Hypotheses $\mathrm{H}_{0(A)}$ - $\mathrm{H}_{0(B)}$ opine the non-existence of main effects of factors A, B and C, respectively, while $\mathrm{H}_{0(AB)}$ - $\mathrm{H}_{0(ABC)}$ suggest the non-existence of the interaction effects in the model.

The test statistic for the three null hypotheses in (3.3) follow the F-distribution.

3.3.3 Three-way Random-Effects Model

A random-effects model is defined when one or more random factors, whose levels were sampled from a pool of many possible levels for analysis purposes, influence the response variable. In this case, the primary objective is to draw inferences about the variations in the response variable over the whole collection of factor levels. Analogous to the fixed-effects model, the three-way random-effects model is expressed as

$$y_{ijkh} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk} + \epsilon_{ijkh}, \tag{3.4}$$

where y$_{ijkh}$ is the response of the hth replicate on the ith level of a random factor A, jth level of a random factor B, and kth level of a random factor C, for $i = 1, 2, ..., a$; and $j = 1, 2, ..., b$; $k = 1, 2, ..., c$; and $h = 1, 2, ..., n_{ijk}$ (all $n_{ijk} = n$ when data is balanced). The term μ is the overall mean; α_i, β_j, and γ_k are the main effects of the i^{th}, j^{th} and k^{th} levels of the random factors A, B and C, respectively. The terms $(\alpha\beta)_{ij}$, $(\alpha\gamma)_{ik}$, $(\beta\gamma)_{jk}$ and $(\alpha\beta\gamma)_{ijk}$, are random interaction effects in the model. It is assumed that $\alpha_i \sim i.i.d.\ N(0; \sigma^2_\alpha)$, $\beta_j \sim i.i.d.\ N(0; \sigma^2_\beta)$, $\gamma_k \sim i.i.d.\ N(0; \sigma^2_\gamma)$, $(\alpha\beta)_{ij} \sim i.i.d.\ N(0; \sigma^2_{\alpha\beta})$, $(\alpha\gamma)_{ik} \sim i.i.d.\ N(0; \sigma^2_{\alpha\gamma})$, $(\beta\gamma)_{jk} \sim i.i.d.\ N(0; \sigma^2_{\beta\gamma})$, $(\alpha\beta\gamma)_{ijk} \sim i.i.d.\ N(0; \sigma^2_{\alpha\beta\gamma})$, and $\epsilon_{ijkh} \sim i.i.d.\ N(0; \sigma^2_e)$, which implies that $E(y_{ijk}) = \mu$, $Var(y_{ijkh}) = \sigma^2_e$ and covariance of the error term with every random effect is zero. Furthermore, $Cov(\alpha_i, \alpha_{i^*}) = 0$ for $i \neq i^*$, $j \neq j^*$, $k \neq k^*$.

3.4 Linear Mixed Models

Linear mixed model methodology has brought statistics to an advanced level. Most often, these models are used to model a broad spectrum of fields, including complex clustered data (data that can be viewed as a sample of samples), repeated measurements, or hierarchical model, longitudinal data (also called panel data, tracked on the same sample at different points in time). Fixed and random effects are allowed to interact and give a better explanation of the differences in the response variable.

3.4.1 Three-way mixed Model

We now consider a unbalanced three-way mixed model, with factors A and B taken as fixed while factor C is deemed as random. The mixed model is similarly expressed as

$$y_{ijkh} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk} + \epsilon_{ijkh}, \tag{3.5}$$

where μ is the overall mean, α_i, β_j, and $(\alpha\beta)_{ij}$ are fixed variables while γ_k, $(\alpha\gamma)_{ik}$, $(\beta\gamma)_{jk}$, $(\alpha\beta\gamma)_{ijk}$ are random variables, for $i = 1, 2, ..., a$; and $j = 1, 2, ..., b$; $k = 1, 2, ..., c$; and $h = 1, 2, ..., n_{ijk}$ (all $n_{ijk} = n$ when data is balanced). It is assumed that $\gamma_k \sim i.i.d.\ N(0; \sigma^2_\gamma)$, $(\alpha\gamma)_{ik} \sim i.i.d.\ N(0; \sigma^2_{\alpha\gamma})$, $(\beta\gamma)_{jk} \sim i.i.d.\ N(0; \sigma^2_{\beta\gamma})$, $(\alpha\beta\gamma)_{ijk} \sim i.i.d.\ N(0; \sigma^2_{\alpha\beta\gamma})$, and $\epsilon_{ijkh} \sim i.i.d.\ N(0; \sigma^2_e)$, are mutually independent variables, with $E(y_{ijk}) = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}$ and $Var(y_{ijk}) = \sigma^2_\epsilon$.

3.4.2 Hypothesis Testing on a Three-way mixed model

Assuming factors A and B are fixed, while factor C is random, the hypotheses of interest on the model (3.5) are given by:

$$\begin{aligned}
H_0^A : \alpha_i = 0 &\qquad \textit{tested} \quad \textit{by} \quad F = \frac{MS_A}{MS_{AC}}, \\
H_0^B : \beta_j = 0 &\qquad \textit{tested} \quad \textit{by} \quad F = \frac{MS_B}{MS_{BC}}, \\
H_0^C : \sigma^2_C = 0 &\qquad \textit{no} \quad \textit{valid} \quad F - \textit{test}, \\
H_0^{AB} : (\alpha\beta)_{ij} = 0 &\qquad \textit{tested} \quad \textit{by} \quad F = \frac{MS_{AB}}{MS_{ABC}}, \\
H_0^{AC} : \sigma^2_{AC} = 0 &\qquad \textit{tested} \quad \textit{by} \quad F = \frac{MS_{AC}}{MS_{ABC}}, \\
H_0^{BC} : \sigma^2_{BC} = 0 &\qquad \textit{tested} \quad \textit{by} \quad F = \frac{MS_{BC}}{MS_{ABC}}, \\
H_0^{ABC} : \sigma^2_{ABC} = 0 &\qquad \textit{tested} \quad \textit{by} \quad F = \frac{MS_{ABC}}{MS_E}.
\end{aligned} \tag{3.6}$$

Provided a valid F-test exists, the main effects of fixed factors A and B in hypotheses (3.6) are tested by the interaction of the fixed factor involved with the random factor, while the main effects the random factor C does not have an exact F-test. The interaction effects of any two factors are tested by the full interaction, whereas the full interaction effect is tested by the error. It is interesting to note that for all sources of variability in a fixed-effects model (when all factors are deemed fixed), valid F-tests for fixed-effects models exist and are tested by the error (i.e., using MSE in the denominator of F). However, for some random-effects and mixed models, this is not always the case. There are situations where no exact F-tests are readily available for random and mixed effects, in which cases, the approximate F-tests are a good alternative (Kuehl, 2000). However, some computer programs, such as R and SAS, do return these F-tests. For instance, if we assume factors A and B are random, while taking factor C as fixed in (3.5), we will have no valid F-test for testing the main effect of fixed factor C, $H_0^C : \gamma_k = 0$.

When the data is unbalanced, the analysis of variance in the hypotheses (3.6) is achieved through equating the mean sum of squares to their expected values. Considering a three-way mixed linear model with factors A and B assumed fixed and factor C assumed random, we define the uncorrected sum of squares found in the model unbalanced model as:

$$\begin{aligned} T_A &= \sum_{i=1}^{a} \frac{y_{i...}^2}{n_{i..}}, \qquad T_B = \sum_{j=1}^{b} \frac{y_{.j..}^2}{n_{.j.}}, \qquad T_C = \sum_{k=1}^{c} \frac{y_{..k.}^2}{n_{..k}}, \\ T_{AB} &= \sum_{i=1}^{a}\sum_{j=1}^{b} \frac{y_{ij..}^2}{n_{ij.}}, \qquad T_{AC} = \sum_{i=1}^{a}\sum_{k=1}^{c} \frac{y_{i.k.}^2}{n_{i.k}}, \qquad T_{BC} = \sum_{j=1}^{b}\sum_{k=1}^{c} \frac{y_{.jk.}^2}{n_{.jk}}, \\ T_{ABC} &= \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c} \frac{y_{ijk.}^2}{n_{ijk}}, \qquad T_\mu = \frac{y_{....}^2}{n_{....}}, \qquad T_\epsilon = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c}\sum_{h=1}^{n_{ijk}} y_{ijkh}^2. \end{aligned} \tag{3.7}$$

By correcting for the mean in (3.7), the sum of squares in the unbalanced linear mixed model, (3.5) are calculated as follows.

$$SS_A = T_A - T_\mu = \sum_{i=1}^{a} \frac{y_{i...}^2}{n_{i..}} - \frac{y_{....}^2}{n_{....}}, \tag{3.8}$$

$$SS_B = T_B - T_\mu = \sum_{j=1}^{b} \frac{y_{.j..}^2}{n_{.j.}} - \frac{y_{....}^2}{n_{....}}, \tag{3.9}$$

$$SS_C = T_C - T_\mu = \sum_{k=1}^{c} \frac{y_{..k.}^2}{n_{..k}} - \frac{y_{....}^2}{n_{....}}, \tag{3.10}$$

$$\begin{aligned} SS_{AB} =& T_{AB} - T_A - T_B + T_\mu \\ =& \sum_{i=1}^{a}\sum_{j=1}^{b} \frac{y_{ij..}^2}{n_{ij.}} - \sum_{i=1}^{a} \frac{y_{i...}^2}{n_{i..}} - \sum_{j=1}^{b} \frac{y_{.j..}^2}{n_{.j.}} + \frac{y_{....}^2}{n_{....}}, \end{aligned} \tag{3.11}$$

$$\begin{aligned} SS_{AC} =& T_{AC} - T_A - T_C + T_\mu \\ =& \sum_{i=1}^{a}\sum_{k=1}^{c} \frac{y_{i.k.}^2}{n_{i.k}} - \sum_{i=1}^{a} \frac{y_{i...}^2}{n_{i..}} - \sum_{k=1}^{c} \frac{y_{..k.}^2}{n_{..k}} + \frac{y_{....}^2}{n_{....}}, \end{aligned} \tag{3.12}$$

$$\begin{aligned} SS_{BC} =& T_{BC} - T_B - T_C + T_\mu \\ =& \sum_{j=1}^{b}\sum_{k=1}^{c} \frac{y_{.jk.}^2}{n_{.jk}} - \sum_{j=1}^{b} \frac{y_{.j..}^2}{n_{.j.}} - \sum_{k=1}^{c} \frac{y_{..k.}^2}{n_{..k}} + \frac{y_{....}^2}{n_{....}}, \end{aligned} \tag{3.13}$$

$$\begin{aligned} SS_{ABC} =& T_{ABC} - T_A - T_B - T_C + T_\mu \\ =& \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c} \frac{y_{ijk.}^2}{n_{ijk}} - \sum_{i=1}^{a} \frac{y_{i...}^2}{n_{i..}} - \sum_{j=1}^{b} \frac{y_{.j..}^2}{n_{.j.}} - \sum_{k=1}^{c} \frac{y_{..k.}^2}{n_{..k}} + \frac{y_{....}^2}{n_{....}}, \end{aligned} \tag{3.14}$$

$$\begin{aligned} SSE =& T_\epsilon - T_{ABC} \\ =& \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c}\sum_{h=1}^{n_{ijk}} y_{ijkh}^2 - \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c} \frac{y_{ijk.}^2}{n_{ijk}}. \end{aligned} \tag{3.15}$$

It is worthy to note that unbalanced data produces different types of sums of squares. Some existing statistical packages, such as R and SAS, provide a number of approaches for computing the sum of squares and testing hypotheses (Shaw and Mitchell-Olds, 1993). In the SAS system, these sum of squares are designated as type I (relevant for balanced data), type II (for factorial designs without interaction), type III (relevant for unbalanced data), and type IV (relevant for factorial designs with missing cells) sums of squares. We illustrate the definition of these sums of squares using a three-way factorial design with full interaction. Type I (sequential) sum of squares measures the amount of additional variation explained by the model when a term is added to the model (i.e. $SS(A|1)$, $SS(B|1,A)$, $SS(C|1,A,B)$, or $SS(AB|1,A,B,C)$. Type II sum of squares measures the amount of variation contributed by a term to the model when all other terms are included except terms that contain the effect being tested (i.e. $SS(A|1,B,C,BC)$ or $SS(BC|1,A,B,C,AB,AC)$). Type III (partial) sum of squares measures the amount of variation contributed by a term to the model when all other terms are included (i.e. $SS(B|1, A, C, AB, AC, BC, ABC)$, or $SS(AB|1, A, B, C, AC, BC, ABC)$). When the data is balanced, type I, II and III sums of squares produce the same result. However, the types of sums of squares are unequal for unbalanced data. When the data are unbalanced and have no missing cells, type III sum of squares is recommended by many statistical authors. Milliken and Johnson (2009) cautioned about the use of type III tests when there are missing cells in the data as this produces invalid results due to a lack of essential information about the missing cell means.

For random and linear mixed models, variance components would be estimated by equating the observed sums of squares to their expected values, which are of the quadratic forms derived using the brute force method (Searle and Gruber, 2017). This involves substituting the equation (3.5) into the mean squares and using the appropriate assumptions of the model to evaluate the expectations. For a random model example, using the definition of SS_A from (3.8), we derive $E(SS_A) = E(T_A) - E(T_\mu)$ by substituting model (3.5) into T_A and T_μ of (3.7), and then take expectations. $E(T_A)$ is derived in (3.16) - (3.17) as follows.

$$\begin{aligned} y_{i...} = \sum_{j=1}^{b}\sum_{k=1}^{c}\sum_{h=1}^{n_{ijk}} y_{ijkh} &= n_{i..}\mu + n_{i..}\alpha_i + \sum_{j=1}^{b} n_{ij.}\beta_j + \sum_{k=1}^{c} n_{i.k}\gamma_k \\ &+ \sum_{j=1}^{b}\sum_{k=1}^{c} n_{ij.}(\alpha\beta)_{ij} + \sum_{j=1}^{b}\sum_{k=1}^{c} n_{i.k}(\alpha\gamma)_{ik} + \sum_{j=1}^{b}\sum_{k=1}^{c} n_{ijk}(\beta\gamma)_{jk} \\ &+ \sum_{j=1}^{b}\sum_{k=1}^{c} n_{ijk}(\alpha\beta\gamma)_{ijk} + \epsilon_{i...} \end{aligned} \tag{3.16}$$

Upon squaring, dividing by $n_{i..}$ and simplifying (3.16) before taking expectation, we have the

$$\begin{aligned} E(T_A) &= \sum_{i=1}^{a} E\left(\frac{y_{i...}^2}{n_{i}..}\right) \\ &= N\mu^2 + N\sigma_\alpha^2 + \sum_{i=1}^{a}\frac{\sum_{j=1}^{b} n_{ij.}^2}{n_i..}\sigma_\beta^2 + \sum_{i=1}^{a}\frac{\sum_{k=1}^{c} n_{i.k}^2}{n_i..}\sigma_\gamma^2 \\ &+ \sum_{i=1}^{a}\frac{\sum_{j=1}^{b}\sum_{k=1}^{c} n_{ij.}^2}{n_i..}\sigma_{\alpha\beta}^2 + \sum_{i=1}^{a}\frac{\sum_{j=1}^{b}\sum_{k=1}^{c} n_{i.k}^2}{n_i..}\sigma_{\alpha\gamma}^2 + \sum_{i=1}^{a}\frac{\sum_{j=1}^{b}\sum_{k=1}^{c} n_{ijk}^2}{n_i..}\sigma_{\beta\gamma}^2 \\ &+ \sum_{i=1}^{a}\frac{\sum_{j=1}^{b}\sum_{k=1}^{c} n_{ijk}^2}{n_i..}\sigma_{\alpha\beta\gamma}^2 + a\sigma_\epsilon^2. \end{aligned} \tag{3.17}$$

Similarly, $E(T_\mu)$ is derived following the steps below:

$$\begin{aligned}
y_{....} &= \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c}\sum_{h=1}^{n_{ijk}} y_{ijkh} \\
&= N\mu + \sum_{i=1}^{a} n_{i..}\alpha_i + \sum_{j=1}^{b} n_{.j.}\beta_j + \sum_{k=1}^{c} n_{..k}\gamma_k \\
&+ \sum_{i=1}^{a}\sum_{j=1}^{b} n_{ij.}(\alpha\beta)_{ij} + \sum_{i=1}^{a}\sum_{k=1}^{c} n_{i.k}(\alpha\gamma)_{ik} + \sum_{j=1}^{b}\sum_{k=1}^{c} n_{.jk}(\beta\gamma)_{jk} \\
&+ \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c} n_{ijk}(\alpha\beta\gamma)_{ijk} + \epsilon_{....} \qquad (3.18)
\end{aligned}$$

$$\begin{aligned}
E(T_\mu) &= E\left(\frac{y^2_{....}}{N}\right) \\
&= N\mu^2 + \frac{\sum_{i=1}^{a} n^2_{i..}}{N}\sigma^2_\alpha + \frac{\sum_{j=1}^{b} n^2_{.j.}}{N}\sigma^2_\beta \\
&+ \frac{\sum_{k=1}^{c} n^2_{..k}}{N}\sigma^2_\gamma + \frac{\sum_{i=1}^{a}\sum_{j=1}^{b} n^2_{ij.}}{N}\sigma^2_{\alpha\beta} + \frac{\sum_{i=1}^{a}\sum_{k=1}^{c} n^2_{i.k}}{N}\sigma^2_{\alpha\gamma} \\
&+ \frac{\sum_{j=1}^{b}\sum_{k=1}^{c} n^2_{.jk}}{N}\sigma^2_{\beta\gamma} + \frac{\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c} n^2_{ijk}}{N}\sigma^2_{\alpha\beta\gamma} + \sigma^2_\epsilon \qquad (3.19)
\end{aligned}$$

Hence, using (3.17) and (3.19), the expected value of $E(SS_A)$ is

$$\begin{aligned}
E(SS_A) &= E(T_A) - E(T_\mu) \\
&= \left(N - \frac{\sum_{i=1}^{a} n^2_{i..}}{N}\right)\sigma^2_\alpha + \left(\sum_{i=1}^{a}\frac{\sum_{j=1}^{b} n^2_{ij.}}{n_{i}..} - \frac{\sum_{j=1}^{b} n^2_{.j.}}{N}\right)\sigma^2_\beta \\
&+ \left(\sum_{i=1}^{a}\frac{\sum_{k=1}^{c} n^2_{i.k}}{n_{i}..} - \frac{\sum_{k=1}^{c} n^2_{..k}}{N}\right)\sigma^2_\gamma + \left(\sum_{i=1}^{a}\frac{\sum_{j=1}^{b}\sum_{k=1}^{c} n^2_{ij.}}{n_{i}..} - \frac{\sum_{i=1}^{a}\sum_{j=1}^{b} n^2_{ij.}}{N}\right)\sigma^2_{\alpha\beta} \\
&+ \left(\sum_{i=1}^{a}\frac{\sum_{j=1}^{b}\sum_{k=1}^{c} n^2_{i.k}}{n_{i}..} - \frac{\sum_{i=1}^{a}\sum_{k=1}^{c} n^2_{i.k}}{N}\right)\sigma^2_{\alpha\gamma} \\
&+ \left(\sum_{i=1}^{a}\frac{\sum_{j=1}^{b}\sum_{k=1}^{c} n^2_{ijk}}{n_{i}..} - \frac{\sum_{j=1}^{b}\sum_{k=1}^{c} n^2_{.jk}}{N}\right)\sigma^2_{\beta\gamma} \\
&+ \left(\sum_{i=1}^{a}\frac{\sum_{j=1}^{b}\sum_{k=1}^{c} n^2_{ijk}}{n_{i}..} - \frac{\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c} n^2_{ijk}}{N}\right)\sigma^2_{\alpha\beta\gamma} + (a-1)\sigma^2_\epsilon \qquad (3.20)
\end{aligned}$$

The rest of the expected sum of squares are derived in the same way. As noted by Searle and Gruber (2017), $E(SS_A)$ contains a non-zero coefficient for every variance component, including the ones without factor A influence, which is not the case when data is balanced.

Suppose we consider a mixed model case, with factors A and B considered as fixed, while factor C is considered random. Since random factors have zero means and covariances, taking expectations of T_A and T_B, as in (3.17), will render some of the terms involving the random effect γ to zero. The expectation of fixed effects will not be zero; for example, $E(n_i\alpha_i^2) = n_i\alpha_i^2$ and $E(n_i\beta_j^2) = n_j\beta_i^2$. As a result, $E(T_A)$ in (3.17) and $E(T_\mu)$ in (3.19) will be modified to, respectively,

$$\begin{aligned} E(T_A) &= \sum_{i=1}^{a} E\left(\frac{y_{i...}^2}{n_{i..}}\right) \\ &= N\mu^2 + \sum_{i=1}^{a} n_{i..}\alpha_i^2 + 2\mu\sum_{i=1}^{a} n_{i..}\alpha_i + \sum_{i=1}^{a}\frac{\sum_{j=1}^{b} n_{ij.}^2\beta_j^2}{n_{i..}} + 2\mu\sum_{i=1}^{a}\sum_{j=1}^{b} n_{ij.}\beta_j \\ &+ 2\mu\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c} n_{ij.}(\alpha\beta)_{ij} + \sum_{i=1}^{a}\frac{\sum_{k=1}^{c} n_{i.k}^2}{n_{i..}}\sigma_\gamma^2 + \sum_{i=1}^{a}\frac{\sum_{j=1}^{b}\sum_{k=1}^{c} n_{i.k}^2}{n_{i..}}\sigma_{\alpha\gamma}^2 \\ &+ \sum_{i=1}^{a}\frac{\sum_{j=1}^{b}\sum_{k=1}^{c} n_{ijk}^2}{n_{i..}}\sigma_{\beta\gamma}^2 + \sum_{i=1}^{a}\frac{\sum_{j=1}^{b}\sum_{k=1}^{c} n_{ijk}^2}{n_{i..}}\sigma_{\alpha\beta\gamma}^2 + a\sigma_\epsilon^2 \end{aligned} \tag{3.21}$$

$$\begin{aligned} E(T_\mu) &= E\left(\frac{y_{....}^2}{N}\right) \\ &= N\mu^2 + \frac{\left(\sum_{i=1}^{a} n_{i..}\alpha_i\right)^2}{N} + 2\mu\sum_{i=1}^{a} n_{i..}\alpha_i + \frac{\left(\sum_{j=1}^{b} n_{ij.}\beta_j\right)^2}{N} + 2\mu\sum_{i=1}^{a}\sum_{j=1}^{b} n_{ij.}\beta_j \\ &+ 2\mu\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c} n_{ij.}(\alpha\beta)_{ij} + \frac{\sum_{k=1}^{c} n_{..k}^2}{N}\sigma_\gamma^2 + \frac{\sum_{i=1}^{a}\sum_{k=1}^{c} n_{i.k}^2}{N}\sigma_{\alpha\gamma}^2 \\ &+ \frac{\sum_{j=1}^{b}\sum_{k=1}^{c} n_{.jk}^2}{N}\sigma_{\beta\gamma}^2 + \frac{\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c} n_{ijk}^2}{N}\sigma_{\alpha\beta\gamma}^2 + \sigma_\epsilon^2 \end{aligned} \tag{3.22}$$

Thus, the expected value of SS_A is

$$\begin{aligned} E(SS_A) &= E(T_A) - E(T_\mu) \\ &= Q_1 + \left(\sum_{i=1}^{a}\frac{\sum_{k=1}^{c} n_{i.k}^2}{n_{i..}} - \frac{\sum_{k=1}^{c} n_{..k}^2}{N}\right)\sigma_\gamma^2 + \left(\sum_{i=1}^{a}\frac{\sum_{j=1}^{b}\sum_{k=1}^{c} n_{i.k}^2}{n_{i..}} - \frac{\sum_{i=1}^{a}\sum_{k=1}^{c} n_{i.k}^2}{N}\right)\sigma_{\alpha\gamma}^2 \\ &+ \left(\sum_{i=1}^{a}\frac{\sum_{j=1}^{b}\sum_{k=1}^{c} n_{ijk}^2}{n_{i..}} - \frac{\sum_{j=1}^{b}\sum_{k=1}^{c} n_{.jk}^2}{N}\right)\sigma_{\beta\gamma}^2 \\ &+ \left(\sum_{i=1}^{a}\frac{\sum_{j=1}^{b}\sum_{k=1}^{c} n_{ijk}^2}{n_{i..}} - \frac{\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c} n_{ijk}^2}{N}\right)\sigma_{\alpha\beta\gamma}^2 + (a-1)\sigma_\epsilon^2 \end{aligned} \tag{3.23}$$

where

$$Q_1 = \sum_{i=1}^{a} n_{i..}\alpha_i^2 - \frac{\left(\sum_{i=1}^{a} n_{i..}\alpha_i\right)^2}{N} + \sum_{i=1}^{a}\frac{\sum_{j=1}^{b} n_{ij.}^2\beta_j^2}{n_{i..}} - \frac{\left(\sum_{j=1}^{b} n_{ij.}\beta_j\right)^2}{N}. \tag{3.24}$$

The other fixed factors are dealt with following the same process as in (3.21) - (3.24). However, the expected values of the sum of square terms in mixed models, in this case, $E(SS_A - SS_C)$ will always include functions of fixed effects that are difficult to manipulate using linear combinations of terms. The two approaches that Searle and Gruber (2017) suggested for this challenge are: (1) reduce the model to purely random by ignoring and eliminating fixed effects; and (2) consider fixed effects as random, and treat the model as entirely random) usually result in biased variance components. Eze and Nwankwo (2016) proposed an alternative approach to circumvent this problem by adjusting the denominators in the F-test for the main effects when the factors are mixed.

Table 3.1 summarises the degrees of freedom, sum of squares and F-ratios for the analysis of variance in an unbalanced three-way linear mixed model (3.5) with fixed factors A and B, random factor C, where $\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c} n_{ijk} = N$.

Table 3.1: ANOVA Table for Unbalanced Three-way Mixed Model

Source	df	Sum of Squares	F
Fixed A	a-1	SS_A	$\frac{MS_A}{MS_{ABC}}$
Fixed B	b-1	SS_B	$\frac{MS_B}{MS_{ABC}}$
Random C	c-1	SS_C	$\frac{MS_C}{MS_E}$
A*B	(a-1)(b-1)	SS_{AB}	$\frac{MS_{A*B}}{MS_E}$
A*C	(a-1)(c-1)	SS_{AC}	$\frac{MS_{A*C}}{MS_E}$
B*C	(b-1)(c-1)	SS_{BC}	$\frac{MS_{B*C}}{MS_E}$
A*B*C	(a-1)(b-1)(c-1)	SS_{ABC}	$\frac{MS_{A*B*C}}{MS_E}$
Error	abc(N-1)	SSE	
Total	abcN-1	SST	

When balanced data is considered, with all $n_{ijk} = n$, the coefficients of variance components in (3.19) and other expected sum of squares summarised in ANOVA Table 3.1 reduce to the balanced kindred displayed in ANOVA Table 3.2.

The ANOVA Table 3.2 is compatible with the mixed model procedure suggested by McLean *et al.* (1991) and also recommended by SAS, to include the interactions between fixed and random factors in estimating the expected mean square errors (EMS) for random effects and analysis of unbalanced data. Based on the expected mean squares, the appropriate F-tests can be performed on the fixed, random and interaction effects.

Table 3.2: Three-way Mixed ANOVA

Source	df	Expected Mean Squares
A	$a-1$	$\sigma_\epsilon^2 + n\sigma_{\alpha\beta\gamma}^2 + b * n\sigma_{\alpha\gamma}^2 + c * n\sigma_{\alpha\beta}^2 + b * c * n \sum \frac{\alpha_i^2}{a-1}$
B	$b-1$	$\sigma_\epsilon^2 + n\sigma_{\alpha\beta\gamma}^2 + a * n\sigma_{\beta\gamma}^2 + c * n\sigma_{\alpha\beta}^2 + a * c * n \sum \frac{\beta_j^2}{b-1}$
C	$c-1$	$\sigma_\epsilon^2 + n\sigma_{\alpha\beta\gamma}^2 + a * n\sigma_{\beta\gamma}^2 + b * n\sigma_{\alpha\gamma}^2 + a * b * n\sigma_\gamma^2$
A*B	$(a-1)(b-1)$	$\sigma_\epsilon^2 + n\sigma_{\alpha\beta\gamma}^2 + c * n\frac{\sum\sum(\alpha\beta)_{ij}^2}{(a-1)(b-1)}$
A*C	$(a-1)(c-1)$	$\sigma_\epsilon^2 + n\sigma_{\alpha\beta\gamma}^2 + b * n\sigma_{\alpha\gamma}^2$
B*C	$(b-1)(c-1)$	$\sigma_\epsilon^2 + n\sigma_{\alpha\beta\gamma}^2 + a * n\sigma_{\beta\gamma}^2$
A*B*C	$(a-1)(b-1)(c-1)$	$\sigma_\epsilon^2 + n\sigma_{\alpha\beta\gamma}^2$
Error (FFR)	$a * b * c(n-1)$	σ_ϵ^2

3.5 Treatment Structure

In experimental design, a treatment structure is defined as a format in which the effects of various treatments, or combinations of treatments or factors under investigation are arranged (Milliken and Johnson, 2009). In a treatment analysis, the experiment deliberately imposes a treatment on a group of experimental units (objects or subjects) to observe the response. In other words, treatment is something that a researcher administers to experimental units. These types of experiments are generally referred to as factorial arrangement treatment structures. For example, assume an experiment involving factors A and B, with a and b levels, respectively, such that each replicate contains all ab treatment combinations. Arranging these factors in a treatment structure results in a crossed-factors scenario.

3.5.1 Two-way Treatment Structure

Suppose we have f_a fixed levels of factor A and f_b fixed levels of factor B. Let r_a be random levels of factor A and r_b be random levels of factor B. In general, if there are n replicates in the experiment, the two-way treatment structure model, as suggested by Njuho and Milliken

(2009), is in the form

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk} \ , \qquad i\text{=}1,\ 2,\ ...,a\ ;j\text{=}1,\ 2,\ ...,b\ ;k = 1,\ 2,\ ...,\ n. \tag{3.25}$$

where α_1, α_2, ..., α_f $(f_a < a)$ are real-valued constants corresponding to the fixed effect levels of factor A and α_{f+1}, α_{f+2}, ..., α_a $(a = f_a + r_a)$ are random effect levels of factor A. Also, β_1, β_2, ..., β_f $(f_b < b)$ are real-valued constants corresponding to the fixed effect levels of factor B and β_{f+1}, β_{f+2}, ..., β_b $(b = f_b + r_b)$ are random effect levels of factor B. μ is the overall mean, and y_{ijk} is the k^{th} observation receiving the i^{th} treatment of factor A and the j^{th} treatment of factor B. We assume that $\alpha_{f+i} \sim i.i.d\ N(0, \sigma^2_\alpha)$, $\beta_{f+j} \sim i.i.d.\ N(0, \sigma^2_\beta)$, $\epsilon_{ijk} \sim i.i.d.\ N(0, \sigma^2_\epsilon)$ and α_{f+i}, β_{f+j} and ϵ_{ijk} are pairwise independent, i = *1, 2, ...,* f_{a+1}*,* f_{a+2}*, ...,* a*,* j = *1, 2, ...,* f_{b+1}*,* f_{b+2}*, ...,* b*,* k*=1, 2, ...,* n*.*

From equation (3.25), the fixed and random levels of factors A and B are used to create partitioned samples in the dataset. Partitioning the model (3.25) in line with the treatment combinations gives us the following four models as shown in Table 3.3.

Table 3.3: Two-way Treatment Structure Models

		Factor B	
		Fixed	Random
Factor A	Fixed	FF	FR
	Random	RF	RR

The models obtained from this treatment structure are fixed-fixed (FF), fixed-random (FR), random-fixed (RF) and random-random (RR) effects. Such models consist of treatment combinations of both fixed and random levels. Two of the four models are given as examples below.

The **Fixed-Fixed (FF) Model** is given by,

$$y_{FF_{ijk}} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{FF_{ijk}}, \tag{3.26}$$

where $\mu = \mu_{FF}$ is the overall mean of the fixed effects; α_1, α_2, ..., α_{fa} and β_1, β_2, ..., β_{fb} are real-valued constants corresponding to the fixed effect levels of factors A and B, respectively; $(\alpha\beta)_{11}$, $(\alpha\beta)_{12}$, ..., $(\alpha\beta)_{fafb}$ are the $A \times B$ interaction effects amongst the fixed levels of the two factors; and the error term $\epsilon_{FF_{ijk}} \sim$ N(0, $\sigma^2_{\epsilon_{FF}}$).

Fixed-Random (FR) Model is expressed as,

$$y_{FR_{ijk}} = \phi_i + b_{FR_j} + c_{FR_{ij}} + \epsilon_{FR_{ijk}}. \tag{3.27}$$

In the Fixed-Random (FR) model, ϕ_i ($i = 1, 2,$ ..., f_a) denotes the mean of the i^{th} fixed effect level of factor A averaged over the random effect levels of factor B;
$b_{FR_j} \sim \ N(0, \sigma^2_{b_{FR}})$ for $j = f_b + 1,\ f_b + 2, ...,\ f_b + r_b = b$, denotes the random effect levels of factor B; $c_{FR_{ij}}$ represents the interaction between the fixed effect levels of A and the random effect levels of B, where $c_{FR_{ij}} \sim \ N(0, \sigma^2_{c_{FR}})$, and $\epsilon_{FR_{ijk}} \sim \ N(0, \sigma^2_{\epsilon_{FR}})$.

The **Random-Fixed (RF) Model** is expressed as,

$$y_{RF_{ijk}} = a_{RF_i} + \omega_j + c_{RF_{ij}} + \epsilon_{RF_{ijk}}. \tag{3.28}$$

In the Random-Fixed (RF) effects model, $a_{RF_i} \sim N(0, \sigma^2_{a_{RF}})$ for $i = f_a{+}1,\ f_a{+}2, ...,\ f_a{+}r_a = a$ denotes the random effect levels of factor A; the mean of the j^{th} fixed effect level of factor B averaged over the random effect levels of factor A is denoted by ω_j ($j = 1, 2,$ ..., f_b); $c_{RF_{ij}}$ represents the interaction between the random effect levels of A and the fixed effect levels of B, where $c_{RF_{ij}} \sim N(0, \sigma^2_{c_{RF}})$; and $\epsilon_{RF_{ijk}} \sim$ N(0, $\sigma^2_{\epsilon_{RF}}$).

Similarly, the **Random-Random (RR) Model** is given by,

$$y_{RR_{ijk}} = \mu_{RR} + a_{RR_i} + b_{RR_j} + c_{RR_{ij}} + \epsilon_{RR_{ijk}}. \tag{3.29}$$

In the Random-Random (RR) model for the factors A and B, μ_{RR} is the overall mean of the random effects, a_{RR_i} for $i = f_a{+}1,\ f_a{+}2,$ ..., $f_a{+}r_a = a$, denotes the effects of the random levels of factor A where $a_{RR_i} \sim$ N(0, $\sigma^2_{a_{RR}}$), $b_{RR_j} \sim N(0, \sigma^2_{b_{RR}})$ for $j = f_b + 1,\ f_b + 2,$ ..., $f_b + r_b = b$, denotes the random effect levels of factor B; $c_{RR_{ij}}$ represents the interaction between the random effect levels of A and the random effect levels of B, and $c_{RR_{ij}} \sim N(0, \sigma^2_{c_{RR}})$, and $\epsilon_{RR_{ijk}} \sim$ N(0, $\sigma^2_{\epsilon_{RR}}$).

3.5.2 Three-way Treatment Structure

Analogous to the two-way treatment structure, a three-way treatment structure is built, culminating into eight partitioned models as summarised in Table 3.4.

Table 3.4: Three-way Treatment Structure Models

Factors A*B					
		Fixed-Fixed	Fixed-Random	Random-Fixed	Random-Random
Factor C	Fixed	FFF	FRF	RFF	RRF
	Random	FFR	FRR	RFR	RRR

The number of models at this stage of the three-way treatment structure is 2^3. Intuitively, we can build the n-way treatment structures and use the same approach to build and interpret the respective partitioned models.

The **Fixed-Fixed-Fixed** (**FFF**) model for the factors A, B and C is constructed as

$$y_{FFF_{ijkh}} = \mu_{FFF} + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk} + \epsilon_{FFF_{ijkh}}, \tag{3.30}$$

where μ_{FFF} is the overall mean of the fixed effects; α_1, α_2, ..., α_{fa}, β_1, β_2, ..., β_{fb} and γ_1, γ_2, ..., γ_{fc} are unknown parameters corresponding to the fixed-effect levels of factors A, B and C, respectively; $(\alpha\beta)_{11}$, $(\alpha\beta)_{12}$, ..., $(\alpha\beta)_{fafb}$ denote the interaction effects among the fixed levels of factors A and B; $(\alpha\gamma)_{11}$, $(\alpha\gamma)_{12}$, ..., $(\alpha\gamma)_{fafc}$ denote the interaction effects among the fixed levels of factor A and C; $(\beta\gamma)_{11}$, $(\beta\gamma)_{12}$, ..., $(\beta\gamma)_{fbfc}$ denote the interaction effects among the fixed levels of factor B and C; $(\alpha\beta\gamma)_{111}$, $(\alpha\beta\gamma)_{112}$, ..., $(\alpha\beta\gamma)_{fafbfc}$ denote the interaction effects among the fixed levels of factors A, B and C; and $\epsilon_{FFF_{ijkh}} \sim i.i.d\ N(0, \sigma^2_{\epsilon_{FFF}})$ is the error term.

The **Random-Random-Random** (**RRR**) model for the factors A, B and C, is similarly constructed. Let a, b and c be levels of factors A, B and C, respectively. Also let m be $(\alpha\beta)$, p be $(\alpha\gamma)$, q be $(\beta\gamma)$ and g be $(\alpha\beta\gamma)$ interpreted as interactions effects.

$$y_{RRR_{ijkh}} = \mu_{RRR} + a_{RRR_i} + b_{RRR_j} + c_{RRR_k} + m_{RRR_{ij}} + p_{RRR_{ik}} + q_{RRR_{jk}} + g_{RRR_{ijk}} + \epsilon_{RRR_{ijkh}}, \tag{3.31}$$

where where μ_{RRR} is the overall mean of the random effects of the three factors; a_{RRR_i} ($i = f_a$+1, $f_a + 2$, ..., $f_a + r_a = a$) denotes the effects of the random levels of factor A with $a_{RRR_i} \sim$ N(0,

$\sigma^2_{a_{RRR}}$); b_{RRR_j} ($j = f_b + 1,\ f_b + 2,\ ...,\ f_b + r_b = b$) denotes the random effect levels of factor B with $b_{RRR_j} \sim N(0, \sigma^2_{b_{RRR}})$; c_{RRR_k} ($k = f_c + 1,\ f_c + 2,\ ...,\ f_c + r_c = c$) represents the random effect levels of factor C with $c_{RRR_k} \sim N(0, \sigma^2_{c_{RRR}})$; $m_{RRR_{ij}}$ ($i = f_a + 1,\ f_a + 2, ...,\ f_a + r_a = a$; $j = f_b + 1,\ f_b + 2,\ ...,\ f_b + r_b = b$) represents the interaction effects of random levels of factor A and factor B, with $m_{RRR_{ij}} \sim N(0, \sigma^2_{m_{RRR}})$; $p_{RRR_{ik}}$ ($i = f_a + 1,\ f_a + 2,\ ...,\ f_a + r_a = a$; $k = f_c + 1,\ f_c + 2,\ ...,\ f_c + r_c = c$) represents the interaction effects of random levels of factor A and factor C, with $p_{RRR_{ik}} \sim N(0, \sigma^2_{p_{RRR}})$; $q_{RRR_{jk}}$ ($j = f_b + 1,\ f_b + 2,\ ...,\ f_b + r_b = b$; $k = f_c + 1,\ f_c + 2,\ ...,\ f_c + r_c = c$), represents the interaction effects of random levels of factor B and factor C, with $q_{RRR_{jk}} \sim N(0, \sigma^2_{q_{RRR}})$; $g_{RRR_{ijk}}$ ($i = f_a + 1,\ f_a + 2,\ ...,\ f_a + r_a = a$; $j = f_b + 1,\ f_b + 2,\ ...,\ f_b + r_b = b$; $k = f_c + 1,\ f_c + 2,\ ...,\ f_c + r_c = c$) represents the interaction effects of random levels of factor A, B and C, with $q_{RRR_{ijk}} \sim N(0, \sigma^2_{q_{RRR}})$; and $\epsilon_{RRR_{ijk}} \sim$ N(0, $\sigma^2_{\epsilon_{RRR}}$) is the error term.

The **Fixed-Fixed-Random (FFR) model** and its components is constructed and interpreted as follows,

$$y_{FFR_{ijkh}} = \phi_{ij} + c_{FFR_k} + m_{FFR_{ij}} + p_{FFR_{ik}} + q_{FFR_{jk}} + g_{FFR_{ijk}} + \epsilon_{FFR_{ijkh}}, \tag{3.32}$$

where ϕ_{ij} ($i = 1, 2,\ ...,\ f_a$; $j = 1, 2,\ ...,\ f_b$) denotes the mean of the i^{th} fixed-effect levels of factor A and the j^{th} fixed-effect levels of factor B averaged over the random-effect levels of factor C ; c_{FFR_k} $k = f_c + 1,\ f_c + 2, ...,\ f_c + r_c = c$ represents the random effect levels of factor C, with $c_{FFR_k} \sim N(0, \sigma^2_{c_{FFR}})$; $m_{FFR_{ij}}$ ($i = 1, 2,\ ...,\ f_a$; $j = 1, 2,\ ...,\ f_b$) represents the interaction effects of fixed levels of factor A, fixed levels of factor B, with $m_{FFR_{ij}} \sim N(0, \sigma^2_{m_{FFR}})$; $p_{FFR_{ik}}$ ($i = 1, 2,$ ..., f_a; $k = f_c + 1,\ f_c + 2, ...,\ f_c + r_c = c$) represents the interaction effects of the fixed levels of factor A and random levels of factor B, with $p_{FFR_{ik}} \sim N(0, \sigma^2_{p_{FFR}})$; $q_{FFR_{jk}}$ ($j = 1, 2,\ ...,\ f_b$; $k = f_c + 1,\ f_c + 2,\ ...,\ f_c + r_c = c$) represents the interaction effects of fixed levels of factor B and random levels of factor C, with $q_{FFR_{jk}} \sim N(0, \sigma^2_{q_{FFR}})$; $g_{FFR_{ijk}}$ ($i = f_a + 1,\ f_a + 2,$ ..., $f_a + r_a = a$; $j = f_b + 1,\ f_b + 2, ...,\ f_b + r_b = b$; $k = f_c + 1,\ f_c + 2,\ ...,\ f_c + r_c = c$) represents the interaction effects of fixed levels of factor A, fixed levels of factor B and random levels of factor C, with $g_{FFR_{ijk}} \sim N(0, \sigma^2_{g_{FFR}})$; and $\epsilon_{FFR_{ijk}} \sim$ N(0, $\sigma^2_{\epsilon_{FFR}}$) is the error term.

The other five models of the three-way treatment structure are similarly constructed and listed below:

$$y_{FRF_{ijkh}} = \phi_{ik} + b_{FRF_j} + m_{FRF_{ij}} + p_{FRF_{ik}} + q_{FRF_{jk}} + g_{FRF_{ijk}} + \epsilon_{FRF_{ijkh}}; \tag{3.33}$$

$$y_{RFF_{ijkh}} = \phi_{jk} + a_{RFF_i} + m_{RFF_{ij}} + p_{RFF_{ik}} + q_{RFF_{jk}} + g_{RFF_{ijk}} + \epsilon_{RFF_{ijkh}}; \tag{3.34}$$

$$y_{RRF_{ijkh}} = a_{RRF_i} + b_{RRF_j} + \mu_k + m_{RRF_{ij}} + p_{RRF_{ik}} + q_{RRF_{jk}} + g_{RRF_{ijk}} + \epsilon_{RRF_{ijkh}}; \tag{3.35}$$

$$y_{FRR_{ijkh}} = \mu_i + b_{FRR_j} + c_{FRR_k} + m_{FRRF_{ij}} + p_{FRR_{ik}} + q_{FRR_{jk}} + g_{FRR_{ijk}} + \epsilon_{FRR_{ijkh}}; \tag{3.36}$$

$$y_{RFR_{ijkh}} = a_{RFR_i} + \mu_j + c_{RFR_k} + m_{RFRF_{ij}} + p_{RFR_{ik}} + q_{RFR_{jk}} + g_{RFR_{ijk}} + \epsilon_{RFR_{ijkh}}. \tag{3.37}$$

The interpretation of these models and their interactions effects is analogous to the Fixed-Fixed (FF) model or the two-way treatment structure models previously given.

3.5.3 Analysis of Variance for Three-way Structure Models

For each of the models (3.30) - (3.37), the analysis of variance (ANOVA) approach will be used to assess the variation in the response variable. Table 3.5 displays the ANOVA model for the Fixed-Fixed-Random (FFR) model for balanced data.

Table 3.5: Fixed-Fixed-Random (FFR) ANOVA

Source	df	Expected Mean Squares
A	$f_a - 1$	$\sigma^2_\epsilon + n\sigma^2_{\alpha\beta\gamma} + f_b * n\sigma^2_{\alpha\gamma} + r_c * n\sigma^2_{\alpha\beta} + f_b * r_c * n \sum \frac{\alpha_i^2}{f_a - 1}$
B	$f_b - 1$	$\sigma^2_\epsilon + n\sigma^2_{\alpha\beta\gamma} + f_a * n\sigma^2_{\beta\gamma} + r_c * n\sigma^2_{\alpha\beta} + f_a * r_c * n \sum \frac{\beta_j^2}{f_b - 1}$
C	$r_c - 1$	$\sigma^2_\epsilon + n\sigma^2_{\alpha\beta\gamma} + f_a * n\sigma^2_{\beta\gamma} + f_b * n\sigma^2_{\alpha\gamma} + f_a * f_b * n\sigma^2_\gamma$
A*B	$(f_a - 1)(f_b - 1)$	$\sigma^2_\epsilon + n\sigma^2_{\alpha\beta\gamma} + r_c * n\frac{\sum\sum(\alpha\beta)^2_{ij}}{(f_a-1)(f_b-1)}$
A*C	$(f_a - 1)(r_c - 1)$	$\sigma^2_\epsilon + n\sigma^2_{\alpha\beta\gamma} + f_b * n\sigma^2_{\alpha\gamma}$
B*C	$(f_b - 1)(r_c - 1)$	$\sigma^2_\epsilon + n\sigma^2_{\alpha\beta\gamma} + f_a * n\sigma^2_{\beta\gamma}$
A*B*C	$(f_a - 1)(f_b - 1)(r_c - 1)$	$\sigma^2_\epsilon + n\sigma^2_{\alpha\beta\gamma}$
Error (FFR)	$f_a * f_b * r_c(n - 1)$	σ^2_ϵ

In this ANOVA table, f_a and f_b are fixed levels of factors A and B, respectively, while r_c denotes the random levels of factor C, and n replications per cell. In the case of unbalanced data, the expected mean squares in Table 3.5 are derived from the reduction sum of squares, $(R(\boldsymbol{\beta}))$, for each corresponding source of variation (Searle and Gruber, 2017). The approach is applied to the rest of the models and thus, can be extended to n-way treatment structure.

3.5.4 Variables for Combined Model

The variables and combinations of effect levels used to build a combined model that has three factors, A, B and C, each having both fixed and random levels. Table 3.6 summarises the variable and partitioned model information required when formulating a combined model.

Table 3.6: Variables in a Combined Model

	Effect levels					
	Fixed parts of factors A, B & C			Random parts of factors A, B & C		
Type of model	FA	FB	FC	RA	RB	RC
FFF	✓	✓	✓	×	×	×
FRF	✓	×	✓	×	✓	×
RFF	×	✓	✓	✓	×	×
FRF	×	×	✓	✓	✓	×
FFR	✓	✓	×	×	×	✓
FRR	✓	×	×	×	✓	✓
RFR	×	✓	×	✓	×	✓
RRR	×	×	×	✓	✓	✓

✓ denotes the presence of the effect;
× denotes the absence of the effect;
F denotes a fixed effect, and
R denotes a random effect.

In all the eight (8) possible partitioned models constructed, the notation FA denotes the fixed effect levels of factor A, whereas RA denotes the random effect levels of factor A. The same interpretation is used for the rest of the factors. A combined model is achieved by syncretising the appropriate pieces of information (degrees of freedom and sum of squares) from the partitioned models as shown in Table 3.6. When constructing the combined model in a three-way treatment structure, the sources of variation, degrees of freedom, and the models supplying the information are as given in Table 3.7.

Table 3.7: Information for the Combined Model ANOVA

Source	**df**	**Models supplying information**
FA (Model)	$4(f_a - 1)$	FFF, FRF, FFR and FRR
FB (Model)	$4(f_b - 1)$	FFF, RFF, FFR and RFR
FC (Model)	$4(f_c - 1)$	FFF, FRF, RFF and RRF
FA*FB*FC	$(f_a - 1)(f_b - 1)(f_c - 1)$	FFF
RA (Model)	$4(r_a - 1)$	RFF, RRF, RFR and RRR
RB (Model)	$4(r_a - 1)$	FRF, RRF, FRR and RRR
RC (Model)	$4(r_a - 1)$	FFR, FRR, RFR and RRR
RA*RB*RC (Model)	$(r_a - 1)(r_b - 1)(r_c - 1)$	RRR
Error	$f_a * f_b * f_c + f_a * f_c * r_b + f_b * f_c * r_a + f_c * r_a * r_b + f_a * f_b * r_c + f_a * r_b * r_c + f_b * r_a * r_c + r_a * r_b * r_c$	FFF, FRF, RFF, RRF, FFR, FRR, RFR and RRR

Similarly, all the three-way structure models can be built and interpreted. This study will attempt to install the partitioning approach in similar experimental designs, such as the completely randomised design (CRD), randomised complete block design (RCBD), and split-split-plot (SSP) design.

Considering the samples Y_{1i}, Y_{2i}, ..., Y_{ki}, to be samples used for the combined model above collected from a normal population with unknown mean μ_1, μ_2, ..., μ_k, respectively. The widely used method for testing $H_0 : \mu_1 = \mu_2 = ... = \mu_k$ is the analysis of variance (ANOVA) under the normality, equality of the variances and independence assumptions. The classical F test statistic of the ANOVA is expressed as a ratio of the between-group and within-group sums of squares of the reduced model under H_0 and the full model. Assuming equal group variances, the F test statistic is written as

$$F = \frac{\frac{SS_{treats}}{k-1}}{\frac{SSE}{N-k}}, \tag{3.38}$$

where SS_{treats}, SSE are the between-group and within-group sum of squares, and k is the number of groups. Under H_0, F follows an F-distribution with $k - 1$ and $N - k$ degrees of freedom.

3.6 Degrees of Freedom Approximation

In statistical experiments involving the analysis of variance, it is important to calculate the degrees of freedom associated with the estimation of uncertainty and hence the F statistic. Degrees of freedom is defined as the pieces of information in a set of data that are free to vary without violating constraints when estimating statistical parameter. Basically, the experimental error degrees of freedom are calculated by subtracting the number of relations from the number of observations $(N - k)$, whereas the total degrees of freedom for the sample mean are given by the number of observations less one $(N - 1)$.

3.6.1 The Satterthwaite Approximation

The total degrees of freedom is not always a simple sum of the independently calculated degrees of freedom (Satterthwaite, 1946). There are many ways to account for sample variances by calculating equivalent degrees of freedom. Three basic approaches are listed below.

- The pooled standard error approach: Assuming that the population standard deviations for the groups are the same, simply pooling the sample standard deviations gives the largest degrees of freedom, i.e. $df = n_1 + n_2 - 2$. If the sample standard deviations substantially differ, the hypotheses test for testing that the standard deviations are the same is not robust, so this method should not be done in practice. This will give the smallest margin of error, and the smallest p-value of the three methods for estimating the degrees of freedom.

- The conservative estimate takes the smallest degrees of freedom, i.e. $min\ (n_1 - 1; n_2 - 1)$. This simple method will give the largest margin of error or the largest p-value of the three methods for estimating the p-value.

- The Satterthwaite (1946) approximation is a formula used to estimate an effective degrees of freedom from linear combinations of mean squares in order to test certain hypotheses where only estimates of the variance are known. A convenient way is to find additive combinations of mean squares by synthesising both the numerator and denominator mean squares, resulting in the Satterthwaite (1946) formula given by,

$$df = \frac{\left(\sum_{i=1}^{n} S_i^2\right)^2}{\sum_{i=1}^{n} \frac{\left(S_i^2\right)^2}{n_i - 1}}, \tag{3.39}$$

where S_i^2 and n_i are the sample variances and sample sizes, respectively.

It can be shown that the Satterthwaite approximation for the degrees of freedom lies between the conservative and pooling estimates. Alternatively, estimates of the components of variance and hence estimation of MS may require subtraction of mean squares; however, as Satterthwaite warned, caution must be exercised when his formula is used.

3.6.2 The Welch Approximation

It is well-known that the F-test ANOVA is robust to the normality but sensitive to violation of other assumptions (Lee and Ahn, 2003), especially the equality of variance. The Behrens-Fisher problem is one example when two normal populations with unequal variances is considered, and the studentised difference of the two sample means no longer follows a t-distribution. Numerous different methods have been proposed and compared for solving the Behrens-Fisher problem. One prominent example is the Welch (1947) approximation of the T-test statistic by means of a student t-distribution with a random number of degrees of freedom given by,

$$DF = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{\left(\frac{S_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{S_2^2}{n_2}\right)^2}{n_2 - 1}}, \tag{3.40}$$

where S_1^2, S_2^2 are the sample variances, and n_1 and n_2 are the respective sample sizes. The Welch approximation is basically identical to the Satterthwaite approximation. Other tests and comparisons involving more than two means with unequal variances also exist (Bartlett, 1937; Levene, 1960; Brown and Forsythe, 1974; Xu *et al.*, 2015; Gokpinar and Gokpinar, 2017).

3.6.3 The Welch-Satterthwaite Approximation

When more than two sample variances are compared, it is generally necessary to estimate the variance by a linear combination of mean squares. For example, when testing the null hypothesis

$H_0 : \sigma_i^2 = 0$, for $i = 1, 2, ..., k$, a linear function of the mean squares $MS = \sum_{i=1}^{k} \alpha_i MS_i$, where the α_i's are known constants, can be computed. If MS is approximately distributed as $\frac{\sigma^2 \chi_v^2}{v}$, with an F-test approximated by MS_i/MS, then the Satterthwaite formula can be used to estimate the degrees of freedom as

$$\hat{df} = \frac{(MS)^2}{\sum_{i=1}^{k} \frac{\alpha_i MS_i}{v_i}}. \tag{3.41}$$

Welch-Satterthwaite is an approximation to the effective degrees of freedom by using the samples' uncertainties (e.g. standard deviation) and degrees of freedoms, as described in Welch (1947) and Satterthwaite (1946). The effective or equivalent degrees of freedom, also known as the pooled degrees of freedom, is a combination of the multiple variances (pooled variance) and their respective degrees of freedom v_i corresponding to the linear combination, $\sum_{i=1}^{n} \alpha_i S_i^2$ $(i = 1, ..., n)$. The Welch-Satterthwaite approximation (Satterthwaite, 1946; Welch, 1947) for effective degrees of freedom is given by,

$$\hat{df} = \frac{\left(\sum_{i=1}^{n} \alpha_i S_i^2\right)^2}{\sum_{i=1}^{n} \frac{\left(\alpha_i S_i^2\right)^2}{v_i}}, \tag{3.42}$$

where the weights α_i are defined as $\alpha_i = \frac{1}{v_i+1}$.

3.6.4 The Kenward-Roger Approximation

Linear mixed model parameters are typically estimated using ML or REML estimates (Patterson and Thompson, 1971; Harville, 1977). Hypothesis testing procedures such as the likelihood ratio test and Wald test are the most widely used methods for drawing inferences on fixed effects of the model. However, the procedures tend to underestimate the variance component of the fixed effect parameter in the mixed model (Zucker *et al.*, 2000) when small samples are involved. The hypothesis, $H_0 : L'\beta = 0$, is tested by a Wald test statistic, which is approximated as an F-distribution with $rank(L)$ numerator degrees of freedom. To account for the variability due to estimation of the variance components in small samples, Kenward and Roger (1997) proposed a scaled Wald statistic, which involves an approximate covariance matrix, ensuring that the modified statistic is asymptotically distributed as an F-distribution for which they provided a method for estimating the denominator degrees of freedom. However, the denominator degrees of freedom can be estimated by the Satterthwaite approximation or any other suitable method. Ignoring the small sample precision when estimating variance components may result in biased confidence intervals. Although its small-sample performance can break down in some circumstances, the Kenward-Roger's (1997) approximation has proven effective for controlling the type I error rate in a variety of contexts.

Both the Satterthwaite (1946) and Kenward-Roger (Kenward and Roger, 1997) approaches are used to estimate denominator degrees of freedom for F-statistics or degrees of freedom for t statistics (Luke, 2017). The Satterthwaite approximation, the default for SAS *PROC MIXED* (Stroup *et al.*, 2018), can be applied to ML or REML models, while the Kenward-Roger approximation is used on REML models only. A simulation study conducted by Luke (2017) shows that both the Satterthwaite and the Kenward-Roger approximation produced highly comparable type I error rates, and were not noticeably anti-conservative and robust for smaller sample sizes. Various statistical software packages are available to provide these approximations. Kuznetsova *et al.* (2017) have developed the ***lmerTest*** package in R, which implements the Satterthwaite's method, and the ***pbkrtest*** package, which generates the Kenward-Roger approximation for approximating degrees of freedom for the $t-$ and F-tests in the construction of type I-III ANOVA tables.

3.7 Model Fit Through Simulation

Simulation is rarely used by researchers, yet it is powerful tool for assessing model fit (Harrison *et al.*, 2018). Simulations samples for this research were generated in R and SAS statistical software. Full details about the simulation codes are provided in Appendices B and C. A sufficiently large number of iterations (10 000, say) can be generated using a set of parameter estimates from the original model and compared to the observed real data used. A good model fit should not show significant deviations (poor model fit) from the observed real data after a considerably large number of iterations. Simulation samples for this study were generated to test both internal and external validity of the approach in each design.

3.8 Conclusion

The fundamental principles and processes of constructing a three-way linear mixed model for the sake of conducting either individual or combined analysis have been presented. The importance of choosing an appropriate research design for analysis of treatment effects cannot be underestimated. The partitioning of factor levels offers a more efficient room for evaluation of treatment effects in different treatment structures. The next Chapter 4 presents the application of the partitioning approach in linear mixed models in the presence of outlier contamination.

CHAPTER 4

LINEAR MIXED MODELS FOR CONTAMINATED DATA

This chapter presents the application of the partitioning approach when analysing contaminated data using linear mixed models. The main focus will be on model construction, hypothesis testing and inference scope when treatments are arranged in completely randomised design (CRD) and randomised complete block design (RCBD).

4.1 Introduction

Real-life data with multiple levels of random variation are often contaminated by outliers or any other contamination in those levels (Koller, 2016). In this chapter, the term "contamination" in experimental data is taken to refer to the presence of outliers in the data. Exploratory studies in various disciplines such as agriculture, biomedical sciences, and physical science, involve factorial designs where combinations of the levels of the factors are investigated (Oliveira *et al.*, 2019). Linear mixed models are a widely used analysis tool for assessing the impact of input factors and their associated interactions on the response variable (Harrar *et al.*, 2019). The traditional F-test plays a crucial role in the testing of these relationships in linear model analysis (Harrar *et al.*, 2019). However, the use of linear mixed models to characterise experimental data requires improved modelling techniques for the specification of both fixed and random effects (Smith and Edwards, 2017).

Practitioners conducting exploratory agricultural and industrial studies often encounter scenarios involving factorial arrangements, which require increased precision and information on testing the main and interaction effects of the factors under investigation (Ott and Longnecker 2016). Model construction, estimation and drawing of inference about the parameters of interest call for proper characterisation of the treatment structure involving different factors. A factor is classified as fixed when the levels of that factor are pre-selected. As such, statistical inference is limited to those specific levels. In contrast, when the levels of a factor are a random sample from a large population of possible levels, statistical inference is desired on the population of levels, and the factor is considered random (Jayalath and Ng, 2018).

According to Robinson (1991), complicated and controversial issues about fixed and random models can be well understood through understanding procedures for estimating random effects. Many computer algorithms have been developed to construct experimental designs that are D-optimum for the fixed parameters of a statistical model (Loeza-Serrano and Donev, 2014). Understanding and characterising treatment effect variation in randomised experiments has become essential for going beyond the "black box" of the average treatment effect (Ding *et al.*, 2019). Correct identification of experimental design and its structure in research prevents the selection of incorrect models and the drawing of misleading inferences (Bate and Chatfield, 2016). In a balanced data case, the analysis of variance (ANOVA) approach is preferred in testing for significance of the treatment means (in case of fixed effects) and in estimating the variance components (for random effects) involved by making use of the linear combination of the mean squares (Crump, 1946; Satterthwaite, 1946; Fisher, 1925).

Improved modelling techniques (Ott and Longnecker, 2016; Smith and Edwards, 2017; Ding *et al.*, 2019) and proper characterisation of the treatment structures, which clearly specify both fixed and random effects (McLean *et al.*, 1991; Robinson, 1991; Jayalath and Ng, 2018; Chaka and Njuho, 2021), are crucial steps to ensure the selection of correct models and inferences (Bate and Chatfield, 2016). However, the precision of estimated linear models is mostly affected by the presence of outliers and other forms of contamination in experimental data (Park and Leeds, 2016; Wang *et al.*, 2020). The presence of outliers influences usually impacts the classic estimates of the fitted model (Koller, 2016). Researchers often face some challenges on how in identifying the best linear mixed model for making valid inferences (Kuran and Özkale, 2021).

4.2 Methodology

A three-way treatment structure with full interaction of factor levels, which are assumed to be either fixed or random in nature, is considered. We demonstrate the model construction and analysis procedures to follow in a completely randomised design when little contamination has been detected in the data.

4.2.1 Model Construction

In a completely randomised design (CRD), the treatment factors might have $t = 1, 2, ..., t_f$ fixed treatment levels and $t_f + 1$, $t_f + 2$,..., t_r random levels where $t = t_f + t_r$ are influencing on the variable of interest. Our approach provides a provision to install these types of factors into the experimental design and investigate the treatment effects based on the treatment combinations formed. A three-way treatment structure where each consists of both fixed and random levels is assumed, with at least one replication per treatment combination has a model give by,

$$y_{ijkl} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk} + \epsilon_{ijkl}, \tag{4.1}$$

where y_{ijkl} is the response variable, $i = 1, 2, ..., a$; $j = 1, 2,..., b$; $k = 1, 2, ..., c$; $l = 1, 2, ..., n$; μ is the overall mean; α_i is the i^{th} treatment effect of factor A; β_j is the j^{th} treatment effect of factor B; γ_k is the k^{th} treatment effect of factor C; $(\alpha\beta)_{ij}$ is the interaction effect of factor A and factor B; $(\alpha\gamma)_{ik}$ is the interaction effect of factor A and factor C; $(\beta\gamma)_{jk}$ is the interaction effect of factor B and factor C; $(\alpha\beta\gamma)_{ijk}$ is the interaction effect of factor A, factor B and factor C; ϵ_{ijkl}'s are assumed independent and uncorrelated random error terms, $\epsilon_{ijkl} \sim N(0; \sigma^2_\epsilon)$.

4.2.2 Partitioning of Factor Levels

We consider a three-way treatment structure in a completely randomised design (CRD) with with full interaction of factors A (with a effect levels), B (with b effect levels), and C (with c effect levels). Suppose, for the same factors, we let f_a be fixed levels and r_a be random levels of factor A ($a = f_a + r_a$); f_b be fixed levels and r_b be random levels of factor B ($b = f_b + r_b$); f_c be fixed levels, and r_c be random levels of factor C ($c = f_c + r_c$). The newly introduced strategies are considered as fixed levels of each factor, and the old existing strategies as random levels selected from the population of strategies which could not be considered in total. Assuming n replicates per treatment, a three-way linear model has the form (4.1), where μ is the overall mean; α_i $(i = 1, 2, \ldots, f_a, f_a + 1, f_a + 2, \ldots, (f_a + r_a) = a)$ is the i^{th} treatment effect of factor A; β_j $(j = 1, 2, \ldots, f_b, f_b + 1, f_b + 2, \ldots, (f_b + r_b) = b)$ is the j^{th} treatment effect of factor B; γ_k $(k = 1, 2, \ldots, f_c, f_c + 1, f_c + 2, \ldots, (f_c + r_c) = c)$ is the k^{th} treatment effect of factor C; $(\alpha\beta)_{ij}$ is the interaction effect between factor A and factor B; $(\alpha\gamma)_{ik}$ is the interaction effect of factor A and factor C; $(\beta\gamma)_{jk}$ is the interaction effect of factor B and factor C; $(\alpha\beta\gamma)_{ijk}$ is the interaction effect of factor A, factor B and factor C; ϵ_{ijkl}'s are assumed independent and uncorrelated random error terms, $\epsilon_{ijkl} \sim N(0; \sigma^2_\epsilon)$.

We construct a partyitioned three-way linear mixed model from (4.1) by partitioning the levels of factors A, B and C as (1) fixed-fixed-fixed (FFF), (2) fixed-fixed-random (FFR), (3) fixed-random-fixed (FRF), (4) fixed-random-random (FRR), (5) random-fixed-fixed (RFF), (6) random-fixed-random (RFR), (7) random-random-fixed (RRF) and (8) random-random-random (RRR). For ease of computation, we code the eight models as displayed in Table 4.1.

Table 4.1: Three-way Treatment Structure with Model Codes (1) - (8)

		Factors A*B			
		Fixed-Fixed	**Fixed-Random**	**Random-Fixed**	**Random-Random**
Factor C	**Fixed**	FFF(1)	FRF(3)	RFF(5)	RRF(7)
	Random	FFR(2)	FRR(4)	RFR(6)	RRR(8)

The model codes (1) - (8) in Table 4.1 are for easy reference and conveneience.

4.2.3 Model Construction

We construct a three-way treatment structure model by partitioning the levels of factors A, B and C as given in Table 4.1. Assuming n_{ijk} replicates per treatment ($n_{ijk} = n$ for the balanced case), each of the partitioned three-way linear model has the form,

$$y_{p_{ijkl}} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk} + \epsilon_{p_{ijkl}}, \qquad (4.2)$$

where the subscript p in y_{p_ijkl} denotes the partition $(p = 1, 2, \ldots, 8)$ as coded in Table 4.1, and all other parameters are as defined in (4.1).

Of the eight partitioned models, we select two to demonstrate the model construction process. The Fixed-Fixed-Fixed (FFF) effects model for the factors A, B and C, is expressed as

$$y_{1_{ijkl}} = \mu_1 + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk} + \epsilon_{1_{ijkl}}, \tag{4.3}$$

where the subscript 1 in $y_{1_{ijkl}}$ denotes the FFF model; μ_1 is the overall mean of the fixed effects, $\alpha_1, \alpha_2, ..., \alpha_{f_a}$; $\beta_1, \beta_2, ..., \beta_{f_b}$ and $\gamma_1, \gamma_2, ..., \gamma_{f_c}$ are real-valued constants corresponding to the fixed effect levels of factors A, B and C, respectively; $(\alpha\beta)_{11}$, $(\alpha\beta)_{12}$, ..., $(\alpha\beta)_{f_af_b}$ are the $A \times B$ interaction effects; $(\alpha\gamma)_{11}$, $(\alpha\gamma)_{12}$, ..., $(\alpha\gamma)_{f_af_c}$ are the $A \times C$interaction effects; $(\beta\gamma)_{11}$, $(\beta\gamma)_{12}$, ..., $(\beta\gamma)_{f_bf_c}$ represent the $B \times C$ interaction effects; $(\alpha\beta\gamma)_{111}$, $(\alpha\beta\gamma)_{112}$, ..., $(\alpha\beta\gamma)_{f_af_bf_c}$ are the $A \times B \times C$interaction effects, and $\epsilon_{1_{ijkl}} \sim i.i.d.N(0, \epsilon^2_{1_\epsilon})$.

The Fixed-Random-Fixed (FRF) model is similarly constructed from fixed levels of factors A and C, and random levels of factor B. For simplicity, let a, b and c be levels of factors A, B and C, respectively, m be the interaction effects of levels of factors A and B, denoted by $(\alpha\beta)$, p be the interaction effects of levels of factors A and C, denoted by $(\alpha\gamma)$, q be the interaction effects of levels of factors B and C, denoted by $(\beta\gamma)$, and g be the interaction effects of factors A, B and C levels denoted by $(\alpha\beta\gamma)$. The Fixed-Random-Fixed (FRF) model is expressed as,

$$y_{3_{ijkl}} = \phi_{i.k} + b_{3_j} + m_{3_{ij}} + p_{3_{ik}} + q_{3_{jk}} + g_{3_{ijk}} + \epsilon_{3_{ijkl}}, \tag{4.4}$$

where the subscript 3 in $y_{3_{ijkl}}$ denotes the FRF model as per the partitions of Table 4.1; $\phi_{i.k}$ denotes the mean of the fixed-effect levels of factor A and the fixed-effect levels of factor C averaged over the random-effect of factor B, b_{3_j} represents the random-effect levels of factor B with $b_{3_j} \sim i.i.d.N(0, \sigma^2_{b_3})$; $m_{3_{ij}}$ represents the random interaction effects of factor A and factor B, with $m_{3_{ij}} \sim i.i.d.N(0, \sigma^2_{m_3})$; $p_{3_{ik}}$ represents the fixed interaction effects of factor A and factor C; $q_{3_{jk}}$ represents the random interaction effects of factor B and factor C, with $q_{3_{jk}} \sim i.i.d.N(0, \sigma^2_{q_3})$; $g_{2_{ijk}}$ represents the random interaction effects of factors A, B and C with $g_{3_{ijk}} \sim i.i.d.N(0, \sigma^2_{g_3})$; and $\epsilon_{3_{ijkl}}$ is the random error term, $\epsilon_{3_{ijkl}} \sim i.i.d.N(0, \sigma^2_{\epsilon_3})$.

We formulate a partitioned models and combined model from the partitions shown in Table 3.6 and 3.7 by integrating the variables involved in each of the eight models.

4.2.4 Model Assumptions

In experimental research involving analysis of variance (ANOVA) as a technique for comparing different treatment means, a set of assumptions which include the usual normality and homogeneity of variance, must be checked before analysis of data (Kotchaporn and Araveeporn, 2018). These assumptions can be checked for each partitioned model using any of the appropriate normality tests (e.g., Q-Q plots or Shapiro Wilk's test) or homogeneity of variance tests (e.g., Levene's test or Bartlett's test). Consider the fixed and random levels of each factor as defined in (4.2). The combined model assumes equality of variance for the error terms and random effects across the partitions, while the same is assumed for groups defined by each variable in each partitioned model.

4.2.4.1 Homogeneity of Variance Assumption

In mixed model hypothesis testing, it is crucial to assess the significance of all or a subset of the random effects included before the analysis of the means is attempted (Hui *et al.*, 2019; Milliken and Johnson, 2009). We consider the equality of variance tests for the variance components first before the treatment effects tests for the fixed part of the combined model. The intention is to establish the rationale for applying the tests that rely on homogeneity of variance assumption, as well as to establish homogeneous variance of the dependent variable exists across multiple groups (Mara and Cribbie, 2018). When more than one hypothesis is simultaneously tested, the probability of committing false statistical inferences would increase considerably. Proper adjustments for multiple comparisons for specific types of tests or situations, such as Bonferroni, Sidak, Dunnett, Holm and others, are required (Thiese *et al.*, 2016). Generally, there is no uniform agreement on when to adjust or what type of adjustment is best, since all these adjustments operate in the same way to lower the likelihood of committing a type I error (Thiese *et al.*, 2016). Testing the assumptions made for the eight models in Table 4.1 requires the Bonferroni's simultaneous test, since it allows for simultaneous comparisons (Westfall and SAS Institute, 1999) using $\frac{\alpha}{8}$ level of significance. The hypotheses of interest are:

$H1$. $H_{A0}: \sigma^2_{\epsilon_1} = \sigma^2_{\epsilon_2} = \sigma^2_{\epsilon_3} = \sigma^2_{\epsilon_4} = \sigma^2_{\epsilon_5} = \sigma^2_{\epsilon_6} = \sigma^2_{\epsilon_7} = \sigma^2_{\epsilon_8}$, against
H_{A1}: At least two are different.

H2. H_{B0}: $\sigma^2_{a_3} = \sigma^2_{a_4} = \sigma^2_{a_7} = \sigma^2_{a_8}$, against H_{B1}: At least two are different.

H3. H_{C0}: $\sigma^2_{b_2} = \sigma^2_{b_4} = \sigma^2_{b_6} = \sigma^2_{b_8}$, against H_{C1}: At least two are different.

H4. H_{D0}: $\sigma^2_{c_5} = \sigma^2_{c_6} = \sigma^2_{c_7} = \sigma^2_{c_8}$, against H_{D1}: At least two are different.

H5. H_{E0}: $\sigma^2_{m_1} = \sigma^2_{m_2} = \sigma^2_{m_3} = \sigma^2_{m_4} = \sigma^2_{m_5} = \sigma^2_{m_6} = \sigma^2_{m_7} = \sigma^2_{m_8}$, against H_{E1}: At least two are different.

H6. H_{F0}: $\sigma^2_{p_1} = \sigma^2_{p_2} = \sigma^2_{p_3} = \sigma^2_{p_4} = \sigma^2_{p_5} = \sigma^2_{p_6} = \sigma^2_{p_7} = \sigma^2_{p_8}$, against H_{F1}: At least two are different.

H7. H_{G0}: $\sigma^2_{q_1} = \sigma^2_{q_2} = \sigma^2_{q_3} = \sigma^2_{q_4} = \sigma^2_{q_5} = \sigma^2_{q_6} = \sigma^2_{q_7} = \sigma^2_{q_8}$, against H_{G1}: At least two are different.

H8. H_{K0}: $\sigma^2_{g_1} = \sigma^2_{g_2} = \sigma^2_{g_3} = \sigma^2_{g_4} = \sigma^2_{g_5} = \sigma^2_{g_6} = \sigma^2_{g_7} = \sigma^2_{g_8}$, against H_{K1}: At least two are different.

Provided the sample size (n_i) is the same for each treatment, the hypotheses H1 - H8 can be tested for homogeneity of variance using Bartlett's (1937) test, whose test statistic is given by

$$T = \frac{1}{C}\left[v.log_e(\hat{\sigma}^2) - \sum_{i=1}^{m} v_i.log_e(\hat{\sigma}^2_i)\right], \tag{4.5}$$

where $C = 1 + \frac{1}{3(m-1)}(\sum_{i=1}^{m}\frac{1}{v_i} - \frac{1}{v})$; m is the number of treatments under consideration ($m = 8$ in this case); v_i represents the degrees of freedom associated with each variance component i, $v = \sum v_i$, and $\hat{\sigma}^2_i$ is the estimate of variance component, $\hat{\sigma}^2 = \sum_{i=1}^{m} v_i(\frac{\hat{\sigma}^2_i}{v})$. We reject the null hypothesis when Bartlett's test statistic $T > \chi^2_{\frac{\alpha}{m}}(m-1)$. However, Bartlett test is sensitive to departure from normality. When one is not certain that the normality assumption is violated in the data set, or the data is nearly normal, the Levene's test or other heteroscedastic alternative testing approaches to the ANOVA F-test (Levene 1960; Parra-Frutos 2013) are recommended. Depending on whether one is using the mean, median or treamed mean, the Levene's test for equal variances across k samples is given by

$$T_L = \frac{N-k}{k-1} \frac{\sum_{i=1}^{n} n_i (Z_{i.} - Z_{..})^2}{\sum_{i=1}^{k} \sum_{j=1}^{n_j} (Z_{ij} - Z_{.i.})}, \tag{4.6}$$

where Z_{ij} is either the mean, trimmed mean or median of the subgroup. Most statistical software populate the p-value of the Levene's test, which is used to reject or retain the null hypothesis of homogeneity of variance. Modifications of the levene's and O'Brien's tests for homogeneity of variance based on median and trimmed mean were exist (Kotchaporn and Autcha, 2018).

Some heteroscedastic alternative testing approaches to the ANOVA F-test (Levene, 1960; Parra-Frutos, 2013) are recommended. Kotchaporn and Araveeporn (2018) proposed some median and trimmed mean to modify the Levene's and O'Brien's tests for testing homogeneity of variance. Sevaral parametric test procedures exist when testing hypotheses of equal variances under various experimental conditions.

In case $n_i \neq n_{i'}$, i.e. when treatment sample sizes are different, various traditional difference-based tests exist that test for equality of variances (Hartley, 1950; Levene, 1960; Brown and Forsythe, 1974). Building from the F-max test (Hartley, 1950), several researchers have developed equivalence-based tests along with several modifications for homogeneity of variances to address the fundamental problems of traditional difference-based tests (Wellek, 2010; Gokpinar and Gokpinar, 2017; Frey, 2010; Mara and Cribbie, 2018). Equivalence-based tests seek to establish if a pre-specified equivalence interval contains the difference of the variances or not (Mara and Cribbie, 2018).

Considering the partitioned models (1) - (8) as defined in Table 4.1, the homogeneity of variance assumption of interest in the combined model might be: $\sigma^2_{\epsilon_1} = \sigma^2_{\epsilon_2} = ... = \sigma^2_{\epsilon_8}$, while the variances across the groups are considered for the main and interaction effects. For example, using the variable combinations illustrated in Table 4.1, equality of variance for random factor A considers $\sigma_A^2 = \sigma_{a_3}^2 = \sigma_{a_4}^2 = \sigma_{a_7}^2 = \sigma_{a_8}^2$.

The F-max test statistic (Hartley, 1950; Frey, 2010) for m samples is given by,

$$F_{max} = \frac{Max(\hat{\sigma}_i^2)}{Min(\hat{\sigma}_i^2)}, \tag{4.7}$$

where H_0 is rejected when $F_{max} > F_{v_{max},v_{min}}(\frac{\alpha}{m})$. Milliken and Johnson (2009) provided an appropriate alternative approache to use in case the null hypothesis is either rejected or not. Different types of homogeneity of variance tests exist, such as the difference-based equality of variance tests when the sample sizes for each treatment are equal (Bartlett, 1937; Hartley, 1950; Levene, 1960; Brown and Forsythe, 1974) and their modifications (equivalence-based tests) built from the F-max test (Hartley, 1950) when treatment sample sizes are different and unequal (Frey, 2010; Wellek, 2010; Gokpinar and Gokpinar, 2017; Mara and Cribbie, 2018).

4.2.4.2 Outlier Contamination

It is generally assumed that experimental data modelled by linear mixed models are normally distributed without outliers (Park and Leeds, 2016), which is highly uncommon in most real-life data structures. In practice, it is difficult to pinpoint the source of contamination in linear mixed models due to the complex nature of the data sets. However, such influence might significantly impact the accuracy of classic estimates (Koller, 2016). Therefore, we use an R package (*robustlmm*) to investigate, detect and flag outliers before generating robust estimates for the linear mixed model in the presence of little contamination in the data set (Koller, 2016). The package functions and syntax are almost similar to the R package $lme4$ that implements classic linear mixed model estimation. However, *robustlmm* has an advantage over $lme4$ in that it is capable of robustly fitting linear mixed models in the presence of mild outlier contamination.

4.2.5 Estimation of the Robust Linear Mixed Model

In order to estimate the fixed and random effects in a classic linear mixed model, we express model (4.2) for each of the p partitions as a general linear mixed model defined in matrix form as given in (2.3), or more conveniently, (2.10). The estimation process of a robust linear mixed model is similar to the traditional approach of estimating a classic linear mixed model.

Consider the matrix $\mathbf{X}$ where $diag(\mathbf{X}) = \mathbf{X}_i$, $(i = 1, 2, ..., 7)$, are the design matrices of full rank associated with each of the first seven models in Table 4.1, and the random intercepts for the fixed effects of the FFR, FRF, FRR, RFF, RFR and RRF models. Denoting the random intercepts for the fixed effects for these models, respectively, by ψ_2, ϕ_3, ϑ_4, δ_5, η_6 and ω_7, respectively, we express the vector of fixed effects components in the combined model is expressed as,

$$\hat{\boldsymbol{\theta}} = \begin{bmatrix} \hat{\mu}_1 \\ \hat{\psi}_2 \\ \hat{\phi}_3 \\ \hat{\vartheta}_4 \\ \hat{\delta}_5 \\ \hat{\eta}_6 \\ \hat{\omega}_7 \end{bmatrix} = \begin{bmatrix} (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{y}_1 \\ (\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'\mathbf{y}_2 \\ (\mathbf{X}_3'\mathbf{X}_3)^{-1}\mathbf{X}_3'\mathbf{y}_3 \\ (\mathbf{X}_4'\mathbf{X}_4)^{-1}\mathbf{X}_4'\mathbf{y}_4 \\ (\mathbf{X}_5'\mathbf{X}_5)^{-1}\mathbf{X}_5'\mathbf{y}_5 \\ (\mathbf{X}_6'\mathbf{X}_6)^{-1}\mathbf{X}_6'\mathbf{y}_6 \\ (\mathbf{X}_7'\mathbf{X}_7)^{-1}\mathbf{X}_7'\mathbf{y}_7 \end{bmatrix}. \tag{4.8}$$

When dealing with real-life data, as is the usual case with unbalanced data, the matrix $\mathbf{X}$ in (2.10) does not necessarily have full column rank, and $(\mathbf{X}'\mathbf{X})^{-1}$ does not exist (Searle and Gruber, 2017). Hence, the normal equations cannot be uniquely solved to $\hat{\boldsymbol{\beta}}$. The generalised inverse of $\mathbf{X}'\mathbf{X}$ can be used to estimate the possible solutions to the mixed model (Njuho and Milliken, 2009; Chaka and Njuho, 2021).

This section presents the estimation procedure for a robust linear mixed model. There are various approaches (packages) available on the Comprehensive R Archive (CRAN) that can be used for the robust estimation of linear mixed models when the assumptions of a classic linear model are not fully satisfied (Koller, 2016). Most of these approaches are limited to grouping structures and correlation of random effects. Following the approach by (Koller, 2016), we consider the classic linear mixed model (2.10), whose regular random effects have been transformed into spherical random effects such that the covariance matrix equals a scaled identity matrix. The transformed linear mixed model becomes

$$\mathbf{y}_p = (\mathbf{X} \quad \mathbf{Z})[\boldsymbol{\beta} \quad \mathbf{U}_b(\boldsymbol{\theta})\mathbf{b}^*]' + \mathbf{U}_\epsilon\boldsymbol{\epsilon}^*, \tag{4.9}$$

where $\mathbf{y}_p$ is an $N \times 1$ vector of observations in the p^{th} partition; $\mathbf{X} : N \times p$ and $\mathbf{Z} : N \times q$ are known incidence matrices associated with the fixed effects vector $\boldsymbol{\beta}$, and spherical random effects vector $\mathbf{b}^*$, respectively. The spherical random effects are related to the classic random effects through a transformation $b = \mathbf{U}_b(\boldsymbol{\theta})\mathbf{b}^*$, where $\mathbf{U}_b(\boldsymbol{\theta})$ is a lower triangular matrix parameterised by the vector $\boldsymbol{\theta}$. Under this model (4.9), the distribution of the observations $\mathbf{y}$ is given by $\mathbf{y} \sim \mathbf{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{V}_y(\boldsymbol{\theta}))$, where $\mathbf{V}_y(\theta) = \mathbf{Z}\mathbf{V}_b(\boldsymbol{\theta})\mathbf{Z}^T + \mathbf{V}_e$. Let $\mathbf{U}_b(\boldsymbol{\theta})$ be the lower triangular Cholesky factor of $\mathbf{V}_b(\boldsymbol{\theta})$, such that $\mathbf{V}_b(\boldsymbol{\theta}) = \mathbf{U}_b(\boldsymbol{\theta})\mathbf{U}_b(\boldsymbol{\theta})^T$ and $\mathbf{V}_e = \mathbf{U}_e\mathbf{U}_e^T$. The spherical random effects ($\mathbf{b}^*$) are related to the classic random effects through a transformation $\mathbf{b} = \mathbf{U}_b(\boldsymbol{\theta})\mathbf{b}^*$. The matrix $\mathbf{U}_e$ is a diagonal matrix of known weights. This model assumes that the spherical random effects and error terms are independently distributed as $\mathbf{b}^* \sim N(\mathbf{0}, \sigma^2\mathbf{I}_q)$ and $\boldsymbol{\epsilon}^* \sim N(\mathbf{0}, \sigma^2\mathbf{I}_q)$, respectively.

Additional assumptions for this model are that (Koller, 2016):

- the model parameters are estimable,
- the covariance matrix of the random effects, $\sum_b(\boldsymbol{\theta}) = \mathbf{U}_b(\boldsymbol{\theta})\mathbf{U}_b(\boldsymbol{\theta})'$, is block-diagonal, with each block of size 2×2 or greater and unstructured, and
- the residual error covariance matrix is diagonal with only one unknown scaling parameter.

The robust estimation process under model (4.9) is achieved by deriving scoring equations through exchanging the residuals and the predicted spherical random effects with bounded functions (Koller, 2013). When the covariance parameters $\boldsymbol{\theta}$ and the scale σ are known, the estimation of fixed and random effects can be done using the iteratively reweighted least squares. Otherwise, the parameters have to be estimated from the data first and the restricted maximum likelihood (REML) is used to robustify the estimating equations (Koller, 2013). The approach produces various estimates of the robust linear mixed model by controlling the parameters in the bounded functions. For example, tuning parameters of the Huber function (a possible choice of a bounded function) to larger values in the robust estimating equation produces estimates comparable to the REML estimates.

4.2.6 Methods of Inference

Considering the model's fixed and random effects, we analyse and compare the model fit information for both the classic and robust linear mixed models fitted using different sets of tuning parameters. Particular attention is given to the effect of contamination on the fit, estimated parameters and their standard errors, as well as variance components in the partitioned models of each scenario.

4.2.6.1 Hypothesis Testing for Fixed Effects

We are interested in testing the main and the interaction effects of the three-factor linear model. Hypotheses HA - HC test the homogeneity of variance; main and interaction effects of fixed factors; and the effect of random effects, respectively, using the appropriate partition covariance matrix $\sum_b = Var(\mathbf{y}_p)$.

$HA.$ $\quad H_0 : \sigma^2_{\epsilon} = \sigma^2_{\epsilon_1} = \sigma^2_{\epsilon_2} = ... = \sigma^2_{\epsilon_8},\ \textit{against}$

$H_1 :\ \sigma^2_{\epsilon_p} = \sigma^2_{\epsilon_{p*}}$ for $p \neq p^*$.

$HB.$ $\quad H_0 : \boldsymbol{\beta}_F = \mathbf{0}$ against

$H_1 : \boldsymbol{\beta}_F \neq \mathbf{0}$ (where $\boldsymbol{\beta}_F$ is a vector of fixed effects).

$HC.$ $\quad H_0 : \sigma^2_{\tau_p} = 0$ against

$H_1 : \sigma^2_{\tau_p} > 0$ (where $\tau = a, b, c$ or the interaction of these).

4.2.6.2 Expected Mean Squares and Variance Components

The estimation of treatment means and the associated F-tests for treatment effects, and variance components depend on the type of hypotheses being tested, targeted inference space, and the replication allowed in the experiment. Following the illustration given in Table 3.11, the analysis of variance and expected mean squares for each partitioned model are obtained based on appropriate sums of squares.

Analysis of variance for an experimental design with unequal cell frequencies is considered differently. In the case of unbalanced data, the reduction sum of squares, $R(\mu, \alpha, \beta, \gamma)$, and their expectations, $E[R(\mu, \alpha, \beta, \gamma)]$, are used to estimate the variance components (Searle and Gruber, 2017). The restricted maximum likelihood estimation procedure, which produces stable estimates irrespective of the unbalanced nature of the data, is preferred (Hartley and Rao, 1967; Dempster *et al.*, 1977). Information from each of the individual models is syncretised based on the targeted model and factor levels. Table 4.2 provides information on how a combined model analysis is achieved, using the degrees of freedom associated with each source of variation and individual models supplying information for each appropriate hypotheses test.

Table 4.2: The Format of a Combined Model

Source of variation	Degrees of freedom	Models supplying information
FA (F.. Model)	$\sum_{i=1,2,5,6}(f_{a_i}-1)$	**F**FF, **F**RF, **F**FR and **F**RR
FB (.F. Model)	$\sum_{i=1,3,5,7}(f_{b_i}-1)$	F**F**F, R**F**F, F**F**R and R**F**R
FC (..F Model)	$\sum_{i=1}^{4}(f_{c_i}-1)$	FF**F**, FR**F**, RF**F** and RR**F**
FA*FB (FF. Model)	$(\sum_{i=1,2,5,6}(f_{a_i}-1))(\sum_{i=1,3,5,7}(f_{b_i}-1))$	**FF**F, and **FF**R
FA*FC (F.F Model)	$(\sum_{i=1,2,5,6}(f_{a_i}-1))(\sum_{i=1}^{4}(f_{c_i}-1))$	**F**F**F** and **F**R**F**
FB*FC (.FF Model)	$(\sum_{i=1,3,5,7}(f_{b_i}-1))(\sum_{i=1}^{4}(f_{c_i}-1))$	F**FF** and R**FF**
FA*FB*FC (FFF Model)	$(\sum_{i=1,2,5,6}(f_{a_i}-1))(\sum_{i=1,3,5,7}(f_{b_i}-1)).$ $(\sum_{i=1}^{4}(f_{c_i}-1))$	**FFF**
RA (R.. Model)	$\sum_{i=3,4,7,8}(r_{a_i}-1)$	**R**RR, **R**FR, **R**RF and **R**FF
RB (.R. Model)	$\sum_{j=2,4,6,8}(r_{b_i}-1)$	R**R**R, F**R**R, R**R**F and F**R**F
RC (..R Model)	$\sum_{i=5}^{8}(r_{c_i}-1)$	RR**R**, RF**R**, FR**R** and FF**R**
RA*RB (RR. Model)	$(\sum_{i=3,4,7,8}(r_{a_i}-1))(\sum_{i=2,4,6,8}(r_{b_i}-1))$	**RR**R and **RR**F
RA*RC (R.R Model)	$(\sum_{i=3,4,7,8}(r_{a_i}-1))(\sum_{i=5}^{8}(r_{c_i}-1))$	**R**R**R** and **R**F**R**
RB*RC (.RR Model)	$(\sum_{i=2,4,6,8}(r_{b_i}-1))(\sum_{i=5}^{8}(r_{c_i}-1))$	R**RR** and F**RR**
RA*RB*RC (RRR Model)	$(\sum_{i=3,4,7,8}(r_{a_i}-1))(\sum_{i=2,4,6,8}(r_{b_i}-1)).$ $(\sum_{i=5}^{8}(r_{c_i}-1))$	**RRR**
Error	$abc(n-1)$	FFF, FRF, FFR, FRR, RFF, RFR, RRF and RRR

For example, the combined FA model is made up of the fixed main effect of factor A, whose pieces of information are obtained from the individual models FFF, FRF, FFR and FRR. Syncretising the four sum of squares from these models provides the degrees of freedom for the FA part in the combined model. Similarly, summing up the degrees of freedom for the four models with fixed effects produces the combined model's degrees of freedom. The same approach is applied to derive the random error term of the combined model, which is calculated by summing up the error terms of all the eight models.

We demonstrate, by means of a numeric example, the application of the partitioning approach using traditional analysis of variance methods, which takes care of the nature of independent factors and the design of experiment.

4.3 Numerical Example 1: Completely Randomised Design (CRD)

An investigation of farmers' experiences from 1997 to 1999 was conducted in Vihiga and Siaya District, Western Kenya. The aim of the study was to estimate the wealth score based on the assets of a farm. A total of 1495 farmers provided their experiences of strategies to improve fallows and rock phosphate during the period. Some of the soil management strategies meant to improve fallows and rock phosphate, which were applied, included fallowing in short rain, use of natural fallow, and application of chemical fertiliser. For demonstration of the partitioning approach, we manipulate the original data set by combining some of the observed variables and summarise them into three main factors, categorised as follows:

- farm category ($farmcat$): with levels, A ($\leq$ 0.5 acres), B (0.5 - 1.0 acres), C (1.1 - 1.5 acres), D (1.6 - 2.0 acres), E (2.0 - 3.0 acres), F ($>$ 3.0 acres),
- soil management approach (*soilmgt*): categorised as 1 (no fallow, no fertiliser); 2 (no fallow, add fertiliser); 3 (one-year fallow, no fertiliser); 4 (one-year fallow, add fertilizer); 5 (two-year fallow, no fertiliser); 6 (two-year fallow, add fertiliser); 7 (three-year fallow, no fertiliser); 8 (three-year fallow, add fertiliser); 9 (three-year fallow, with or without fertiliser),
- cattle management (*loccow*): categorised as 0 (none), 1 (one cow), 2 (two cows), 3 (three cows), 4 (four cows), 5 (five and more cows).

4.3.1 The CRD Classic Model Results

Consider farm size ($farmcat$) as the first factor with three fixed levels (A, B and F) and three random levels C, D and E selected from a population of possible farm sizes; soil management practice (*soilmgt*) as the second factor with three fixed levels (6, 8 and 9) and six random levels (1-5 and 7) selected from numerous possible soil management approaches; and cattle management (*loccow*) as the third factor with two fixed levels (4 and 5) and four random levels (0-3). We apply the partitioning approach on the analysis of farmers' experiences and wealth scores based on the assets accumulated on the farm. A three-way completely randomised design (CRD) is proposed for the experiment. Table 4.3 summarises the mean squares, degrees of freedom, and variance components (in parentheses) for the partitioned models FFF, FRF, FFR, FRR, RFF, RRF, RFR and RRR.

Table 4.3: The Analysis of Variance for CRD Partitioned Models

Model Type	**Source of variation**	**DF**	**Mean Square**	**Model Type**	**Source of variation**	**DF**	**Mean Square**
FFF	FA	2	153.747	RFF	RA	2	3.054
	FB	2	6.480		FB	2	0.488
	FC	1	3.800		FC	1	16.067
	FA*FB	4	9.096		RA*FB	4	19.634
	FA*FC	2	32.741		RA*FC	2	7.273
	FB*FC	2	21.410		FB*FC	2	1.960
	FA*FB*FC	4	11.899		RA*FB*FC	4	14.519
	ERROR	442	1.354		ERROR	436	1.815
FFR	FA	2	75.906	RFR	RA	2	13.212
	FB	2	7.096		FB	2	49.284
	RC	3	31.334		RC	3	35.149
	FA*FB	4	7.884		RA*FB	4	8.783
	FA*RC	6	11.740		RA*RC	6	12.143
	FB*RC	6	5.352		FB*RC	6	20.625
	FA*FB*RC	12	7.212		RA*FB*RC	12	14.101
	ERROR	853	1.372		ERROR	843	1.326
FRF	FA	2	93.338	RRF	RA	2	0.269
	RB	5	23.020		RB	5	26.735
	FC	1	13.814		FC	1	3.309
	FA*RB	10	9.771		RA*RB	10	28.829
	FA*FC	2	6.342		RA*FC	2	10.272
	RB*FC	5	7.595		RB*FC	5	16.940
	FA*RB*FC	10	11.860		RA*RB*FC	10	26.965
	ERROR	584	1.284		ERROR	870	1.281
FRR	FA	2	16.567	RRR	RA	2	15.935
	RB	5	35.449		RB	5	135.525
	RC	3	29.776		RC	3	72.449
	FA*RB	10	10.367		RA*RB	10	10.700
	FA*RC	6	5.145		RA*RC	6	13.108
	RB*RC	15	11.049		RB*RC	15	10.416
	FA*RB*RC	30	9.854		RA*RB*RC	30	8.464
	ERROR	1921	1.171		ERROR	1708	1.225

We are interested in testing the main and interaction effects of the fixed part of each of the partitioned models first, using the variance-covariance matrix, $\boldsymbol{\sigma}_m^2 = Var(\mathbf{Y}_m)$, for $m = 1, 2, ..., 8$, and μ_{ijk}, the expected mean response of treatment combination of factors A ($farmcat$), B ($soilmgt$) and C ($loccow$). The fixed effects model (1) (FFF), with $i = 1, 2, ..., f_a$; $j = 1, 2, ..., f_b$; and $k = 1, 2, ..., f_c$ fixed levels, respectively, had significant main and interaction effects ($p - value < 0.05$) except for the third factor ($loccow$) main effect which was not ($p - value = 0.095$).

In each of the partitioned mixed effects models (2)-(8), the interaction effects of the three factors (A*B*C) were significant ($p - value < 0.05$). Factor A ($farmcat$) significantly contributed to the wealth scores in models FFR (2) and FFR (3) ($p - value < 0.032$ and 0.005, respectively), whereas factor B ($soilmgt$) was significant ($p - value = 0.003$) in the random model RRR (8).

4.3.2 Results for the Combined CRD Classic Model

The combined model treatment structure is achieved by syncretising the sum of squares from and degrees of freedom obtained from the individual models (1) to (8). Similarly, the random error term for the combined model is derived from the sum of error terms of all the eight models. Table 4.4 provides the analysis of variance F- tests for the main and interaction effects in the combined model factors. Consider the fixed factor levels as new farm categories (FA), new soil management approaches (FB) and new categories of cattle management practices (FC), whilst the old factor levels are a sample from a population of existing levels that are similarly denoted by RA, RB and RC.

Table 4.4: Combined CRD Model ANOVA Table Based on Type 3 Sum of Squares

Source of variation	**Degrees of freedom**	**Sum of Squares**	**Mean Square**	**F**
farmcat (FA)	8	3349.744	418.718	38.670
soilmgt (FB)	14	886.872	63.348	5.850
loccow (FC)	4	147.96	36.99	3.416
$(FA \times FB)$	112	1897.28	16.94	1.564
$(FA \times FC)$	32	1250.656	39.083	3.609
$(FB \times FC)$	56	1308.72	23.37	2.158
$(FA \times FB \times FC)$	448	5330.752	11.899	1.099
farmcat (RA)	8	259.76	32.47	2.999
soilmgt (RB)	20	4414.58	220.729	20.385
loccow (RC)	12	2024.496	168.708	15.581
$(RA \times RB)$	160	6324.64	39.529	3.651
$(RA \times RC)$	96	2424.096	25.251	2.332
$(RB \times RC)$	240	5051.76	21.049	1.944
$(RA \times RB \times RC)$	1920	16250.88	8.464	0.782
Error	7927	85833.556	10.828	
Total	**11057**	136755.752		

Almimi *et al.* (2009) suggested some approaches to calculating model accuracy metrics in factorial experiments, such as the coefficient of determination (R^2) measure, and the adjusted coefficient of determination (R^2-adjusted) measure equation. Table 4.5 provides the combined model analysis information needed to calculate the model adequacy metrics.

Table 4.5: Combined CRD Model Adequacy

Source of variation	**Degrees of freedom**	**Sum of Squares**	**Mean Square**	**F**
Model	762	29340.564	1106.185	71.924
Residual	10295	158337.384	15.380	
Pure error	7927	136755.752	10.828	
Lack of fit	2368	21581.632	20.363	
Total	**11057**	**187677.948**		

The F-ratio in Table 4.5 is obtained by dividing the mean square of the combined model (1106.185) by its residual mean square (15.380). Subsequently, the overall model accuracy metrics were obtained as $R^2 = 15.6\%$ and $R^2 - adj = 9.4\%$. We conclude that both the R^2 and R^2-adjusted coefficients of determinations signify a poor model fit despite the fact that most of the main and interaction effects of the factors in the model significantly contributed to the variation in wealth scores.

4.4 Simulation Results for the CRD Classic Model

In order to establish both internal and external validity as well as the appropriateness of the partitioning approach and methods, simulation samples were used, and the simulation results compared for consistence with those from the original data set. Based on the model parameters and random effects of each partitioned model, 10 000 samples were simulated, and an analysis of variance was performed. A detailed analysis of variance for the FFF, RFR, and RRR partitioned models is presented. The analysis of variance for the other partitioned models is summarised at the end of this section.

4.4.1 The FFF Model Simulation Results

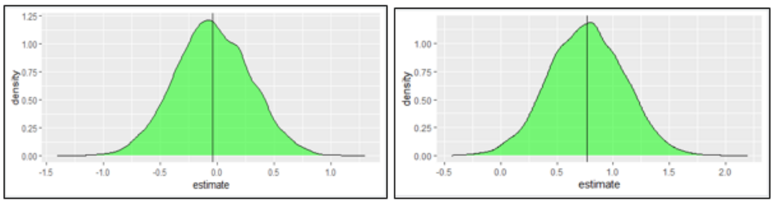
The factor of interest was *wealth* of a farm, explained by $farmcat$, a factor with three fixed levels (A, B and F); *soilmgt* factor with three fixed levels (6, 8 and 9) and *loccow* factor with two fixed levels (4 and 5). Table 4.6 summarises the estimated model parameters, standard errors, and significance p-values of the fixed-fixed-fixed (FFF) model fitted based on 10 000 simulated samples.

Table 4.6: The ANOVA Table for Simulated FFF Model in CRD

Coeficient	**Estimate**	**Standard Error**	**t value**	Pr(>|t|)
(Intercept)	2.33748	0.22079	10.587	<2e-16
farmcatB	-0.10626	0.31225	-0.340	0.7338
farmcatF	0.68474	0.31225	2.193	0.0288
soilmgt8	-0.37854	0.31225	-1.212	0.2260
soilmgt9	-0.39535	0.30923	-1.278	0.2017
loccow5	-0.43621	0.31225	-1.397	0.1631
farmcatB:soilmgt8	0.22195	0.44159	0.503	0.6155
farmcatF:soilmgt8	0.25886	0.43748	0.592	0.5544
farmcatB:soilmgt9	-0.06467	0.43946	-0.147	0.8831
farmcatF:soilmgt9	0.14482	0.43732	0.331	0.7407
farmcatB:loccow5	-0.36328	0.44159	-0.823	0.4111
farmcatF:loccow5	0.06890	0.43748	0.157	0.8749
soilmgt8:loccow5	0.44757	0.44159	1.014	0.3114
soilmgt9:loccow5	-0.06945	0.43946	-0.158	0.8745
farmcatB:soilmgt8:loccow5	0.10571	0.62160	0.170	0.8650
farmcatF:soilmgt8:loccow5	-0.52312	0.61717	-0.848	0.3971
farmcatB:soilmgt9:loccow5	0.49389	0.62299	0.793	0.4283
farmcatF:soilmgt9:loccow5	-0.41438	0.61706	-0.672	0.5022

The FFF model had a residual standard error of 1.104. Only the farm category F had a significant difference with the reference category A (p-value = 0.0288) at 5% significance level. This is supported by very low model fit measures, Multiple R-squared 0.165 (17%) and the Adjusted R-squared 0.1328 (13%).

We picked the farm category factor for analysis since it significantly contributed to the mean wealth of a farm. The distribution of differences in the mean wealth between farm categories B and F against the reference farm category A are displayed in Figure 4.1.



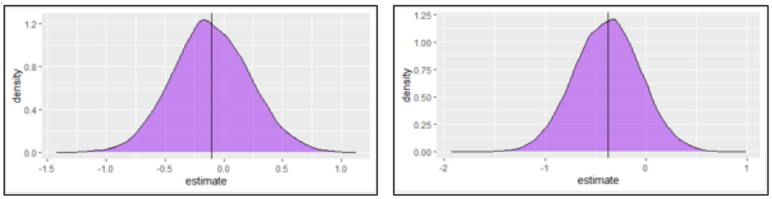
Farm Size B

Farm Size F

Figure 4.1: Mean Wealth Distribution in Farm Categories in the FFF Model

The peaks of the distribution of the coefficient of both farm categories B and F are roughly around the true values of -0.04252 and 0.76604, respectively. However, even though the coefficient of *farmcat F* is slightly overestimated, the slight discrepancy is not uncommon. There is a considerable range in the estimated *farmcat* coefficients across the 10 000 simulations, with *farmcat B* in the range (-1.25 , 1.25) and *farmcat F* in the range (-0.5 , 2.0).

Similarly, the distributions of the coefficient of *soilmgt* 8 and 9 differences in Figure 4.2. are roughly around the true values of -0.10692 and -0.3782, respectively. Even though there is a slight underestimation in soilmgt 8 coefficient and a slight overestimation in soilmgt 9, the accuracy is within acceptable terms.



Soilmgt 8

soilmgt 9

Figure 4.2: Mean Wealth Distribution in Soil Management Methods in the FFF Model

Similarly, the distribution of standard deviations across the simulation samples is displayed in Figure 4.3.

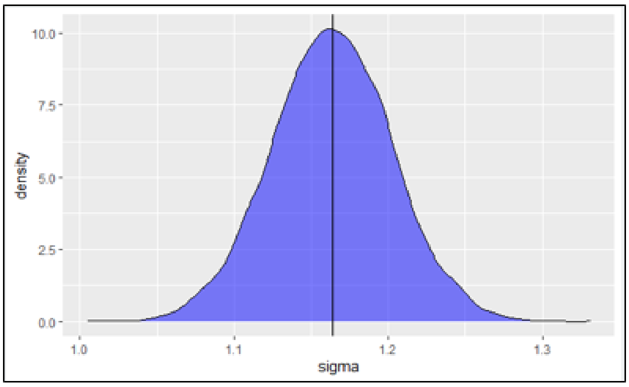


Figure 4.3: Distribution of Wealth Standard Deviation in the FFF Model

The distribution of the estimated variation is roughly centred on the true value of 1.104, with an estimated range of (1.0 , 1.30). The estimated variation ($1.104^2 = 1.219$) compares well with the variation (1.354) in the original FFF data set in Table 4.3. This shows that, on average, the model performs very well in estimating the standard deviation. This is further supported by the proportion of models that correctly rejected the null hypothesis (H_0: *farmcat F* effect = 0), given that we know the null hypothesis is not true, which was approximately 0.6334 (63%). A statistical power of 63% is reasonably fair.

4.4.2 The RFR Model Simulation Results

The RFR model involves a random factor $farmcat$, with three random levels (C, D and E); a fixed *soilmgt* factor, with three fixed levels (6, 8 and 9) and a random *loccow* factor, with four random levels (0, 1, 2 and 3). Table 4.7 summarises the variance contributed by the random effects as well as the estimated model parameters, standard errors, and significance p-values of the random-fixed-random (RFR) model fitted based on 10 000 simulated samples.

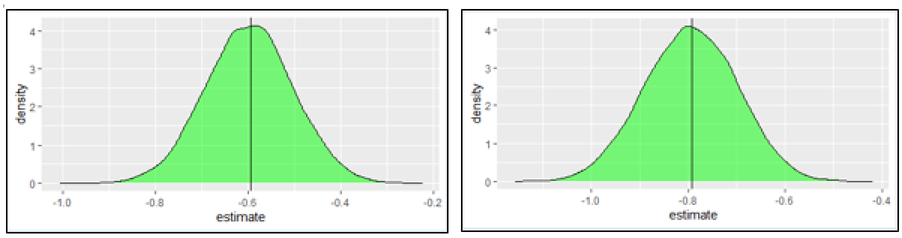
Table 4.7: The ANOVA Table for Simulated RFR Model in CRD

Random effects:			
Groups	Name	Variance	Std. Dev.
farmcat:soilmgt:loccow	(Intercept)	0.00000	0.000
soilmgt:loccow	(Intercept)	0.00000	0.000
farmcat:loccow	(Intercept)	0.00000	0.000
farmcat:soilmgt	(Intercept)	0.00000	0.000
loccow	(Intercept)	0.06654	0.258
farmcat	(Intercept)	0.00000	0.000
Residual		1.21473	1.102

Fixed effects:					
Coefficients	Estimate	Std. Error	df	t value	Pr(>|t|)
(Intercept)	3.34810	0.14421	3.99481	23.217	2.06e-05
soilmgt8	-0.51794	0.09061	873.00752	-5.716	1.49e-08
soilmgt9	-0.82895	0.09162	873.02376	-9.048	< 2e-16

The RFR model had a residual error variance of 1.21473. The *soilmgt* category differences with the reference category had a significant effect (p-value < 0.0001) on the response variable. On the other hand, only the *loccow* random factor had a noticeable variance contribution in isolation (variance = 0.06654), whereas the rest of the main and interaction effects yielded zero variance.

The distribution of differences in mean wealth between soil management strategies 8 and 9 against the reference category (*soilmgt 6*) are displayed in Figure 4.4.



Soilmgt 8

Soilmgmt 9

Figure 4.4: Mean Wealth Distribution of Soil Management in the RFR Model

The peak of the distributions of the coefficient of both soil management practices 8 and 9 are right around their true values (-0.5947 and -0.7932, respectively). The estimated *soilmgt 8* coefficient ranges from -1.0 to -0.2, whereas the coefficient of *soilmgt 9* falls in the range (-1.2 , -0.4) across the 10 000 simulations. The statistical power, or proportion of models that correctly rejected the null hypothesis (H_0: *soilmgt 8* effect = *soilmgt 8* effect = 0), given that we know the null hypothesis is not true, was 0.9964 and 1, respectively. The distributions and high statistical power suggest that the model estimated the fixed effects coefficients very well.

The distribution of standard deviations across the simulation samples, displayed in Figure 4.5, shows that the standard deviation was underestimated about 54.4% of the times, with an estimated range of 1.05 to 1.25. The estimated variation (1.21473) compares well with the error variance (1.326) in the original RFR analysis in Table 4.2. The model performs fairly well in estimating the error variance.

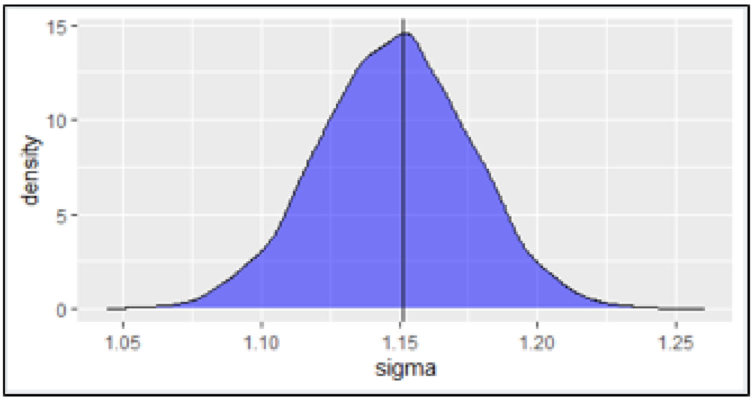


Figure 4.5: Distribution of Wealth Standard Deviation in RFR model

The distributions of the variances of the random factor *farmcat* on different sample sizes could not be displayed due to the fact that there was very little variation (0.01915) in the farm categories.

The variances of the random factor *farmcat* on different sample sizes ($n = 5, 20, 30$), where n represents the number of levels of the explanatory factor, is displayed in Figure 4.6.

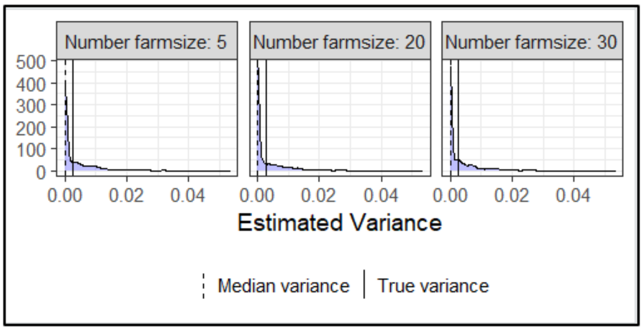


Figure 4.6: Distribution of Variances of Farm Categories in the RFR model

Although the variances obtained were very small, the three sample sizes of the *farmcat* factor levels showed a small difference between the median variance and the true variance. The estimated variance is close to the true variance, with a range of 0.000 to 0.0532 across the samples.

A further examination of the distributions of the variances of the random factor *loccow* on different sample sizes ($n = 5, 20, 30$), where n represents the number of levels of the explanatory factor, is displayed in Figure 4.7.

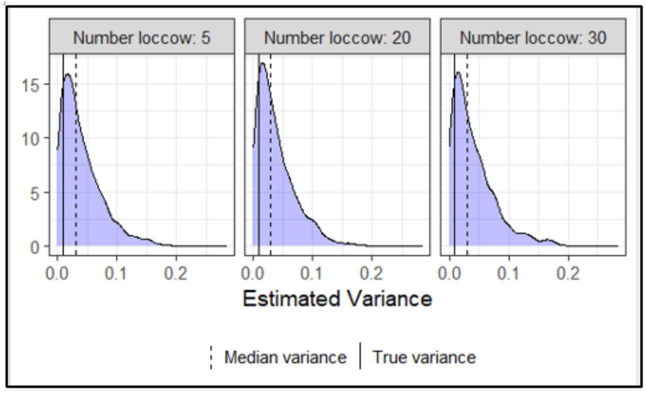


Figure 4.7: Distribution of Variances of Cattle Management Practices in the FRR Model

The three sample sizes of the *loccow* factor levels showed a considerably better variation than the *farmsize* variances, with the median variance almost equal to the true variance especially when $n = 20$. The range of variance was from 0.000 to 0.555 across the samples.

4.4.3 The RRR Model Simulation Results

The RRF model involves three random factors, $farmcat$ with three random levels (C, D and E); *soilmgt* factor, with six random levels (1 - 5 and 7), and *loccow* factor, with four random levels levels (0 - 3). Table 4.8 summarises the variance contributed by the random effects of the random-random-random (RRR) model fitted based on 10 000 simulated samples.

Table 4.8: The ANOVA Table for the Simulated RRR Model in CRD

Random effects:					
Groups			Name	Variance	Std. Dev.
farmcat:soilmgt:loccow			(Intercept)	1.184e-09	3.441e-05
soilmgt:loccow			(Intercept)	1.231e-02	1.110e-01
farmcat:soilmgt			(Intercept)	3.228e-03	5.682e-02
farmcat:loccow			(Intercept)	0.000e+00	0.000e+00
soilmgt			(Intercept)	5.777e-02	2.404e-01
loccow			(Intercept)	2.746e-01	5.240e-01
farmcat			(Intercept)	2.263e-10	1.504e-05
Residual				1.273e+00	1.128e+00
Fixed effects:					
Coefficients	Estimate	Std. Error	df	t-value	Pr(>|t|)
(Intercept)	2.8785	0.2823	3.8439	10.2	0.000637

The RRR model had a residual error variance of 1.273. The distribution of standard deviations across the simulation samples, displayed in Figure 4.8, shows that the standard deviation was slightly underestimated about 54.2% of the times, with an estimated range of 1.02 to 1.17. The estimated variation (1.273) compares well with the error variance (1.225) in the original RRR analysis in Table 4.3. The model performs very well in estimating the error variance.

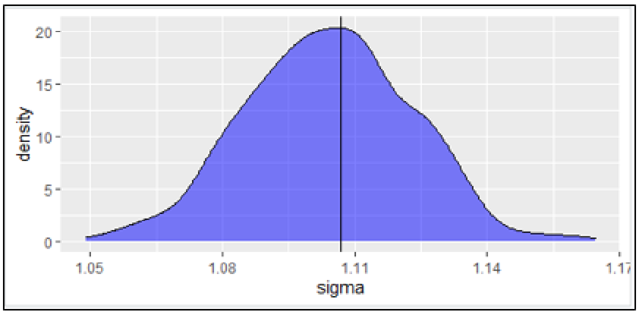


Figure 4.8: Distribution of Wealth Standard Deviation in the RRR Model

The distributions of the variances of the random factor *farmcat* on different sample sizes could not be displayed due to the fact that there was very little variation (<0.0001) in the farm categories. The distributions of the variances of the random factor *soilmgt* on different sample size ($n = 5, 20, 30$), where n represents the number of levels of the explanatory factor is displayed in Figure 4.9.

Figure 4.9: Distribution of Variances of Soil Management Practices in the RRR model

The three sample sizes of the *soilmgt* factor levels showed a considerably large variation ranging from 0.01 to 1.88 across the samples. The median variance gets approximately equal to the true variance when $n = 20$.

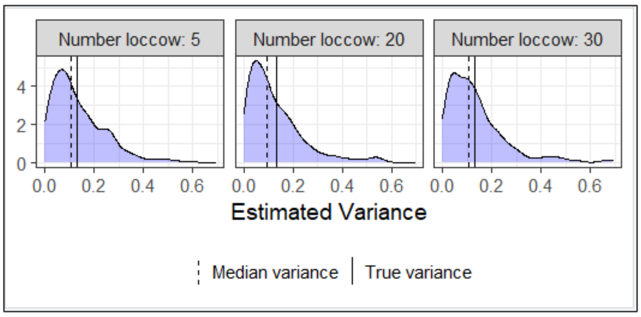


Figure 4.10: Distribution of Variances of *loccow* Levels in RRR Model

Although the variances obtained were very small, the three sample sizes of the *loccow* factor levels showed a small difference between the median variance and the true variance. The best approximation of the *loccow* true variance is when $n = 30$, with a range of 0.000 to 0.692 across the samples.

4.4.4 Summary of CRD Simulation Results

Based on 10000 samples, the simulation analysis results for the CRD linear mixed models in partitions (2) - (8) are summarised in Table 4.9.

Table 4.9: Summary of Simulation Sample Results for the CRD Models

Model	**Random Effects**	**Variance**	**Significant Fixed Effects**	**Statistical Power**
FFR	loccow residual	0.0186 1.241	farmcatB farmcatF soilmgt	1.0000 0.9995 0.9352
FRF	soilmgt residual	0.1398 1.3695	farmcat loccow	1.0000 0.9295
FRR	loccow residual	0.00918 1.2025	farmcatF	0.9998
RFF	farmcat residual	0.0192 1.759	None	
RRF	farmcat soilmgt residual	0.0000 0.0311 1.279	None	

Table 4.8 shows very little effect of farm categories ($farmcat$) and cattle management practices (*loccow*) on wealth variation across the simulation samples. In contrast, a fairly higher wealth variation due to random categories of soil management practices (*soilmgt*) was detected. Furthermore, the proportion of models that correctly rejected the $farmcat$ differences null hypothesis and *loccow* fixed-effects null hypothesis, given that we know the null hypothesis is not true, was just over 90% (statistical power > 0.9). Generally, the analysis results from the simulation samples are comparable to those of the original data. Thus, the simulated CRD partitioned models performed well in estimating the model parameters.

4.5 Numerical Example 2: Randomised Complete Block Design (RCBD)

From the experiment in Example 1, it is suspected that the ethnicity of the village where the farm is located influences the variation in the wealth score of a farm due to differences among the farmers. Although not a factor of primary interest, we consider ethnicity (*ethnic*), with four levels randomly selected from a population of numerous ethnic groups in Kenya, as a random blocking factor that explains the variation in the responses. We maintain the other factors as before, factor A ($farmcat$) with three fixed levels (A, B and F) and three random levels (C, D and E) selected from a population of possible farm sizes; factor B (*soilmgt*) with three fixed levels (6, 8 and 9) and six random levels (1-5 and 7) selected from numerous possible soil management approaches; factor C (*loccow*) with two fixed levels (4 and 5) and four random levels (0-3). A three-way randomised complete block design (RCBD) is proposed for the experiment. We apply the new approach to analyse the effect of the nuisance factor (*ethnic*) and the three explanatory factors on the wealth scores of farms. The partitioning in Table 4.1 for the models FFF, FRF, FFR, FRR, RFF, RRF, RFR and RRR were maintained, with an additional blocking factor (*ethnic*) included in the analysis. Table 4.10 shows the analysis of variance for FFF, FRF, FFR, FRR, RFF, RRF, RFR and RRR models.

The restricted maximum likelihood estimation (REML) approach was used to estimate the variance components for the unbalanced mixed models FRF, FFR, FRR, RFF, RRF, RFR and RRR models as coded in Table 4.1. From the RCBD models, there was a noticeable variation due to (*ethnic*) blocks in the FFR ($\sigma^2_{blocks} = 0.0022$) and RRR ($\sigma^2_{blocks} = 0.0025$). However, in

Table 4.10: The Analysis of Variance for Sub-models in RCBD

Model Type	**Source of variation**	**DF**	**Mean Square**	**Model Type**	**Source of variation**	**DF**	**Mean Square**
FFF	ethnic	3	1.426	RFF	ethnic	3	0.050
	FA	2	153.523		RA	2	3.058
	FB	2	6.527		FB	2	0.489
	FC	1	3.809		FC	1	16.081
	FA*FB	4	9.091		RA*FB	4	19.650
	FA*FC	2	32.767		RA*FC	2	7.261
	FB*FC	2	21.288		FB*FC	2	1.956
	FA*FB*FC	4	47.581		RA*FB*FC	4	14.491
	ERROR	439	594.292		ERROR	433	1.828
FFR	ethnic	3	1.860	RFR	ethnic	3	1.391
	FA	2	76.119		RA	2	13.261
	FB	2	6.95		FB	2	49.606
	RC	3	31.565		RC	3	35.146
	FA*FB	4	7.834		RA*FB	4	8.753
	FA*RC	6	11.830		RA*RC	6	12.161
	RB*RC	6	5.326		FB*RC	6	20.542
	FA*FB*RC	12	7.251		RA*FB*RC	12	14.103
	ERROR	850	1.371		ERROR	840	1.326
FRF	ethnic	3	0.775	RRF	ethnic	3	1.407
	FA	2	93.267		RA	2	0.290
	RB	5	22.947		RB	5	26.649
	FC	1	14.109		FC	1	3.325
	FA*RB	10	9.768		RA*RB	10	28.855
	FA*FC	2	6.431		RA*FC	2	10.248
	RB*FC	5	7.671		RB*FC	5	16.961
	FA*RB*FC	10	11.796		RA*RB*FC	10	26.950
	ERROR	851	1.286		ERROR	867	1.281
FRR	ethnic	3	0.020	RRR	ethnic	3	2.325
	FA	2	95.671		RA	2	15.578
	RB	5	35.448		RB	5	135.381
	RC	3	29.790		RC	3	72.860
	FA*RB	10	10.373		RA*RB	10	10.591
	FA*RC	6	5.139		RA*RC	6	12.928
	RB*RC	15	11.047		RB*RC	15	10.341
	FA*RB*RC	30	9.856		RA*RB*RC	30	8.524
	ERROR	1918	1.173		ERROR	1705	1.223

all the partitioned models, the variation due to ethnicity was not significant ($p - value > 0.05$).

This resulted in a minor reduction in the experimental errors due to *ethnic* blocks. Hence, the blocking factor (*ethnic*) had an insignificant contribution to the variation in wealth score, therefore it can be dropped from the analysis. Furthermore, factor A ($farmsize$), with three fixed levels, had a significant effect ($p - value < 0.0001$).

4.6 Simulation Sample Results for the RCBD Model

We present the simulation sample analysis results for the mixed model (FFR) in RCBD, and compile a summary of simulation results analysis for the rest of the partitioned models.

4.6.1 The FFR Model Simulation Results

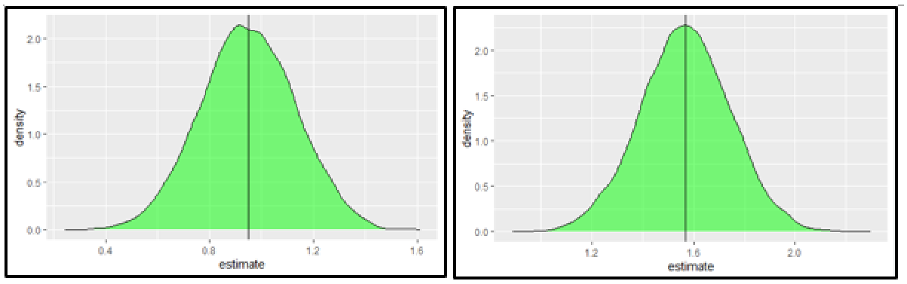
We considered two fixed factors, $farmcat$ with three fixed levels (A, B and F) and *soilmgt* factor with three fixed levels (6, 8 and 9), and one random factor, *loccow* with four random levels (0, 1, 2 and 3), randomly assigned to participants in ethnicity (*ethnic*) blocks, with four levels. Table 4.11 summarises the variation contributed by the random effects as well as the estimated model parameters, standard errors, and significance p-values of the fixed-fixed-random (FFR) model fitted based on 10 000 samples.

Table 4.11: The ANOVA Table for Simulated FFR Model in CRD

Random effects:					
Groups			Name	Variance	Std. Dev.
farmcat:soilmgt:loccow			(Intercept)	0.0483	0.2197
soilmgt:loccow			(Intercept)	0.0318	0.1782
farmcat:loccow			(Intercept)	0.00000	0.0000
block			(Intercept)	0.0134	0.1158
loccow			(Intercept)	0.1543	0.3928
Residual				1.289	1.135
Fixed effects:					
Coefficients	Estimate	Std. Error	df	t-value	Pr(>|t|)
(Intercept)	1.866960	0.273550	8.133928	6.825	0.000124
farmcatB	0.797313	0.223470	17.437356	3.568	0.002291
farmcatF	1.545278	0.226202	18.275184	6.831	1.98e-06
soilmgt8	0.213974	0.257115	18.176746	0.832	0.416091
soilmgt9	0.144171	0.257128	18.180638	0.561	0.581847
farmcatB:soilmgt8	-0.001440	0.315528	17.317329	-0.005	0.996411
farmcatF:soilmgt8	-0.005998	0.319049	18.090877	-0.019	0.985206
farmcatB:soilmgt9	0.153971	0.315709	17.361770	0.488	0.631864
farmcatF:soilmgt9	0.082372	0.319498	18.186580	0.258	0.799442

The FFR model had a residual error variance of 1.289, which is comparable to the classic model estimate. The fixed factor *farmcat* had significant category differences (p-value < 0.001) at a 5% significance level with respect to the reference category, and non-significant interactions with other factors. The model could capture considerably few random effects from the *loccow* (variance of 0.1543), *block* (variance 0.0134) factors and a residual error variance of 1.289.

We consider the distribution of farm categories in contribution to the mean wealth of a farm across the samples. The distribution of the differences in mean wealth between farm category *B* and *F* against the reference category farm category A are displayed in Figure 4.11.



Farmsize B

Farmsize F

Figure 4.11: Distribution of Mean Wealth Between Farms

The peaks of the distribution of both farm categories B and F are slightly shifted to the left of the true values of 0.95 and 1.5678, respectively. This is an indication that the coefficients of farm categories are slightly underestimated in the model, but the slight discrepancy is not a cause for alarm. The estimated *farmcat B* coefficient ranges from 0.4 to 1.6, whereas the coefficient of *farmcat F* falls in the range (1.0, 2.2) across the 10 000 simulations.

The distribution of standard deviations across the simulation samples is displayed in Figure 4.12. The distribution of the estimated variation is roughly centred on the true standard deviation (1.135), with an estimated range of (0.0 , 1.30).

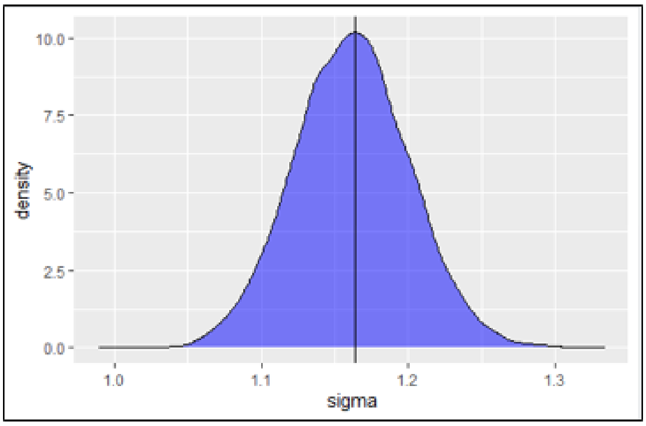


Figure 4.12: Distribution of Wealth Standard Deviation in the FFR Model

The distribution shows that the variance is slightly overestimated 52.2% of the times. Although the estimated variation (1.289) is slightly less than the variation (1.375) in the original FFR data set in Table 4.2, the two compare well. The model performs very well in estimating the standard deviation. This is further supported by the high proportion (statistical power = 1.0 and 0.9995) of models that correctly rejected the null hypothesis (H_0: *farmcat* effect = 0), given that we know the null hypothesis is not true.

4.6.2 Summary of RCBD Simulation Results

Based on 10000 samples, the simulation analysis results for the RCBD linear mixed models in partitions (2) - (8) are summarised in Table 4.12. There is almost no significant effect of block and random farm categories in wealth variation across the simulation samples. However, a moderate wealth variation due to soil management practices (*soilmgt*) and the cattle management practices (*loccow*) was detected in the random effects (RRR) model.

Table 4.12: Summary of Simulation Sample Results for RCBD Models

Model	Random Effects	Variance	Significant Fixed Effects	Statistical Power
FFF	block residual	0.0000 1.241	farmcatF	0.6406
FRF	block soilmgt residual	0.0000 0.0838 1.2840	farmcatF	0.9343
FRR	block soilmgt loccow residual	0.0000 0.0287 0.0085 1.1540	farmcatF	0.9999
RFF	block farmcat residual	0.0000 0.0000 1.9760	None	
RFR	block farmcat loccow residual	0.0000 0.0190 0.0000 1.4464	soilmgt8 soilmgt9	0.9965 1.0000
RRF	block farmcat soilmgt residual	0.0000 0.0023 0.0044 1.2981	None	
RRR	block farmcat soilmgt loccow residual	0.0002 0.0000 0.2860 0.1940 1.2050		

Furthermore, the proportion of models that correctly rejected the farm category differences null hypothesis and soil management practices fixed-effects null hypothesis, given that we know the null hypothesis is not true, was just over 64% (statistical power range 0.64 - 0.999) and 99% (statistical power > 0.99), respectively. Generally, the analysis results from the simulation samples are comparable to those of the original data. Thus, the simulated partitioned models performed well in estimating the model parameters.

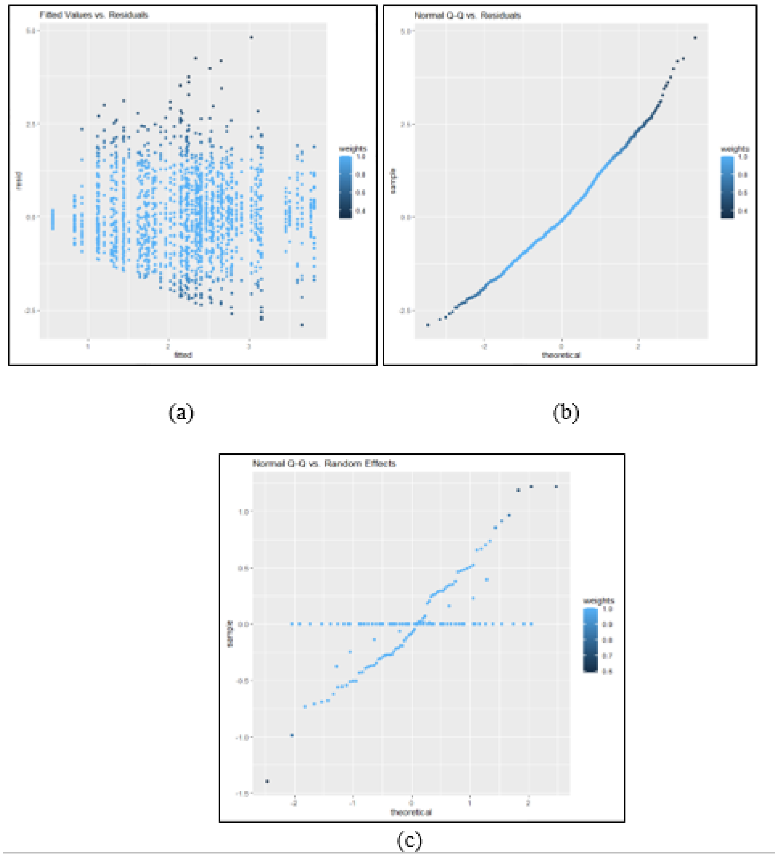
4.7 Linear mixed Model Results for Contaminated Data

Using Example 1 data set, we apply the new approach on the analysis of farmers' experiences and wealth scores based on the assets accumulated on the farms. Consider the first factor ($farmcat$) to have three fixed levels (A, B and F) and three random levels (C, D and E) selected from a population of possible farm sizes; the second factor (*soilmgt*) to have three fixed levels (6, 8 and 9) and six random levels (1-5 and 7) selected from numerous possible soil management approaches; and the third factor (*loccow*) to have two fixed levels (4 and 5) and four random levels (0-3). An unbalanced three-way completely randomised design (CRD) is proposed for the experiment. First, we use the R package (*robustlmm*) to assess and detect outlier contamination in the random effects of each of the partitioned models 2 - 8 of Table 3.9. We then fit a robust linear mixed model using the *robustlmm* package when outliers have been detected and confirmed to have little influence on the response variable.

4.7.1 Contamination Analysis for Random Effects

For easy reference and detailed contamination analysis, model FRR (4) was selected. Figure 4.13 displays the residual analysis plots and the normal Q-Q plot for the predicted random effects of model (4).

The dark-coloured plots indicate the observations with low robustness weights and hence possible outliersobservations with low robustness weights and hence possible outliers. The Q-Q plot shows very minimal deviation from normality since the plots are fairly following the diagonal line. The deviating feature in Figure 4.13 pronounces the variance-mean relationship. Interested readers could possibly consider applying simple data transformations (e.g., square



(a)

(b)

(c)

Figure 4.13: Residual Analysis Plots Showing (a) Fitted Values vs. Residuals, (b) Normal Q-Q vs. Residuals, (c) Normal Q-Q vs. Random Effects for FRR (4) Model Robust Fit

root or logarithm) to stabilise the variance. However, for the purpose of illustrating the robust linear mixed model methodology, we skip other alternatives and proceed to conduct the contamination analysis and fitting the robust linear mixed model.

A summary of the contamination analysis of model (4) is given in Table 4.13. A detailed comparison of the estimated coefficients (with the corresponding standard deviation given in parentheses) and variance components from the classic (clm) and robust ($rlm1$ and $rlm2$) fits for the FRR (4) model. The contamination analysis results confirm the presence of outliers and little contamination in the data set. The analysis detected a total of 417 out of 1576 residuals (i.e., approximately 26%) and a total of 8 out of 128 estimated random effects (i.e., approximately 6%) as possible outliers. The minimum robust weights for residuals and random effects were around 0.3, while the maximum weight was 0.99 for both robust models rlm1 and rlm2.

Table 4.13: Contamination Analysis for the FRR Model

Model		Fit			Robust Weights	
		clm	*rlm*1	*rlm*2	≈ 1	Rem
FFR(4)	**Coefficient:**				Residuals:	
	intercept	2.0095	1.9476	1.9463	1576	417
		(0.237)	(0.237)	(0.215)		
	FarmcatB	0.0723	0.0723	0.0718		
		(0.192)	(0.192)	(0.191)		
	Farmcat	0.7293	0.7293	0.7306	Random effects:	
		(0.192)	(0.192)	(0.191)	128	8
	Variance Component:					
	Farmcat:Soilmgt	0.0110	0.0000	0.0000		
	Farmcat:Loccow	0.0000	0.0000	0.0000		
	Soilmgt:Loccow	0.0223	0.0000	0.0000		
	Farmcat:Soilmgt:Loccow	0.3192	0.3551	0.3515		
	Soilmgt	0707	0.1119	0.0898		
	Loccow	0.0396	0.0698	0.0467		
	Residual	1.1710	1.1793	1.1594		

The robust estimates of the variance of random effects and the random errors were slightly inflated in the $rlm2$ (higher efficiency) model, which proves that robust model $rlm2$ performed better than $rlm1$ (lower efficiency) in this model. The same results are confirmed by the residual analysis plots in Figure 4.13.

4.7.2 Robust Linear Mixed Model Results

With the little contamination due to the presence of detected outliers, we fitted two robust linear mixed models ($rlm1$ and $rlm2$) using different tuning parameters of the smoothed Huber function. Table 4.14 gives the options for other possible tuning constants for the smoothed Huber Ψ-function (Koller and Stahel, 2011), where parameters k and parameter s determine the interval $(+/-k)$ and the smoothness of the bend of the ψ-function, respectively. The term "efficiency" is the percentage measure of the efficiency of the regression estimator. For high robustness and low efficiency, rho-functions (smoothed Huber) that were used for fitting random effects, and residuals in robust fit $rlm1$ were set at default parameters ($k = 1.345$, $s = 10$) and Proposal II ($k = 1.345$, $s = 10$), respectively. On the other hand, the rho-function parameters for the robust model $rlm2$ were tuned as $k = 1.345$, $s = 10$, and Proposal II ($k = 2.28$, $s = 10$), in order to achieve increased efficiency.

Table 4.14: Tuning Parameters for Scale Estimates for the Huber Ψ-Function

Efficiency	k for $\hat{\mu}$	k for $\hat{\sigma}$	k for $\hat{\sigma}$, Prop. II
0.80	0.53	0.50	1.49
0.85	0.73	0.71	1.69
0.90	0.98	1.08	1.94
0.95	1.345	1.66	2.28

The rest of the contamination analysis for the other partitioned models was summarised in Table 4.15. A similar pattern is seen in most of the partitioned models; that is, the robust method slightly inflates the estimates of the variance components as expected (Koller, 2016).

Higher efficiency robust fit $rlm2$ produced better estimates of variance components closer to the classic fit in models 2, 3, 6 and 8, while the lower efficiency robust fit $rlm1$ performed better for models 4, 5, and 7. On average, around 6% contamination (ratio of robust weight far from 1 to robust weights close to 1) was detected in the estimated random effects, and around 26% in the residual errors. The robust fits with lower efficiency were fairly comparable to those with higher efficiency even though different results in different partitions were realised.

4.7.3 Contaminated Simulation Sample Results

Simulation samples of size 10000 were generated using the robust fit estimated parameters and standard deviation to validate robust methods on partitioned linear mixed models 1 - 8. The simulation results for the robust fits indicate that the simulated robust estimates of coefficients and variance components were comparable with both the classic and robust estimates. For example, the simulated robust estimates of error variances: $\hat{\sigma}^2_{\epsilon_1} = 1.219$; $\hat{\sigma}^2_{\epsilon_2} = 1.417$; $\hat{\sigma}^2_{\epsilon_3} = 1.190$; $\hat{\sigma}^2_{\epsilon_4} = 1.145$; $\hat{\sigma}^2_{\epsilon_5} = 1.682$; $\hat{\sigma}^2_{\epsilon_6} = 1.080$; $\hat{\sigma}^2_{\epsilon_7} = 1.335$; and $\hat{\sigma}^2_{\epsilon_8} = 1.329$, were consistent with the robust estimates counterparts. Similar results for hypotheses tests HA - HC were realised in most of the partitions. This indicates that the robust methods used returned consistent results in the simulation samples.

4.8 Conclusion

The partitioning approach allows for improved precision in both particularised and combined analysis of the experimental data for the targeted inference space (Chaka and Njuho, 2021). However, the method is not immune to the influence of contamination from sources such as outliers. Contamination analysis on random effects in each partitioned data set was a neces-

Table 4.15: Summary of Contamination Analysis for Models

Model		Fit			Robust Weights	
		clm	*rlm*1	*rlm*2	≈ 1	Rem
FFR(2)	**Coefficient:**				Residuals:	
	intercept	1.7417	1.7359	1.7363	695	194
	Variance Component:				Random effects:	
	Loccow	0.0856	0.0834	0.0744	60	4
	Residual	1.3726	1.3314	1.3493		
FRF(3)	**Coefficient:**				Residuals:	
	intercept	2.3164	2.4460	2.4195	711	179
	Variance Component:				Random effects:	
	Soilmgt	0.0844	0.0000	0.0304	66	6
	Residual	1.2840	1.1950	1.2417		
RFF(5)	**Coefficient:**				Residuals:	
	intercept	2.9327	2.9678	2.9692	355	99
	Variance Component:				Random effects:	
	Farmcat	0.0000	0.0000	0.0000	34	2
	Residual	1.8153	1.8153	1.7903		
RFR(6)	**Coefficient:**				Residuals:	
	intercept	3.2070	3.0850	3.0782	687	192
	Variance Component:				Random effects:	
	Farmcat	0.0026	0.0000	0.0000	72	4
	Loccow	0.0683	0.0649	0.0000		
	Residual	1.3260	1.0439	1.2291		
RRF(7)	**Coefficient:**				Residuals:	
	intercept	2.8080	2.8080	2.8080	718	188
	Variance Component:				Random effects:	
	Farmcat	0.0000	0.0000	0.0000	73	8
	Soilmgt	0.0321	0.0000	0.0000		
	Residual	1.2810	1.2840	1.2867		
RRR(8)	**Coefficient:**				Residuals:	
	intercept	2.5480	2.5350	2.5220	1420	360
	Variance Component:				Random effects:	
	Farmcat	0.0010	0.0000	0.0000	128	11
	Soilmgt	0.4128	0.5976	0.4817		
	Loccow	0.1307	0.2403	0.1742		
	Residual	1.2249	1.2476	1.2249		

sary process that successfully identified prevalent outliers and their expected impact on the analysis. Table 4.15 confirms that the robust method flagged a considerably small number of observations as outliers in each partition, whose influence was ultimately neutralised by the robust estimation process. As expected in each robust fit, the slightly inflated estimates of random effects and random error variance confirm that the robust estimates were better than the classic estimates (Koller, 2016). Furthermore, the standard deviation of the estimates and the variance components from the simulation samples confirm that the simulated estimates match the counterparts from both the classic and robust methods. The partitioned analysis, coupled with robust estimation procedures, has proven to be an essential approach to control for the influence of outlier contamination in experimental data. Therefore, the robust methods worked consistently well to warrant both internal and external validity. Chapter 5 deals with the application of the partitioning approach in multi-stratum experimental designs, such as split-split-plot experiments, that require multiple randomisation processes to generate different levels of precision.

CHAPTER 5

LINEAR MIXED MODELS IN SPLIT-SPLIT-PLOT DESIGN

This chapter presents the results from a publication by Chaka and Njuho (2021), appearing in the International Journal of Agricultural and Statistical Sciences, Volume 17, Issue Number 2, 2021.

5.1 Introduction

Linear mixed models find application as statistical tools for analysing factorial structure data. These models have become more attractive due to their essential property of being able to accommodate both fixed and random effects (Pan and Shang, 2018). Application of linear mixed models on studies that involve multi-step processes requires care when selecting and specifying factors. The construction of an adequate statistical model that explains particular relationships amongst the variables involved is of interest to research scientists (Smith and Edwards, 2017). Failure to select an appropriate structure of a linear mixed model for a complex process can greatly impact the treatment effects and error estimates of the model.

Split-plot designs have taken the centre stage in optimal designs for response surface experiments which usually involve complex multi-step processes that focus on determining the predictive capability of the design (Jones and Goos, 2012a). When designing experiments for complex multi-step processes, researchers need the skills to combine process and a mixture factors, and specify a model in order to obtain an optimal design to with an optimality criterion (Raminez *et al.*, 2010). The methods for constructing multi-stratum response-surface designs, of which split-plot and split-split-plot designs structures are special cases, include the D-optimality criterion, which aims to minimise the variance of the factor effect estimates in an omnibus sense (Jones and Goos, 2012b; Trinca and Gilmour, 2017); the I-optimality criterion, which minimises the average variance of prediction over the region of experimentation (Goos *et al.* 2020); the new Bayesian compound D-optimal design criterion, which pays attention to both the variance components and the fixed treatment effects (Mylona *et al.*, 2020), among others. In cases where prediction aspects of the system are of interest, the I-optimal split-plot design is recommended over the D-optimality criterion for generating response surface designs since it has low prediction variance in much of the design space and also gives reasonably precise parameter estimates (Njoroge *et al.*, 2017; Jones and Goos, 2012b; Nguyen and Pham, 2015).

Constructing a linear mixed model for a split-split plot structure in RCBD considers intra-block analysis for the fixed effects and inter-block analysis for the random effects (Dixon, 2016). Fitting block effects as fixed and proceeding to estimate treatment effects eliminates block effects, leading to intra-block analysis. In many situations, blocks contain influential effects on treatment combinations, which may be essential for estimating treatment effects. In order to gain more efficient results, inter-block information lost through intra-block analysis can be recovered through fitting blocks as a random effect. Recovery of the inter-block information and combining the same with the intra-block information provide precise information (Shah, 1970; Möhring *et al.*, 2015). The amount of inter-block information to be recovered depends on the blocking factor's contribution to the variation in experimental units, as explained by the reduction in experimental error. In other words, if the block effects are significant, then the amount of information which may be recovered from the inter-block analysis will be small, and vice-versa. Depending on the precision of variance estimates, recovery of inter-block information may be worthwhile (Möhring *et al.*, 2015), but the challenge is in deciding when to pursue the recovery.

The choice of modelling blocks as fixed or random effects has some implications analysis of results (Dixon, 2016). The combination of inter- and intra-block information produces the generalised least squares estimator, and the corresponding analysis is called mixed model analysis (Shah, 1970). When using the SAS *PROC MIXED* procedure, the least squares means for treatments correspond to the combined intra- and inter-block estimates of the treatment effects. Other studies involving factors that are mostly quantitative in nature, require the use of optimal split-plot designs and response-surface models (Nguyen and Pham, 2015; Jones and Goos, 2012a; Macharia and Goos, 2010). Complications in interpretation and analysis from such designs can be experienced (Piepho and Edmondson, 2018).

Most agricultural and industrial experiments involve multi-stratum experimental designs. Split-plot designs are the most widely used experimental designs for studies which apply restricted randomisation on some hard-to-vary factors when practical limitations and issues related to time and cost prevail (Trinca and Gilmour, 2017). The conventional approach considers a factor as either fixed or random in a distinct form. Classification of factors used in an experiment into either fixed or random takes place at the onset of the experiment. As dictated by the selected experimental design, the application of the levels of these factors to the experimental units follow a randomisation procedure, which leads to different levels of precision. After that, the estimation process and hypotheses testing consider model assumptions and attainable precision. Some situations arise when either a hard-to-change factor or an easy-to-change factor consists of both fixed and random levels in a basic split-plot design. An example is when new strategies or improvements are introduced to be tested against the numerous existing techniques. In this case, the researcher needs to conceptualise a situation where each of the factors has some levels that are fixed (new strategies) and others that are random (old and existing strategies). The combination of these factors at all possible levels results in a split-plot or split-split-plot design arranged in CRD or RCBD. Under these conditions, we extend the concept by Njuho and Milliken (2005, 2009) to three factors and formulate a linear mixed model considering the model construction, estimation, hypotheses testing and inferential space processes.

5.2 Methodology

An experimental design with a basic split-plot structure exists when, due to conditions that complicate the complete randomisation of treatment combinations, the levels of some hard-to-set treatment factors are kept the same for the runs within the same whole-plot, while the easy-to-change treatment factor levels are subsequently varied on a run-to-run basis. The nested blocking structure in a split-plot design is such that subplots are nested within the whole-plots, which may also be nested within blocks for a blocked split-plot design. Extension of splitting each plot to accommodate additional factors gives rise to split-split-plot design, among others.

5.2.1 Model Construction

We consider a replicated factorial experiment in a split-split-plot design with random blocks, one whole-plot factor A, subplot factor B and sub-subplot factor C, where each of the experimental factors consists of both fixed and random levels. Let a, b and c be levels of factors A, B and C, respectively, where the whole-plot factor A has f_a fixed levels and r_a random levels ($a = f_a + r_a$); the subplot factor B has f_b fixed levels and r_b random levels ($b = f_b + r_b$); the sub-subplot factor C has f_c fixed levels and r_c random levels ($c = f_c + r_c$); with r replicates per treatment.

5.2.1.1 The General Linear Mixed Model

We consider a conventional linear model where the factors are not classified into either fixed or random for a split-split-plot treatment arrangement with a whole-plot factor laid in an RCBD, where each treatment combination is replicated r times. Define the conventional linear mixed model for the split-split-plot treatment structure in an RCBD as

$$y_{ijkl} = \mu + \pi_l + \alpha_i + \epsilon_{il}^{(1)} + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijl}^{(2)} + \gamma_k + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk} + \epsilon_{ijkl}^{(3)}, \quad (5.1)$$

where y_{ijkl} ($i = 1, 2, ..., a$; $j = 1, 2, ..., b$; $k = 1, 2, ..., c$; $l = 1, 2, ..., r$) is the experimental unit in the i^{th} whole-plot receiving the main-plot factor A, j^{th} subplot factor B and k^{th} sub-subplot factor C in the l^{th} block; $\epsilon_{il}^{(1)}$ is the whole-plot error associated with the factor A, $\epsilon_{il}^{(1)} \sim N(0, \sigma_{\alpha\pi}^2)$; $\epsilon_{ijl}^{(2)}$ is the subplot error associated with the factor B, $\epsilon_{ijl}^{(2)} \sim N(0, \sigma_{\alpha\pi}^2)$; α_i the whole-plot effects; β_j is the subplot effects; γ_k is the split-split-plot effects; π_l is the effect of the l^{th} block; $(\alpha\beta)_{ij}$, $(\alpha\gamma)_{ik}$, $(\beta\gamma)_{jk}$, and $(\alpha\beta\gamma)_{ijk}$ are the interaction effects of the three factors A, B and C; and $\epsilon_{ijkl}^{(3)}$ is the split-split-plot random error term, $\epsilon_{ijkl}^{(3)} \sim N(0, \sigma_{\epsilon}^2)$.

5.2.1.2 Proposed Linear Mixed Model

Suppose we have information on the factors allowing for the fixed and random levels for each factor to be known. Model (5.1) can be partitioned based on the combinations of the levels of factors A, B and C resulting in eight sub-models which are coded as in Table 4.1. With i, j

and k as defined in (5.1), we express each of the partitioned models (1-8) in matrix and vector form as

$$\mathbf{y}_p = \mathbf{X}_p\boldsymbol{\beta} + \boldsymbol{\epsilon}_p, \tag{5.2}$$

where, $\mathbf{y}_p : N \times 1$ $(p = 1, ..., 8)$ is the response vector of the p^{th} partitioned model, $\mathbf{X}_p = (\mathbf{1}\ \mathbf{X}_1\ \mathbf{X}_2 ...\ \mathbf{X}_m)$ is the design matrix of known coefficients of the fixed and/or random effects parts of the model, $\boldsymbol{\beta} = [\mu,\ \pi_l,\ \alpha_i,\ \beta_j,\ \gamma_k,\ (\alpha\beta)_{ij},\ (\alpha\gamma)_{ik},\ (\beta\gamma)_{jk},\ (\alpha\beta\gamma)_{ijk})]$ is a vector of coefficients corresponding to block effects, factors A, B, C and their respective interaction effects, $\boldsymbol{\epsilon}_p = (\epsilon^{(1)}_{il},\ \epsilon^{(2)}_{ijl},\ \epsilon^{(3)}_{ijkl})$ is a vector of whole-plot, split-plot and experimental error terms in the p^{th} partition.

We illustrate the partitioning approach using the random-fixed-random (RFR) partitioned model, which is from the partition with r_a random levels of whole plot factor A, f_b fixed levels of subplot factor B and r_c random levels of sub-subplot factor C. The random-fixed-random (RFR) split-split-plot mixed model for the factors A, B and C is expressed as

$$y_{6_{ijkl}} = \mu_6 + \pi_{6_l} + \alpha_{6_i} + \epsilon^{(1)}_{6_{il}} + \beta_{6_j} + (\alpha\beta)_{6_{ij}} + \epsilon^{(2)}_{6_{ijl}} + \gamma_{6_k} + (\alpha\gamma)_{6_{ik}} + (\beta\gamma)_{6_{jk}} + (\alpha\beta\gamma)_{6_{ijk}} + \epsilon^{(3)}_{6_{ijkl}}, \tag{5.3}$$

where the subscript 6 in the model denotes the 6^{th} partition as defined in Table 4.1, and all other components are as defined in (5.1). The model assumptions and estimation techniques have been discussed in detail in Section 3.8.4. The partitioning approach conforms with Henderson's (1953) procedure of estimating variance components for random effects in (5.3).

Henderson's (1953) approach requires expressing the general linear mixed (5.3) in matrix form as

$$\mathbf{y}_6 = \mathbf{X}_6\boldsymbol{\beta} + \boldsymbol{\epsilon}_6, \tag{5.4}$$

where $\boldsymbol{\beta}$ is partitioned as $\boldsymbol{\beta'} = (\mu, \boldsymbol{\beta}'_1, \boldsymbol{\beta}'_2, ..., \boldsymbol{\beta}'_m,)$ is a vector of all the effects in the model, be they fixed, random or mixed. We assume $E(\boldsymbol{\epsilon}) = \mathbf{0}$, $Var(\boldsymbol{\epsilon}) = \sigma^2_\epsilon\mathbf{I}$. Similarly, the incidence matrix $\mathbf{X}_6$ conforms with the partitions of $\boldsymbol{\beta}$ and is partitioned as $\mathbf{X}_6 = (\mathbf{1}\ \mathbf{X}_{6_1}\ \mathbf{X}_{6_2} ...\ \mathbf{X}_{6_m})$. The reduction in the sum of squares due to fitting the linear model (5.4) is given by Searle and

Gruber (2017),

$$R(\boldsymbol{\beta}) = \hat{\boldsymbol{\beta}}'\mathbf{X}_6'\mathbf{y}_6 = \mathbf{y}_6'\mathbf{X}_6(\mathbf{X}_6'\mathbf{X}_6)^{-1}\mathbf{X}_6'\mathbf{y}_6, \tag{5.5}$$

provided $(\mathbf{X}_6'\mathbf{X}_6)$ has full rank, the absence of which the generalised inverse $(\mathbf{X}_6'\mathbf{X}_6)^{-}$ is used. The reduction in sum of squares is considered in light of the expected value of the quadratic form $\mathbf{y}'\mathbf{Q}\mathbf{y}$. According to Henderson's (1953) methods I and II, the expectation of a random sampling from model (5.4) is given by

$$E(\mathbf{y}_6'\mathbf{Q}\mathbf{y}_6) = E(\mathbf{y}_6')\mathbf{Q}E(\mathbf{y}_6) + tr\left\{\mathbf{Q}Var(\mathbf{y}_6)\right\}, \tag{5.6}$$

where $\mathbf{Q} = \mathbf{X}_6(\mathbf{X}_6'\mathbf{X}_6)^{-}$, the notation "tr" symbolises the trace operation, and $Var(\mathbf{y}_6)$ is the variance covariance matrix of $\mathbf{y}_6$. A case when the model in (5.6) has $\boldsymbol{\beta}$ as a vector of fixed effects, with $E(\mathbf{y}_6) = \mathbf{X}_6\boldsymbol{\beta}$ and $Var(\mathbf{y}_6) = \sigma_\epsilon^2\mathbf{I}_{N_6}$, yields equation (5.8) as,

$$E(\mathbf{y}_6'\mathbf{Q}\mathbf{y}_6) = (\boldsymbol{\beta}'\mathbf{X}_6')\mathbf{Q}(\mathbf{X}_6\boldsymbol{\beta}) + \boldsymbol{\epsilon}_{\epsilon_6}^2 tr(\mathbf{Q}). \tag{5.7}$$

Letting $\mathbf{Q} = \mathbf{X}_6(\mathbf{X}_6'\mathbf{X}_6)^{-}\mathbf{X}_6'$ in (5.7) reduces the quadratic form to,

$$R(\boldsymbol{\beta}) = \boldsymbol{\beta}'\mathbf{X}_6'\mathbf{X}_6\boldsymbol{\beta} + \boldsymbol{\epsilon}_{\epsilon_6}^2 tr\left\{\mathbf{X}_6(\mathbf{X}_6'\mathbf{X}_6)^{-}\mathbf{X}_6'\right\} = \boldsymbol{\beta}'\mathbf{X}_6'\mathbf{X}_6\boldsymbol{\beta} + \boldsymbol{\epsilon}_{\epsilon_6}^2 r(\mathbf{X}_6) \tag{5.8}$$

When (5.4) is a mixed model, we have $\boldsymbol{\beta}$ partitioned as $\boldsymbol{\beta}' = (\boldsymbol{\beta}_1', \boldsymbol{\beta}_2', ..., \boldsymbol{\beta}_k')$, and the incidence matrix $\mathbf{X}_6$ conforms with the partitions of $\boldsymbol{\beta}$ and is partitioned as $\mathbf{X}_6 = (\mathbf{X}_{6_1}\ \mathbf{X}_{6_2} ...\ \mathbf{X}_{6_k})$. Let $\boldsymbol{\beta}_1$ (including the overall mean μ) be the fixed effects of the model, we have $E(\mathbf{y}_6) = \mathbf{X}_{6_1}\boldsymbol{\beta}_1$; and $Var(\mathbf{y}) = \mathbf{X}_{6_2}Var(\boldsymbol{\beta}_2)\mathbf{X}_{6_2}' + \mathbf{X}_{6_3}Var(\boldsymbol{\beta}_3)\mathbf{X}_{6_3}' + ... + \mathbf{X}_{6_k}Var(\boldsymbol{\beta}_k)\mathbf{X}_{6_k}' + \sigma_\epsilon^2\mathbf{I}$, where $Var(\boldsymbol{\beta}_3) = \sigma_\epsilon^2\mathbf{I}$, for $i = 2, 3, ..., k$. The quadratic form for the mixed model (5.7) becomes

$$E(\mathbf{y}_6'\mathbf{Q}\mathbf{y}_6) = (\boldsymbol{\beta}_1'\mathbf{X}_{6_1}')\mathbf{Q}(\mathbf{X}_{6_1}\boldsymbol{\beta}_1) + \sum_{i=2}^{k}\sigma_i^2 tr(\mathbf{Q}\mathbf{X}_i\mathbf{X}_i') + \boldsymbol{\epsilon}_{\epsilon_6}^2 tr(\mathbf{Q}). \tag{5.9}$$

However, when all effects in (5.6) are random, except for the overall mean, μ, we substitute $\boldsymbol{\beta}_1$ by a scalar μ and $\mathbf{X}_{6_1}$ by a vector of 1's in (5.9) to get the quadratic form,

$$E(\mathbf{y}_6'\mathbf{Q}\mathbf{y}_6) = (\mu^2\mathbf{1}'\mathbf{1})\mathbf{Q}(\mathbf{1}'\mathbf{1}) + \sum_{i=2}^{k}\sigma_i^2 tr(\mathbf{Q}\mathbf{X}_i\mathbf{X}_i') + \boldsymbol{\epsilon}_{\epsilon_6}^2 tr(\mathbf{Q}), \tag{5.10}$$

where μ is a scalar and $\mathbf{1}$ is a vector of 1's.

These results can be used to find the expectations of any quadratic form $\mathbf{y}'\mathbf{Q}\mathbf{y}$ of the model $\mathbf{y}$ that involves the partitioning of a vector of all the effects, $\boldsymbol{\beta}$, into sub-vectors of main an interaction terms, be they fixed (5.8), random (5.10) or mixed (5.9). Henderson's (1953) methods I, II and III as well as Searle and Gruber's (2017) method IV apply these results to derive expected mean squares and their subsequent expectations that are used to decide the denominators for testing each component in linear models. As echoed by Searle and Gruber (2017), the most preferred method for linear mixed models is Henderson's (1953) method III, which involves computing mean squares by conventional least squares analysis of non-orthogonal data, equating the mean squares to their expectations and solve for the unknown variances.

Following Searle and Gruber's (2017) approach, we estimate the variance components of a linear mixed model in a split-split plot design, which is based on Henderson's (1953) method III, since it is convenient for calculating generalised inverses of large matrices. Moreso, Henderson's method III yields variance components not affected by fixed effects. First, we modify equation (5.6) for the quadratic form to cater for $E(\boldsymbol{\beta})$ of any nature.

$$\begin{aligned} E(\mathbf{y}_6'\mathbf{Q}\mathbf{y}_6) &= E(\boldsymbol{\beta}')\mathbf{X}_6'\mathbf{Q}\mathbf{X}_6E(\boldsymbol{\beta}) + tr\left\{\mathbf{Q}[\mathbf{X}_6Var(\boldsymbol{\beta})\mathbf{X}_6' + \sigma_\epsilon^2\mathbf{I}]\right\} \\ &= tr[\mathbf{X}_6'\mathbf{Q}\mathbf{X}_6E(\boldsymbol{\beta}\boldsymbol{\beta}')] + \sigma_\epsilon^2 tr(\mathbf{Q}). \end{aligned} \tag{5.11}$$

Hence the reduction in the sum of squares due to fitting the full model (5.4), irrespective of whether $\boldsymbol{\beta}$ is fixed, random or mixed, is given by

$$R(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, ..., \boldsymbol{\beta}_k) = \hat{\boldsymbol{\beta}}'\mathbf{X}_6'\mathbf{y}_6 = \mathbf{y}_6'\mathbf{X}_6(\mathbf{X}_6'\mathbf{X}_6)^-\mathbf{X}_6'\mathbf{y}_6. \tag{5.12}$$

Taking the differences of the reductions in the sum of squares due to fitting appropriate submodels from the full model, reduction yields unbiased variance components of the full model. For example, with the reduction due to fitting the full model in (5.5), the reduction due to fitting a reduced model $R(\boldsymbol{\beta}_1|\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, ..., \boldsymbol{\beta}_k)$, is supplied by

$$\begin{aligned} E(\boldsymbol{\beta}_1|\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, ..., \boldsymbol{\beta}_k) &= E[R(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, ..., \boldsymbol{\beta}_k) - R(\boldsymbol{\beta}_2, ..., \boldsymbol{\beta}_k)] \\ &= tr\left\{\mathbf{X}_{6_1}'[\mathbf{I} - \mathbf{X}_*(\mathbf{X}_*'\mathbf{X}_*)^-\mathbf{X}_*']\mathbf{X}_{6_1}E(\boldsymbol{\beta}_1\boldsymbol{\beta}_1')\right\} + \sigma_\epsilon^2[r(\mathbf{X}) - r(\mathbf{X}_*)] \\ &= tr\left\{\mathbf{X}_{6_1}'\mathbf{X}_{6_1} - \mathbf{X}_{6_1}'\mathbf{X}_*(\mathbf{X}_*'\mathbf{X}_*)^-\mathbf{X}_*'\mathbf{X}_{6_1}\sigma_{\boldsymbol{\beta}_1}^2\mathbf{I}\right\} + \sigma_\epsilon^2[r(\mathbf{X}) - r(\mathbf{X}_*)] \end{aligned} \tag{5.13}$$

where $\mathbf{X} = (\mathbf{X}_{6_1},\ \mathbf{X}_{6_2}, ..., \mathbf{X}_{6_k})$ and $\mathbf{X}_* = (\ \mathbf{X}_{6_2}, ..., \mathbf{X}_{6_k})$ are partitioned in the fashion of $\boldsymbol{\beta}' = (\boldsymbol{\beta}_1', \boldsymbol{\beta}_2', ..., \boldsymbol{\beta}_k')$.

5.2.2 Model Adequacy

The adequacy of a linear model in split-split-plot experimental design has three experimental errors which can be used to diagnose its fitness. Common measures of adequacy of fit for split-plot models in literature include the coefficient of determination (R^2), adjusted coefficient of determination (R^2-adjusted), prediction error sum of squares (PRESS), R^2-prediction statistics (Almimi *et al.*, 2009) and other graphical approaches.

Following the approach proposed by Almimi *et al.* (2009) on a split-plot model, we extend the procedure for calculating R^2 and R^2-adjusted for the split-split-plot model as follows:

- Identify the significant and negligible WP, SP and SSP effects from the ANOVA table.
- Extract the sum of squares for the significant effects and negligible effects, and then separate these sums of squares into three sections: one for the WP, one for the SP and the other for the SSP effects.
- Create a new ANOVA table with three detached sections for the WP, SP and SSP subdivisions.
- Calculate and include, in each subdivision, the model sum of squares (i.e., add all the significant sum of squares), the residual sum of squares (i.e., negligible sum of squares or lack of fit plus pure error), and the total sum of squares (i.e., model sum of squares plus residual sum of squares).
- From the WP subdivision, divide the model sum of squares by the total sum of squares to obtain R^2 for the WP sub-model,

$$R^2_{WP} = \frac{SS_{model}(WP)}{SS_{total}(WP)}. \tag{5.14}$$

- Repeating the same process above for the SP section, i.e., dividing the model sum of squares by the total sum of squares to obtain R^2 for the SP sub-model,

$$R^2_{SP} = \frac{SS_{model}(SP)}{SS_{total}(SP)}. \tag{5.15}$$

- The SSP sub-division gives R^2 is given by,

$$R^2_{SSP} = \frac{SS_{model}(SSP)}{SS_{total}(SSP)}. \tag{5.16}$$

- The R^2-adjusted for each sub-division is calculated by dividing the residual sum of squares and the total sum of squares by their degrees of freedom before subtracting the quotient from 1.

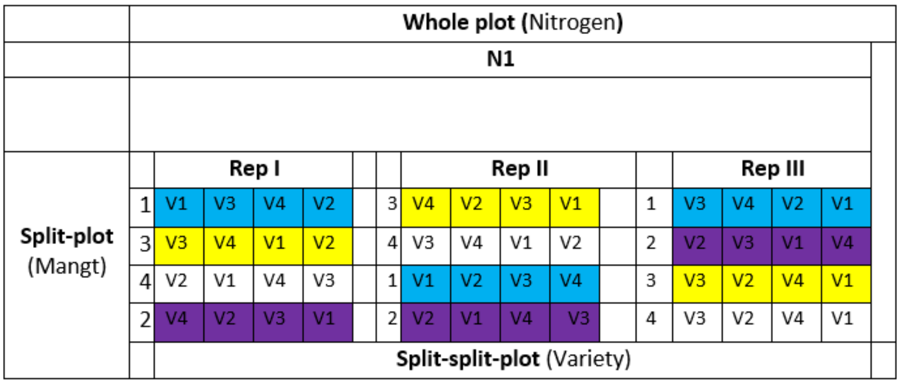
$$R^2 - adj = 1 - \frac{(SS_{residual} \div df(residual)}{SS_{total} \div df(total)}. \tag{5.17}$$

- Analogously, the adequacy of fit for the combined split-split-plot model is similarly determined by considering R^2 and R^2-adjusted from the combined analysis of variance.

5.3 Numerical Example

An experiment reported by Gomez and Gomez (1984) was conducted to investigate the effect on grain yield (tonnes per hectare) of three rice varieties (V1, V2 and V3) randomly grown under three management practices (minimum M1, optimum M2 and intensive M3) and randomly treated with five Nitrogen Levels (N1 0 kg N/ha, N2 50 kg N/ha, N3 80 kg N/ha, N4 110 kg N/ha and N5 140 kg N/ha), in a split-split-plot design, with three replications. The variable $Nitrogen$ was assigned to the main-plots (factor A) since it is hard to vary fertilizer levels in a confined space, management practice ($Management$) was taken as a subplot factor B, while rice variety ($Variety$) was considered as a sub-subplot factor C.

Suppose, for illustration purposes, an additional factor level for rice variety (V4) and a management practice (extreme M4) are suggested while the five nitrogen levels remain the same. The grain yield data for these additional factor levels were simulated from the original data statistics. We then consider the whole-plot factor ($Nitrogen$) to have new levels, N3 and N5 (considered as fixed), tested against the old levels, N1, N2 and N4 (considered as random), the sub-plot factor ($Management$) to have fixed levels (M3 and M4) tested against old and random levels (M1 and M2), while the sub-sub-plot factor ($Variety$) has new levels (V2 and V4) also tested against the old and existing levels (V1 and V3). Figure 5.1 displays a single plot for the whole-plot factor ($Nitrogen$) of the proposed 5 × 4 × 4 split-split-plot design with three replications arranged in completely randomised design (CRD).



	Whole plot (Nitrogen)														
	N1														
	Rep I					**Rep II**					**Rep III**				
Split-plot (Mangt)	1	V1	V3	V4	V2	3	V4	V2	V3	V1	1	V3	V4	V2	V1
	3	V3	V4	V1	V2	4	V3	V4	V1	V2	2	V2	V3	V1	V4
	4	V2	V1	V4	V3	1	V1	V2	V3	V4	3	V3	V2	V4	V1
	2	V4	V2	V3	V1	2	V2	V1	V4	V3	4	V3	V2	V4	V1
	Split-split-plot (Variety)														

Figure 5.1: Split-Split-Plot Design in CRD

Using the partitioning approach, we construct each of the eight partitioned split-split-plot models defined as (5.3), with similar codes (1-8) given in Table 4.1.

5.3.1 Checking Model Assumptions

A normality check on partitioned and combined data was conducted using the Shapiro-Wilk normality test. Partitions 1-3, 5-8, and the original (combined) data had non-significant Shapiro-Wilk test statistic ($p{-}value > 0.05$), implying that the samples were from normal distributions. Only model 4 (FRR model) had a significant Shapiro-Wilk test p-value (0.03251) less than alpha (0.05), signifying a violation of the normality assumption in this sample.

Homogeneity of variance test in the combined model data (WP, SP and SSP subdivisions) as well as in the partitioned models was done using Bartlett's homogeneity of variance test for data assumed normal (models 1-3, 5-8 and combined), while Levene's test was used for the non-normal data sample from model 4. Results from Bartlett's test showed that models 2, 3, and the combined one had non-constant error variances ($p - value < 0.05$), whereas the rest of the data samples satisfied the homogeneity of variance assumption ($p - value > 0.05$). Furthermore, Levene's test established that model 4, combined model WP, SP and SSP samples satisfied the homogeneity of variance assumption as well ($p - value > 0.05$). Figure 5.2 shows the normal probability plots of the WP, SP and the SSP residuals, respectively, confirming the normality assumption as indicated by the plots closely following the diagonal line in each of the Q-Q plot.

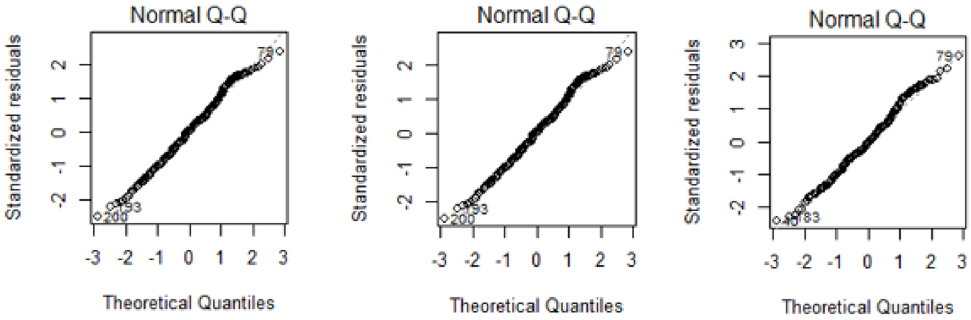


Figure 5.2: Normal Probability Plots of the WP, SP and SSP Residuals

The independence of the WP, SP and the SSP residuals is guaranteed by the three-stage randomisation process in a split-split plot design.

Figure 5.3 displays the possible outliers plotted in the WP, SP and SSP, respectively. There were no outstanding outliers extreme enough to warrant exclusion from the three sample as all extreme values were within the range of other data points.

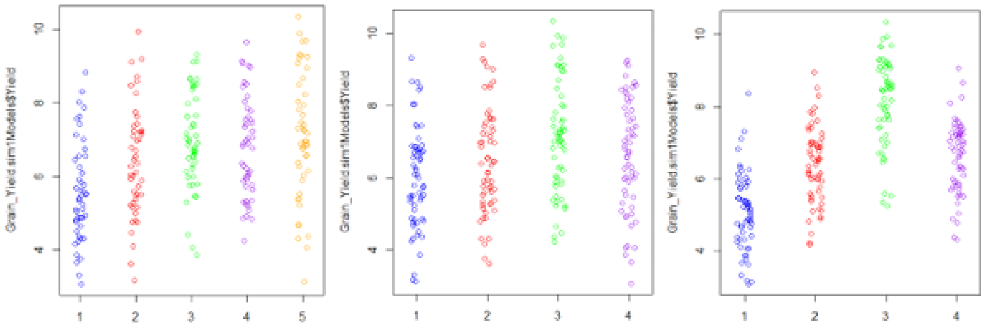


Figure 5.3: Outliers for the Combined WP, SP and SSP Models

All possible outliers were retained to maintain the required sample sizes.

5.3.2 Model Estimation

The linear model for the split-split-plot design with fixed WP, SP and SSP factors α, β, and γ respectively, and variances of $\epsilon_{il}^{(1)} = \sigma_{WP}^2$, $\epsilon_{ijl}^{(2)} = \sigma_{SP}^2$ and $\epsilon_{ijkl}^{(3)} = \sigma_{SSP}^2$, is given in (5.1). Each of the partitioned models 1-4 of Table 4.1, expressed in general linear form (5.2), is conveniently expanded and expressed in vector form as

$$\mathbf{y} = \mu\mathbf{1} + \mathbf{X}_1 \begin{bmatrix} \pi_1 \\ \pi_2 \\ \pi_3 \end{bmatrix} + \mathbf{X}_2 \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix} + \mathbf{X}_3 \begin{bmatrix} \beta_1 \\ \beta_2 \\ \pi_3 \end{bmatrix} + \mathbf{X}_4 \begin{bmatrix} \gamma_1 \\ \gamma_2 \end{bmatrix} + \mathbf{X}_5 \begin{bmatrix} \alpha\beta_{11} \\ \alpha\beta_{12} \\ \alpha\beta_{21} \\ \alpha\beta_{22} \end{bmatrix} + \mathbf{X}_6 \begin{bmatrix} \alpha\gamma_{11} \\ \alpha\gamma_{12} \\ \alpha\gamma_{21} \\ \alpha\gamma_{22} \end{bmatrix}$$

$$+\mathbf{X}_7 \begin{bmatrix} \beta\gamma_{11} \\ \beta\gamma_{12} \\ \beta\gamma_{21} \\ \beta\gamma_{22} \end{bmatrix} + \mathbf{X}_8 \begin{bmatrix} \alpha\beta\gamma_{111} \\ \alpha\beta\gamma_{112} \\ \alpha\beta\gamma_{121} \\ \alpha\beta\gamma_{122} \\ \alpha\beta\gamma_{211} \\ \alpha\beta\gamma_{212} \\ \alpha\beta\gamma_{221} \\ \alpha\beta\gamma_{222} \end{bmatrix} + \boldsymbol{\epsilon}, \tag{5.18}$$

whereas models 5-8 will have an additional $\mathbf{X}_2$ random level and the interactions thereof. We demonstrate the derivations of expected values for the fixed-fixed-random (FFR) model from the given numerical example. The model (5.18) becomes

$$\mathbf{y}_{FFR} = \mathbf{X}_{FFR}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{5.19}$$

where $\mathbf{y}_{FFR} : 24 \times 1$ is a vector variable, $\mathbf{X}_{FFR} : 24 \times 30 = (\mathbf{1} \quad \mathbf{X}_1 \quad \mathbf{X}_2 \quad ... \quad \mathbf{X}_8)$ and $\boldsymbol{\beta} : 30 \times 1 = (\mu, \beta_1', \beta_2', ..., \beta_8')'$. The normal equations for the $\mathbf{y}_{FFR}$ model are given by

$$(\mathbf{X}_{FFR}'\mathbf{X}_{FFR})\boldsymbol{\beta} = \mathbf{X}_{FFR}'\mathbf{y}_{FFR}. \tag{5.20}$$

Using equation (5.12), we illustrate Henderson's (1953) method III to derive the reduction in sum of squares and their expectations of the model $\mathbf{y}_{FFR}$. With $r(\mathbf{X}_{FFR}) = 11$ and the generalised inverse $(\mathbf{X}_{FFR}'\mathbf{X})_{FFR}^{-}$ determined accordingly, the uncorrected reduction in sum of squares for the full model $\mathbf{y}_{FFR}$ are as follows.

Whole-plot effect:

$R(\beta_1, ..., \beta_k) = \hat{\boldsymbol{\beta}}'\mathbf{X}_{FFR}'\mathbf{y}_{FFR} = \mathbf{y}_{FFR}'(\mathbf{X}_{FFR}'\mathbf{X}_{FFR})^{-}\mathbf{X}_{FFR}'\mathbf{y}_{FFR} = 1380.593$

$R(\beta_1) = \mathbf{y}_1'\mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-}\mathbf{X}_1'\mathbf{y}_1 = 1305.684$

$R(\beta_2) = \mathbf{y}_2'\mathbf{X}_2(\mathbf{X}_2'\mathbf{X}_2)^{-}\mathbf{X}_2'\mathbf{y}_2 = 1302.014$

$R(\beta_1, \beta_2) = \mathbf{y}_*'\mathbf{X}_*(\mathbf{X}_*'\mathbf{X}_*)^{-}\mathbf{X}_*'\mathbf{y}_* = 1307.419$, where $\mathbf{X}_* = (\mathbf{X}_1, \mathbf{X}_2)$ and $\mathbf{y}_* = (\mathbf{y}_1, \mathbf{y}_2)$

$SSE_{WP} = \mathbf{y}'_{FFR}\mathbf{y}_{FFR} - R(\beta_1, \beta_2) = 1394.768 - 1307.419 = 87.349$

Split-plot effect:

$R(\beta_3) = \mathbf{y}'_3\mathbf{X}_3(\mathbf{X}'_3\mathbf{X}_3)^{-}\mathbf{X}'_3\mathbf{y}_3 = 1405.139$

$R(\beta_5) = \mathbf{y}'_5\mathbf{X}_5(\mathbf{X}'_5\mathbf{X}_5)^{-}\mathbf{X}'_5\mathbf{y}_5 = 1309.581$

$R(\beta_1, \beta_2, \beta_3, \beta_5) = \mathbf{y}'_{\odot}\mathbf{X}_{\odot}(\mathbf{X}'_{\odot}\mathbf{X}_{\odot})^{-}\mathbf{X}'_{\odot}\mathbf{y}_{\odot} = 1313.420,$
where $\mathbf{X}_{\odot} = (X_1, X_2, X_3, X_5)$ and $\mathbf{y}_{\odot} = (\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3, \mathbf{y}_5)$

$SSE_{SP} = \mathbf{y}'_{FFR}\mathbf{y}_{FFR} - R(\beta_1, \beta_2, \beta_3, \beta_5) = 1394.768 - 1313.420 = 81.348$

Split-split-plot effect:

$R(\beta_4) = \mathbf{y}'_4\mathbf{X}_4(\mathbf{X}'_4\mathbf{X}_4)^{-}\mathbf{X}'_4\mathbf{y}_4 = 1364.702$

$R(\beta_6) = \mathbf{y}'_6\mathbf{X}_6(\mathbf{X}'_6\mathbf{X}_6)^{-}\mathbf{X}'_6\mathbf{y}_6 = 1347.494$

$R(\beta_7) = \mathbf{y}'_7\mathbf{X}_7(\mathbf{X}'_7\mathbf{X}_7)^{-}\mathbf{X}'_7\mathbf{y}_7 = 1372.260$

$R(\beta_8) = \mathbf{y}'_8\mathbf{X}_8(\mathbf{X}'_8\mathbf{X}_8)^{-}\mathbf{X}'_8\mathbf{y}_8 = 1374.008$

$SSE = \mathbf{y}'_{FFR}\mathbf{y}_{FFR} - R(\boldsymbol{\beta}) = 1394.768 - 1380.593 = 14.175$

$\hat{\sigma}_{\epsilon_{FFR}} = MSE = 14.175 \div 8 = 1.7719$

We make use of equation (5.13) to derive the expectations of the reduction in sum of squares calculated for the random effects in the $\mathbf{y}_{FFR}$ model.

$E(\beta_1|\boldsymbol{\beta}) = tr\left\{\mathbf{X}'_1\mathbf{X}_1 - \mathbf{X}'_1\mathbf{X}_*(\mathbf{X}'_*\mathbf{X}_*)^{-}\mathbf{X}'_*\mathbf{X}_1\sigma^2_{\beta_1}\mathbf{I}\right\} + \sigma^2_{\epsilon}\left[r(\mathbf{X}) - r(\mathbf{X}_*)\right] = 15\sigma^2_{\beta_1} + 2\sigma^2_{\epsilon}$

$E(\beta_4|\boldsymbol{\beta}) = E(\beta_7|\beta) = 0$

$E(\beta_6|\boldsymbol{\beta}) = 1.1667\sigma^2_{\beta_6} + \sigma^2_{\epsilon}$

$E(\beta_8|\boldsymbol{\beta}) = 2.8\sigma^2_{\beta_8} + \sigma^2_{\epsilon}$

5.4 Results and Discussion

Table 5.1 summarises the procedure for analysis of variance of the partitions in a combined model. The table displays the partitioned models involved in the processes of calculating the sum of squares and degrees of freedom for both main and interaction effects of fixed and the random models (1 - 8).

Table 5.1: Degrees of Freedom and Sub-models for the Combined Model

Source of variation	Degrees of freedom	Sub-models involved
Replication	$8(l-1)$	1 - 8
Nitrogen (A): (FA)	$\sum_{i=1}^{4}(f_{a_i}-1)$	1 - 4
(RA)	$\sum_{i=5}^{8}(r_{a_i}-1)$	5 - 8
Whole-plot E(a)	$(l-1)[\sum_{i=1}^{4}(f_{a_i}-1)+\sum_{i=5}^{8}(r_{a_i}-1)]$	1 - 8
Management (B): (FB)	$\sum_{i=1,2,5,6}(f_{b_i}-1)$	1,2,5 and 6
(RB)	$\sum_{i=3,4,7,8}(r_{b_i}-1)$	3,4,7 and 8
A × **B**: $(FA\times FB)$	$(\sum_{i=1}^{4}(f_{a_i}-1))(\sum_{i=1,2,5,6}(f_{b_i}-1))$	1 and 2
$(RA\times RB)$	$(\sum_{i=5}^{8}(r_{a_i}-1))(\sum_{i=3,4,7,8}(r_{b_i}-1))$	7 and 8
Sub-plot E(b)	$(l-1)[f_a * \sum_{i=1,2,5,6}(f_{b_i}-1)$ $+r_a * \sum_{i=3,4,7,8}(r_{b_i}-1)]$	1 - 8
Variety (C): (FC)	$\sum_{i=1,3,5,7}(f_{c_i}-1)$	1,3,5 and 7
(RC)	$\sum_{i=2,4,6,8}(r_{c_i}-1)$	2,4,6 and 8
A × **C**: $(FA\times FC)$	$(\sum_{i=1}^{4}(f_{a_i}-1))(\sum_{i=1,3,5,7}(f_{c_i}-1))$	1 and 3
$(RA\times RC)$	$(\sum_{i=5}^{8}(r_{a_i}-1))(\sum_{i=2,4,6,8}(r_{c_i}-1))$	6 and 8
B × **C**: $(FB\times FC)$	$(\sum_{i=1,2,5,6}(f_{b_i}-1))(\sum_{i=1,3,5,7}(f_{c_i}-1))$	1 and 3
$(RB\times RC)$	$(\sum_{i=3,4,7,8}(r_{b_i}-1))(\sum_{i=2,4,6,8}(r_{c_i}-1))$	6 and 8
A × **B** × **C**: $(FA\times FB\times FC)$	$(\sum_{i=1}^{4}(f_{a_i}-1))(\sum_{i=1,2,5,6}(f_{b_i}-1))$ $(\sum_{i=1,3,5,7}(f_{c_i}-1))$	1
$(RA\times RB\times RC)$	$(\sum_{i=5}^{8}(r_{a_i}-1))(\sum_{i=3,4,7,8}(r_{b_i}-1))$ $(\sum_{i=2,4,6,8}(r_{c_i}-1))$	8
Sub-sub-plot E(C)	$(l-1)[f_a * f_b \sum_{i=1,3,5,7}(f_{c_i}-1)$ $+r_a * r_b \sum_{i=2,4,6,8}(r_{c_i}-1)]$	1 - 8
Total	N-1	

Depending on the targeted main factor or interaction effect as indicated by the source of variation column in Table 5.1, the degrees of freedom and sums of squares are obtained from the partitioned sub-models given in the last column. We make use of the partition FRF, fit the partitioned linear mixed model, and and present a summary of the analysis of variance in Table 5.2. original FRF model results

Of the new farming strategies considered, the *Nitrogen* categories were contributing significantly to the grain yield (p-value = 0.0107) at a 5% level of significance. In addition, noticeable

Table 5.2: The ANOVA Table for Simulated FRF Model in RCBD

Random effects:					
Groups			Name	Variance	Std. Dev.
Nitrogen:Management:rep			(Intercept)	0.00000	0.0000
Variety:Nitrogen:Management			(Intercept)	0.00000	0.0000
Nitrogen:rep			(Intercept)	0.00000	0.0000
Variety:Management			(Intercept)	0.01769	0.1330
Management:Nitrogen			(Intercept)	0.00000	0.0000
rep			(Intercept)	0.11680	0.3418
Management			(Intercept)	0.00000	0.0000
Residual				1.7150	0.4141
Fixed effects:					
Coefficients	Estimate	Std. Error	df	t-value	Pr(>|t|)
(Intercept)	6.23983	0.27633	4.37599	22.581	0.0000107
Nitrogen5	0.69017	0.23910	16.00004	2.887	0.0107
Variety4	0.52850	0.27359	4.99601	1.932	0.1113
Nitrogen5:Variety4	0.01933	0.33813	16.00004	0.057	0.9551

variance components were detected for the $Variety : Management$ interaction and the plot (rep) effect. Therefore, the combination of new and old farming strategies had little impact on boosting the grain yield in this case since there is no significant interaction of these factors displayed.

The analysis of variance for the main and interaction effects of the combined model is displayed in Table 5.3. The newly invented farming strategies are the fixed levels of the WP, SP and SSP factors, symbolised by FA (Nitrogen amounts), FB (management practices) and FC (varieties of rice), respectively, whilst the old strategies selected from a population of existing strategies are similarly denoted by RA, RB and RC.

Based on the ANOVA Table 5.3 results displayed, the main effects of the new management practices (FB), old rice varieties (RC), interaction of old fertiliser amounts old rice varieties ($RA \times RC$) and interaction of old management practices old rice varieties ($RB \times RC$), contributed significantly to the rice yield at 5% level of significance. In addition, the interaction of the three old farming methods ($RA \times RB \times RC$) had a significant effect at 10% alpha level. Hence the hypotheses H2 - H4 were addressed by these ANOVA F-tests. The overall combined model estimates of the WP, SP, and SSP error variance were $\sigma^2_{WP} = 18.2976$, $\sigma^2_{WP} = 10.1585$

Table 5.3: Combined Model ANOVA Table Based on Reductions in Sum of Squares

Source of variation	**Degrees of freedom**	**Sum of Squares**	**Mean Square**	**F**
Rep	16	21.8666	1.3667	0.0747
Nitrogen (FA)	4	2.9512	0.7378	0.0403
(RA)	8	40.5146	5.0643	0.2768
Whole plot E(a)	**24**	**439.1420**	**18.2976**	
Management (FB)	4	110.1580	27.5395	2.7110**
(RB)	4	10.0790	2.5198	0.2480
$(FA \times FB)$	16	13.4360	0.8398	0.0827
$(RA \times RB)$	16	36.3010	2.2688	0.2233
Subplot E(b)	**40**	**406.339**	**10.1585**	
Variety (FC)	4	2.6960	0.6740	0.6595
(RC)	4	293.0339	73.2585	71.6885**
$(FA \times FC)$	16	6.4609	0.4038	0.3951
$(RA \times RC)$	16	177.7800	11.1113	10.8732**
$(FB \times FC)$	16	6.7342	0.4209	0.4119
$(RB \times RC)$	16	146.4580	9.1536	8.9574**
$(FA \times FB \times FC)$	64	9.2128	0.1440	0.1409
$(RA \times RB \times RC)$	64	82.871	1.2949	1.2671*
Sub-subplot E(c)	**80**	**81.7504**	**1.0219**	
Total	**412**	**1887.7846**		

"**" Significant at 0.05 "*" Significant at 0.1

and $\sigma_\epsilon^2 = 1.0219$, respectively. The associated mean and covariance estimates were:

$$\hat{\boldsymbol{\mu}} = \begin{bmatrix} 7.3655 \\ 7.3655 \\ 6.8540 \\ 6.8423 \\ 6.4895 \\ 6.8347 \\ 5.9290 \\ 5.8058 \end{bmatrix} ; \boldsymbol{\sigma}_\epsilon^2 = \sum = \begin{bmatrix} \sum_1 & \cdots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \sum_8 \end{bmatrix} .$$

Table 5.4 displays the analysis of variance for the combined model corrected for the mean. The contribution of all factors in the model to the yield is depicted from the significant F-ratio.

Table 5.4: Corrected Model ANOVA Table

Source of variation	**Degrees of freedom**	**Sum of Squares**	**Mean Square**	**F**
Model	152	515.0279	3.3883	4.4956**
Error	80	60.2968	0.7537	
Total	**232**	**575.3247**		

"**" Significant at 0.05

The estimates of error variance from the partitioned models 1-8 used to test hypotheses H1, calculated using the Henderson (1953) approach, were recorded as follows:

$\sigma^2_{WP_1}$=8.2490; $\sigma^2_{WP_2}$=43.6745; $\sigma^2_{WP_3}$=2.5190; $\sigma^2_{WP_4}$=45.775; $\sigma^2_{WP_5}$=9.5145; $\sigma^2_{WP_6}$=28.7378; $\sigma^2_{WP_7}$=1.8803; $\sigma^2_{WP_8}$=19.5443;

$\sigma^2_{SP_1}$=2.6573; $\sigma^2_{SP_2}$=20.3370; $\sigma^2_{SP_3}$=1.1933; $\sigma^2_{SP_4}$=21.9528; $\sigma^2_{SP_5}$=5.4001; $\sigma^2_{SP_6}$=18.4515; $\sigma^2_{SP_7}$=1.1000; $\sigma^2_{SP_8}$=12.0113;

$\sigma^2_{\epsilon_1}$=0.3694; $\sigma^2_{\epsilon_2}$=1.7719; $\sigma^2_{\epsilon_3}$=0.2710; $\sigma^2_{\epsilon_4}$=0.8155; $\sigma^2_{\epsilon_5}$=2.3100; $\sigma^2_{\epsilon_6}$=1.1838; $\sigma^2_{\epsilon_7}$=0.3974; $\sigma^2_{\epsilon_8}$=0.7695.

We use equations (5.14) - (5.16) to calculate the coefficient of determination (R^2) measures and equation (5.17) for the adjusted coefficient of determination (R^2-adjusted) measures based on the combined model analysis. Table 5.5 summarises the ANOVA for the whole-plot, subplot and split-split-plot sub-divisions.

Table 5.5: Combined Model Subdivisions for Model Accuracy

ANOVA for the Whole-plot Subdivision				
Source of variation	Degrees of freedom	Sum of Squares	Mean Square	F
Model	0	0	0	0
Residual	52	105.847	2.0355	
Pure error	24	40.5146	1.6881	
Lack of fit	28	65.3324	2.3333	
Total	**52**	**105.847**		
ANOVA for the Subplot Subdivision				
Source of variation	Degrees of freedom	Sum of Squares	Mean Square	F
Model	4	110.1580	27.5395	4.4899^{**}
Residual	76	466.1550	6.1336	
Pure error	40	406.3390	10.1585	
Lack of fit	36	59.8160	1.6616	
Total	**80**	**576.3130**		
ANOVA for the Split-split-plot Subdivision				
Source of variation	Degrees of freedom	Sum of Squares	Mean Square	F
Model	100	700.1429	7.0015	11.7950^{**}
Residual	180	106.8543	0.5936	
Pure error	80	81.7504	1.0219	
Lack of fit	100	25.1039	0.2510	
Total	**280**	**806.9972**		

"**" Significant at 0.05

The statistics obtained were:

$R^2(WP) = 0.00$; $R^2(SP) = 0.1911$; $R^2(SSP) = 0.8676$.

$R^2 - adj(WP) = 0.00$; $R^2 - adj(SP) = 0.1486$; $R^2 - adj(SSP) = 0.7940$.

One can safely conclude that the adjusted coefficients of determinations for the WP (0%), SP (15%) signify, on average, a poor model fit despite the high model fit for the SSP (79%) subdivision.

Based on the partitions given in Table 4.1, we consider FFF model (1) as fixed model, and models (2-8) as random models. Thus, inference space is considered only for the random models (2-8). We select the FFR model (2) to illustrate inference space and calculate of standard errors. REML estimated variance components for random variables replication, variety, variety × nitrogen, variety × management, variety × nitrogen × management, and error are 0.2158, 5.1025, 0.07946, 0, 0, 0, 0, and 0.9776, respectively.

Table 5.6 displays the predictable functions for the mean and differences, with the associated scope of inference (narrow, broad and intermediate) from the FFR model (2). Each estimable function derives information for the estimate from the fixed effects, while the uncertainty derives from the combined contribution of all of the random effects (Stroup, 2016). Function F_1 and F_3 estimate the marginal mean yield of new Nitrogen level (N3) averaged over a broad spectrum of rice varieties (broad inference space), while function F_2 estimates the mean yield Nitrogen level (N5) averaged over rice varieties V1 and V3 (i.e. narrow inference space). F_4 estimates the difference between the whole plot factor levels N3 and N5 for narrow inference space.

Table 5.6: Predictable Functions ($\mathbf{K}'\boldsymbol{\beta} + \mathbf{M}'\mathbf{u}$) for the FFR Model

Predictable function											
	$\hat{Y}_{\alpha_1}$	$\hat{Y}_{\alpha_2}$	$\hat{Y}_{\alpha_1}$	$\hat{Y}_{\alpha_1}$ -$\hat{Y}_{\alpha_2}$	$\hat{Y}_{\beta_1}$	$\hat{Y}_{\beta_2}$	$\hat{Y}_{\beta_2}$	$\hat{Y}_{\beta_1}$ -$\hat{Y}_{\beta_2}$	$\hat{Y}_{\gamma_1}$	$\hat{Y}_{\gamma_2}$	$\hat{Y}_{\gamma_1}$ -$\hat{Y}_{\gamma_2}$
Function	F_1	F_2	F_3	F_4	F_5	F_6	F_7	F_8	R_9	R_{10}	$R_{d_{11}}$
Scope*	B	N	I	N	N	I	B	B	N	B	N
Effect											
μ	1	(2	(2	(0	(2	(2	(2	(0	1	(2	0
π_1	0	1	1	0	1	1	0	0	1	0	0
π_2	0	1	1	0	1	1	0	0	1	0	0
π_3	0	1	1	0	1	1	0	0	1	0	0
α_1	1	0	2	2	0	0	0	0	0	0	1
α_2	0	2	0	-2	0	0	0	0	1	0	-1
β_1	0	0	0	0	2	2	0	2	0	0	1
β_2	0	0	0	0	0	0	2	-2	1	0	-1
γ_1	0	1	1	0	1	1	0	0	0	2	1
γ_2	0	1	1	0	1	1	0	0	1	0	-1
$\alpha\beta_{11}$	0	0	0	0	0	0	0	0	0	0	1
$\alpha\beta_{12}$	0	0	0	0	0	0	0	0	0	0	0
$\alpha\beta_{21}$	0	0	0	0	0	0	0	0	0	0	0
$\alpha\beta_{22}$	0	0	0	0	0	0	0	0	1	0	-1
$\alpha\gamma_{11}$	0	0	0	1	0	0	0	0	0	0	1
$\alpha\gamma_{12}$	0	0	0	1	0	0	0	0	0	0	0
$\alpha\gamma_{21}$	0	1	0	-1	0	0	0	0	0	0	0
$\alpha\gamma_{22}$	0	1	0	-1	0	0	0	0	1	0	-1
$\beta\gamma_{11}$	0	0	0	0	1	0	0	0	0	0	1
$\beta\gamma_{12}$	0	0	0	0	1	0	0	0	0	0	0
$\beta\gamma_{21}$	0	0	0	0	0	0	0	0	0	0	0
$\beta\gamma_{22}$	0	0	0	0	0	0	0	0	1	0	-1
$\alpha\beta\gamma_{111}$	0	0	0	0	0	0	0	0	0	0	1
$\alpha\beta\gamma_{112}$	0	0	0	0	0	0	0	0	0	0	0
$\alpha\beta\gamma_{121}$	0	0	0	0	0	0	0	0	0	0	0
$\alpha\beta\gamma_{122}$	0	0	0	0	0	0	0	0	0	0	0
$\alpha\beta\gamma_{211}$	0	0	0	0	0	0	0	0	0	0	0
$\alpha\beta\gamma_{212}$	0	0	0	0	0	0	0	0	0	0	0
$\alpha\beta\gamma_{221}$	0	0	0	0	0	0	0	0	0	0	0
$\alpha\beta\gamma_{222}$	0	0)/2	0)/2	0)/2	0)/2	0)/2	0)/2	0)/2	1	0)/2	-1
Estimate	7.362	7.369	7.362	-0.006	7.925	7.925	6.807	1.118	8.429	5.787	-2.070
S.E	1.657	0.315	0.373	0.429	0.345	0.345	1.651	0.404	0.071	0.438	0.697
p-value	0.123	$< .001$	0.015	0.988	$< .001$	$< .001$	0.136	0.014	0.003	0.009	0.009

"*" Inference space: N - narrow, B - broad, I - intermediate

SAS *PROC MIXED* procedure generated estimates of the standard errors by substituting the REML estimates of the variance components associated with the random variables in the predictable function. Interestingly, the predictable function involving narrow inference space of nitrogen level N5 and intermediate inference space difference between the nitrogen levels are significant ($p - value < 0.0001$ and $p - value = 0.015$, respectively), leading rejection of null hypotheses: $H_0 : \mathbf{K}'\boldsymbol{\beta} = 0$ where $\mathbf{K}'$=[1 0 0 0 1 0 ... 0] and $H_0 : \mathbf{K}'\boldsymbol{\beta} = 0$ where $\mathbf{K}'$=[0 0 0 0 -1 -1 0 ... 0], at 0.05 alpha level of significance in the respective inference spaces.

Similarly, functions F_5 through to F_8 estimate the marginal mean yield of new Management strategies (M3 and M4) averaged over the two old Variety levels (V1 and V3) in the three inference spaces (narrow, intermediate and broad). The results show a significant contribution by the subplot variable in all inference spaces with the exception of broad inference ($p-value = 0.136$). There is a significant difference between the two management practices (M3 and M4) for the broad inference space, as shown by a significant estimate in function F_8.

The last three columns of Table 5.6 display the estimates and standard errors of predictable functions, R_9 and $R_{d_{11}}$, which estimate the marginal mean of the old and random Variety levels (V1 and V3), averaged over the new Management strategies and Nitrogen levels. The variety levels are contributing significantly to the rice yield ($p - value = 0.003$ and 0.0085, respectively) when considered in both narrow and broad inference scope, thus confirming the hypothesis: $H_0 : \sigma^2_{\gamma_k} > 0$ at 0.05 alpha level of significance in both cases. The difference between the two varieties is significant ($p - value = 0.009$) for the narrow inference space, with standard error of 0.6973 for the difference estimate.

Agricultural experiments often involve large populations of crop levels, whose main effects are independent of their interactions with these varieties. The same applies to the main effects of management strategies, when different managerial skills independently impact different types of crops. These scenarios require inference for fixed effect marginal means to be applied to the whole population of random effects and their respective random interactions. Thus, broad inference space is of primary, if not exclusive, interest in the vast majority of practical applications (Littell *et al.*, 2006), and has to be incorporated in statistical experiments. As expected, the results in Table 5.5 clearly confirm that the standard errors increase with inference space.

5.5 Simulation Results for the Split-split-plot Model

We present the simulation sample analysis results for the mixed model (FRF) in a split-split-plot structure arranged in an RCBD, and compile a summary of the simulation analysis for the rest of the partitioned models.

5.5.1 Simulation Results for the FRF Model

We considered two fixed factors: $Nitrogen$, with two fixed levels (N3, and N5), and $Variety$, with two fixed levels (V2 and V4); and one random factor, $Management$, with two random levels (M1 and M2). The treatments were randomly assigned to three blocks or plots (*rep*). Table 5.7 summarises the variation contributed by the random effects as well as the estimated model parameters, standard errors, and significance p-values of the FRF model fitted based on 10 000 samples.

Table 5.7: The ANOVA Table for Simulated FRF Model in RCBD

Random effects:					
Groups			Name	Variance	Std. Dev.
Nitrogen:Management:rep			(Intercept)	0.00000	0.0000
Variety:Nitrogen:Management			(Intercept)	0.00000	0.0001
Nitrogen:rep			(Intercept)	0.00944	0.09716
Variety:Management			(Intercept)	0.00000	0.0000
Management:Nitrogen			(Intercept)	0.004891	0.06993
rep			(Intercept)	0.0382	0.1176
Management			(Intercept)	0.00000	0.0000
Residual				0.09635	0.3104
Fixed effects:					
Coefficients	Estimate	Std. Error	df	t-value	Pr(>|t|)
(Intercept)	6.23983	0.27633	4.37599	22.581	0.0000107 ***
Nitrogen5	0.69017	0.23910	16.00004	2.887	0.0107 *
Variety4	0.52850	0.27359	4.99601	1.932	0.1113
Nitrogen5:Variety4	0.01933	0.33813	16.00004	0.057	0.9551

Significance codes: 0; $'***'$ 0.001; $'**'$ 0.01; $'*'$ 0.05

The simulated FRF model had a residual error variance of 0.09635, which is comparable to the original model estimate (0.17150). The fixed factor $Nitrogen5$ had significant category differences (p-value = 0.0107) at a 5% significance level with respect to the reference category ($Nitrogen3$), and non-significant interactions effect in both the simulated and original data. The model could capture considerably little random effects from the blocks (*rep* variance of 0.01382), $Management : Nitrogen$ (variance 0.004891), and $Nitrogen : rep$ variance of 0.00944.

We consider the distribution of the differences in grain yield due to the effect of nitrogen categories across the samples. The distribution of the differences in grain yield between $Nitrogen5$ against the reference category $Nitrogen3$ is displayed in Figure 5.4.

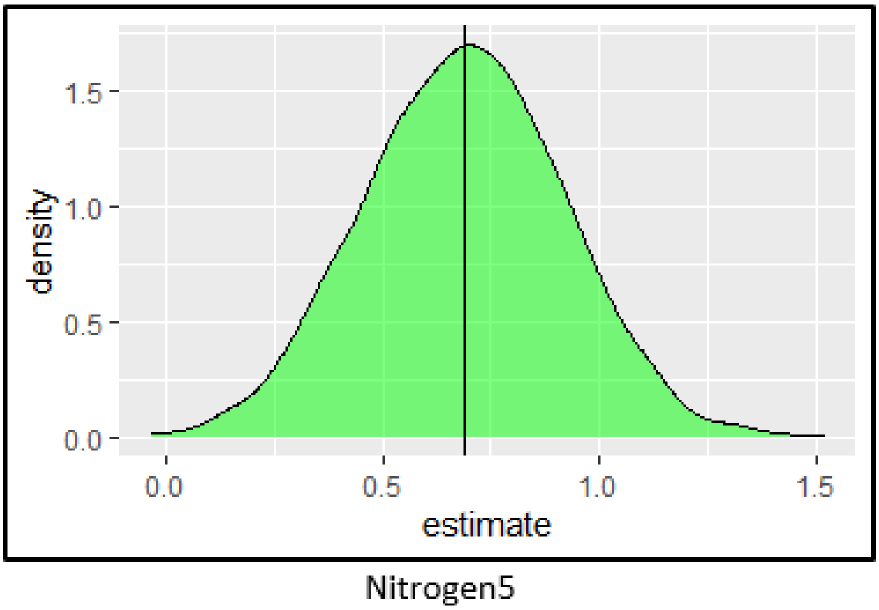


Figure 5.4: Distribution of Grain Yield Between Nitrogen Categories

The peak of the distribution is around the true estimate value of 0.69017. This is an indication that the coefficient of $Nitrogen5$ category is well estimated in the model. The peak of the distribution is around the true estimate value of 0.52850. The same applies to the $Variety4$ coefficient estimate, whose distribution is displayed in Figure 5.5.

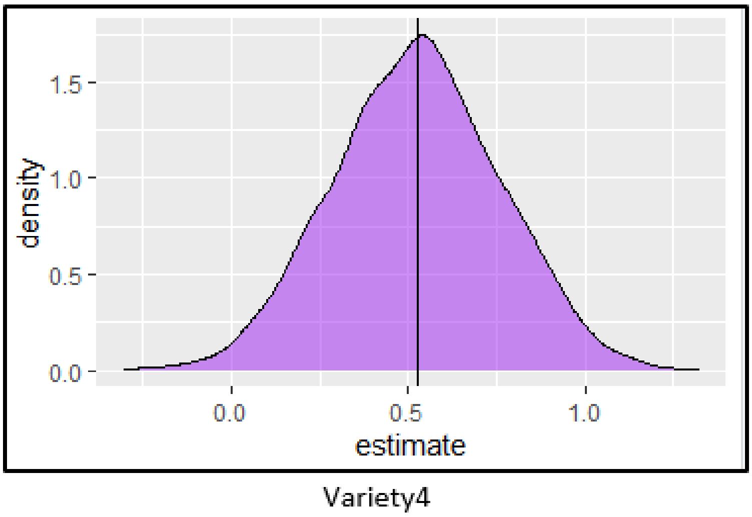


Figure 5.5: Distribution of Grain Yield Between Variety Categories

The distribution of standard deviations across the simulation samples is displayed in Figure 5.6. The distribution of the estimated variation is slightly underestimated 65.5% of the times, when compared to the true standard deviation (0.4141).

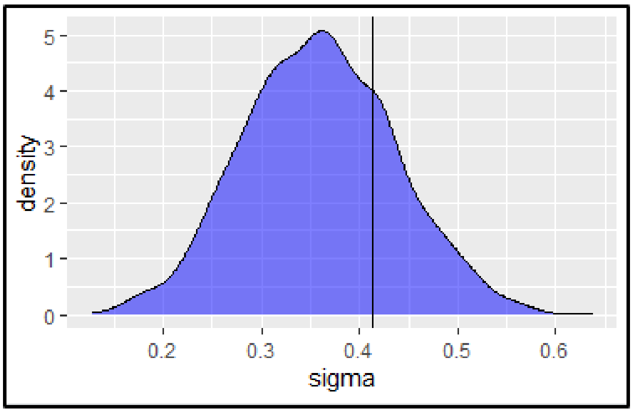


Figure 5.6: Distribution of Grain Yield Standard Deviation in the FRF Model

As shown in Table 5.7, there was insignificant amount of variation (<0.000001) in the grain yield due to *Management* factor categories that warrants a graphical display for the distribution. Although the estimated variation (0.09635) is slightly less than the true variation (0.17150) in the original FRF data set, the two estimates compare moderately well. The FRF model performs fairly well in estimating the standard deviation. This is further supported by a moderate proportion (statistical power = 0.529) of models that correctly rejected the null hypothesis (H_0: *Nitrogen5* effect = 0), given that we know the null hypothesis is not true.

5.5.2 Summary of Split-split-plot Simulation Results

Based on 10000 samples, the simulation analysis results for the rest of the partitioned split-split-plot linear mixed models (1-2 and 4- 8) are summarised in Tables 5.8 and 5.9. With the exception of the partition model FRF, which displays a significant effects of new *Nitrogen* categories, Tables 5.8 and 5.9 show that none of the new (fixed) farming methods had significant effects of on the grain yield variation across the simulation samples. However, some noticeable amounts of variation due to the random levels of *Nitrogen* and *Management* were detected in the RFR, RRF and RRR models. Based on the simulation sample results, only nitrogen levels and management practices were significantly contributing to the grain yield.

A variance estimate of zero in Table 5.8 indicates non-significant contribution by the random

Table 5.8: Simulation Results for the Split-split-plot Models

Model	Random Effects	Variance	Significant Fixed Effects	Statistical Power
FFF	Nitrogen: Management:rep	0.21372	None	
	Nitrogen:rep	0.01512		
	rep	0.07398		
	residual	0.21406		
FFR	Nitrogen: Management:rep	0.1381	None	
	Variety:Nitrogen:Management	0.0000		
	Nitrogen:rep	0.0000		
	Variety:Management	0.2967		
	Variety:Nitrogen	0.0000		
	rep	0.0000		
	Variety	0.1260		
	residual	0.7649		
FRR	Nitrogen: Management:rep	0.04625	None	
	Variety:Nitrogen:Management	0.0000		
	Nitrogen:rep	0.1891		
	Variety:Management	0.1180		
	Variety:Nitrogen	0.07098		
	Management:Nitrogen	0.0000		
	rep	0.0000		
	Variety	3.271		
	Management	3.237		
	residual	0.1703		
RFF	Nitrogen: Management:rep	0.0000	None	
	Variety:Nitrogen:Management	0.0000		
	Nitrogen:rep	0.0000		
	Nitrogen:Variety	0.0000		
	Nitrogen:Management	0.0000		
	rep	0.0000		
	Nitrogen	0.0000		
	residual	1.637		

effect in the model, indicating zero variation.

Table 5.9: Simulation Results for the Split-split-plot Models

Model	Random Effects	Variance	Significant Fixed Effects	Statistical Power
RFR	Nitrogen: Management:rep	0.02118	None	
	Variety:Nitrogen:Management	0.0000		
	Nitrogen:rep	0.0000		
	Nitrogen:Variety	0.0000		
	Nitrogen:Management	0.0000		
	Variety:Management	0.0000		
	rep	0.0000		
	Nitrogen	0.71217		
	Variety	0.40286		
	residual	0.34051		
RRF	Nitrogen: Management:rep	0.03589	None	
	Variety:Nitrogen:Management	0.0006		
	Nitrogen:rep	0.0000		
	Nitrogen:Variety	0.0000		
	Nitrogen:Management	0.0000		
	Variety:Management	0.0000		
	rep	0.3319		
	Nitrogen	0.6397		
	Management	0.0000		
	residual	0.1231		
RRR	Nitrogen: Management:rep	0.0000	None	
	Variety:Nitrogen:Management	0.0000		
	Nitrogen:rep	0.0000		
	Nitrogen:Variety	0.03028		
	Nitrogen:Management	0.0000		
	Variety:Management	0.0000		
	rep	0.0000		
	Nitrogen	0.4467		
	Variety	2.661		
	Management	0.3792		
	residual	0.4034		

Similarly, a variance estimate of zero in Tables 5.9 indicates non-significant contribution by the random effect in the model, indicating zero variation.

This confirms the results displayed in Table 5.2, where the two factors, $Nitrogen$ and $Management$, had significant interaction effects on the grain yield. Thus, the simulated partitioned models performed fairly well in estimating the variance components and the model estimates.

5.6 Conclusion

Conceptualising split-split-plot factors as sharing both fixed and random levels, representing new and old strategies, for the purpose of evaluating the difference in effectiveness in the methods leads to a linear mixed model scenario. Implicit to mixed model analysis is the scope of inference (broad, narrow or intermediate inference space), which is often ignored or wrongly interpreted by many statistical data analysts. Targeting wrong inference space leads to biased point estimates, interval estimates, and ultimately misleading hypothesis test conclusions for the entire population represented by the random factors. Depending on the context and objectives of the experiment, predictable functions can be manipulated to cater for population-wide or broad inference scope on the treatment effects (Littell *et al.*, 2006). We have demonstrated the application of the new analysis approach on an agricultural experiment in split-split-plot design when factors consisting of both fixed and random levels are involved. The usefulness of the approach can be explore in other related and complex designs. Chapter 6 presents the application of the partitioning approach in repeated measures design.

CHAPTER 6

PARTITIONING APPROACH IN REPEATED-MEASURES DESIGN

This chapter presents results from a publication by Chaka and Njuho (2022), appearing in the Stats Journal, Volume 5, Issue Number 2, 2022.

6.1 Introduction

Repeated-measures designs consists of between-subjects and within-subject factors that influence a response variable whose observations are recorded over time. When each of the predictors is conceptualised as having both fixed and random levels, and the structure of the variance-covariance matrix of the error terms is not the identity, it might be convenient to partition the observed data set based on the factor levels of interest. The partitioning gives room for individual analysis of subsets of the observed data that assume the same covariance structure. We consider this setting as the partitioning approach on repeated-measures design under a linear mixed model, with factors sharing both fixed and random components of the model.

Design and analysis of experiments which involve factors each consisting of both fixed and random levels require the application of linear mixed models. Linear mixed models have gained popularity in analysing Gaussian data due to their essential property of being able to handle both fixed and random effects simultaneously. The assumed linear mixed model design matrix takes either a full-rank or less than full-rank form. In addition, linear mixed models linear mixed models are convenient for modelling not only the means of the data but also the covariances (Pan and Shang, 2018). The fundamental consideration in the estimation process of linear models is the special case where the elements of the error vector are assumed equal and uncorrelated. However, in other fields such as psychology and medicine, longitudinal or correlated outcomes are common. Correlated data have properties that do not usually conform to the generality of a mixed model (Muller *et al.*, 2007). The complexity in data structures of such experiments calls for consideration in model selection and parameter estimation process. Approaches such as partitioning of fixed and random effects, which allow for analysis of factorial and correlated data, are necessary to explore (Chaka and Njuho, 2021).

Repeated measures studies are defined when multiple responses or measurements are observed on a set of independent sampling units on longitudinal (across time), spatial (across location), or multivariate (on different scales) setting. Missing observations are commonly to encounter in longitudinal outcome studies, leading to balanced and unbalanced measurements which require appropriate statistical methodologies. Repeated measures data are often dependent. Linear mixed models are one of the most convenient statistical approaches that can account for this dependency. However, setting them up for data analysis requires some care, especially in choosing the most appropriate covariance structure to keep the type I error down (Matuschek *et al.*, 2017). Estimation of parameters assuming linear mixed model might consider different assumptions on the structure of the variance-covariance matrix, other than the special case. Unlike the independent data scenario with the traditional homogenous variance structure, numerous candidate covariance structures for correlated data are available in statistical software packages such as SAS through the PROC MIXED procedure. The residual maximum likelihood (REML) (Patterson and Thompson, 1971) is one of the most famous methods for estimating the covariance parameters associated with linear mixed models among other alternatives (Dempster *et al.*, 1977).

6.2 Materials and Methods

We consider a repeated measures design with multiple between-subjects factors where each of these factors has both fixed and random levels. We illustrate the partitioning approach using a three-factor linear mixed model data fitted to a longitudinal data. The partitioning approach enriches in exploring the fixed and random levels of the same factor and the subsequent interaction of levels of factors of interest. We get to assess the differences between levels of the same factor and understand the variation within the same factor. We present the construction and analysis of a three-factor linear mixed model for repeated-measures designs when the between-subjects factors consist of both fixed and random levels, and the structure of the variance-covariance matrix of the error terms is not the identity, $\sigma_\epsilon^2\mathbf{I}$. The fixed levels allow for comparison of specific levels within the factor, whereas the random levels allow for assessment of variation within the same factor. In addition, we introduce heterogeneity of error terms, selection of the most appropriate covariance structure and assessment of the changes that occur in the estimation and when drawing the inferences.

6.2.1 An Illustrative Data Structure

We motivate the approach using data collected from a study that investigates the impact of combining carbon tetrachloride ($CCI4$) with four levels (0, 1.0, 2.5 and 5.0 mM), and chloroform ($CHCI3$) with four levels (0, 5 10 and 25 mM) on toxicity of cells on in vitro toxicity of isolated hepatocyte suspensions (Gennings, Chinchilli and Carter, 1989). Four flasks were assigned to each of the 16 treatments. Cell toxicity was measured by the amount of lactic dehydrogenase (LHD) enzyme percentage leakage from each of the 64 flasks after 0.01, 0.25, 0.5, 1, 2, and 3 hours after applying the treatment. For illustration purposes, we consider the between-subjects factor $CCI4$ levels 2.5 and 5.0 as fixed and new, while levels 0 and 1.0 as the existing random levels. Similarly, we consider the between-subjects factor $CHCI3$ levels 10 and 25 as fixed and new while the levels 0 and 5 as the random old levels. Of interest in the analysis is the percentage leakage observed on times 1, 2 and 3. We demonstrate the model construction procedure under certain assumptions in a completely randomized design (CRD) and a repeated-measures design (RMD).

6.2.2 Construction of a Linear mixed Model in CRD

Consider a three-way treatment structure in a balanced completely randomized design (CRD) with full interaction of factors A, B, and C, each consisting of f fixed and r random levels. Assume we have f_A, f_B and f_C fixed levels, and r_A, r_B and r_C random levels of factor A, respectively. We partition the dataset based on the combinations of factor levels and construct a partitioned model in each partition. For example, the FRF partitioned model is built from the f_A, r_B and f_C levels. Similarly, other possible partitions are FFF, FFR, RFF, RRF, RFR, FRR and RRR. We illustrate the model construction using the FRF linear mixed model in CRD, having at least one replication per treatment combination, and expressed as

$$y_{FRF_{ijkl}} = \mu_{FRF} + \phi_{A_i} + \phi_{B_j} + \phi_{C_k} + \pi_1 + ... + \pi_t + \epsilon_{FRF_{ijkl}}, \tag{6.1}$$

where $y_{FRF_{ijkl}}$ is the l^{th} observation in the $(ijk)^{th}$ treatment cell of the FRF partition, $l = 1, ..., r_h$ are the replicates (where all $r_h = r$ for balanced data); μ_{FRF} is the overall mean, ϕ_{A_i}, ϕ_{B_j}, ϕ_{C_k} are the main effects of the three factors; $\pi_1, ..., \pi_t$ are the interaction effects; ϕ_{A_i} $(i = 1, 2, ..., f_A, f_A + 1, f_A + 2, ..., a\,(a = f_A + r_A))$, ϕ_{B_j} $(j = r_B + 1, r_B + 2, ..., b\,(b = f_B + r_B))$, and ϕ_{C_k} $(k = 1, 2, ..., f_C, f_C + 1, f_C + 2, ..., c\,(c = f_C + r_C))$, are an unknown parameter corresponding to fixed factor A, random factor B and fixed factor C, respectively. Defining the random main effect as ϕ_R and the random interaction effect as π_R in (6.1), the random effects and the random error term $\epsilon_{FRF_{ijkl}}$, are commonly assumed to have a zero mean and variance, i.e., $\phi_R \sim N(0, \sigma^2_{\phi_R})$, $\pi_R \sim N(0, \sigma^2_{\pi_R})$, and $\epsilon_{FRF_{ijkl}} \sim N(0, \sigma^2_{\epsilon_{FRF}})$.

For a balanced data scenario, with r replications per cell, say, the general linear mixed model equation (6.1) is normally expressed in matrix form as

$$\mathbf{y}_{FRF} = \mathbf{X}_{FRF}\boldsymbol{\beta} + \mathbf{Z}_{FRF}\mathbf{u} + \boldsymbol{\epsilon}_{FRF}, \tag{6.2}$$

where $\mathbf{y}_{FRF} : N \times 1$ is a vector of response observations in the FRF partition; matrix $\mathbf{X}_{FRF} : N \times p$ is a known incidence matrix associated with the vector of p fixed-effects $\boldsymbol{\beta} : p \times 1$ in the model; matrix $\mathbf{Z}_{FRF} : N \times q$ is a known incidence matrix associated with the vector of q random-effects $\mathbf{u} : q \times 1$ in the model; and $\boldsymbol{\epsilon}_{FRF} : N \times 1$ is a vector of random errors. The usual assumption under this model is that the random effects $\mathbf{u} \sim N(\mathbf{0}, \mathbf{G})$, and the random residuals $\boldsymbol{\epsilon}_{FRF} \sim N(\mathbf{0}, \mathbf{R})$, where $\mathbf{R} = \sigma^2\mathbf{I}_N$, and $\mathbf{G}$ is a diagonal matrix of variance components (i.e., different variances, and all zero covariances):

$$\mathbf{G} = \begin{bmatrix} \sigma_1^2\mathbf{I}_r & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \sigma_2^2\mathbf{I}_r & \mathbf{0} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \cdots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \sigma_1^2\mathbf{I}_t \end{bmatrix} \text{ and } \mathbf{R} = \begin{bmatrix} \sigma_\epsilon^2 & 0 & 0 & \cdots & 0 \\ 0 & \sigma_\epsilon^2 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \cdots & \vdots \\ 0 & 0 & 0 & \cdots & \sigma_\epsilon^2 \end{bmatrix}.$$

The total variance-covariance is the structured matrix $\mathbf{V} = \mathbf{Z}\mathbf{G}\mathbf{Z}' + \mathbf{R}$, a structure that guarantees independence and homogeneity of residual errors. This implies that the variance of $\mathbf{y}$ is modelled through $\mathbf{Z}$, $\mathbf{G}$ and $\mathbf{R}$. The simple total variance-covariance, $\mathbf{V}$, has a block-diagonal structure given by the matrix

$$\mathbf{V} = \begin{bmatrix} \sigma_1^2\mathbf{I}_r + \sigma_\epsilon^2 & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \sigma_2^2\mathbf{I}_r + \sigma_\epsilon^2 & \mathbf{0} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \cdots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \sigma_1^2\mathbf{I}_t + \sigma_\epsilon^2 \end{bmatrix}.$$

6.2.3 Linear Mixed Model in RMD

Traditionally, between-subjects and within-subject factors in repeated measures experiments are designated as either fully fixed effects or random effects. Depending on the objectives of the experiment, some factors in linear mixed models might exist with both fixed and random levels (Njuho and Milliken 2005, 2009; Chaka and Njuho, 2021). The same scenario is common with a repeated-measures experiment where either the between-subjects factor or the repeated measures consists of both fixed and random levels. For instance, for improved results, a researcher might decide to consider additional levels of a between-subjects factor, in addition to the old and existing levels. In that case, the improved analysis would need to consider the new factor levels as fixed levels while the old and existing levels are considered as random. The approach creates an opportunity to compare and evaluate the effectiveness of new factor levels (fixed) against the existing ones (random), and/or compile a combined analysis of both. In addition, the random levels allow for assessment of variation between and within the factor levels in the entire population.

6.2.3.1 Construction of a Linear mixed Model in RMD

We consider a repeated measures experiment with n experimental units (EU) that are randomly assigned to each of the a levels of the between-subjects factor A, b levels of the between-subjects factor B, and then t observations taken across time on each experimental unit. The three-factor repeated measures ANOVA model (RM-ANOVA) with two between-subjects factors and one within-subject factor is expressed as

$$y_{ijkl} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \gamma_{k(ij)} + \tau_l + (\alpha\tau)_{il} + (\beta\tau)_{jl} + (\alpha\beta\tau)_{ijl} + \epsilon_{ijkl}, \qquad (6.3)$$

where y_{ijkl} is the l^{th} measurement ($l = 1, 2, ..., t$) of the k^{th} experimental unit ($k = 1, 2, ..., n$) in the i^{th} level ($i = 1, 2, ..., a$) of factor A and the j^{th} level ($j = 1, 2, ..., b$) of factor B; μ is the overall population mean; α_i is the effect of the i^{th} level of factor A; β_j is the effect of the j^{th}level of factor B; $(\alpha\beta)_{ij}$ is the interaction effect of the i^{th}level of factor A and the j^{th} level of factor B; $\gamma_{k(ij)}$ is the effect of the k^{th} experimental unit in the i^{th} level of factor A and the j^{th} level of factor B; τ_l is the effect of the l^{th} period; $(\alpha\tau)_{il}$, $(\beta\tau)_{jl}$, and $(\alpha\beta\tau)_{ijl}$ are the interaction effects of the levels of factors A and B with the l^{th} period; ϵ_{ijkl} are the random error terms which are assumed to be independent and distributed as $\epsilon_{ijkl} \sim N(0, \sigma^2)$. Depending on the purpose of inference, the three factor effects α_i, β_j and τ_l may be considered as fixed or random effect. The primary objective of a repeated-measures analysis of variance is to compare the factor and period effects with respect to differences in the response, as well as to understand the factor by period interaction effects.

6.2.3.2 Partitioning the Factors

For illustration purposes, we consider a three-factor repeated measures experiment with n experimental units (EU) that are randomly assigned to each of the a levels of the between-subjects factor A (with f_a fixed levels and r_a random levels, $a = f_a + r_a$, say), the between-subjects factor B (with f_b fixed levels and r_b be random levels, $b = f_b + r_b$, say), and the within-subjects factor C with t measurements (all considered to be fixed in this case) taken on each of the experimental units (EU). Partitioning the between-subjects factors A and B, and the fixed levels of the within-subject factor C gives us the fixed-fixed-fixed (FFF), fixed-random-fixed (FRF), random-fixed-fixed (RFF), and random-random-fixed (RRF) subsets. The repeated measures linear mixed model from each of the four partitions will have the form,

$$y_{p_{ijkl}} = \mu_p + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \gamma_{k(ij)} + \tau_l + (\alpha\tau)_{il} + (\beta\tau)_{jl} + (\alpha\beta\tau)_{ijl} + \epsilon_{p_{ijkl}}, \tag{6.4}$$

where the subscript p in $y_{p_{ijkl}}$ and $\epsilon_{p_{ijkl}}$ denotes the partition $(p = 1, 2, ..., 4)$; $y_{p_{ijkl}}$ is the l^{th} measurement $(l = 1, 2, ..., t)$ of the k^{th} experimental unit $(k = 1, 2, ..., n)$ in the i^{th} level $(i = 1, 2, ..., a)$ of factor A and the j^{th} level $(j = 1, 2, ..., b)$ of factor B in the p^{th} partition. Depending on whether the effects are fixed or random, the model parameters are as defined in (6.3).

6.2.4 Model Assumptions

In experimental research that involves analysis of variance (ANOVA) as a technique for comparing different treatment means, a set of assumptions which include the usual normality and homogeneity of variance must be checked before analysis of data (Kotchaporn and Araveeporn, 2018). These traditional normality tests such as the Q-Q plots or Shapiro Wilk's test, and outlier detection approaches (e.g., box plots) are appropriate for diagnosing violation of assumptions.

6.2.4.1 Sphericity (Circularity) Assumption

Sphericity, or homogeneity of variances over time, is one of the most important assumptions in repeated-measures analysis (Sullivan, 2008). Similar to homogeneity of variances in a between-subjects analysis of variance (ANOVA), sphericity or circularity holds in ANOVA for repeated measures designs when the variances of the differences among all possible pairs of related groups (within-subject factor levels) means are equal (Armstrong, 2017), i.e., when a fixed variability exists amongst the repeated measurements. The assumption is usually unrealistic in repeated-measures designs where observations are correlated since a random factor that causes a measurement in one subject to be a bit high (or low) should have an effect on the next measurement in the same subject.

Univariate tests for within-subjects effect apply when sphericity holds, otherwise some possible alternatives such as the multivariate test (which does not assume sphericity), the Greenhouse-Geisser (Geisser and Greenhouse, 1958) correction or the Huynh-Feldt (Huynh and Feldt, 1976) correction would be appropriate since they provide a corrected degrees of freedom for the treatment and error terms, thus enabling a more accurate adjusted p-value. Failure to address the problem of sphericity when conducting analysis of variance often leads to inflated F-ratios and type I errors. In addition, the conclusions drawn from post-hoc tests for group mean differences will be biased and inaccurate (Armstrong, 2017). The tests for sphericity are offered in most statistical computing packages (SAS, R, SPSS, etc).

6.2.4.2 Compound Symmetry Assumption

An overly restrictive assumption but closely related to sphericity is the compound symmetry assumption, which states that the variance and correlation of the difference scores from the same subject must be constant. The assumption implies that there is a constant correlation between observations regardless of the time between the observations, which is not always realistic in many repeated measures applications (Ott and Longnecker, 2016). Compound symmetry implies sphericity, but not vice-versa.

The compound symmetry is simplified in the Huynh-Feldt (Huynh and Feldt, 1976) condition which states that the variances of the differences between any pair of observations on the same experimental unit are equal. Univariate tests for within-subjects effect apply when compound symmetry holds. This condition is tested in many software packages by the Mauchly's test (Mauchly, 1940) for compound symmetry. If the compound symmetry (and hence sphericity) holds, then a split-plot analysis will be a more accurate approximation to the repeated measures experiment since it provides relatively more accurate p-values for testing treatment effects (Ott and Longnecker, 2016). Therefore, it follows that if sphericity is violated, then compound symmetry may not hold as well. However, the Mulchly's test for sphericity has been criticised for its over sensitivity and tendency to reject compound symmetry (UCLA: Statistical Consulting Group, n.d.).

6.2.5 Estimation Techniques

We generalise the linear mixed model (6.2) to describe data from partitioned repeated measurements, wherein the fixed-effects component, $\mathbf{X}_p\boldsymbol{\beta}$, consists of the design matrix $\mathbf{X}_p$ and the fixed-effects coefficients $\boldsymbol{\beta}$ are as defined in the general linear model; the random-effects component, $\mathbf{Z}_p\mathbf{u}$, contains the block-diagonal random-effects design matrix $\mathbf{Z}_p$ with the design matrices for the individual subjects (Z_i), $\mathbf{u}$ is a vector of random coefficients (the between-subjects variance-covariance components), and $\boldsymbol{\epsilon}_p$ denotes the within-subjects error vector (the within-subjects variance-covariance components). The random effects and the residuals follow the distribution $\mathbf{u} \sim N(\mathbf{0}, \mathbf{G})$ and $\boldsymbol{\epsilon}_p \sim N(\mathbf{0}, \mathbf{R})$, respectively, where $\mathbf{G}$ is a block-diagonal covariance matrix of the random effects and $\mathbf{R}$ is a diagonal covariance matrix with partitions corresponding to each subject (within-subjects errors) in the analysis. The covariance matrix for the repeated measures data is composed of matrices $\mathbf{Z}_p$, $\mathbf{G}$ and $\mathbf{R}$, which is a block-diagonal $\boldsymbol{\Sigma} = Var(y) = \mathbf{Z}_p\mathbf{G}\mathbf{Z}_p' + \mathbf{R}$. The non-singular components $\mathbf{G}$ and $\mathbf{R}$ are usually estimated by two principal likelihood methods for estimating variance components (Moskowitz *et al.*, 2002), i.e., the maximum likelihood (ML) method, and the restricted maximum likelihood (REML). These procedures are available in various mixed model statistical software such as the SAS *PROC MIXED* procedure) and R, with the REML estimates generally preferred unless the data sets are quite large (Moskowitz *et al.*, 2002).

Assuming that both the random effects and the error terms are normally distributed, the likelihood function for the repeated measures mixed model is given by (Hocking, 1985):

$$\begin{aligned} l = log\left[L(\mathbf{y}_p)\right] &= \frac{-N}{2}log(2\pi) - \frac{1}{2}log|\boldsymbol{\Sigma}_p| - \frac{1}{2}(\mathbf{y}_p - \mathbf{X}_p\boldsymbol{\beta})'(\boldsymbol{\Sigma}_p^{-1})(\mathbf{y}_p - \mathbf{X}_p\boldsymbol{\beta}) \\ &= C - \frac{1}{2}log|\boldsymbol{\Sigma}_p| - \frac{1}{2}(\mathbf{y}_p - \mathbf{X}_p\boldsymbol{\beta})'(\boldsymbol{\Sigma}_p^{-1})(\mathbf{y}_p - \mathbf{X}_p\boldsymbol{\beta}), \end{aligned} \tag{6.5}$$

where $\mathbf{y}_p$ and $\mathbf{V} = \boldsymbol{\Sigma}_p$ are as defined in (6.2). Similarly, a modification of the ML procedure through factorisation of the likelihood function was proposed as alternative method of estimating covariance parameters is the restricted maximum likelihood function (Harville, 1977):

$$\begin{aligned} l_{re} = log\left[L(\mathbf{y}_p)\right] &= \frac{-N}{2}log(2\pi) - \frac{1}{2}log|\boldsymbol{\Sigma}_p| - \frac{1}{2}(\mathbf{y}_p - \mathbf{X}_p\boldsymbol{\beta})'(\boldsymbol{\Sigma}_p^{-1})(\mathbf{y}_p - \mathbf{X}_p\boldsymbol{\beta}) \\ &= C - \frac{1}{2}log|\mathbf{X}_p'\boldsymbol{\Sigma}_p^{-1}\mathbf{X}_p| - \frac{1}{2}(\mathbf{y}_p - \mathbf{X}_p\hat{\boldsymbol{\beta}})'(\boldsymbol{\Sigma}_p^{-1})(\mathbf{y}_p - \mathbf{X}_p\hat{\boldsymbol{\beta}}), \end{aligned} \tag{6.6}$$

where the available covariance matrix $\mathbf{V}$ is used to estimate the fixed effects parameters, $\hat{\boldsymbol{\beta}} = (\mathbf{X}_p'\boldsymbol{\Sigma}_p^{-1}\mathbf{X}_p)^{-1}\mathbf{X}_p'\boldsymbol{\Sigma}_p^{-1}\mathbf{y}_p$.

The main challenge in repeated measures analysis of variance is to determine the adequate correlation structure, since the constant variance assumption for the distribution of the error terms is likely not reasonable for the distribution of error terms within subjects. There are various possible choices of covariance structure for repeated measures within each subject depending on the chosen parameterisation for $\mathbf{G}$ and $\mathbf{R}$. The choices are usually guided by the limitation of the software and the insight of the researcher. The commonest covariance structures include variance components, compound symmetry (common covariance plus diagonal), unstructured (general covariance), and autoregressive (SAS Institute Inc., 2017). With the $PROC\ MIXED$ statement in SAS, one can specify any repeated measurements covariance structure for $\mathbf{G}$ by using the $RANDOM$ statement, and specify the form of $\mathbf{R}$ by the $REPEATED$ statement, in conjunction with the $TYPE$ option (Milliken and Johnson, 2002). Excluding the $REPEATED$ statement specifies the classical $\mathbf{R}$ which is assumed to be equal to be $\sigma_\epsilon^2\mathbf{I}_N$.

There are numerous ways of identifying the most appropriate covariance structure amongst a set candidate structures (SAS Institute Inc., 2017). The most recommended approach is to select the structure that gives the smallest Akaike's Information Criterion (AIC), (Akaike, 1974), a statistic that is defined by the model and the maximum likelihood estimates of the parameters from specifying the variance-covariance as

$$AIC = (-2)\mathcal{L}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\Sigma}}) + 2(k), \tag{6.7}$$

where k is effective number of independently adjusted parameters in the covariance matrix , and $\mathcal{L}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\Sigma}}) = log(ML)$ is the value of the likelihood function evaluated at $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\Sigma}})$. A better model is the one with the smallest AIC value. Different forms of the $\mathbf{R}$ can be compared for adequacy using the likelihood ratio test statistic (Milliken and Johnson, 2002). The hypotheses involve are, H_0 : $\mathbf{R}_1$ is as adequate as $\mathbf{R}_2$ against H_1 : $\mathbf{R}_1$ is not as adequate as $\mathbf{R}_2$, where $\mathbf{R}_1$ is a special case of $\mathbf{R}_2$. Suppose $\mathbf{R}_1$ and $\mathbf{R}_2$ have k_1 and k_2 parameters, respectively, with $(k_1 < k_2)$. The test statistic is $Q = (-2)[\mathcal{L}(\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\Sigma}}_1) - \mathcal{L}(\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\Sigma}}_1)]$ which is distributed as $\chi^2(k_2 - k_1)$. We reject H_0 when $Q \geq \chi^2_{\frac{\alpha}{2}}(k_2 - k_1)$.

6.2.6 Methods of Inference

We present the algorithm for obtaining expected mean squares using the ANOVA approach, for the FRF model.

6.2.6.1 Algorithm for Deriving Expected Mean Squares

The following steps are used to derive the expected mean squares of the effects in the model:

(a) Based on the model involved, construct a two-way table with column headings corresponding to the source of variation, effect labels, each of the subscripts included in the model, and row headings corresponding to each source of variation in the ANOVA table.

(b) Above each subscript, write the associated number of factor levels, and insert on top either an "F" if the factor levels are fixed, or an "R" if the factor levels are random.

(c) Create an extra column on the extreme right for the variance components corresponding to the source of variation, and insert the appropriate random variance component ($\sigma^2_{.}$) or fixed variance component ($\theta_{.}$) for each source of variation.

(d) Compare the column subscript and the factor effect in each row and write the number of levels corresponding to that subscript if the column subscript is not included in the factor effect label. Otherwise leave blank.

(e) For rows that have an effect which contains bracketed subscripts, write a "1" under the column if the subscript is included in the bracket.

(f) For each row that has a fixed variance component ($\theta_{.}$), put a zero in the cell headed by an "F" when the subscript is included in the effect label.

(g) Enter a "1" in all remaining blank cells.

(h) To get the expected mean squares for each effect, identify all the variance components associated with that effect label. Cover the column(s) headed by the effect subscript(s) in that effect, and obtain the coefficient of each of the identified components from the product of the entries in the column(s) headed by the uncovered subscript(s). Include the variance component σ^2_ϵ with the coefficient of 1 in the list.

Table 6.1 summarises the variance components obtained for a three-factor repeated measures design, with two between-subjects factors (A and B) on one within-subject factor (C) using Algorithm 6.2.6.1. We use the model (6.4) to build an FRF model, which assumes factors A and C as fixed, factor B as random, and experimental units (EU) as random.

Table 6.1: Variance Components for a Three-factor Repeated Measures Model

		F	**R**	**F**	**R**	
		a	**b**	**t**	**n**	
Source	**Effect**	**i**	**j**	**l**	**k**	**Components**
A	α_i	0	b	t	n	θ_α
B	β_j	a	1	t	n	σ^2_β
AB	$(\alpha\beta)_{ij}$	1	1	t	n	$\sigma^2_{\alpha\beta}$
EU	$\gamma_{k(ij)}$	1	1	t	1	$\sigma^2_{\gamma(\alpha\beta)}$
P	τ_l	a	b	0	n	θ_τ
PA	$(\alpha\tau)_{il}$	0	b	0	n	$\theta_{\alpha\tau}$
PB	$(\beta\tau)_{jl}$	a	1	1	n	$\sigma^2_{\beta\tau}$
PAB	$(\alpha\beta\tau)_{ijl}$	1	1	1	n	$\sigma^2_{\alpha\beta\tau}$
Error	$\epsilon_{l(ijk)}$	1	1	1	1	σ^2_ϵ

For example, $E(MSA)$, with effect α_i, is composed by the variance components θ_α, $\sigma^2_{\alpha\beta}$, $\sigma^2_{\gamma(\alpha\beta)}$, $\theta_{\alpha\tau}$, $\sigma^2_{\alpha\beta\tau}$ and σ^2_ϵ as follows,

$$E(MSA) = \sigma^2_\epsilon + t\sigma^2_{\gamma(\alpha\beta)} + tn\sigma^2_{\alpha\beta} + n\sigma^2_{\alpha\beta\tau} + btn\theta_\alpha.$$

Table 6.2 displays the ANOVA layout and the expected mean squares for a three-factor repeated measures design when one of the factors is a within-subject factor. The FRF model is considered for illustration, with experimental units assumed to be random.

Table 6.2: Expected Mean Squares for a Three-factor Repeated Measures Design

Source	**Sum of Squares**	**Degrees of Freedom**	**E[MS]**
A	SSA	$a-1$	$\sigma^2_\epsilon + t\sigma^2_{\gamma(\alpha\beta)} + tn\sigma^2_{\alpha\beta} + n\sigma^2_{\alpha\beta\tau} + btn\theta_\alpha$
B	SSB	$b-1$	$\sigma^2_\epsilon + n\sigma^2_{\alpha\beta\tau} + an\sigma^2_{\beta\tau} + t\sigma^2_{\gamma(\alpha\beta)} + tn\sigma^2_{\alpha\beta}$ $+atn\sigma^2_\beta$
$A \times B$	$SS(A \times B)$	$(a-1)(b-1)$	$\sigma^2_\epsilon + t\sigma^2_{\gamma(\alpha\beta)} + n\sigma^2_{\alpha\beta\tau} + tn\sigma^2_{\alpha\beta}$
$Unit(A \times B)$	$SSU(A \times B)$	$ab(n-1)$	$\sigma^2_\epsilon + t\sigma^2_{\gamma(\alpha\beta)}$
$Period$	SSP	$t-1$	$\sigma^2_\epsilon + n\sigma^2_{\gamma(\alpha\beta)} + an\sigma^2_{\alpha\beta\tau} + abn\theta_\tau$
$Period \times A$	$SSP \times A$	$(a-1)(t-1)$	$\sigma^2_\epsilon + n\sigma^2_{\alpha\beta\tau} + bn\theta_{\alpha\tau}$
$Period \times B$	$SSP \times B$	$(b-1)(t-1)$	$\sigma^2_\epsilon + n\sigma^2_{\alpha\beta\tau} + an\theta_{\alpha\tau}$
$Period \times A \times B$	$SSP \times A \times B$	$(a-1)(b-1)(t-1)$	$\sigma^2_\epsilon + n\sigma^2_{\alpha\beta\tau}$
$Residual$	SSE	$ab(t-1)(n-1)$	σ^2_ϵ
$Total$	SST	$abtn-1$	

$$\theta_\alpha = \tfrac{1}{(a-1)}\textstyle\sum_{i=1}^{a}\alpha_i^2; \quad \theta_{\alpha\tau} = \tfrac{1}{(a-1)(t-1)}\textstyle\sum_{i=1}^{a}\sum_{l=1}^{t}(\alpha\tau)^2_{il}; \quad \theta_\tau = \tfrac{1}{(t-1)}\textstyle\sum_{l=1}^{t}\tau_l^2$$

6.2.6.2 Hypothesis Testing for Fixed Effects

We are interested in testing the main and the interaction effects of the between-subjects factor and the within-subjects factor in both the partitioned and the combined repeated measures linear mixed model. In addition to checking model assumptions, the following hypotheses are of interest for each partitioned model:

H1: Between-subjects main and interaction effects (e.g., $H_0 : \alpha_i = 0$)

H2: Within-subjects main and interaction effects (e.g., $H_0 : \tau_i = 0$)

The test statistic for $H1$ is given by $F = \frac{MSA}{(MSU(A \times B)} \sim F_{(a-1,ab(n-1))}(\alpha)$, while the test statistic for $H2$ is given by $F = \frac{MSP}{MSE} \sim F_{(t-1,ab(t-1)(n-1))}(\alpha)$. The interaction effects among the between-subjects and the interaction effects among within-subjects are tested by the $MSU(A \times B)$ and the MSE on the denominator, respectively.

6.2.6.3 Hypothesis Testing for Random Effects

Variance components are estimated by equating mean square to expected mean squares derived in Table 2. Where there are no valid F-tests, approximate F-tests are constructed for the sources of variability in random effects (Kuehl, 2000). For random factor B, the hypothesis of interest might be,

H3. Random effects (e.g., $H_0 : \sigma^2_\beta = 0$, against $H_1 : \sigma^2_\beta > 0$).

Computer software programs such as R and SAS can be used to generate these statistical tests with p-values at predetermined level of significance.

6.2.6.4 Combined Analysis

The individual partitioned models provide the pieces of information which is needed for an integrated analysis. The combined model, is built from combining the degrees of freedom and sum of squares associated with each source of variation for each appropriate hypotheses test. For example, the combined effect of the within-subjects factor C ($Period$) in the FC model is obtained from the partitions where the factor $Period$ is fixed, i.e., the pieces of information is supplied by the partitioned models $FF\mathbf{F}$, $FR\mathbf{F}$, $RF\mathbf{F}$ and RRF. Similarly, the other main and interaction effects for the combined model are obtained by summing up the associated degrees of freedom and sums of squares. Alternatively, the combined analyis can be implemented in th SAS mixed model package, $PROC\ GLIMMIX$ using proper coding for the fixed and random factor levels (Piepho *et al.*, 2006) to generate comparisons among the means of the partitioned subsets of data.

6.3 Results

The three-factor repeated-measures experiment aimed at investigating the impact of combining carbon tetrachloride (CCI4), with four levels (0, 1.0, 2.5 and 5.0 mM), and chloroform (CHCI3), with four levels (0, 5 10 and 25 mM), on the percentage leakage observed over *Time* (observations considered at times 1, 2 and 3). Table 6.3 shows the multivariate data (wide format) layout for a three-factor repeated measures experiment.

Table 6.3: Data layout for a Three-factor Repeated-measures Experiment

			Time Period			
CCI4	CHCI3	Flask	1	2	$\cdots$	t
1	1	1	y_{1111}	y_{1112}	$\cdots$	y_{111t}
		⋮	⋮	⋮	⋮	⋮
		n	y_{11n1}	y_{1n2}	$\cdots$	y_{11nt}
	2	1	y_{1211}	y_{1212}	$\cdots$	y_{121t}
		⋮	⋮	⋮	⋮	⋮
		n	y_{12n1}	y_{12n2}	$\cdots$	y_{12nt}
	⋮	⋮	⋮	⋮	⋮	⋮
	b	1	y_{1b11}	y_{1b12}	$\cdots$	y_{1b1t}
		⋮	⋮	⋮	⋮	⋮
		n	y_{1bn1}	y_{1bn2}	$\cdots$	y_{1bnt}
2	1	1	y_{2111}	y_{2112}	$\cdots$	y_{211t}
		⋮	⋮	⋮	⋮	⋮
		n	y_{21n1}	y_{2n2}	$\cdots$	y_{21nt}
	2	1	y_{2211}	y_{2212}	$\cdots$	y_{221t}
		⋮	⋮	⋮	⋮	⋮
		n	y_{22n1}	y_{22n2}	$\cdots$	y_{22nt}
	⋮	⋮	⋮	⋮	⋮	⋮
	b	1	y_{2b11}	y_{2b12}	$\cdots$	y_{2b1t}
		⋮	⋮	⋮	⋮	⋮
		n	y_{2bn1}	y_{2bn2}	$\cdots$	y_{2bnt}
⋮	⋮	⋮	⋮	⋮	⋮	⋮
a	1	1	y_{a111}	y_{a112}	$\cdots$	y_{a11t}
		⋮	⋮	⋮	⋮	⋮
		n	y_{a1n1}	y_{an2}	$\cdots$	y_{a1nt}
	2	1	y_{a211}	y_{a212}	$\cdots$	y_{a21t}
		⋮	⋮	⋮	⋮	⋮
		n	y_{a2n1}	y_{a2n2}	$\cdots$	y_{a2nt}
	⋮	⋮	⋮	⋮	⋮	⋮
	b	1	y_{ab11}	y_{ab12}	$\cdots$	y_{ab1t}

6.3.1 Checking Model Assumptions

In order to test the suitability of using repeated measures design on the experiment, the normality, outliers and sphericity assumptions were tested for each partitioned data subset. Q-Q plots were used to simultaneously check the normality and outlier assumptions. Figure 6.1 shows the normal Q-Q plots for the four data subsets.

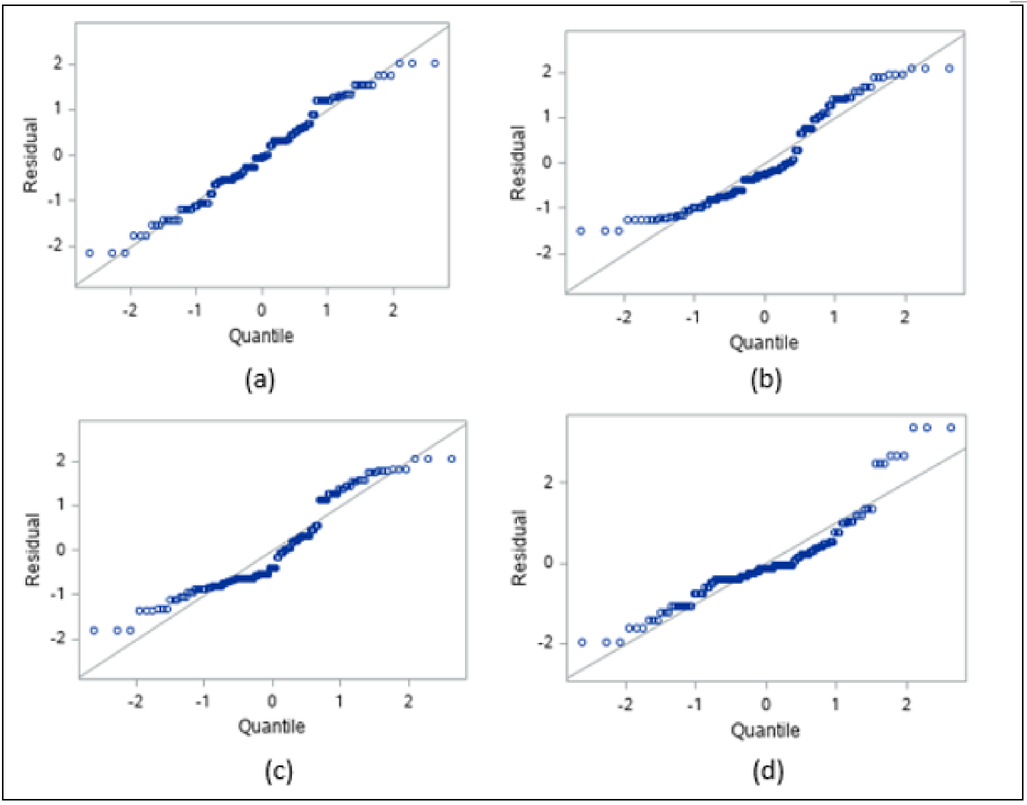


Figure 6.1: Q-Q Plots for (a) FFF (b) FRF (c) RFF (d) RRF Data Sets

The FFF, FRF, RFF and RRF data sets did not show any serious deviations from normality. Furthermore, the Q-Q plots do not show any influential point (outlier) that warrants exclusion since all plots were not very far from the diagonal.

The SAS reports on Mauchly's W test was used to test the sphericity assumption in each partitioned data subset. The *PROC MIXED* procedure in SAS was used to fit two models: one that specifies an unrestricted covariance structure, and the other with a less conservative Huynh-Feldt (H-F) adjustment in the *Type* option. Based on the hypothesis,

$$H_0 : Sphericity\ assumption\ is\ satisfied,$$

where the difference between the -2log-likelihoods of the two compared models follows a Chi-square distribution with degrees of freedom equal to the difference in the number of parameters in them, the test results were as follows:

FFF: $\chi^2_{35}(0.05) = 55.76$, $D = 1722.6$, significant;

FRF: $\chi^2_{38}(0.05) = 55.76$, $D = 184.54$, significant;

RFF: $\chi^2_{37}(0.05) = 55.76$, $D = 288.8$, significant;

RRF: $\chi^2_{36}(0.05) = 55.76$, $D = 384.1$, significant.

The Mauchly's test statistics for the partitioned data FFF, FRF, RFF and RRF were significant at 5% level of significance, which implies that the sphericity assumption was not satisfied in these data subsets.

6.3.2 Analysis of Results

We use the restricted maximum likelihood estimation (REML) approach to estimate these variance components. Table 6.4 contains the estimated AIC from each of the covariance structures as well as the number of covariance parameters estimated with a non-zero value.

Table 6.4: Akaike's Information Criteria (AIC) for the Partitioned Models

Model FFF				**Model FRF**		
Covariance Structure	Number of Parameters	AIC AIC		Covariance Structure	Number of Parameters	AIC AIC
CS	2	-282.6		CS	4	-361.3
AR(1)	2	-443.1		AR(1)	3	-492.4
ARH(1)	10	-525.6		ARH(1)	13	-506.3
CSH	10	-356.7				
Model RFF				**Model RRF**		
Covariance Structure	Number of Parameters	AIC AIC		Covariance Structure	Number of Parameters	AIC AIC
CS	4	-450.1		CS	4	-665.8
AR(1)	4	**-522.4**		AR(1)	4	-782.4
ARH(1)	13	-509.3		ARH(1)	13	-**785.3**

The covariance structure ARH(1) had minimum AIC values in models FFF, FRF, and RRF, while covariance structure AR(1) had a smallest AIC value in model RFF. Therefore, based on AIC, the covariance structure ARH(1) was chosen as the most adequate covariance structure for the partitioned models. The *PROC MIXED* procedure is traditionally robust and flexible to handle cases where the sphericity assumption is not satisfied.

Table 6.5 gives a summary of the F-tests, for the main and interaction effects of the between-subjects and the within-subjects factors. The F-tests for the portioned model RFF were obtained from the split-plot analysis of repeated measures since the sphericity condition was valid in the experiment. The CHCI3 factor had significant effect on leakage percentage ($p - value < 0.001$) over time. Time effect was significant in the FFF partition.

Table 6.5: Akaike's Information Criteria (AIC) for the Partitioned Models

Model	Effect	Numerator Degrees of Freedom	Denominator Degrees of Freedom	F	P-value
FFF	CCI4	1	12	0.24	0.6363
	CHCI3	1	12	17.33	0.0013
	CCI4*CHCI3	2	12	1.35	0.2678
	Time	2	24	95.99	<0.0001
	Time*CCI4	2	24	2.66	0.0908
	Time*CHCI3	2	24	20.91	<0.0001
	Time*CCI4*CHCI3	2	24	1.26	0.3023
FRF	CCI4	1	1	0.5	0.6079
	Time	2	2	2.17	0.3156
	Time*CCI4	2	2	2.83	0.2610
RFF	CHCI3	1	1	7.07	0.2290
	Time	2	2	5.42	0.1557
	Time*CHCI3	2	2	3.26	0.2348
RRF	Time	2	2	1.01	0.4978

Considering the fixed-effects (FFF) repeated measures study, the factor $CHCI3$ has a significant effect on the response ($p - value < 0.001$). Hence, we conclude that chloroform ($CHCI3$) had a significant impact on the amount of lactic dehydrogenase (LHD) enzyme percentage leakage (toxicity of cells) over time, whereas neither carbon tetrachloride ($CCI4$) in isolation nor the interaction thereof had non-significant influence. Furthermore, the *Time* factor played an important role in determining the amount of leakage as well. However, the interaction of old (random) and new (fixed) levels of the between-subjects factor levels had non-significant effects except in RFF when $CHCI3$ fixed levels were involved.

Table 6.6 gives a summary of the estimated covariance parameters in each of the partitioned model. There was a zero variance for the random levels of chloroform ($CHCI3$), and very small estimates in other factors, which resulted in very low estimates of covariance parameters in the models.

Table 6.6: Covariance Parameter Estimates of the Partitioned Models

Model	Covariance Parameter	Estimate	Standard Error	Proportion of Variation Accounted for
FRF	CHCI3	0	0	0
	CCI4*CHCI3	0.04249	0.04470	22.5
	Time*CHCI3	0	0	0
	Time*CCI4*CHCI3	0.000621	0.000759	0.3
RFF	CCI4	0.000607	0.003056	0.05
	CCI4*CHCI3	0	0	0
	Time*CCI4	0.000149	0.000809	0.02
	Time*CCI4*CHCI3	0.000806	0.000984	0.08
RRF	CCI4	0.0011	0.00401	2.6
	CHCI3	0.00306	0.00677	7.1
	CCI4*CHCI3	0.00232	0.00396	5.4
	Time*CCI4	0.000002	0.00002	0.0
	Time*CHCI3	0.000065	0.000096	0.1
	Time*CCI4*CHCI3	0.000013	0.000025	0.0

The $CHCI3$, $CCI4$ random levels and their interaction had a noticeable contribution to the proportion of variation in the amount of lactic dehydrogenase (LHD) enzyme percentage leakage in RRF partitions. Generally, the variable *Time* had very little interaction effect with the between-subjects factors FRF in determining the proportion of variation toxicity of cells.

Due to some limitations in most statistical packages such as the R CRAN *lme*4 package for linear mixed models, which currently does not have options of other covariance structures to cater for correlated error variances, generating a combined analysis may not a straightforward exercise. However, the mixed model methodology (SAS *PROC MIXED* procedure) has options for other covariance structures which accommodate correlated error variances even though it does not provide for the computation of sums of squares or F-statistics from the ratio of mean squares. We scrapped the targeted data subset based on the effects of interest before analysing using the *PROC MIXED* procedure (if sphericity is violated) or the split-plot approximation if sphericity holds. Analogous to the partitioned analyses, comparison of model fit via the AIC approach was conducted. For convenience purposes, let the factors $CCI4$, $CHCI3$ and *Time* be labelled as factor A, B and C, respectively. The *PROC MIXED* procedure was used to fit the repeated measures linear mixed models for the intended narrow inference space (McLean *et al.*, 1991; Chaka and Njuho, 2021). Table 6.7 shows the results for the combined analysis generated using $PROC\ GLIMMIX$ procedure (see Annexture D for the SAS code).

Table 6.7: Fixed Effects F-Tests for the Combined Models

Type III Tests of Fixed Effects in Combined Models						
Model	AIC [CS]	Effect	Num DF	Den DF	F	Pr > F
FA	-127.6 [ARH(1)]	A	1	24	3.36	0.0794
		B	3	24	5.81	0.0039
		A*B	3	24	9.10	0.0003
		C	2	48	12.33	0.0001
		A*C	2	48	2.52	0.0908
		B*C	6	48	2.33	0.0468
		A*B*C	6	48	0.76	0.6027
FB	-127.6 [AR(1)]	A	3	24	0.53	0.6631
		B	1	24	16.3	0.0005
		A*B	3	24	0.22	0.8801
		C	2	48	24.82	<0.0001
		A*C	6	48	1.09	0.3795
		B*C	2	48	7.82	0.0012
		A*B*C	6	48	0.50	0.8064
$FA \times FB$	-69.1 [AR(1)]	A	1	12	0.14	0.7106
		B	1	12	10.63	0.0068
		A*B	1	12	0.83	0.3807
		C	2	24	19.75	<0.0001
		A*C	2	24	0.58	0.5667
		B*C	2	24	4.65	0.0196
		A*B*C	2	24	0.26	0.7725

Of the possible candidate covariance structures (CS, CSH, AR(1) and ARH(1)), structure ARH(1) was selected as the most appropriate covariance structure for the combined fixed-effects models FA, while AR(1) was appropriate for FB and $FA \times FB$. The factors B ($CHCI3$) and C (Time) had significant effects ($p-value < 0.05$) in the combined models. The broad inference scope results for the combined models (assuming random factor A or B effects) were similarly analysed.

6.4 Discussion

Based on the illustrative example results, the approach managed to isolate the effects of new and old factor levels over time. The combined analysis confirmed the results of the partitioned analysis on the percentage leakage in cells. The partitioning approach conforms to the model construction and the analysis procedures in repeated-measures design. It can be used as a planning tool where factor combination and time are of interest in designing of experiments which involve repeated measures. In such experiments, blindly adopting the assumption of homogeneous of error terms without exploring on possible candidate covariance structures may compromises the ability of an experiment to detect sufficient variation in response variable. In addition, approach enhances the accuracy of inferences by providing partitioned analysis of heterogeneous variances and covariance structures, which sometimes are not identical in the data subsets.

Given the increased complexity of research data in the various research fields, the application of linear mixed model methodology has to be in line with the data covariance structures for accurate results to be achieved. One of the approaches that has proved to be a reliable tool for managing big data complexity issues is the partitioning approach (Njuho and Milliken, 2005, 2009; Chaka and Njuho, 2021) when the traditional homogeneous error variance structure is assumed. The current study extends the new approach to a three-factor treatment structure in repeated-measures design where linear mixed models are applicable. In essence, the approach can be extended to cater for the repeated measures experiments where any number of between-subjects and within-subjects factors are involved. In most cases, repeated measures experiments do not assume equal and uncorrelated error vectors since regularly timed measurements taken on the same subject over time are usually correlated (Moskowitz *et al.*, 2002).

For the fixed-effects partitions, the linear mixed models for a repeated-measures design are fit by PROC GLM procedure, and combined analysis of these would be obtained by syncretising the sum of squares and degrees of freedom from the fit models. However, obtaining a combined analysis using the SAS PROC MIXED procedure is impossible using the sum of squares approach since the PROC MIXED procedure uses a likelihood-based estimation scheme instead of least squares method. A comparable alternative for the reduction sum of squares for the fit model in PROC MIXED is to consider the amount of information retained by the fit model when compared to the null model, i.e. the difference in AIC information ($AIC_{model} - AIC_{null}$).

A more convenient and easier approach of generating combined analyis for comparisons among the means of the partitioned subsets of data is achievable using the SAS *PROC GLIMMIX* procedure and proper coding for the fixed and random factor levels (Piepho *et al.*, 2006).

6.5 Conclusion

The main purpose of this chapter was to demonstrate how the partitioning approach can be implemented on a three-factor linear mixed model for correlated data when some of the factors involved have both fixed and random levels. The partitioning approach was useful in model construction and hypothesis testing in repeated-measures data when heterogeneous error structure is assumed. Approaches to generating combined analyses were discussed. Although the MLE method used in SAS *PRO MIXED* does not estimate sums of squares, data scrapping based on the targeted factor levels and the SAS *PROC GLIMMIX* with proper coding for the fixed and random factor levels, proved to be useful alternatives to obtain the combined analyses. The partitioning approach can be adopted as an essential tool for comparison of new inventions against the existing strategies and equipment. The partitioning approach enriches in exploring the fixed and random levels of the same factor and the subsequent interaction of levels of factors of interest. We get to assess the differences between levels of the same factor and understand the variation within the same factor. In addition, the modelling allows for assessment of various covariance structures.

CHAPTER 7

EXTENTION TO MULTIPLE FACTORS

7.1 Introduction

The analysis of factors with both fixed and random levels has been illustrated and explained using two-way and three-way mixed linear models. We present the extension of the concept to a general p-factor linear mixed model with interaction in completely randomised design (CRD) and other experimental designs. We demonstrate how a p-factor linear mixed model is constructed under such conditions, and extend the analysis approach to other complex situations.

7.1.1 Construction of a P-factor Linear Mixed Model in CRD

Analogous to the three-way linear mixed model scenario in the previous chapters, suppose we have p factors, each with a combination of f fixed and r random levels, whose main and interaction effects are considered to predict a single response variable. Assume we have $f_A, f_B, ..., f_P$ fixed levels and $r_A, r_B, ..., r_P$ random levels of factor 1 through factor P, respectively. The proposed concept suggests that, the p-way linear mixed model is constructed as

$$y_{ijk...hn} = \mu + \varphi_{Ai} + \varphi_{Bj} + ... + \varphi_{Ph} + \pi_1 + ... + \pi_t + \epsilon_{ijk...hn}, \tag{7.1}$$

where μ is the overall mean, $\varphi_{Ai}, \varphi_{Bj}, ..., \varphi_{Ph}$ are the main effects of p factors; $\pi_1, ..., \pi_t$ are the interaction effects; and $n = 1, ..., r_s$ are the replicates (where all $r_s = r$ for balanced data). Suppose that φ_{Ai} $(i = 1, 2, ..., f_A, f_A + 1, f_A + 2, ..., a(a = f_A + r_A))$ an unknown parameter corresponding to factor A, with f_A fixed and r_A random levels, the same explanation is true for the P^{th} main factor. Defining the random main effect as φ_R and the random interaction effect as π_R in (7.1), the random effects and the usual random error term $\epsilon_{ijk...hn}$, are assumed to have a zero mean and variance, i.e. $\varphi_R \sim N(0, \sigma^2_{\varphi_R})$, $\pi_R \sim N(0, \sigma^2_{\pi_R})$, and $\epsilon_{ijk...hn} \sim N(0, \sigma^2_\epsilon)$.

7.1.2 Partitioning a P-factor Linear Mixed Model in CRD

Intuitively, we can build the p-way treatment structures by partitioning (7.1) using the definitions of fixed and random effect levels, i.e. $a = f_A + r_A$ levels of factor A through $p = f_P + r_P$ levels of factor P. The number of models for the p-way treatment structure can be generalised as 2^p. We use the same approach to build and interpret the respective partitioned models. We demonstrate the partitioning approach using a p-way fixed, random and mixed partitioned model.

The Fixed Partitioned model $(FF...F)$ model for the factors $A, B, ..., P$ is expressed as

$$y_{FF...F_{ijk...hn}} = \mu_{FF...F} + \varphi_{Ai} + \varphi_{Bj} + ... + \varphi_{Ph} + \pi_1 + ... + \pi_t + \epsilon_{FF...F_{ijk...hn}}, \tag{7.2}$$

where $i = 1, 2, ..., f_A$; $j = 1, 2, ..., f_B$ through $p = 1, 2, ..., f_P$; are the fixed levels of the p main factors, respectively, and $\pi_{F_1}, ..., \pi_{F_t}$ are the fixed interaction effects in the model. We consider the p-factor classification in (7.2) to be a balanced model with r replications per cell, having f_p fixed levels for each p^{th} factor. Model (7.2) is expressed in matrix form as

$$\mathbf{y}_{FF...F} = \mathbf{1}_N \mu_{FF...F} + \mathbf{X}_1 \boldsymbol{\beta}_1 + \mathbf{X}_2 \boldsymbol{\beta}_2 + ... + \mathbf{X}_m \boldsymbol{\beta}_m + \boldsymbol{\epsilon}_{FF...F}, \tag{7.3}$$

where $\mu_{FF...F}$ is the overall mean, N is the total number of observations, $\boldsymbol{\beta}_m$ is a vector of either the main effect or interaction effects, and $\boldsymbol{\epsilon}_{FF...F} \sim N(\mathbf{0}, \sigma^2_{FF...F}\mathbf{I}_N)$. The incidence matrix $\mathbf{X}_m$ in a p-factor classification model is made up of Kronecker products of $(p+1)$ matrices, $\mathbf{Q}_p$, which are either $\mathbf{I}_{n_p}$ $(n_p \times n_p)$ identity matrix or $\mathbf{1}_{n_p}$ (a vector with all n_p components equal to 1), provided the model (7.3) is balanced. That is, $\mathbf{X}_m = \mathbf{Q}_1 \otimes \mathbf{Q}_2 \otimes ... \otimes \mathbf{Q}_p \otimes \mathbf{1}_r$, where $\mathbf{Q}_p = \mathbf{I}_{n_p}$ if the index p corresponds to the p^{th} factor in the model, or $\mathbf{Q}_p = \mathbf{1}_{n_p}$ if not. The last submatrix, $\mathbf{1}_r$ (a vector with all r components equal to 1), represents the replications per cell in the case of a balanced model. For a balanced model, with r replications per cell, the Kronecker products of the intercept, main and interaction effect incidence submatrices are given by:

$$\begin{aligned}
\mathbf{1}_N &= \mathbf{1}_{f_1} \otimes \mathbf{1}_{f_2} \otimes ... \otimes \mathbf{1}_{f_p} \otimes \mathbf{1}_r = \mathbf{J}_N, \\
\mathbf{X}_2 &= \mathbf{I}_{f_1} \otimes \mathbf{1}_{f_2} \otimes ... \otimes \mathbf{1}_{f_p} \otimes \mathbf{1}_r = \mathbf{D}_{f_1}(\mathbf{J}_r), \\
&\vdots \\
\mathbf{X}_{12} &= \mathbf{I}_{f_1} \otimes \mathbf{I}_{f_2} \otimes \mathbf{1}_{f_3} \otimes ... \otimes \mathbf{1}_{f_p} \otimes \mathbf{1}_r = \mathbf{D}_{f_1 f_2}(\mathbf{J}_r), \\
&\vdots \\
\mathbf{X}_{123...p} &= \mathbf{I}_{f_1} \otimes \mathbf{I}_{f_2} \otimes \mathbf{I}_{f_3} \otimes ... \otimes \mathbf{I}_{f_p} \otimes \mathbf{1}_r = \mathbf{D}_{f_1 f_2 f_3 ... f_p}(\mathbf{J}_r),
\end{aligned}$$

where $\mathbf{D}_{f_1}(\mathbf{J}_r)$ is a diagonal matrix of f_1 column vectors of ones, each of length r (i.e. $\mathbf{J}_r$ is a vector of r ones), and $\mathbf{X}_{123...p}$ is the Kronecker product for the interaction effects of factors $A, B, C, ..., P$. Combining the incidence submatrices for an p-factor linear mixed model with full interaction gives a single incidence matrix $\mathbf{X} = (\mathbf{1}_N \ \mathbf{X}_1 \ \mathbf{X}_2 \ \mathbf{X}_3 \ ... \ \mathbf{X}_{12} \ \mathbf{X}_{13} \ ... \ \mathbf{X}_{123...p})$.

In the case of an unbalanced data set with unequal observations per cell or treatment combination, the intercept, main and interaction effect incidence submatrices of an p-factor unbalanced linear mixed model would be expressed as (Hocking, 1985):

$$\begin{aligned}
\mathbf{1}_N &= \mathbf{J}_N, \\
\mathbf{X}_2 &= \mathbf{D}_{f_1}(\mathbf{J}_{n_i}), \\
&\vdots \\
\mathbf{X}_{12} &= \mathbf{D}_{f_1 f_2}(\mathbf{J}_{n_i n_j}), \\
&\vdots \\
\mathbf{X}_{123...p} &= \mathbf{D}_{f_1 f_2 f_3 ... f_p}(\mathbf{J}_{n_i n_j n_k ... p_h}),
\end{aligned}$$

where $\mathbf{D}_{f_1}(\mathbf{J}_r)$ is a diagonal matrix of f_1 column vectors of ones, each of length n_i, $\mathbf{D}_{f_1 f_2 f_3 ... f_p}(\mathbf{J}_{n_i n_j n_k ... p_h})$ is a diagonal matrix of order $f_1.f_2.f_3....f_p$ column vectors with strings of ones of unequal lengths $n_i, n_j, n_k, ..., p_h$, respectively, with $\mathbf{J}_{n_i n_j n_k ....p_h}$ being a vector of $n_i.n_j.n_k....p_h$ ones.

Similarly, the p-way random-effects model is expressed as

$$y_{RR...R_{ijk...hn}} = \mu_{RR...R} + \varphi_{R_Ai} + \varphi_{R_Bj} + ... + \varphi_{R_Ph} + \pi_{R1} + ... + \pi_{Rt} + \epsilon_{RR...R_{ijk...hn}}, \tag{7.4}$$

where $i = f_A + 1, f_A + 2, ..., a$; $j = f_B + 1, f_B + 2, ..., b$ through $j = f_P + 1, f_P + 2, ..., p$; are the random levels of the p main factors respectively, and $\varphi_{R1}, ..., \varphi_{Rt}$ are the random interaction effects in the model. Depending on specified fixed and random levels of each factor, the partitioned p-way mixed model with one random factor and $(p-1)$ fixed factors $y_{FF...FR_{ijk...hn}}$ is similarly expressed as

$$\begin{aligned} y_{FF...FR_{ijk...hn}} =&\mu_{FF...FR} + \varphi_{F_Ai} + \varphi_{F_Bj} + ... + \varphi_{F_{(P-1)}m}+ \\ &\varphi_{R_Ph} + \pi_{F1} + ... + \pi_{Ft} + \pi_{R1} + ... + \pi_{Rs} + \epsilon_{FF...R_{ijk...hn}}, \end{aligned} \tag{7.5}$$

where $\varphi_{F_Ai}, ..., \varphi_{F_{(P-1)}m}$ are effects of fixed factors and φ_{R_Ph} is the effect of a random factor. Similarly, the interaction effects $\pi_{F1}, ..., \pi_{Ft}$ are the fixed interaction effects while $\pi_{R1}, ..., \pi_{Rs}$ are the random interaction effects in the model, with $\varphi_{R_Ph} \sim N(0, \sigma^2_{\varphi_{R_Ph}})$, $\pi_{R.} \sim N(0, \sigma^2_{\pi_{R.}})$, and $\epsilon_{FF...R_{ijk...hn}} \sim N(0, \sigma^2_{\epsilon_{FF...R}})$.

Thus, with the appropriate definition of diagonal submatrices for balanced or unbalanced linear mixed model, the incidence matrix $\mathbf{X}$ for a p-factor linear mixed model is generally expressed as $\mathbf{X} = (\mathbf{1}_N\ \mathbf{X}_1\ \mathbf{X}_2\ \mathbf{X}_3\ ...\ \mathbf{X}_{12}\ \mathbf{X}_{13}\ ...\ \mathbf{X}_{123...p})$, where matrices with multi-digit subscripts are incidence matrices for the interaction effects.

7.2 Covariance Structure of a P-factor Linear Mixed Model

Following the three-way structure and the guideline proposed by Moser and Sawyer (1998), we construct a covariance matrix for a p-factor balanced, infinite linear mixed model using Kronecker products. We use model $y_{FF...FR_{ijk...hn}}$ as defined in (7.5) to derive the covariance structure. Consider a p-factor experiment modelled by (7.5), with f_A fixed levels of factor A, f_B fixed levels of factor B, through $f_{(P-1)}$ fixed levels of the $(P-1)^{th}$ factor, r_P random levels of the P^{th} factor, and r replicates in each treatment combination of the p factors. Expressing (7.5) as linear mixed model in matrix form gives

$$\mathbf{y}_{FF...FR} = \mathbf{X}_{FF...FR}\boldsymbol{\beta} + \mathbf{Z}_{FF...FR}\mathbf{u} + \boldsymbol{\epsilon}_{FF...R}, \tag{7.6}$$

where $\mathbf{y}_{FF...FR} : N \times 1$ is a vector of response observations with mean vector $\mathbf{X}_{FF...FR}\hat{\boldsymbol{\beta}}$; known fixed-effects incidence matrix $\mathbf{X}_{FF...FR} : N \times (f_A.f_B....f_{(P-1)})$; known random-effects incidence matrix $\mathbf{Z}_{FF...FR} : N \times r_P$; $\boldsymbol{\beta} : (f_A.f_B....f_{(P-1)}) \times 1$ and $\mathbf{u} : r_P \times 1$ are unknown vectors fixed and random effects, respectively; $\boldsymbol{\epsilon}_{FF...R}$ is a vector of random errors. The covariance structure of (7.6) is $\mathbf{V} = \mathbf{Z}_{FF...FR}\mathbf{G}\mathbf{Z}'_{FF...FR} + \mathbf{R}$, where $\mathbf{G} = Cov(\mathbf{u})$ and $\mathbf{R} = Cov(\boldsymbol{\epsilon}_{FF...R})$. Defining $\mathbf{J}_r$ and $\mathbf{J}_N$ as vectors of r ones and N ones, respectively, the incidence matrix becomes $\mathbf{X} = (\mathbf{1}_N \;\; \mathbf{X}_A \;\; \mathbf{X}_B \;\; \mathbf{X}_C \;\; ... \;\; \mathbf{X}_{(N-1)} \;\; \mathbf{X}_{AB} \;\; ... \;\; \mathbf{X}_{ABC...(P-1)})$, where:

$$\begin{aligned}
\mathbf{1}_N &= \mathbf{1}_{f_A} \otimes \mathbf{1}_{f_B} \otimes ... \otimes \mathbf{1}_{f_{(P-1)}} \otimes \mathbf{1}_r = \mathbf{J}_N,\\
\mathbf{X}_2 &= \mathbf{I}_{f_A} \otimes \mathbf{1}_{f_B} \otimes ... \otimes \mathbf{1}_{f_{(P-1)}} \otimes \mathbf{1}_r = \mathbf{D}_{f_A}(\mathbf{J}_r),\\
&\vdots\\
\mathbf{X}_{AB} &= \mathbf{I}_{f_A} \otimes \mathbf{I}_{f_B} \otimes \mathbf{1}_{f_C} \otimes ... \otimes \mathbf{1}_{f_{(P-1)}} \otimes \mathbf{1}_r = \mathbf{D}_{f_A f_B}(\mathbf{J}_r),\\
&\vdots\\
\mathbf{X}_{ABC...(P-1)} &= \mathbf{I}_{f_A} \otimes \mathbf{I}_{f_B} \otimes \mathbf{I}_{f_C} \otimes ... \otimes \mathbf{I}_{f_{(P-1)}} \otimes \mathbf{1}_r = \mathbf{D}_{f_A f_B f_C \cdots f_{(P-1)}}(\mathbf{J}_r),
\end{aligned}$$

and the vector of fixed effects $\boldsymbol{\beta} = [\mu, \varphi_{F_A i}, \varphi_{F_B j}, ..., \varphi_{F_{(P-1)}m}, \pi_{(f_A f_B)ij}, ..., \pi_{(f_A f_B \cdots f_{(P-1)ij...(p-1)}}]'$ for $i = 1, 2, ..., f_A$; $j = 1, 2, ..., f_B$. The incidence matrix $\mathbf{Z}_{FF...FR} = (\mathbf{Z}_P \;\; \mathbf{Z}_{AP} \;\; \mathbf{Z}_{BP} ... \;\; \mathbf{Z}_{ABC...(P-1)P})$, where

$$\begin{aligned}
\mathbf{Z}_P &= \mathbf{1}_{f_A} \otimes \mathbf{1}_{f_B} \otimes ... \otimes \mathbf{1}_{f_{(P-1)}} \otimes \mathbf{I}_{R_P} \otimes \mathbf{1}_{f_r} = \mathbf{J}_N,\\
\mathbf{Z}_{AP} &= \mathbf{I}_{f_A} \otimes \mathbf{1}_{f_B} \otimes ... \otimes \mathbf{1}_{f_{(P-1)}} \otimes \mathbf{I}_{R_P} \otimes \mathbf{1}_{f_r} = \mathbf{D}_{f_A R_P}(\mathbf{J}_r),\\
&\vdots\\
\mathbf{Z}_{ABP} &= \mathbf{I}_{f_A} \otimes \mathbf{I}_{f_B} \otimes \mathbf{1}_{f_C} \otimes ... \otimes \mathbf{1}_{f_{(P-1)}} \otimes \mathbf{I}_{R_P} \otimes \mathbf{1}_{f_r} = \mathbf{D}_{f_A f_B R_P}(\mathbf{J}_r),\\
&\vdots\\
\mathbf{Z}_{ABC...(P-1)P} &= \mathbf{I}_{f_A} \otimes \mathbf{I}_{f_B} \otimes \mathbf{I}_{f_C} \otimes ... \otimes \mathbf{I}_{f_{(P-1)}} \otimes \mathbf{I}_{R_P} \otimes \mathbf{1}_{f_r} = \mathbf{D}_{f_A f_B f_C \cdots f_{(P-1)} R_P}(\mathbf{J}_r),
\end{aligned}$$

where $\mathbf{I}_{R_P}$ is the identity matrix corresponding to the vector of random effects, $\mathbf{u} = [\varphi_{R_P}, \pi_{(F_A R_P)ih}, ..., \pi_{(F_A F_B ... F_{(P-1)} R_P)ij...mh}]'$.

Let $\mathbf{1}_m\mathbf{1}_m' = \boldsymbol{\mathcal{J}}_M$, and the simple structures of $\mathbf{G}$ and $\mathbf{R}$ in (7.6) are expressed respectively as

$$\mathbf{G} = \begin{bmatrix} \sigma_P^2\mathbf{I}_P & \cdots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \sigma^2_{(ABC...P)}\mathbf{I}_{f_A} \otimes \mathbf{I}_{f_B} ... \otimes \mathbf{I}_{f_{(P-1)}} \otimes \mathbf{I}_{R_P} \end{bmatrix},$$

and

$$\mathbf{R} = \begin{bmatrix} \sigma^2_{R_{(ABC...P)}} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma^2_{R_{(ABC...P)}} \end{bmatrix}.$$

Using the definitions of $\mathbf{Z}_{FF...FR}$, $\mathbf{G}$ and $\mathbf{R}$ in the covariance matrix $\mathbf{V}$, we obtain

$$\begin{aligned}
\mathbf{V} &= \mathbf{Z}_{FF...FR}G\mathbf{Z}'_{FF...FR} + \mathbf{R} \\
&= (\mathbf{1}_{f_A} \otimes \mathbf{1}_{f_B} \otimes ... \otimes \mathbf{1}_{f_{(P-1)}} \otimes \mathbf{I}_{R_P} \otimes \mathbf{1}_r)(\sigma_P^2\mathbf{I}_P)(\mathbf{1}_{f_A} \otimes \mathbf{1}_{f_B} \otimes ... \otimes \mathbf{1}_{f_{(P-1)}} \otimes \mathbf{I}_{R_P} \otimes \mathbf{1}_r)' \\
&+ (\mathbf{I}_{f_A} \otimes \mathbf{1}_{f_B} \otimes ... \otimes \mathbf{1}_{f_{(P-1)}} \otimes \mathbf{I}_{R_P} \otimes \mathbf{1}_r)(\sigma_{AP}^2\mathbf{I}_{f_A} \otimes \mathbf{I}_{R_P})(\mathbf{I}_{f_A} \otimes \mathbf{1}_{f_B} \otimes ... \otimes \mathbf{1}_{f_{(P-1)}} \otimes \mathbf{I}_{R_P} \otimes \mathbf{1}_r)' \\
&+ (\mathbf{1}_{f_A} \otimes \mathbf{I}_{f_B} \otimes ... \otimes \mathbf{1}_{f_{(P-1)}} \otimes \mathbf{I}_{R_P} \otimes \mathbf{1}_r)(\sigma_{BP}^2\mathbf{I}_{f_B} \otimes \mathbf{I}_{R_P})(\mathbf{1}_{f_A} \otimes \mathbf{I}_{f_B} \otimes ... \otimes \mathbf{1}_{f_{(P-1)}} \otimes \mathbf{I}_{R_P} \otimes \mathbf{1}_r)' \\
&+ ... + (\mathbf{I}_{f_A} \otimes \mathbf{I}_{f_B} \otimes ... \otimes \mathbf{I}_{f_{(P-1)}} \otimes \mathbf{I}_{R_P} \otimes \mathbf{1}_r)(\sigma^2_{(ABC...P)}\mathbf{I}_{f_A} \otimes \mathbf{I}_{f_B} ... \otimes \mathbf{I}_{f_{(P-1)}} \otimes \mathbf{I}_{R_P}) \\
&(\mathbf{I}_{f_A} \otimes \mathbf{I}_{f_B} \otimes ... \otimes \mathbf{I}_{f_{(P-1)}} \otimes \mathbf{I}_{R_P} \otimes \mathbf{1}_r)' \\
&+ (\sigma^2_{R_{(ABC...P)}})(\mathbf{I}_{f_A} \otimes \mathbf{I}_{f_B} \otimes ... \otimes \mathbf{I}_{f_{(P-1)}} \otimes \mathbf{I}_{R_P} \otimes \mathbf{1}_r) \\
&= \sigma_P^2\boldsymbol{\mathcal{J}}_{f_A} \otimes \boldsymbol{\mathcal{J}}_{f_B} \otimes ... \otimes \boldsymbol{\mathcal{J}}_{f_{(P-1)}} \otimes \mathbf{I}_{R_P} \otimes \boldsymbol{\mathcal{J}}_r + \sigma_{AP}^2\mathbf{I}_{f_A} \otimes \boldsymbol{\mathcal{J}}_{f_B} \otimes ... \otimes \boldsymbol{\mathcal{J}}_{f_{(P-1)}} \otimes \mathbf{I}_{R_P} \otimes \boldsymbol{\mathcal{J}}_r \\
&+ \sigma_{BP}^2\boldsymbol{\mathcal{J}}_{f_A} \otimes \mathbf{I}_{f_B} \otimes ... \otimes \boldsymbol{\mathcal{J}}_{f_{(P-1)}} \otimes \mathbf{I}_{R_P} \otimes \boldsymbol{\mathcal{J}}_r + ... \\
&+ \sigma^2_{(ABC...P)}\mathbf{I}_{f_A} \otimes \mathbf{I}_{f_B} \otimes ... \otimes \mathbf{I}_{f_{(P-1)}} \otimes \mathbf{I}_{R_P} \otimes \boldsymbol{\mathcal{J}}_r \\
&+ (\sigma^2_{R_{(ABC...P)}})(\mathbf{I}_{f_A} \otimes \mathbf{I}_{f_B} \otimes ... \otimes \mathbf{I}_{f_{(P-1)}} \otimes \mathbf{I}_{R_P} \otimes \mathbf{1}_r) \qquad (7.7)
\end{aligned}$$

An algorithm for the covariance structure, $\mathbf{V} = \mathbf{Z}_{FF...FR}G\mathbf{Z}'_{FF...FR} + \mathbf{R}$, for a p-factor linear mixed model, is similarly summarised as (Moser and Sawyer, 1998):

Step 1: Create rows of random main and interaction effects.

Step 2: Create column headings of factor letters and subscript letters on the variance.

Step 3: If the letter in the column heading is part of the variance subscript letter-combination, write $\mathbf{I}_d$.

Step 4: Otherwise, write $\boldsymbol{\mathcal{J}}_d$ elsewhere.

Table 7.1 summarises the four steps of the algorithm of constructing the covariance matrix, $\mathbf{V} = \mathbf{\Sigma}$, for the model (7.7). We assume that the p^{th} factor is random with p levels , while the rest of the factors are fixed with levels $a, b, c, d, ...(p-1)$, respectively.

Table 7.1: Constructing Covariance Matrices

Factor	A	B	C	$\cdots$	P	R	
Subscript d	a	b	c	$\cdots$	p	r	
σ^2_P	$\boldsymbol{\mathcal{J}}_a\otimes$	$\boldsymbol{\mathcal{J}}_b\otimes$	$\boldsymbol{\mathcal{J}}_c\otimes$	$\cdots$	$\mathbf{I}_p\otimes$	$\boldsymbol{\mathcal{J}}_r\otimes$	$+$
σ^2_{AP}	$\mathbf{I}_a\otimes$	$\boldsymbol{\mathcal{J}}_b\otimes$	$\boldsymbol{\mathcal{J}}_c\otimes$	$\cdots$	$\mathbf{I}_p\otimes$	$\boldsymbol{\mathcal{J}}_r\otimes$	$+$
σ^2_P	$\boldsymbol{\mathcal{J}}_a\otimes$	$\mathbf{I}_b\otimes$	$\boldsymbol{\mathcal{J}}_c\otimes$	$\cdots$	$\mathbf{I}_p\otimes$	$\boldsymbol{\mathcal{J}}_r\otimes$	$+$
$\vdots$				$\vdots$			$\vdots$
$\sigma^2_{(ABC...P)}$	$\mathbf{I}_a\otimes$	$\mathbf{I}_b\otimes$	$\mathbf{I}_c\otimes$	$\cdots$	$\mathbf{I}_p\otimes$	$\boldsymbol{\mathcal{J}}_r\otimes$	$+$
$\sigma^2_{R(ABC...P)}$	$\mathbf{I}_a\otimes$	$\mathbf{I}_b\otimes$	$\mathbf{I}_c\otimes$	$\cdots$	$\mathbf{I}_p\otimes$	$\mathbf{I}_r\otimes$	$+$

The sum of row elements in Table 7.1 gives the same covariance matrix result as in (7.7). The algorithm for covariance matrix is extended to build an algorithm for determining sums of squares that researchers would use when conducting the analysis of variance tests. Let $\mathbf{Y}'\mathbf{M}_x\mathbf{Y}$ be the sum of squares associated with the overall mean and factors $A, B, ..., P$ in the p-factor linear mixed model with full interactions. Analogous to the covariance matrix algorithm (Moser and Sawyer, 1998), we derive the algorithm for determining matrices ($\mathbf{M}_x$) associated with sums of squares ($\mathbf{Y}'\mathbf{M}_x\mathbf{Y}$) in a p-factor linear mixed model as follows:

Step S1: Create the first row heading for the letters of the overall mean, factors and interactions in the model, and the second row heading of the associated matrices ($\mathbf{M}_x$).

Step S2: Create two column headings, one for the factor letters and the second for the number of levels (d) of the factor.

Step S3a: If the first-row heading letter does not match the column heading letter, write $d^{-1}J_d$ in the Kronecker product.

Step S3b: If the first-row heading letter of a non-nested factor matches the column heading letter, write $I_d - d^{-1}J_d$ in the Kronecker product.

Step S4: Place I_d elsewhere.

Table 7.2 summarises the algorithm steps **S1-S4** for constructing matrices associated with the sum of squares $\mathbf{Y}'\mathbf{M}_x\mathbf{Y}$ from a p-factor linear mixed model.

Table 7.2: Constructing Matrices Associated with the Sum of Squares

	Factor	A	B	$\cdots$	P	R
	Level d	a	b	$\cdots$	p	r
μ	$\mathbf{M}_\mu =$	$a^{-1}\boldsymbol{\mathcal{J}}_a\otimes$	$b^{-1}\boldsymbol{\mathcal{J}}_b\otimes$	$\cdots$	$n^{-1}\boldsymbol{\mathcal{J}}_p\otimes$	$r^{-1}\boldsymbol{\mathcal{J}}_r$
A	$\mathbf{M}_A =$	$(\mathbf{I}_a - a^{-1}\boldsymbol{\mathcal{J}}_a)\otimes$	$b^{-1}\boldsymbol{\mathcal{J}}_b\otimes$	$\cdots$	$p^{-1}\boldsymbol{\mathcal{J}}_p\otimes$	$r^{-1}\boldsymbol{\mathcal{J}}_r$
B	$\mathbf{M}_B =$	$a^{-1}\boldsymbol{\mathcal{J}}_a\otimes$	$(\mathbf{I}_b - b^{-1}\boldsymbol{\mathcal{J}}_b)\otimes$	$\cdots$	$p^{-1}\boldsymbol{\mathcal{J}}_p\otimes$	$r^{-1}\boldsymbol{\mathcal{J}}_r$
$\vdots$				$\vdots$		
P	$\mathbf{M}_P =$	$a^{-1}\boldsymbol{\mathcal{J}}_a\otimes$	$b^{-1}\boldsymbol{\mathcal{J}}_b\otimes$	$\cdots$	$(\mathbf{I}_p - p^{-1}\boldsymbol{\mathcal{J}}_p)\otimes$	$r^{-1}\boldsymbol{\mathcal{J}}_r$
AB	$\mathbf{M}_{AB} =$	$(\mathbf{I}_a - a^{-1}\boldsymbol{\mathcal{J}}_a)\otimes$	$(\mathbf{I}_b - b^{-1}\boldsymbol{\mathcal{J}}_b)\otimes$	$\cdots$	$p^{-1}\boldsymbol{\mathcal{J}}_p\otimes$	$r^{-1}\boldsymbol{\mathcal{J}}_r$
$AB...P$	$\mathbf{M}_{AB...P} =$	$(\mathbf{I}_a - a^{-1}\boldsymbol{\mathcal{J}}_a)\otimes$	$(\mathbf{I}_b - b^{-1}\boldsymbol{\mathcal{J}}_b)\otimes$	$\cdots$	$(\mathbf{I}_p - p^{-1}\boldsymbol{\mathcal{J}}_p)\otimes$	$r^{-1}\boldsymbol{\mathcal{J}}_r$
ϵ	$\mathbf{M}_\epsilon =$	$\mathbf{I}_a\otimes$	$\mathbf{I}_b\otimes$	$\cdots$	$\mathbf{I}_p\otimes$	$(\mathbf{I}_r - r^{-1}\boldsymbol{\mathcal{J}}_r)$

To construct the sum of squares $\mathbf{Y}'\mathbf{M}_x\mathbf{Y}$ of the corresponding factor of factor interactions, read across Table 7.2. For example, the sum of squares matrices for the main factor A and the random error are given by, respectively:

$\mathbf{Y}'\mathbf{M}_x\mathbf{Y} = (\mathbf{I}_a - a^{-1}\boldsymbol{\mathcal{J}}_a) \otimes b^{-1}\boldsymbol{\mathcal{J}}_b \otimes \cdots \otimes p^{-1}\boldsymbol{\mathcal{J}}_p \otimes r^{-1}\boldsymbol{\mathcal{J}}_r$, and

$\mathbf{M}_\epsilon = \mathbf{I}_a \otimes \mathbf{I}_b \otimes \cdots \otimes \mathbf{I}_p \otimes (\mathbf{I}_r - r^{-1}\boldsymbol{\mathcal{J}}_r)$.

7.3 Generalised Inverse

The estimation of parameters for the linear mixed model equation,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{7.8}$$

poses a lot of challenges, especially when the model involved is over-parameterised and has no unique solution for the parameter estimates (Saeed *et al.*, 2014). The popularly used least squares method tries to minimise the sum of squares of the residuals, $(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})$, which leads to the corresponding normal equation. The theory of linear models involves a great deal of finding solutions to a set of equations called the least squares equations or the normal equations given by

$$\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y},$$

where $\hat{\boldsymbol{\beta}}$ is the least squares estimator of $\boldsymbol{\beta}$ to satisfy the equations. When the square matrix $\mathbf{X}'\mathbf{X}$ is of full rank (and hence non-singular), the inverse $(\mathbf{X}'\mathbf{X})^{-1}$ exists, and the least squares estimator, $\hat{\boldsymbol{\beta}}$, is unique. The solution to the normal equations is therefore given by

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{Y})$$

Computing the inverse of $\mathbf{X}'\mathbf{X}$ is generally not an easy task (Milliken and Johnson, 2009), especially when large matrices are involved. However, with the advancement in technology and statistical software, large matrices can be inverted without much effort. For example, the Comprehensive R Archive Network (CRAN), a free software environment for statistical computing and graphics, provides the *inv*() function in *matlib* package for computing inverses of square matrices. Alternatively, by recognising a certain pattern or structure in the matrix $\mathbf{X}'\mathbf{X}$, one can easily exploit the pattern to compute its inverse (Graybill, 1983; Milliken and Johnson, 2009). One of the most useful theorems on patterned matrices popularly used by statisticians is the inverse of a diagonal matrix D, where entries $d_{ii} \neq 0$, which is $D^{-1} = [d_{ii}^{-1}]$. Among the patterned matrices that are frequently encountered in experimental design is a matrix with the following structure (Greenberg and Sarhan, 1959; Graybill, 1983):

$$\mathbf{C} = \begin{pmatrix} \alpha_1 & \alpha_2\mathbf{1}' & \alpha_3\mathbf{1}' & \cdots & \alpha_t\mathbf{1}' \\ \alpha_2\mathbf{1} & \beta_2\mathbf{I}+\gamma_2\boldsymbol{\mathcal{J}} & \beta_3\mathbf{I}+\gamma_3\boldsymbol{\mathcal{J}} & \cdots & \beta_t\mathbf{I}+\gamma_t\boldsymbol{\mathcal{J}} \\ \alpha_3\mathbf{1} & \beta_3\mathbf{I}+\gamma_3\boldsymbol{\mathcal{J}} & \delta_3\mathbf{I}+\theta_3\boldsymbol{\mathcal{J}} & \cdots & \delta_t\mathbf{I}+\theta_t\boldsymbol{\mathcal{J}} \\ \vdots & \vdots & \vdots & \ddots & \\ \alpha_t\mathbf{1} & \beta_t\mathbf{I}+\gamma_t\boldsymbol{\mathcal{J}} & \delta_t\mathbf{I}+\theta_t\boldsymbol{\mathcal{J}} & \cdots & \epsilon_t\mathbf{I}+\omega_t\boldsymbol{\mathcal{J}} \end{pmatrix}.$$

where $\alpha_i, \beta_i, ..., \omega_i$ are scalars or constants, $\mathbf{I}$ is a $k \times k$ identity matrix, $\boldsymbol{\mathcal{J}}$ is a $k \times k$ matrix whose elements are all unity, and $\mathbf{1}$ is a $k \times 1$ column vector whose elements are all unity. The inverse matrix $\mathbf{C}^{-1}$, whose structure is the same as $\mathbf{C}$ with constants $\alpha_i^*, \beta_i^*, ..., \omega_i^*$, is found by solving for the constants in the equation $\mathbf{C}\mathbf{C}^{-1} = \mathbf{I}^*$, where

$$\mathbf{I}^* = \begin{pmatrix} 1 & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{I} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{I} & \cdots & \mathbf{0} \\ \vdots & & & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{I} \end{pmatrix}.$$

For more theorems on computing inverses and determinants of diagonal matrices, diagonal matrix of type 2, variance-covariance matrices, and other special patterned matrices, see Graybill (1983).

The maximum number of linearly independent rows or columns of the design matrix $\mathbf{X} : m \times n$, with $m \geq n$, gives the rank of $\mathbf{X}$, and hence the rank of $\mathbf{X'X}$ in the normal equations. Ideally, $rank(\mathbf{X}) \leq n$. If $rank(\mathbf{X}) = n$, we say $\mathbf{X}$ is full-rank, and the normal equations have a unique solution for $\hat{\boldsymbol{\beta}}$. However, if $rank(\mathbf{X}) < n$, then $\mathbf{X}$ is rank-deficient, $\mathbf{X'X}$ is singular, and the normal equations have infinitely many solutions. As is common with various experimental data in applied statistics and other fields, the matrix $\mathbf{X'X}$ may not necessarily be full rank. When the matrix $\mathbf{X}$ is rank-deficient, $\mathbf{X'X}^{-1}$ does not exist. Thus the estimation procedure uses a generalised inverse or pseudo-inverse $\mathbf{X'X}^{-}$ in the normal equation, which satisfies the estimator, $\hat{\boldsymbol{\beta}}$, is unique. The solution to the normal equations is therefore given by

$$\mathbf{X'X}(\mathbf{X'X})^{-}\mathbf{X'X} = \mathbf{X'X}.$$

Depending on the type of restrictions used on the parameters, various possible solutions exist to the same normal equations. The solution to the normal equation solution is therefore given by

$$\boldsymbol{\beta}^0 = (\mathbf{X'X})^{-}\mathbf{X'Y} = \mathbf{GX'Y},$$

which is one of the infinitely many solution vectors corresponding to the generalised inverse used. The matrix $\mathbf{X}$ is rank-deficient; hence $\mathbf{X'X}$ is not invertible, and the normal equations have no unique solution. Therefore, it cannot be over-emphasised that $\boldsymbol{\beta}^0$ is always referred to as a solution amongst infinitely many, and not as an estimator (Searle and Gruber, 2017).

7.3.1 Over-parameterised Model

The singular model scenario usually results from the fact that the model has more parameters than can be uniquely estimated from the data collected (Milliken and Johnson, 2009). Hence the effects model of this nature is known as an over-specified or over-parameterised model. There are several approaches used to solve the corresponding normal equations of over-specified effects models, which yield infinitely many least squares solutions. One of the commonly used approaches is the use of generalised inverses (Graybill, 1983). As defined in (Garybill, 1983:106), for each $m \times n$ matrix $\mathbf{X}$, there is a matrix $\mathbf{X}^{-}$, called the generalised inverse (or simply, g-inverse) of $\mathbf{X}$, satisfying the condition:

- $\mathbf{XX}^{-}$ is symmetric,

- $\mathbf{X}^{-}\mathbf{X}$ is symmetric,
- $\mathbf{X}\mathbf{X}^{-}\mathbf{X} = \mathbf{X}$,
- $\mathbf{X}^{-}\mathbf{X}\mathbf{X}^{-} = \mathbf{X}^{-}$.

A useful algorithm for computing a generalised inverse based on knowing or first finding the rank of the matrix exists (Searle and Gruber, 2017). The algorithm for obtaining a reflexive generalised inverse of $\mathbf{X}$ works as follows:

- In a matrix $\mathbf{X}$ of rank r, find any non-singular minor of order r. Call it $\mathbf{M}$.
- Invert $\mathbf{M}$ and transpose the inverse to obtain $(\mathbf{M}^{-1})'$.
- In $\mathbf{X}$, replace each element of $\mathbf{M}$ with the corresponding element of $(\mathbf{M}^{-1})'$.
- Replace all other elements of $\mathbf{X}$ with zero.
- Transpose the resulting matrix.

Different generalised inverses of $\mathbf{X}$ are obtained from different choices of the minor of rank r. Other types of generalised inverses use different algorithms. An example is the Moore-Penrose inverse, which is obtained using the singular value decomposition.

Several other theorems about the g-inverses, including some patterned matrices, have been suggested in literature. One useful theorem that applies to any $m \times n$ matrix $\mathbf{X}$ of either full row rank or full column rank has been documented (Graybill, 1983):

- If $\mathbf{X}$ is an $m \times n$ matrix of rank m, then $\mathbf{X}^{-} = \mathbf{X}'(\mathbf{X}\mathbf{X}')^{-1}$ and $\mathbf{X}\mathbf{X}^{-} = \mathbf{I}$.
- If $\mathbf{X}$ is an $m \times n$ matrix of rank n, then $\mathbf{X}^{-} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ and $\mathbf{X}^{-}\mathbf{X} = \mathbf{I}$.
- If $\mathbf{X}$ is an $m \times n$ matrix of whatever rank, then $\mathbf{X}^{-} = \mathbf{X}'(\mathbf{X}\mathbf{X}')^{-} = (\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'$.

7.3.2 Re-parameterised Model

Other approaches to solving the corresponding normal equations involve placing restrictions on the parameters in the model, which eventually produce generalised inverse solutions as well. Examples include (Milliken and Johnson, 2009; Saeed *et al.*, 2014):

- the sum-to-zero restrictions: which require the sums of certain parameters to be set to zero; solve for some of the parameters in terms of others with the restrictions being taken into account, and then substitute the expressions back into the model.
- the set-to-zero restrictions: which require that one of the parameters (first, second, last or any other parameter) in each treatment level be set to zero and reduce the design matrix $\mathbf{X}$ before solving the normal equations.

The result of incorporating these restrictions is a re-parameterised model, where the set-to-zero restrictions, deleting the columns corresponding parameters set to zero, produce a much simpler model than the sum-to-zero restrictions. The whole concept of over-parameterised model and non-unique least squares solutions revolves around the issue of estimability.

7.4 Estimable Functions

Different approaches to imposing restrictions usually produce parameters or functions of the parameters that have different estimates for the re-parameterised models. If it happens that the parameters or functions of parameters of these approaches have the same estimates, then we have what is called estimable functions of the parameters. Milliken and Johnson (2009) define a parameter β_i or function of the parameters $f(\boldsymbol{\beta})$ as estimable if and only if the estimate of the parameter or function of parameters is invariant with respect to the choice of a least squares solution. The definition suggests that the value of the estimate of the parameter or function of parameters remains the same regardless of which solution to the normal equations is used. The implication is that an estimate of an estimable function of the parameters produces the same value, leading to the same decisions about estimable functions of the parameters, even if different least squares solutions were used (Milliken and Johnson, 2009). Due to this invariance property, estimable functions become the primary focus of interest in estimating the parameters of a linear model (Searle and Gruber, 2017).

7.4.1 Estimable Functions of Full-rank Case

Consider the linear model (7.8) expressed in matrix form, where $\mathbf{X}$ has full column rank. The linear estimable functions $\mathbf{K}'\boldsymbol{\beta}$ are defined as linear combinations of the parameter vector $\boldsymbol{\beta}$ where $\mathbf{K}'$ is a $q \times p$ matrix of full-row rank (i.e., all of the rows of $\mathbf{K}'$ are linearly independent).

Each of the rows of $\mathbf{K}'\boldsymbol{\beta}$ is in the form $\mathbf{k}'\boldsymbol{\beta}$, where $\mathbf{k}'$ is a $1 \times p$ row vector of constants. Ideally, the linear function $\mathbf{k}'\boldsymbol{\beta}$ is defined as estimable if and only if it is identically equal to some linear function of the expected value of the vector of observation (Searle and Gruber, 2017). Technically, $\mathbf{k}'\boldsymbol{\beta}$ is estimable if some vector $\mathbf{t}'$, whose value is not as important as its existence, is such that $\mathbf{t}'E(\mathbf{y}) = \mathbf{k}'\boldsymbol{\beta}$. This implies that each row x_i of the design matrix $\mathbf{X}$ is a vector of constants necessary for any linear combination $x_i'\boldsymbol{\beta}$ to be an estimable function.

7.4.2 Estimable functions of Non-full-rank Case

When matrix $\mathbf{X}$ is rank-deficient, the inverse $(\mathbf{X}'\mathbf{X})^{-1}$ does not exist, and the normal equations have no unique solution. In that case, the generalised inverse $(\mathbf{X}'\mathbf{X})^{-}$ is used in the estimation procedure of the normal equation, the same way $(\mathbf{X}'\mathbf{X})^{-1}$ would be used had it existed.

7.4.3 Some Properties of Estimable Functions

Following the definition of an estimable function, we derive the properties (Searle and Gruber, 2017):

- the expected value of any observation is estimable,
- linear combinations of estimable functions are estimable,
- when $\mathbf{k}'\boldsymbol{\beta}$ is estimable, $\mathbf{k}'\boldsymbol{\beta}^0$ is invariant to whatever solution of the normal equation is used for $\boldsymbol{\beta}^0$,
- the least squares estimators $\hat{\boldsymbol{\beta}}$ are the best linear unbiased estimator for a full-rank model, while the estimable linear combinations of solutions to the normal equation are the best linear unbiased estimators for the less than full-rank model,
- for some vector $\mathbf{t}'$, we have $\mathbf{k}'\boldsymbol{\beta} = \mathbf{t}'E(\mathbf{y}) = \mathbf{t}'E(\mathbf{X}\boldsymbol{\beta}) = \mathbf{t}'\mathbf{X}\boldsymbol{\beta}$, which reduces to $\mathbf{k}' = \mathbf{t}'\mathbf{X}$ since estimability does not depend on the value of $\boldsymbol{\beta}$.

Since different least squares solutions provide different estimates, these estimators should never be considered valid estimates of the treatment effects since they are all biased (Saeed *et al.*, 2014). We use estimable functions to overcome this challenge since they have parameter estimates that do not depend on the choice of least squares solution used. The implication is that, only functions of the parameters that are estimable are best linear unbiased estimates (BLUE) which should be considered when making inferences from linear models because they have the smallest variances (Searle and Gruber, 2017). Otherwise, no estimate should be provided for a parameter or function of the parameters that is not estimable (Milliken and Johnson, 2009).

7.4.4 Distributional Properties of Estimable Functions

The distribution properties of a generalised inverse are crucial for deriving of test statistics used for testing hypotheses involving estimable functions. Unlike the $\hat{\boldsymbol{\beta}}$ estimator of a full-rank model, which is distributed as $N(\boldsymbol{\beta}, (\mathbf{X}'\mathbf{X})^{-1}\sigma_\epsilon^2)$, the definition of $\boldsymbol{\beta}^0$ solution shows that it is a function of a matrix of observations, $\mathbf{y}$, with the expected value given by

$$E(\boldsymbol{\beta}^0) = \mathbf{G}\mathbf{X}'E(\mathbf{Y}) = \mathbf{G}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \mathbf{H}\boldsymbol{\beta},$$

where $\mathbf{H} = \mathbf{G}\mathbf{X}'\mathbf{X}$, which shows that $\boldsymbol{\beta}^0$ is not an estimator of $\boldsymbol{\beta}$ but of $\mathbf{H}\boldsymbol{\beta}$. Similarly, the variance of $\boldsymbol{\beta}^0$ is given by

$$Var(\boldsymbol{\beta}^0) = \mathbf{G}\mathbf{X}'Var(\mathbf{Y})\mathbf{X}\mathbf{G}' = \mathbf{G}\mathbf{X}'\mathbf{X}\mathbf{G}'\sigma_\epsilon^2,$$

Hence, $\boldsymbol{\beta}^0 \sim N(\mathbf{G}\mathbf{X}'\mathbf{X}\boldsymbol{\beta}, \mathbf{G}\mathbf{X}'\mathbf{X}\mathbf{G}'\sigma_\epsilon^2)$. Clearly, the variance of $\boldsymbol{\beta}^0$ depends on the choice of g-inverse used. For example, the $Var(\boldsymbol{\beta}^0) = \mathbf{G}\sigma_\epsilon^2$ when the reflexive inverse G is chosen, or $Var(\boldsymbol{\beta}^0) = \mathbf{G}_{mp}\sigma_\epsilon^2$ when a Moore-Penrose inverse G_{mp} is the used to solve the normal equation.

7.4.5 Estimability

We can determine whether $\mathbf{k}'\boldsymbol{\beta}$ is estimable or not by checking if it satisfies the equation $\mathbf{k}'\mathbf{H} = \mathbf{k}'$ where $\mathbf{H} = \mathbf{G}\mathbf{X}'\mathbf{X}$ as before. The proof of this condition follows from the fact that, if $\mathbf{k}'\boldsymbol{\beta}$ is estimable, there exists some vector $\mathbf{t}'$ such that $\mathbf{k}' = \mathbf{t}'\mathbf{X}$. Hence,

$$\mathbf{k}'\mathbf{H} = \mathbf{t}'\mathbf{X}\mathbf{H} = \mathbf{t}'\mathbf{X}\mathbf{G}\mathbf{X}'\mathbf{X} = \mathbf{t}'\mathbf{X} = \mathbf{k}'.$$

Conversely, if $\mathbf{k}'\mathbf{H} = \mathbf{k}'$, then

$$\mathbf{k}' = \mathbf{k}'\mathbf{H} = \mathbf{k}'\mathbf{G}\mathbf{X}'\mathbf{X} = \mathbf{k}'\mathbf{t}'\mathbf{X} = \mathbf{t}'\mathbf{X},$$

for $\mathbf{t}' = \mathbf{G}\mathbf{X}'$ so that $\mathbf{k}'\boldsymbol{\beta}$ is estimable. Thus, we conclude that if $\mathbf{k}'\boldsymbol{\beta}$ does not satisfy the equation $\mathbf{k}'\mathbf{H} = \mathbf{k}'$, then it is not estimable.

The theory of linear models is centred on the different kinds of hypotheses that might be of interest to researchers in various fields of application. Some of the hypotheses that are of interest in most research applications include:

- Hypothesis about the linear model parameters ($H_0 : \beta_i = b_0$), with a special case when $b_0 = 0$,
- Some linear combination of $\boldsymbol{\beta}$ elements assumed to be equal to a constant value ($H_0 : \mathbf{k}'\boldsymbol{\beta} = c$),
- Reduced model hypothesis, ($H_0 : \beta_q = 0$), where β_q is a subset of β_i's.

All the above hypotheses and other linear combinations are special cases of the general hypothesis $H_0 : \mathbf{K}'\boldsymbol{\beta} = \mathbf{m}$, where $\boldsymbol{\beta} : p \times 1$, is the vector of parameters of the model; $\mathbf{K}' : s \times p$ is any matrix of linear estimable functions of the parameter vector, $\boldsymbol{\beta}$, such that $\mathbf{K}'$ is of full-row rank, that is, $r(\mathbf{K}') = s$; and $\mathbf{m} : s \times 1$ is a vector of specified constants. Each row $\mathbf{k}' : 1 \times p$ of $\mathbf{K}'$ is basically a row vector of constants, and $\mathbf{k}_i'\boldsymbol{\beta}$ is a linear combination of β_i's. This suggests that $\mathbf{K}'\boldsymbol{\beta}$ is a set of estimable functions.

Keeping in mind the invariance property of the estimable functions, i.e., $\mathbf{k}_i'\boldsymbol{\beta}$ yield identical results regardless of the solution vector used; we would be interested in testing the hypotheses $H_0 : \mathbf{k}_i'\boldsymbol{\beta} = \mathbf{m}$ versus $H_1 : \mathbf{k}_i'\boldsymbol{\beta} \neq \mathbf{m}$, where $\mathbf{k}_i$ is an estimable vector, and $\hat{\boldsymbol{\beta}}$ is any least squares solution vector (from set-to-zero or sum-to-zero restrictions). The linear combination $\mathbf{k}_i'\boldsymbol{\beta}$ is distributed as $N[\mathbf{k}_i'\boldsymbol{\beta}, \mathbf{k}_i'(\mathbf{X}'\mathbf{X})^{-}\mathbf{k}_i]$. A set of these estimable functions tested simultaneously will lead to the following hypotheses:

$$H_0 : \mathbf{K}'\boldsymbol{\beta} = \mathbf{m} \quad versus \quad H_1 : \mathbf{K}'\boldsymbol{\beta} \neq \mathbf{m},$$

where K is a matrix of estimable constants (built by combining the estimable vectors in question), and $\hat{\boldsymbol{\beta}}$ is any least squares solution vector. The estimability of $\mathbf{K}'\boldsymbol{\beta}$ is confirmed by any one of the following ways (Searle and Gruber, 2017) otherwise, $\mathbf{K}'\boldsymbol{\beta}$ is not estimable if none of these conditions are met:

- if $\mathbf{K}'$ satisfies $\mathbf{K}' = \mathbf{K}'\mathbf{X}$;
- if a matrix $\mathbf{T}'$ exists that satisfies $\mathbf{K}' = \mathbf{T}'\mathbf{X}$;

- if matrix $\mathbf{C}'$ exists such that $\mathbf{K}' = \mathbf{C}'\mathbf{U}$, where $\mathbf{U}$ is a column orthogonal matrix in the singular value decomposition of $\mathbf{X}'\mathbf{X}$ and/or
- if $\mathbf{K}'$ satisfies either $\mathbf{K}' = \mathbf{U}\mathbf{C}' = \mathbf{K}'$ or $\mathbf{K}'\mathbf{V}\mathbf{V}' = \mathbf{0}$, where $\mathbf{V}$ is the normalised eigenvector of $\mathbf{0}$ for $\mathbf{X}'\mathbf{X}$.

The concept of estimable functions is also fundamental to inference in the linear mixed models of the general form

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\epsilon}, \tag{7.9}$$

where $\mathbf{y}$ is the vector of observations, $\mathbf{X}$ is a matrix of known constants associated with the vector of fixed effects, $\boldsymbol{\beta}$, matrix $\mathbf{Z}$ is of known constants associated with the vector of random effects, $\mathbf{u}$, and $\boldsymbol{\epsilon}$ is a vector of random errors. Traditionally, it is assumed that the random effects and errors of the general linear mixed model have a joint distribution given by

$$\begin{bmatrix} \mathbf{u} \\ \boldsymbol{\epsilon} \end{bmatrix} \sim N\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \mathbf{G} & \mathbf{0} \\ \mathbf{0} & \mathbf{R} \end{bmatrix} \right),$$

whose non-singular components $\mathbf{G}$ and $\mathbf{R}$ are usually estimated by the restricted maximum likelihood (REML) or the method of moments in various mixed model statistical software. There are numerous methods of estimating these variance components, which ultimately lead to the $\boldsymbol{\beta}$ solution (BLUE) and the $\mathbf{u}$ solution (BLUP). Fundamental to linear mixed model inference is the concept of estimable functions, $\mathbf{K}'\boldsymbol{\beta}$, such that the estimability criteria explained above are satisfied, or the predictable functions, $\mathbf{K}'\boldsymbol{\beta}+\mathbf{M}'\mathbf{u}$, such that $\mathbf{K}'\boldsymbol{\beta}$ is estimable (Stroup and Kachman, 1994).

7.4.6 Hypothesis Testing of Estimable Functions

In order to test hypotheses about linear functions of the parameters (or estimable functions) of a linear model, appropriate test statistics need to be developed. The principle of conditional error and the likelihood ratio statistic are the most commonly used approaches used in this regard. Consider the general hypothesis

$$H_0 : \mathbf{K}'\boldsymbol{\beta} = \mathbf{m} \quad versus \quad H_1 : \mathbf{K}'\boldsymbol{\beta} \neq \mathbf{m},$$

where the matrix $\mathbf{K}'\boldsymbol{\beta}$ of estimable functions of $\boldsymbol{\beta}$ and $\mathbf{K}'$ is of full-row rank, i.e., $r(\mathbf{K}') = s$. The appropriate test statistic for H_0 is given by

$$F_{cal} = \frac{\frac{Q}{s}}{\hat{\sigma}^2}, \tag{7.10}$$

where $Q = (\mathbf{K}'\boldsymbol{\beta} - \mathbf{m})'[\mathbf{K}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{K}](\mathbf{K}'\boldsymbol{\beta} - \mathbf{m})$ is the sum of squares due to deviations from the null hypothesis, $s = r(\mathbf{K}')$ is the rank of $\mathbf{K}'$, and $\hat{\sigma}^2$ is the estimate of the population variance based on the least squares solution for $\boldsymbol{\beta}$ which is given by

$$\sigma^2 = \frac{1}{N - r(\mathbf{X})}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = \frac{SSE}{N - r(\mathbf{X})}.$$

Equivalently, when $\mathbf{X}'\mathbf{X}$ is singular, that is, when $\mathbf{X}$ is not of full rank, and $\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$, the test statistic about the estimable functions $\mathbf{K}'\boldsymbol{\beta}$ is developed in terms of a re-parameterised model using the generalised inverse $(\mathbf{X}'\mathbf{X})^{-}$. The test statistic is therefore given by

$$F^*_{cal} = \frac{\frac{Q^*}{s}}{\hat{\sigma}^{2*}}, \tag{7.11}$$

where $Q^* = (\mathbf{K}'\boldsymbol{\beta}^0 - \mathbf{m})'[\mathbf{K}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{K}](\mathbf{K}'\boldsymbol{\beta}^0 - \mathbf{m})$ is the sum of squares due to deviations from the null hypothesis. In either case, we consider the assumption that the error vector $\boldsymbol{\epsilon} = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$ are i.i.d $N(\mathbf{0}, \sigma^2\mathbf{I})$, which implies that the test statistic has an F-distribution with degrees of freedom $s = r(\mathbf{K}')$ and $N - r(\mathbf{X})$, i.e., $F^*_{cal} \sim F_{s,N-r(\mathbf{X})}(\alpha)$. The estimate of the population variance $\hat{\sigma}^2$ is based on the solution $\boldsymbol{\beta}^0$, which is given by

$$\sigma^{2*} = \frac{1}{N - r(\mathbf{X})}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}^0)'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}^0) = \frac{SSE}{N - r(\mathbf{X})}.$$

7.4.7 Testability of Estimable Functions

The theory for testing the general linear hypothesis $H_0 : \mathbf{K}'\boldsymbol{\beta} = \mathbf{m} \quad versus \quad H_1 : \mathbf{K}'\boldsymbol{\beta} \neq \mathbf{m}$ when the design matrix $\mathbf{X}$ is of full rank is constrained on the condition that $\mathbf{K}'$ has full-row rank. Under $H_0 : \mathbf{K}'\boldsymbol{\beta} = \mathbf{m}$ the F- statistic defined in (7.10) provides the required test. The theory can be extended to cases where the design matrix $\mathbf{X}$ is non-full-rank. However, some hypotheses are testable in such cases, while others are not. The condition under which a linear hypothesis $H_0 : \mathbf{K}'\boldsymbol{\beta} = \mathbf{m}$ is testable is that $\mathbf{K}'\boldsymbol{\beta}$ should be made up of estimable functions as defined in the previous sections. Analogous to the full-rank case, the F-test statistic for the linear hypothesis $H_0 : \mathbf{K}'\boldsymbol{\beta} = \mathbf{m}$ for the non-full-rank case would be expected to have a component $(\mathbf{K}'\boldsymbol{\beta}^0 - \mathbf{m})$. Ideally, $\mathbf{K}'\boldsymbol{\beta}^0$ is expected to be invariant to any choice of the general solution $\boldsymbol{\beta}^0$ to the normal equation. As a result, the invariance condition can only be met when $\mathbf{K}'\boldsymbol{\beta}$ consists of estimable functions; otherwise the linear hypothesis $H_0 : \mathbf{K}'\boldsymbol{\beta} = \mathbf{m}$ will not be testable.

In formal terms, a linear hypothesis $H_0 : \mathbf{K}'\boldsymbol{\beta} = \mathbf{m}$ is testable if each linear combination $\mathbf{k}_i'\boldsymbol{\beta}$ $(i = 1, 2, ..., m)$ of $\mathbf{K}'\boldsymbol{\beta}$ is estimable. From the fundamental definition of an estimable function, this implies that $\mathbf{k}_i' = \mathbf{t}_i'\mathbf{X}$ for some $\mathbf{t}_i'$. Hence, we have $\mathbf{K}_i' = \mathbf{T}_i'\mathbf{X}$ for some matrix $\mathbf{T}$ of order $s \times N$, where $s = r(\mathbf{K}')$. In addition, $\mathbf{K}'$ should always be of full-row rank since the hypothesis $H_0 : \mathbf{K}'\boldsymbol{\beta} = \mathbf{m}$ is considered only in terms of its linearly independent components (Searle and Gruber, 2017).

The general implication of estimability of $\mathbf{K}'\boldsymbol{\beta}$ is the existence of the error sum of squares based on the null hypothesis, $Q^* = (\mathbf{K}'\boldsymbol{\beta}^0 - \mathbf{m})'[\mathbf{K}'\mathbf{G}\mathbf{K}]^{-1}(\mathbf{K}'\boldsymbol{\beta}^0 - \mathbf{m})$, where $\mathbf{G} = (\mathbf{X}'\mathbf{X})^-$ is the generalised inverse used, which in particular implies the existence of $[\mathbf{K}'\mathbf{G}\mathbf{K}]^{-1}$. However, the condition for testability is not solely focused on the estimability of $\mathbf{K}'\boldsymbol{\beta}$ because the error sum of squares, Q^*, can be calculated even when $\mathbf{K}'\boldsymbol{\beta}$ is not estimable, provided $[\mathbf{K}'(\mathbf{X}'\mathbf{X})^-\mathbf{K}]^{-1}$ exists. It is therefore important to check if $\mathbf{K}'\boldsymbol{\beta}$ is estimable before proceeding to calculate either Q^* or the test statistic F^*_{cal}. In summary, the concepts of estimability and testability only apply to a non-full-rank model and not a full-rank model. The reason is that, for a full-rank model, all linear functions are testable, and all linear hypotheses are testable (Searle and Gruber, 2017).

7.5 Variance of Error Terms in Linear Models

Most standard analytical models which use the analysis of variance (ANOVA) technique usually require a restrictive assumption that all pairs of effects have homogeneous variance and covariance, which in many circumstances is not always realistic (Hu and Spilke, 2011). We first consider the general linear model, also known as the Gauss-Markov model,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}. \tag{7.12}$$

The error vector is defined as $\hat{\boldsymbol{\epsilon}} = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$. The normal equations corresponding to this equation are derived using least squares under the traditional assumptions that $E(\boldsymbol{\epsilon}) = \mathbf{0}$ (assumption I), and that the elements of the error vector are all equal and uncorrelated (i.e. $Cov(e_i, e_j) = 0$ for all $i \neq j$), then $Var(\boldsymbol{\epsilon}) = \mathbf{V} = \sigma^2\mathbf{I}$ (assumption II). Thus, the normal equations of this model,

$$\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \mathbf{X}'\mathbf{y},$$

have a unique solution when matrix $\mathbf{X}$ is full-rank, that is, $\mathbf{X}'\mathbf{X}$ is non-singular or has infinitely many solutions if matrix $\mathbf{X}$ is rank-deficient and $\mathbf{X}'\mathbf{X}$ is not invertible. The traditional approach in statistical analysis involving either full-rank or non-full-rank linear models considers a special case where the error terms are assumed to have $Var(\boldsymbol{\epsilon}) = \mathbf{V} = \sigma^2\mathbf{I}$. If the errors are assumed to be independently distributed with the first four moments equal to the first four moments of a normal distribution, then $\hat{\sigma}^2$ is the best quadratic unbiased estimate of σ^2. If the errors are also normally distributed, then $\hat{\sigma}^2$ is the best unbiased estimator of σ^2, and the sampling distribution of $(n-r)\frac{\hat{\sigma}^2}{\sigma^2}$ is a central chi-square distribution with $n - r$ degrees of freedom (Milliken and Johnson, 2009).

However, the assumption of equal and uncorrelated error terms may not be true in many cases, such as repeated measures design where experimental units are the same or clustered data scenarios where experimental units are drawn from the same neighbourhood. The presence of unequal variances across these observations in different neighbourhoods or the presence of covariance among the observations of the response variable is the basis for generalising ordinary least squares. The homoscedasticity assumption used on the Gauss-Markov model may not be satisfied in such cases. Instead, it is appropriate to assume that $E(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$ and $Cov(\mathbf{y}) = Cov(\boldsymbol{\epsilon}) = \sigma^2\mathbf{V}$ (Assumption III), where $\mathbf{V}$ is a known symmetric positive definite matrix.

From assumption III, matrix $\mathbf{V}$ is a symmetric positive definite matrix, which implies that there exists an invertible matrix $\mathbf{H}$ such that $\mathbf{V} = \mathbf{H}\mathbf{H}'$. We pre-multiply the OLS Gauss-Markov model by H^{-1} and express the model as

$$\mathbf{y}^* = \mathbf{X}^*\boldsymbol{\beta} + \boldsymbol{\epsilon}^*, \tag{7.13}$$

where $\mathbf{y}^* = \mathbf{H}^{-1}\mathbf{y}$, $\mathbf{X}^* = \mathbf{H}^{-1}\mathbf{X}$, and $\boldsymbol{\epsilon}^* = H^{-1}\boldsymbol{\epsilon}$. The general linear model under assumptions I and III is referred to as Aitken's model (Aitken, 1935) or the generalised least squares (GLS) model where $E(\boldsymbol{\epsilon}) = \mathbf{0}$, and $Var(\boldsymbol{\epsilon}) = \mathbf{V} = \sigma^2\mathbf{I}$, thus satisfying all assumptions of Gauss-Markov model. The corresponding normal equations

$$\mathbf{X}^{*\prime}\mathbf{X}^*\boldsymbol{\beta} = \mathbf{X}^{*\prime}\mathbf{y}^*$$

have a solution (commonly known as Aitken's estimator)

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{GLS} &= (\mathbf{X}^{*\prime}\mathbf{X}\mathbf{X})^g\mathbf{X}^{*\prime}\mathbf{y}^* \\ &= [(\mathbf{H}^{-1}\mathbf{X})'(\mathbf{H}^{-1}\mathbf{X})]^g(\mathbf{H}^{-1}\mathbf{X})'(\mathbf{H}^{-1}\mathbf{y}) \\ &= [\mathbf{X}'(\mathbf{H}^{-1})'(\mathbf{H}^{-1}\mathbf{X})]^g\mathbf{X}'(\mathbf{H}^{-1})'\mathbf{H}^{-1}\mathbf{y} \\ &= [\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}]^g\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}. \end{aligned} \tag{7.14}$$

Hence, $\hat{\boldsymbol{\beta}}_{GLS}$ is a generalised least squared (GLS) estimator of $\boldsymbol{\beta}_{GLS}$. Similarly, the generalised estimator of $\mathbf{y}$ is given by

$$\begin{aligned}\mathbf{y}_{GLS} &= \mathbf{X}\hat{\boldsymbol{\beta}}_{GLS} \\ &= \mathbf{X}[\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}]^g\mathbf{X}'\mathbf{V}^{-1}\mathbf{y} \\ &= \mathbf{L}_{GLS}\mathbf{y},\end{aligned}$$

where $\mathbf{L}_{GLS} = \mathbf{X}[\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}]^g\mathbf{X}'\mathbf{V}^{-1}$. When $\mathbf{X}$ is of full-rank, $[\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}]^{-1}$ exists and thus the estimator, $\hat{\boldsymbol{\beta}}_{GLS}$, is an unbiased estimator,

$$\hat{\boldsymbol{\beta}}_{GLS} = [\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}]^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y},$$

and so is $\hat{\mathbf{y}}_{GLS}$ since

$$E(\mathbf{y}_{GLS}) = \mathbf{X}[\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}]^{-1}\mathbf{X}'\mathbf{V}^{-1}E(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}.$$

A crucial step in developing statistical models that adequately make use of data information is specifying the appropriate covariance structure (Guo and Tang, 2021), which describes the nature of the correlation among data points within a given category. For example, studies involving clustered, repeated measurement or correlated data analysis cannot assume independence in residual errors among all observations. This is particularly essential to ensure accurate parameter estimates, the overall model fit, and standard errors, which tend to be sensitive to the model covariance structure.

7.5.1 Covariance Structures in Linear Models

One of the major strengths of linear models that gives them leverage over other analytical models is that they do not have any restrictions on random effects and residual errors, thereby allowing them to assume various structures of variance-covariance matrices, which mirror the characteristics of random effects and residual errors (Hu and Spilke, 2011). For example, in studies involving longitudinal or non-longitudinal clustered data, various covariance structures exist for various types of assumptions about the associations between responses from the same cluster. The basic variance-covariance matrix of responses, which is generally assumed to be the same for all clusters, is defined by an $n \times n$ symmetric matrix

$$\mathbf{V} = \begin{pmatrix} \sigma_{11}^2 & \sigma_1\sigma_2 & \cdots & \sigma_1\sigma_n \\ \sigma_2\sigma_1 & \sigma_{22}^2 & \cdots & \sigma_2\sigma_n \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_n\sigma_1 & \sigma_n\sigma_2 & \cdots & \sigma_{nn}^2 \end{pmatrix}.$$

where the diagonal elements σ_{ii}^2 are variances, off-diagonal elements $\sigma_i\sigma_j$ are covariances, and n is the number of observations per cluster.

The common patterns of variance-covariance matrix appropriate for different scenarios have been summarised (Barnett *et al.*, 2010) as:

- an independent covariance, which is appropriate when the variance is homogeneous but none of the corresponding effects are correlated ($\sigma_{ii}^2 = \sigma^2$ and $\sigma_i\sigma_j = 0$ for all $i \neq j$);
- an unstructured covariance, which is an appropriate choice when variance and covariance are not homogeneous ($\sigma_{ii}^2 \neq \sigma_{i^*i^*}^2$ and $\sigma_i\sigma_j = \sigma_{i^*}\sigma_{j^*}$ for $i \neq i^*$, $j \neq j^*$);
- an exchangeable covariance, which is appropriate when the responses from the same cluster are equally correlated ($\sigma_{ii}^2 = \sigma^2$ and $\sigma_i\sigma_j = \sigma^2\rho$ for all i and j) or

- an autoregressive covariance, which is an appropriate choice when the correlation between responses decreases with increasing time or distance.

Other forms may emerge as modifications or improvements to the basic four patterns above. An example is when heterogeneous variance (unequal diagonal elements) and homogeneous covariance (equal off-diagonal elements) are assumed. Hu and Spilke (2011) recorded several other forms of variance-covariance structures that allow heterogeneity of variances.

7.5.2 Covariance Structure in Linear Mixed Models

Data from experiments and studies that involve treatment structures with fixed and random effects are usually described using mixed models with more than one variance-covariance parameter. There are numerous types of mixed models, ranging from randomised complete or incomplete blocks models, split-plot models, repeated measures type models, and other hierarchical models. Analysis in mixed models basically focuses on the fixed effects, random effects and residual parts of the model. The distribution of the response variable is based on some assumptions considered on the random and residual effects, which are ingredients for the modelling covariance structure of the random effects in the model.

We consider the general linear mixed model defined in (7.9), whose treatment structure consists of fixed effects and random effects. The classical mixed model is essential for modelling the response measurements in any type of grouped data, be it correlated or repeated measures within subjects in randomised block and split-plot designs. For a classical linear mixed model (7.9), which can also be expressed as (7.12), where the variance-covariance matrix consists of a component for the random effects $\mathbf{u} \sim MVN(\mathbf{0}, \mathbf{G})$, and the component for the random residuals $\boldsymbol{\epsilon} \sim MVN(\mathbf{0}, \mathbf{R})$. The response $\mathbf{y}$ is also multivariate normal with mean $\mathbf{X}\boldsymbol{\beta}$ and total variance-covariance $\mathbf{V} = \mathbf{Z}\mathbf{G}\mathbf{Z}' + \mathbf{R}$, which implies that the variance of $\mathbf{y}$ is modelled through $\mathbf{Z}$, $\mathbf{G}$ and $\mathbf{R}$, where $\mathbf{G}$ represents the covariance structure on the random effect terms and $\mathbf{R}$ represents the covariance structure of the residuals.

Henderson (1975) provided solutions to the mixed model (7.12), part of which is the most crucial parameter estimate, $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$. With known parameters of $\mathbf{V}$, the estimated

BLUE (best linear unbiased estimator) and BLUP (best linear unbiased predictor) are obtained, respectively (Milliken and Johnson, 1984). However, in most cases, the parameters of $\mathbf{V}$ are unknown and have to be estimated. In simple linear models, it is commonly assumed that $\mathbf{R} = \sigma^2\mathbf{I}$, where $\mathbf{I}$ is an $n \times n$ identity matrix, a structure that guarantees independence and homogeneity of residual errors; and that $\mathbf{G}$ is a diagonal matrix of variance components (i.e., each different variances, and all zero covariances). This structure of the variance-covariance matrix is a special case of the general mixed model, which can be extended to heterogeneous error models which allow various arbitrary parameterised covariance structures in $\mathbf{G}$ or $\mathbf{R}$ or both. Some of the common parameterised covariance structures for both $\mathbf{G}$ and $\mathbf{R}$ in the general normal-theory linear mixed model framework include the diagonal, time series $AR(1)$, unstructured, and compound symmetry. The following sections briefly explain some of these covariance structures in different treatment designs.

7.5.2.1 Covariance Structure of a P-factor Linear Mixed Model in CRD

Consider a p-way treatment structure in a completely randomised design (CRD), where only factors A and B are fixed, and the rest of the $(p-2)$ factors are random. The linear mixed model, in this case, is given by

$$y_{ijk\ldots ph} = \mu + \alpha_i + \beta_j + \gamma_k + \ldots + \omega_p + g_{ij} + \ldots + g_{ijk\ldots p} + \epsilon_{ijk\ldots ph}, \tag{7.15}$$

where μ denotes the mean response; α_i, β_j, and γ_k denote the effects of the i^{th} level of fixed factor A, the j^{th} level of factor B and the k^{th} level of fixed factor C, respectively; terms in g denotes the interaction effects between the factors in the model; and $\epsilon_{ijk\ldots ph}$ denotes the residual effect. We assume that the random effects $g_{ij} \sim i.i.d\, N(0, \sigma_g^2)$ and residual parts $\epsilon_{ijk\ldots ph} \sim i.i.d\, N(0, \sigma_\epsilon^2)$.

For repeated measures design, we express the model (7.15) in a general linear mixed model in matrix notation as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}_1 + \mathbf{Z}\mathbf{u}_2 + \ldots + \mathbf{Z}\mathbf{u}_t + \boldsymbol{\epsilon}, \tag{7.16}$$

where $\mathbf{y}$ is an $N \times 1$ vector of observations, $\mathbf{X}\boldsymbol{\beta}$ is the fixed-effects part of the model, $\mathbf{Z}\mathbf{u}_1, \mathbf{Z}\mathbf{u}_2, \ldots, \mathbf{Z}\mathbf{u}_t$, and $\boldsymbol{\epsilon}$ are the random effects and the residual parts of the model, respectively, where $\mathbf{u}_t \sim$

$N(\mathbf{0}, \sigma_t^2 \mathbf{I}_{n_t})$ and $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma_\epsilon^2 \mathbf{I}_N)$ are independent random variables. The simple covariance structure of the random effects and of the residual parts of the model would be based on the classical assumptions that $\mathbf{u} \sim N(\mathbf{0}, \mathbf{G})$ and $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \mathbf{R})$, respectively, where

$$\mathbf{G} = \begin{pmatrix} \sigma_1^2 \mathbf{I}_{n_1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \sigma_2^2 \mathbf{I}_{n_2} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \sigma_n^2 \mathbf{I}_{n_n} \end{pmatrix}.$$

and

$$\mathbf{R} = \begin{pmatrix} \sigma_\epsilon^2 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \sigma_\epsilon^2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \sigma_\epsilon^2 \end{pmatrix},$$

or equivalently, the total variance-covariance, $\mathbf{V} = \mathbf{ZGZ}' + \mathbf{R}$, given by the block-diagonal matrix

$$\mathbf{V} = \begin{pmatrix} \sigma_1^2 \mathbf{I}_{n_1} + \sigma_\epsilon^2 \mathbf{I}_{n_1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \sigma_2^2 \mathbf{I}_{n_2} + \sigma_\epsilon^2 \mathbf{I}_{n_2} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \sigma_n^2 \mathbf{I}_{n_n} + \sigma_\epsilon^2 \mathbf{I}_{n_n} \end{pmatrix}.$$

7.5.2.2 Other Covariance Structures in Mixed Model Analysis

The mixed model is a vital and flexible tool, especially in the analysis experiments with repeated measures data, due to its ability to embed the structure and relationships among the errors. The traditional and rigid approach of only considering a very simple structure of dependence among errors $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma_\epsilon^2 \mathbf{I}_N)$ may not be an attractive pattern in some studies with diverse relationships among errors. Examples of common possible candidates of covariance structures for parameterising in mixed models, which are defined by the structure of matrix $\mathbf{R}$, include the following:

- **Variance Components (VC) Covariance Structure**: also known as Independent or Simple Covariance Structure. This structure has equal or constant variances on the main diagonal and zero covariance (independent residuals) on the off-diagonals. This is the classic covariance structure for the standard fixed-effects ANOVA model but not appropriate for repeated measures. Thus, the VC covariance matrix is given by

$$\mathbf{V} = \begin{pmatrix} \sigma_\epsilon^2 & 0 & \cdots & 0 \\ 0 & \sigma_\epsilon^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_\epsilon^2 \end{pmatrix}.$$

This is a parameterised covariance structure, with only one parameter, σ_ϵ^2. However, for variables that are completely independent of each other and measured on different scales, the ideal VC covariance matrix will assume the pattern

$$\mathbf{V} = \begin{pmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_N^2 \end{pmatrix}.$$

- **First-Order Autoregressive $AR(1)$ Covariance Structure**: is appropriate when a correlation between any two adjacent observations is assumed. This is typical for a classic repeated measures model or time series analysis centred on the idea that the current observation depends on its previous value (i.e., a first-order autoregressive model). Assume that the correlation between any two adjacent is ρ for $-1 \leq \rho \leq 1$, and the correlation between any two observations separated by $n-1$ other observations be ρ^n. Then, the AR(1) covariance matrix is expressed as

$$\mathbf{V} = \sigma^2 \begin{pmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{k-1} \\ \rho & 1 & \rho & \cdots & \rho^{k-2} \\ \rho^2 & \rho & 1 & \cdots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \rho \\ \rho^{k-1} & \rho^{k-2} & \cdots & \rho & 1 \end{pmatrix},$$

  where k is the number of repeated measurements per experimental subject or unit. The covariance structure has two parameters, σ^2 and ρ. An important feature of this type of covariance structure is that, as the distance between two observations increases, their correlation decreases while the variances remain constant in the main diagonal.

- **Heterogeneous First-order Autoregressive ARH(1) Covariance Structure**: which allows the variances to differ in the main diagonal, thereby attracting more parameters. The variance-covariance matrix of ARH(1) is expressed as

$$\mathbf{V} = \begin{pmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho & \sigma_1\sigma_3\rho^2 & \cdots & \sigma_1\sigma_k\rho^{k-1} \\ \sigma_2\sigma_1\rho & \sigma_2^2 & \sigma_2\sigma_3\rho & \cdots & \sigma_2\sigma_k\rho^{k-2} \\ \sigma_3\sigma_1\rho^2 & \sigma_3\sigma_2\rho & \sigma_3^2 & \cdots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \sigma_{k-2}\sigma_k\rho \\ \sigma_k\sigma_1\rho^{k-1} & \sigma_k\sigma_2\rho^{k-2} & \cdots & \sigma_k\sigma_{k-2}\rho & \sigma_k^2 \end{pmatrix},$$

  with $k+1$ covariance parameters and heterogeneous variances in the main diagonal, where k is the number of repeated measurements per experimental subject or unit.

- **Heterogeneous Compound Symmetry (CSH) Covariance Structure**: a covariance structure which does not require variances to be homogeneous, as does the compound symmetry (CS) structure. The CSH variance-covariance structure is expressed as

$$\mathbf{V} = \begin{pmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho & \sigma_1\sigma_3\rho & \cdots & \sigma_1\sigma_k\rho \\ \sigma_2\sigma_1\rho & \sigma_2^2 & \sigma_2\sigma_3\rho & \cdots & \sigma_2\sigma_k\rho \\ \sigma_3\sigma_1\rho & \sigma_3\sigma_2\rho & \sigma_3^2 & \cdots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \sigma_{k-2}\sigma_k\rho \\ \sigma_k\sigma_1\rho & \sigma_k\sigma_2\rho & \cdots & \sigma_k\sigma_{k-2}\rho & \sigma_k^2 \end{pmatrix},$$

  with $k+1$ covariance parameters and heterogeneous variances in the main diagonal, where k is the number of repeated measurements per experimental subject or unit. The difference between CS and CSH covariance structures is analogous to the difference between the AR(1) and ARH(1) covariance structures.

- **Unstructured (UN) Covariance Structure**: refers to the covariance structure which places no condition on the covariance structure but allows both the variance and covariance terms to be heterogeneous. The UN variance-covariance structure is expressed as

$$\mathbf{V} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} & \cdots & \sigma_{1k} \\ \sigma_{21} & \sigma_2^2 & \sigma_{23} & \cdots & \sigma_{2k} \\ \sigma_{31} & \sigma_{32} & \sigma_3^2 & \cdots & \sigma_{3k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sigma_{k1} & \sigma_{k2} & \sigma_{k3} & \cdots & \sigma_k^2 \end{pmatrix},$$

  The UN covariance structure requires fitting $\frac{k(k+1)}{2}$ variance-covariance parameters.

- **Toeplitz (TOEP) Covariance Structure**: is analogous to the AR(1), although the two do not necessarily have the same pattern. In the Toeplitz covariance structure, we have equal correlations and covariances within each off-diagonal band, and different correlations and covariances among bands; i.e., correlations of the first two adjacent measurements are homogeneous, the measurements two apart have the same correlation different from the first, measurements three apart have the same correlation different from the first two, etc. This pattern technically makes the AR(1) a special case of the Toeplitz. The Toeplitz variance-covariance structure is expressed as

$$\mathbf{V} = \begin{pmatrix} \sigma^2 & \sigma_1 & \sigma_2 & \cdots & \sigma_k \\ \sigma_1 & \sigma^2 & \sigma_1 & \cdots & \sigma_{k-1} \\ \sigma_2 & \sigma_1 & \sigma^2 & \cdots & \sigma_{k-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sigma_k & \sigma_{k-1} & \cdots & \sigma_1 & \sigma^2 \end{pmatrix},$$

with k covariance parameters and homogeneous variances in the main diagonal, where k is the number of repeated measurements per experimental subject or unit.

- **Heterogeneous Toeplitz (TOEPH) Covariance Structure**: is the covariance structure that has unequal variances in the main diagonal of the matrix, resulting in additional parameters to be estimated, one for every measurement. The TOEPH variance-covariance structure is expressed as

$$\mathbf{V} = \begin{pmatrix} \sigma_1^2 & \sigma_1 & \sigma_2 & \cdots & \sigma_k \\ \sigma_1 & \sigma_2^2 & \sigma_1 & \cdots & \sigma_{k-1} \\ \sigma_2 & \sigma_1 & \sigma_3^2 & \cdots & \sigma_{k-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sigma_k & \sigma_{k-1} & \cdots & \sigma_1 & \sigma_k^2 \end{pmatrix},$$

with $2k - 1$ covariance parameters and heterogeneous variances in the main diagonal, where k is the number of repeated measurements per experimental subject or unit.

7.5.2.3 Covariance Structure in Repeated-measures Type Model

The term "repeated measures design" usually refers to a completely randomised design with multiple, or repeated, measurements taken on the same experimental subject or unit, which is observed sequentially over time (Littell *et al.*, 2006). Measurements from a repeated measures study are often correlated, especially for two measurements taken closer together than those taken further apart. Therefore, it is crucial to identify an appropriate covariance structure of the errors when analysing data from a repeated measures design (Littell *et al.*, 2000). The most important issue in repeated measures analysis of covariance is to determine an appropriate parameterisation for $\mathbf{R}$ before estimating the resulting parameters.

Analogous to the split-plot design, where whole plots have a completely randomised design with a repeated measures structure to the subplots, the repeated measures variance-covariance structure is obtained assuming compound symmetry. The covariance matrix within a subject of a repeated measures model with compound symmetry expressed as,

$$\mathbf{V} = \sigma^2 \begin{pmatrix} 1 & \rho & \rho & \cdots & \rho \\ \rho & 1 & \rho & \cdots & \rho \\ \rho & \rho & 1 & \cdots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \rho \\ \rho & \rho & \cdots & \rho & 1 \end{pmatrix},$$

However, it is not always the case that the simple compound symmetry assumption holds when modelling the covariance structure in the split-plot and repeated measures data. In the event that the compound symmetry assumption is not satisfied, other approaches in the mixed model, that do not require such structure must be explored. The development of mixed models software such as *lme*4 package in R (Bates *et al.*, 2014) and SAS *PROCMIXED* procedure (Littell *et al.*, 2006) has seen an increase the use of mixed models.

7.5.3 Selecting an Appropriate Covariance Structure

The mixed model (7.9) is generally used to analyse mixed experimental data in various contexts when $\mathbf{u}$ and $\boldsymbol{\epsilon}$ are assumed to have $N(\mathbf{0}, \mathbf{G})$ and $N(\mathbf{0}, \mathbf{R})$, respectively. When using the SAS *PROCMIXED* procedure, the two variance-covariance components ($\mathbf{G}$ and $\mathbf{R}$) are specified by the *RANDOM* and *REPEATED* statements, respectively. The "$TYPE =$ " option used in conjunction with the *REPEATED* statement in the SAS *PROCMIXED* system gives options to select the desired covariance structure for repeated measures. There are various ways of identifying the appropriate covariance structure amongst a set of candidate structures (SAS Institute Inc., 1999). However, the most famous approach is to select the structure that gives the smallest Akaike's Information Criterion (AIC) (Akaike, 1974). AIC is a statistic that is defined by the model and the maximum likelihood estimates of the parameters from the specified variance-covariance as

$$AIC = (-2)\mathcal{L}(\hat{\boldsymbol{\beta}}\hat{\mathbf{V}}) + 2(p), \tag{7.17}$$

where p is an effective number of independently adjusted parameters in the covariance matrix , and $\mathcal{L}(\hat{\boldsymbol{\beta}}\hat{\mathbf{V}}) = log(ML)$ is the value of the likelihood function evaluated at $(\hat{\boldsymbol{\beta}}\hat{\mathbf{V}})$. A better model is the one with the smallest AIC value. Different forms of the matrix $\mathbf{R}$ can be compared for adequacy using the likelihood ratio test statistic (Milliken and Johnson, 1984). The hypotheses involved are $H_0 : \mathbf{R}_1$ is as adequate as $\mathbf{R}_2$ against $H_1 : \mathbf{R}_1$ is not as adequate as $\mathbf{R}_2$, where $\mathbf{R}_1$ is a special case of $\mathbf{R}_2$. Suppose $\mathbf{R}_1$ and $\mathbf{R}_2$ have p_1 and p_2 parameters, respectively, with $p_1 < p_2$. The test statistic is $Q = (-2)\mathcal{L}[(\hat{\boldsymbol{\beta}}_1\hat{\mathbf{V}}_1) - \mathcal{L}(\hat{\boldsymbol{\beta}}_2\hat{\mathbf{V}}_2)]$, which is distributed as $\chi^2(p_2 - p_1)$. We reject H_0 when $Q \geq \chi^2_{\frac{\alpha}{2}}(p_2 - p_1)$.

7.6 Conclusion

When applying the new approach of partitioning experimental data based on the factor levels and the desired inference space, we may encounter different relationships among errors in each of the partitioned data subsets, which subsequently implies certain structures of the variance-covariance matrices. The classic approach linear mixed model analysis approach is traditionally centred around the assumption that the error terms are pairwise uncorrelated with zero means and variance σ^2, a necessary assumption for point estimation, which leads to the normal distribution assumption, $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma_\epsilon^2 \mathbf{I}_N)$, for the purposes of hypothesis testing and confidence interval estimation. However, it is critical to choose an appropriate covariance structure that best suits the particular relationships among errors in the data observations. Therefore, it is always important to understand the covariance structure of the dataset first before blindly assuming the classical structure. This can be achieved by fitting a possible covariance structure, starting with the unstructured, and examining the pattern to have a brief idea about its suitability before trying other possible candidate structures. Proper hypothesis tests for selecting the most appropriate covariance structure were discussed by Milliken and Johnson (1984). Ideally, an adequate covariance structure is the one with the least number of parameters, which eventually provides a larger number of degrees of freedom for the tests and estimates. The mixed model approach, in particular the SAS $PROCMIXED$ procedure, has the ability to accommodate different kinds of variance-covariance structures as well and take care of missing data and unequal time spacing.

CHAPTER 8

CONCLUSIONS, RECOMMENDATIONS AND FUTURE WORKS

8.1 Introduction

The recent advancements in technology has brought the need to evaluate the newly invented methods and contrast them against the old and existing ones to evaluate their worthiness. The evaluation and comparison processes are conveniently handled in the mixed model methodology framework. The proposed linear mixed model would regard the new methods as fixed, while the old and existing methods are considered random. Implicit to the mixed model analysis is the scope of inference (broad, narrow or intermediate inference space), which forms an integral part of statistical data analysis. Targeting the wrong inference space leads to biased point estimates, interval estimates, and ultimately misleading hypothesis test conclusions for the entire population represented by the random factors. Depending on the context and objectives of the experiment, predictable functions can be manipulated to cater for population-wide or broad inference scope on the treatment effects (Littell *et al.* 2006). In addition, the partitioning approach allows for the alternative variance-covariance structures other than the homogeneous error variance structure, which is a default assumption in traditional linear mixed model approaches.

8.2 Conclusion

Situations where new strategies or interventions are needed either to complement or replace the old and existing methods are increasing in agricultural, industrial and engineering fields. The studies in such cases are conveniently handled when these factors are conceptualised as each having both fixed (new) and random (old) levels. The use of the traditional notion of factors as fully fixed or random may not yield sufficient analysis results that capture all the variations present in the research data. The current study proposes an analysis approach that allows for a specific comparison of certain fixed levels and assessment of variability (random levels) within the same factor. Depending on the number of factors, factor levels and the selected design of the experiment, this arrangement poses some complexities in determining the possible partitions of the data, model construction, parameter-estimation and variance estimation processes, hypothesis testing, combined analysis, and controlling for experimental error. These processes are tackled in different ways depending on the selected design of the experiment.

The approach was extended to complex cases where three factors, each having both fixed and random levels, are arranged in a completely randomised design (CRD). These scenarios are commonly encountered by agricultural and industrial analysts, who might be interested in measuring some of the factors with a higher precision than the others. Appropriate classification and partitioning of factors based on the targeted factor levels results in partitioned linear mixed models, whose design matrices are either full-rank or less-than-full-rank form. Attempts to apply this approach leaves the partitioned data subsets vulnerable to outlier contamination, which might subsequently compromise the level of accuracy and precision of the selected partitioned models. This study proposes the use of robust estimation methods that require the use of linear mixed models with considerably little outlier contamination. The conclusion was that the fusion of the partitioning approach and the use of robust estimation methods led to improved precision in the model estimates.

In multi-stratum experimental designs, where factors are randomised at each level in order to assess the model precision at different levels of the experiment, a need to conceptualise each factor as having both fixed and random levels might be necessary. An example is a split-split-plot treatment structure, where all the three factors would be laid in a randomised complete block design (RCBD). The current study demonstrated how the partitioning approach could be used to construct a linear mixed model, estimate the model parameters, test hypotheses and assess the model adequacy in this scenario. The particularised analysis proposed by the new approach provides room for researchers to manipulate the appropriate factor combinations in order to allow for narrow, intermediate and broad inferential space on the levels of each of the factors as well as their associated interactions. The study established that researchers need to consider the choice of inference space as it directly impacts the magnitude of standard errors of estimates.

The fundamental consideration in the estimation process of linear models is the special case in which elements of the error vector are assumed equal and uncorrelated, which is not always appropriate for real-life data. The proposed analysis approach could accommodate heteroscedastic error terms in repeated measures data when factors were conceptualised as having both fixed and random levels. In such cases, the estimation of partitioned linear mixed model parameters considers assumptions other than the identity case on the structure of the variance-covariance matrix of the error vector. Accessing the options for other covariance structures that are appropriate for correlated error variances and computing the sums of squares and F-ratios via the usual maximum likelihood method was a hurdle since most of the available statistical software do not have such options. The current study proposes that partitioned analyses can be obtained, which will then be syncretised to a combined analysis.

An extension of the proposed analysis approach to a general linear mixed model with p-factors is possible, which disintegrates to 2^p partitioned models. The current study proposed some useful procedures for constructing covariance matrices, sums of squares and variance components which are needed when fitting a linear mixed model from either balanced or unbalanced data involving factors that consist of both fixed and random levels. The approach is recommended for complex real-life factorial experiments which require consideration of different inference spaces and various covariance structures when constructing partitioned linear mixed models when factors involved consist of both fixed and random levels. Of importance is the need to select the most appropriate covariance structure before estimating model parameters and drawing inferences. We believe that the partitioning approach can be extended to other experimental designs and data analyses that conform to the principles of linear mixed model methodology.

8.3 Recommendations

Based on the illustrations discussed in different experimental designs and structures, we, therefore, recommend that researchers:

- apply the partitioned analysis approach in modelling experiments with large datasets where factors need to be conceptualised as having both fixed and random levels,
- consider and explore alternative error covariance structures, especially where the traditional equal error variance is not practically reasonable,
- use the SAS $PROC\ MIXED$ and $PROC\ GLIMMIX$ procedures, which provide various options for the covariance structure for the partitioned and combined analysis of linear mixed models through the likelihood-based estimation scheme,
- use the partitioning approach coupled with appropriate robust estimation methods when experimental data is subject to minimal outlier contamination.

8.4 Limitations and Weaknesses of the Study

In this study, we considered only the general linear model methodology for quantifying the effects of predictor variables on a single response variable when data is normally distributed. Most classical statistical analysis techniques for univariate data analysis usually rely on normally distributed data. Some cases exist, when non-normal (binary, multinomial or count response) data is involved. Faced with such cases, researchers often resort to robust estimation techniques and other statistical approaches that shoehorn their data into classical statistical frameworks to satisfy the general linear model assumptions. However, it is not always guaranteed that the use of these approaches will achieve normality.

Secondly, the current study was limited to the scenario when the associations between the response and the predictors are assumed to remain constant at different levels. Although some efforts were made to use other estimation methods other than the ordinary least squares (OLS) method in Chapter 6 to address the issue of heterogeneous error variance-covariance matrices, the interest remained in measuring the differences in outcome variables between populations at the mean. Often, researchers may be interested in group differences across the distribution of a given response variable rather than only at the mean, leading to quantile regression methods.

The third limitation is that we considered experiments with a considerably fewer number of factors for illustration purposes. There are high-dimensional environments which naturally create possible complex interactions and unexpected heterogeneity (Chen *et al.*, 2015). Inferences in such scenarios often pose challenges that common approaches may not be able to handle.

Lastly, the study focused on repeated-measures designs under a limited scope of variance-covariance structures that fall into a category of diagonal covariance structures. There is room to explore the application of the partitioning approach to other designs of experiments that incorporate more complex and non-diagonal covariance structures.

8.5 Future Areas of Research

We close with a list of possible research areas we intend to explore in the future. As indicated earlier, the partitioned approach is appropriate for investigating and comparing the new methods and strategies against the old and existing ones in various research settings. Therefore, we propose the extension of the proposed analysis approach to the following research areas:

- Linear mixed models have been used routinely to model scenarios involving both fixed and random effects. All the scenarios we considered in our study involve the general linear model methodology, which assumes that the associations between the response and the predictors remain constant at different levels (common regression slope assumption). The possibility of using the partitioned analysis approach to experiments involving non-linear mixed models needs to be explored. This may include, but not limited to, generalised linear models (GLMs) and quantile regression, which allow for different associations between the response and the predictors.

- Central to linear mixed model inference is the estimation of fixed effects and variance components (Ferreira *et al.*, 2020). The least squares method is the most popular method used for estimating model parameters. The commonly used methods for the estimation of variance components in linear mixed models, such as the maximum likelihood (ML), restricted maximum likelihood (REML) (Harville, 1977), and Bayesian methods (Box and Tiao, 1992; Agresti, 2015), work effectively well in orthogonal linear mixed models (when the normality assumption is assumed) and balanced data. However, when the normality assumption is not considered, nonorthogonal mixed models are involved, and inference becomes challenging. We, therefore, pose an open problem of exploring the possibility of applying the partitioning approach to orthogonal linear mixed models.

- There are situations when the number of fixed effects in a study is large (high-dimensional case), and the number of fixed effects diverges as the sample size goes to infinity (Chen *et al.*, 2015). High-dimensional data often require the existing linear mixed model tests to be modified in order to handle the problem with a sparse model structure. Most classical approaches used to test fixed effects linear mixed models (Kenward and Roger, 1997; Wang and Dai, 2014) are robust in small datasets, but they tend to break down in high-dimensional data (Bradic *et al.*, 2020). We pose an open problem for future research that can be investigated using the partitioning approach.

- We pose an open research problem on the application of the partitioning approach to other designs of experiments that incorporate more complex and non-diagonal covariance structures.

REFERENCES

1. Aghamohammadi, A., and Meshkani, M. R. (2017). Bayesian quantile regression for skew-normal linear mixed models. *Communications in Statistics - Theory and Methods*, **46**(22), 10953-10972.

2. Agostinelli, C., and Yohai, V. J. (2016). Composite Robust Estimators for Linear Mixed Models. *Journal of the American Statistical Association*, **111**(516), 1764-1774.

3. Agresti, A. (2015). *Foundations of linear and generalized linear models.* John Wiley and Sons Inc. Aitken, A. C. (1935). On Least Squares and Linear Combination of Observations. *Proceedings of the Royal Society of Edinburgh*, **55**, 42-48.

4. Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, **19**(6), 716-723.

5. Almimi, A. A., Kulahci, M., and Montgomery, D. C. (2009). Checking the Adequacy of Fit of Models from Split-Plot Designs. *Journal of Quality Technology*, **41**(3), 272-284.

6. Anderson, V. L., and McLean, R. A. (2019). *Design of Experiments: A realistic approach.* CRC PRESS.

7. Armstrong, R. A. (2017). Recommendations for analysis of repeated-measures designs: Testing and correcting for sphericity and use of MANOVA and mixed model analysis. *Ophthalmic and Physiological Optics*, **37**(5), 585-593.

8. Baayen, R. H., Davidson, D. J., and Bates, D. M. (2008). mixed modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, **59**(4), 390-412.

9. Barnett, A. G., Koper, N., Dobson, A. J., Schmiegelow, F., and Manseau, M. (2010). Using information criteria to select the correct variance-covariance structure for longitudinal data in ecology: Selecting the correct variance-covariance. *Methods in Ecology and Evolution*, **1**(1), 15-24.

10. Barr, D. J., Levy, R., Scheepers, C., and Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, **68**(3), 255-278.

11. Bartlett, M. S. (1937). Properties of sufficiency and statistical tests. *Proceedings of the Royal Society of London. Series A - Mathematical and Physical Sciences*, **160**(901), 268-282.

12. Bate, S. T., and Chatfield, M. J. (2016). Identifying the Structure of the Experimental Design. *Journal of Quality Technology*, **48**(4), 343-364.

13. Bates, D. M. (2010). *lme4: mixed Modelling with R.* https://lme4.R-Forge.R-project.org/book/

14. Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting Linear mixed Models Using lme4. *Journal of Statistical Software*, **67**(1).

15. Blouin, D. C., and Riopelle, A. J. (2005). On Confidence Intervals for Within-Subjects Designs. *Psychological Methods*, **10**(4), 397-412.

16. Boisgontier, M. P., and Cheval, B. (2016). The anova to mixed model transition. *Neuroscience and Biobehavioral Reviews*, **68**, 1004-1005.

17. Box, G. E. P., and Tiao, G. C. (1992). *Bayesian inference in statistical analysis* (Wiley classics library ed). Wiley.

18. Bradic, J., Claeskens, G., and Gueuning, T. (2020). Fixed Effects Testing in High-Dimensional Linear Mixed Models. *Journal of the American Statistical Association*, **115**(532), 1835-1850.

19. Brauer, M., and Curtin, J. J. (2018). Linear mixed models and the analysis of nonindependent data: A unified framework to analyze categorical and continuous independent variables that vary within-subjects and/or within-items. *Psychological Methods*, **23**(3), 389-411.

20. Brown, M. B., and Forsythe, A. B. (1974). Robust Tests for the Equality of Variances. *Journal of the American Statistical Association*, **69**(346), 364-367.

21. Chaka, L., and Njuho, P. (2022). Repeated-measures analysis in the context of heteroscedastic error terms with factors having both fixed and random levels. *Stats*, **5**(2), 458-476.

22. Chaka, L., and Njuho, P. (2021). Construction of a linear mixed model with each factor

having both fixed and random levels: A case of split-split-plot structure in a RCBD. *International Journal of Agricultural and Statistical Sciences*, **17**(2), 501-518.

23. Chen, F., Li, Z., Shi, L., and Zhu, L. (2015). Inference for mixed models of ANOVA type with high-dimensional data. *Journal of Multivariate Analysis*, **133**, 382-401.

24. Christian, L. E. (1980). *Modification and simplification of symmetric differences squared procedure for estima- tion of genetic variances and covariances* [Ph.D. Dissertation]. The Ohio State University.

25. Crump, S. L. (1946). The Estimation of Variance Components in Analysis of Variance. *Biometrics Bulletin*, **2**(1), 7.

26. Cui, H., Ng, K. W., and Zhu, L. (2004). Estimation in mixed effects model with errors in variables. *Journal of Multivariate Analysis*, **91**(1), 53-73.

27. Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B* (Methodological), **39**(1), 1-38.

28. Dempster, A. P., Rubin, D. B., and Tsutakawa, R. K. (1981). Estimation in Covariance Components Models. *Journal of the American Statistical Association*, **76**(374), 341-353.

29. Ding, P., Feller, A., and Miratrix, L. (2019). Decomposing Treatment Effect Variation. *Journal of the American Statistical Association*, **114**(525), 304-317.

30. Dixon, P. (2016). Should Blocks be Fixed or Random? *Conference on Applied Statistics in Agriculture*. https://doi.org/10.4148/2475-7772.1474.

31. Du, H., and Wang, L. (2020). Testing Variance Components in Linear Mixed Modeling Using Permutation. *Multivariate Behavioral Research*, **55**(1), 120-136.

32. Eze, F. C., and Nwankwo, E. U. (2016). Analysis of Variance in an Unbalanced Two-Way Mixed Effect Interactive Model. *Open Journal of Statistics*, **06**(02), 310-319.

33. Ferreira, D., Ferreira, S. S., Nunes, C., and Mexia, J. T. (2017). Estimation in mixed models through three step minimization. *Communications in Statistics - Simulation and Computation*, **46**(2), 1156-1166.

34. Ferreira, D., Ferreira, S. S., Nunes, C., and Mexia, J. T. (2020). Inference in nonorthogonal mixed models. *Mathematical Methods in the Applied Sciences*, mma.6866.

35. Fisher, R. A. (1925). *Statistical methods for research workers* London: Oliver and Boyd.

36. Frey, J. (2010). Testing for equivalence of variances using Hartley's ratio. *Canadian Journal of Statistics*, **38**(4), 647-664.

37. Gad, A. M., and EL-Zayat, N. I. (2018). Fitting Multivariate Linear Mixed Model for Multiple Outcomes Longitudinal Data with Non-ignorable Dropout. *International Journal of Probability and Statistics*, **7**(4), 97-105.

38. Geisser, S., and Greenhouse, S. W. (1958). An Extension of Box's Results on the Use of the F Distribution in Multivariate Analysis. *The Annals of Mathematical Statistics*, **29**(3), 885-891.

39. Gelman, A. (2005). Analysis of variance-Why it is more important than ever. *The Annals of Statistics*, **33**(1), 1-53.

40. Gelman, A. (2016). The problems with p-values are not just with p-values. *The American Statistician*, **70**, 1-2.

41. Gennings, C., Chinchilli, V. M., and Carter, W. H. (1989). Response Surface Analysis with Correlated Data: A Nonlinear Model Approach. *Journal of the American Statistical Association*, **84**(407), 805-809.

42. Ghidey, W., Lesaffre, E., and Eilers, P. (2004). Smooth Random Effects Distribution in a Linear Mixed Model. *Biometrics*, **60**(4), 945-953.

43. Gokpinar, E., and Gokpinar, F. (2017). Testing equality of variances for several normal populations. *Communications in Statistics - Simulation and Computation*, **46**(1), 38-52.

44. Goldberger, A. (1962). Best Linear Unbiased Prediction in the Generalized Linear Regression Model. *Journal of the American Statistical Association*, **57**(298), 369-375.

45. Gomez, K. A., and Gomez, A. A. (1984). *Statistical procedures for agricultural research* (2nd ed). Wiley.

46. Goos, P., Syafitri, U., Sartono, B., and Vazquez, A. R. (2020). A nonlinear multidimen-

sional knapsack problem in the optimal design of mixture experiments. *European Journal of Operational Research*, **281**(1), 201-221.

47. Graybill, F. A. (1983). *Matrices with applications in statistics* (2. ed). Wadsworth.

48. Greenberg, B. G., and Sarhan, A. E. (1959). Matrix Inversion, Its Interest and Application in Analysis of Data. *Journal of the American Statistical Association*, **54**(288), 755-766.

49. Grimes, L. W., and Harvey, W. R. (1980). Estimation of Genetic Variances and Covariances Using Symmetric Differences Squared. *Journal of Animal Science*, **50**(4), 634-644.

50. Guo, X., and Tang, C. Y. (2021). Specification tests for covariance structures in high-dimensional statistical models. *Biometrika*, **108**(2), 335-351.

51. Güven, G., Gürer, Ö., Şamkar, H., and Şenoğlu, B. (2019). A fiducial-based approach to the one-way ANOVA in the presence of nonnormality and heterogeneous error variances. *Journal of Statistical Computation and Simulation*, **89**(9), 1715-1729.

52. Halekoh, U., and H∅jsgaard, S. (2014). A Kenward-Roger Approximation and Parametric Bootstrap Methods for Tests in Linear Mixed Models - The R Package pbkrtest. *Journal of Statistical Software*, **59**(9), 1-30.

53. Harrar, S. W., Ronchi, F., and Salmaso, L. (2019). A comparison of recent nonparametric methods for testing effects in two-by-two factorial designs. Journal of Applied Statistics, **46**(9), 1649-1670.

54. Harrison, X. A., Donaldson, L., Correa-Cano, M. E., Evans, J., Fisher, D. N., Goodwin, C. E. D., Robinson, B. S., Hodgson, D. J., and Inger, R. (2018). A brief introduction to mixed effects modelling and multi-model inference in ecology. *PeerJ*, **6**, e4794.

55. Hartley, H. O. (1950). The Maximum F-Ratio as a Short-Cut Test for Heterogeneity of Variance. *Biometrika*, **37**(3/4), 308.

56. Hartley, H. O., and Rao, J. N. K. (1967). Maximum-Likelihood Estimation for the Mixed Analysis of Variance Model. *Biometrika*, **54**(1/2), 93.

57. Harville, D. A. (1976). Extension of the Gauss-Markov Theorem to Include the Estimation of Random Effects. *The Annals of Statistics*, **4**(2), 384-395.

58. Harville, D. A. (1977). Maximum Likelihood Approaches to Variance Component Estimation and to Related Problems. *Journal of the American Statistical Association*, **72**(358), 320-338.

59. Harville, D. A. (1978). Alternative Formulations and Procedures for the Two-Way Mixed Model. *Biometrics*, **34**(3), 441.

60. Henderson, C. R. (1953). Estimation of Variance and Covariance Components. Biometrics, **9**(2), 226-252.

61. Henderson, C. R. (1975). Best Linear Unbiased Estimation and Prediction under a Selection Model. *Biometrics*, **31**(2), 423.

62. Henderson, C. R. (1984). Estimation of Variances and Covariances Under Multiple Trait Models. *Journal of Dairy Science*, **67**, 1581-1589.

63. Hocking, R. R. (1985). *The analysis of linear models*. Brooks/Cole Pub. Co.

64. Hohenstein, S., Matuschek, H., and Kliegl, R. (2017). Linked linear mixed models: A joint analysis of fixation locations and fixation durations in natural reading. *Psychonomic Bulletin and Review*, **24**(3), 637-651.

65. Hu, X., and Spilke, J. (2011). Variance-covariance structure and its influence on variety assessment in regional crop trials. *Field Crops Research*, **120**(1), 1-8.

66. Hui, F. K. C., Müller, S., and Welsh, A. H. (2019). Testing random effects in linear mixed models: Another look at the F-test (with discussion). *Australian and New Zealand Journal of Statistics*, **61**(1), 61-84.

67. Huynh, H., and Feldt, L. S. (1976). Estimation of the Box Correction for Degrees of Freedom from Sample Data in Randomized Block and Split-Plot Designs. *Journal of Educational Statistics*, **1**(1), 69.

68. Janssen, D. P. (2012). Twice random, once mixed: Applying mixed models to simultaneously analyze random effects of language and participants. *Behavior Research Methods*, **44**(1), 232-247.

69. Jayalath, K. P., and Ng, H. K. T. (2018). Analysis of means approach for random factor

analysis. *Journal of Applied Statistics*, **45**(8), 1426-1446.

70. Jones, B., and Goos, P. (2012a). I-Optimal Versus D-Optimal Split-Plot Response Surface Designs. *Journal of Quality Technology*, **44**(2), 85-101.

71. Jones, B., and Goos, P. (2012b). An Algorithm for Finding D-Efficient Equivalent-Estimation Second-Order Split-Plot Designs. *Journal of Quality Technology*, **44**(4), 363-374.

72. Kackar, R. N., and Harville, D. A. (1984). Approximations for Standard Errors of Estimators of Fixed and Random Effect in Mixed Linear Models. *Journal of the American Statistical Association*, **79**(388), 853.

73. Keele, J. W., and Harvey, W. R. (1989). Estimation of components of variance and covariance by symmetric differences squared and minimum nom quadratic unbiased estimation: A comparison. *Journal of Animal Science*, **67**, 348.

74. Keele, J. W., Long, T. E., and Jonnson, R. K. (1991). Comparison of methods of estimating variance components in pigs. *Faculty Papers and Publications in Animal Science*, 49.

75. Kenward, M. G., and Roger, J. H. (1997). Small Sample Inference for Fixed Effects from Restricted Maximum Likelihood. *Biometrics*, **53**(3), 983.

76. Kéry, M. (2010). *Introduction to WinBUGS for ecologists: A Bayesian approach to regression, ANOVA, mixed models and related analyses* (1st ed). Elsevier.

77. Koller, M. (2013). *Robust Estimation of Linear Mixed Models* [Dissertation, ETH Zürich]. http://e-collection.library.ethz.ch/eserv/eth:6670/eth-6670-02.pdf?pid=eth:6670&dsID=eth-6670-02.pdf

78. Koller, M. (2016). robustlmm: An R Package for Robust Estimation of Linear mixed Models. *Journal of Statistical Software*, **75**(6).

79. Koller, M., and Stahel, W. A. (2011). Sharpening Wald-type inference in robust regression for small samples. *Computational Statistics and Data Analysis*, **55**(8), 2504-2515.

80. Kotchaporn, S., and Araveeporn, A. (2018). Modifications of Levene's and O'Brien's

Tests for Testing the Homogeneity of Variance Based on Median and Trimmed Mean. *Thailand Statistician*, **16**(2), 106-128.

81. Kuehl, R. O. (2000). *Design of experiments: Statistical principles of research design and analysis* (2. ed). Brooks/Cole, Cengage Learning.

82. Kuran, Ö., and Özkale, M. R. (2021). Improvement of mixed predictors in linear mixed models. *Journal of Applied Statistics*, **48**(5), 924-942.

83. Kuznetsova, A., Brockhoff, P. B., and Christensen, R. H. B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, 82(13).

84. La Motte, L. R. (1971). *Locally best quadratic estimators of variance components.* (Technical Report 22). Univiversitty of Kentucky.

85. Lachos, V. H., Cabral, C. R. B., and Abanto-Valle, C. A. (2012). A non-iterative sampling Bayesian method for linear mixed models with normal independent distributions. *Journal of Applied Statistics*, **39**(3), 531-549.

86. Laird, N. M., and Ware, J. H. (1982). Random-Effects Models for Longitudinal Data. Biometrics, **38**(4), 963.

87. Lee, S., and Ahn, C. H. (2003). Modified ANOVA for Unequal Variances. *Communications in Statistics - Simulation and Computation*, **32**(4), 987-1004.

88. Levene, H. (1960). *Robust tests for equality of variances.* In: Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling, I. Olkin, eds. Palo Alto. Stanford University Press., 278-292.

89. Littell, R. C., Milliken, G. A., Stroup, W. W., Wolfinger, R. D., and Schabenberger, O. (Eds.). (2006). *SAS for mixed models* (2nd ed). SAS Institute, Inc.

90. Littell, R. C., Pendergast, J., and Natarajan, R. (2000). Modelling covariance structure in the analysis of repeated measures data. *Statistics in Medicine*, **19**(13), 1793-1819.

91. Littell, R. C., Stroup, W. W., and Freund, R. J. (2010). *SAS for linear models* (4. ed). SAS Inst.

92. Liu, X., Huang, M., Fan, B., Buckler, E. S., and Zhang, Z. (2016). Iterative Usage of

Fixed and Random Effect Models for Powerful and Efficient Genome-Wide Association Studies. *PLOS Genetics*, **12**(2), e1005767.

93. Loeza-Serrano, S., and Donev, A. N. (2014). Construction of Experimental Designs for Estimating Variance Components. *Computational Statistics and Data Analysis*, **71**: 1168-77.

94. Luepsen, H. (2018). Comparison of nonparametric analysis of variance methods: A vote for van der Waerden. *Communications in Statistics - Simulation and Computation*, **47**(9), 2547-2576.

95. Luke, S. G. (2017). Evaluating significance in linear mixed models in R. *Behavior Research Methods*, **49**(4), 1494-1502.

96. Macharia, H., and Goos, P. (2010). D-Optimal and D-Efficient Equivalent-Estimation Second-Order Split-Plot Designs. *Journal of Quality Technology*, **42**(4), 358-372.

97. Mara, C. A., and Cribbie, R. A. (2018). Equivalence of Population Variances: Synchronizing the Objective and Analysis. *Journal of Experimental Education*, **86**(3), 442-457.

98. Martinez, M.-J., and Holian, E. (2014). An Alternative Estimation Approach for the Heterogeneity Linear Mixed Model. *Communications in Statistics - Simulation and Computation*, **43**(10), 2628-2638.

99. Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., and Bates, D. (2017). Balancing Type I error and power in linear mixed models. *Journal of Memory and Language*, **94**, 305-315.

100. Mauchly, J. W. (1940). Significance Test for Sphericity of a Normal n-Variate Distribution. *The Annals of Mathematical Statistics*, **11**(2), 204-209.

101. McCulloch, C. E., Searle, S. R., and Neuhaus, J. M. (2008). *Generalized, linear, and mixed models* (2nd ed). Wiley.

102. McCulloch C.E. and Searle S.R. (2001). *Generalized, Linear and Mixed Models*. Wiley.

103. McLean, R. A., and Sanders, W. L. (1988). Approximating degrees of freedom for standard errors in mixed linear models. In: *Proceedings of the Statistical Computing Section,*

American Statistical Association, New Orleans, LA., 50-59.

104. McLean, R. A., Sanders, W. L., and Stroup, W. W. (1991). A Unified Approach to Mixed Linear Models. *The American Statistician*, **45**(1), 54.

105. Milliken, G. A., and Johnson, D. E. (1984). *Analysis of messy data: Vol. VOLUME III: ANALYSIS OF COVARIANCE*. Lifetime Learning Publications.

106. Milliken, G. A., and Johnson, D. E. (2002). *Analysis of messy data. 3: Analysis of covariance*. Chapman and Hall/CRC.

107. Milliken, G. A., and Johnson, D. E. (2009). *Analysis of messy data* (2nd ed). CRC Press.

108. Möhring, J., Williams, E., and Piepho, H.-P. (2015). Inter-block information: To recover or not to recover it? *Theoretical and Applied Genetics*, **128**(8), 1541-1554.

109. Molenberghs, G., and Verbeke, G. (2000). *Linear Mixed Models for Longitudinal Data.* https://doi.org/10.1007/978-1-4419-0300-6

110. Montgomery, D. C. (2013). *Design and analysis of experiments* (Eighth edition). John Wiley and Sons, Inc.

111. Moser, B. K., and Sawyer, J. K. (1998). Algorithms for Sums of Squares and Covariance Matrices Using Kronecker Products. *The American Statistician*, **52**(1), 54-57.

112. Moskowitz, D. S., Hershberger, S. L., and American Psychological Association (Eds.). (2002). *Modeling intraindividual variability with repeated measures data: Methods and applications*. L. Erlbaum Associates.

113. Muller, K. E., Edwards, L. J., Simpson, S. L., and Taylor, D. J. (2007). Statistical tests with accurate size and power for balanced linear mixed models. *Statistics in Medicine*, **26**(19), 3639-3660.

114. Munyakazi, L., Hintz, R. L., and Selby, B. D. (1994). Applications of estimable functions in Agricultural research with special emphasis on the GLM procedure of SAS. *Conference on Applied Statistics in Agriculture.* https://doi.org/10.4148/2475-7772.1365

115. Mutlu, H. T., Gökpinar, F., Gökpinar, E., Gül, H. H., and Güven, G. (2017). A new com-

putational approach test for one-way ANOVA under heteroscedasticity. *Communications in Statistics - Theory and Methods*, **46**(16), 8236-8256.

116. Mylona, K., Gilmour, S. G., and Goos, P. (2020). Optimal Blocked and Split-Plot Designs Ensuring Precise Pure-Error Estimation of the Variance Components. *Technometrics*, **62**(1), 57-70.

117. Nguyen, N.-K., and Pham, T.-D. (2015). Searching for D-Efficient Equivalent-Estimation Second-Order Split-Plot Designs. *Journal of Quality Technology*, **47**(1), 54-65.

118. Njoroge, G. G., Simbauni, J. A., and Koske, J. A. (2017). An Optimal Split-Plot Design for Performing a Mixture-Process Experiment. *Science Journal of Applied Mathematics and Statistics*, **5**(1), 15.

119. Njuho, P. M., and Milliken, G. A. (2005). Analysis of Linear Models with One Factor Having Both Fixed and Random Levels. *Communications in Statistics: Theory and Methods*, **34**(9/10), 1979-1989.

120. Njuho, P. M., and Milliken, G. A. (2009). Analysis of Linear Models with Two Factors Having Both Fixed and Random Levels. *Communications in Statistics: Theory and Methods*, **38**(14), 2348-2365.

121. Oliveira, S., Nunes, C., Moreira, E., Fonseca, M., and Mexia, J. T. (2019). Balanced prime basis factorial fixed effects model with random number of observations. *Journal of Applied Statistics*, **47**(13-15), 2737-2748.

122. Ott, L., and Longnecker, M. (2016). *An introduction to statistical methods and data analysis* (Seventh edition). Cengage Learning.

123. Özkale, M. R., and Kuran, Ö. (2020). A further prediction method in linear mixed models: Liu prediction. *Communications in Statistics - Simulation and Computation*, **49**(12), 3171-3195.

124. Pan, J., and Shang, J. (2018). A simultaneous variable selection methodology for linear mixed models. *Journal of Statistical Computation and Simulation*, **88**(17), 3323-3337.

125. Park, C., and Leeds, M. (2016). A highly efficient robust design under data contamination.

Computers and Industrial Engineering, **93**, 131-142.

126. Parra-Frutos, I. (2013). Testing homogeneity of variances with unequal sample sizes. *Computational Statistics*, **28**(3), 1269-1297.

127. Patterson, H. D., and Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika*, **58**(3), 545-554.

128. Patterson, H. D., and Thompson, R. (1975). Maximum likelihood estimation of components of variance. *Proceedings of the 8th International Biometric Conference*, 197-207.

129. Piepho, H. P., and Edmondson, R. N. (2018). A tutorial on the statistical analysis of factorial experiments with qualitative and quantitative treatment factor levels. *Journal of Agronomy and Crop Science*, **204**(5), 429-455.

130. Piepho, H.P., Williams, E.R. and Fleck, M. (2006). A note on the analysis of Designed Experiments with complex treatment structure. *HortScience*, **41**(2): 446-52.

131. Pinheiro, J. C., and Bates, D. M. (2000). *mixed models in S and S-PLUS*. Springer.

132. Pinheiro, J. C., Liu, C., and Wu, Y. N. (2001). Efficient Algorithms for Robust Estimation in Linear mixed Models Using the Multivariate t Distribution. *Journal of Computational and Graphical Statistics*, **10**(2), 249-276.

133. Purcell, S. W., Lalavanua, W., Cullis, B. R., Cocks, N., and Blasiak, H. editor: R. (2018). Small-scale fishing income and fuel consumption: Fiji's artisanal sea cucumber fishery. *ICES Journal of Marine Science / Journal Du Conseil*, **75**(5), 1758-1767.

134. Pusponegoro, N. H., Rachmawati, R. N., Notodiputro, K. A., and Sartono, B. (2017). Linear Mixed Model for Analyzing Longitudinal Data: A Simulation Study of Children Growth Differences. *Procedia Computer Science*, **116**, 284-291.

135. Raminez, J. G., Gore, W. L., and Associates Inc. (2010). *The Best of Both Worlds: Designing Complex Experiments in the JMP-SAS Environment*. [Paper 293-2010]. SAS Global Forum 2010.

136. Rao, C. R. (1971). Minimum variance quadratic unbiased estimation of variance components. *Journal of Multivariate Analysis*, **1**(4), 445-456.

137. Robinson, D. L. (1987). Estimation and Use of Variance Components. *The Statistician*, **36**(1), 3.

138. Robinson, G. K. (1991). That BLUP is a Good Thing: The Estimation of Random Effects. *Statistical Science*, **6**(1), 15-32.

139. Rogers, G. S. (1984). Kronecker Products in ANOVA-A First Step. The American Statistician, **38**(3), 197.

140. Runcie, D. E., and Crawford, L. (2019). Fast and flexible linear mixed models for genome-wide genetics. *PLOS Genetics*, **15**(2), e1007978.

141. Saeed, B. I. I., Appiah, S. K., Nuamah, N. N. N. N., Munyakazi, L., and Musah, A. A. I. (2014). Model Equivalence in General Linear Models: Set-to-Zero, Sum-to-Zero Restrictions, and Extra Sum of Squares Method. *International Journal of Statistics and Probability*, **3**(4), p42.

142. SAS Institute Inc. (1999). *SAS/STAT user's guide, Version 8.* SAS Institute Inc.

143. SAS Institute Inc. (2017). *SAS/STAT® 14.3 User's Guide.* SAS Institute Inc.

144. Satterthwaite, F. E. (1946). An Approximate Distribution of Estimates of Variance Components. *Biometrics Bulletin*, **2**(6), 110.

145. Saw, S. L. C. (1992). ANOVA Sums of Squares as Quadratic Forms. *The American Statistician*, **46**(4), 288-290.

146. Schaeffer, L. R. (1986). Pseudo Expectation Approach to Variance Component Estimation. *Journal of Dairy Science*, **69**(11), 2884-2889.

147. Schielzeth, H., and Forstmeier, W. (2009). Conclusions beyond support: Overconfident estimates in mixed models. *Behavioral Ecology*, **20**(2), 416-420.

148. Schnell, P. M., and Bose, M. (2019). Spectral parameterization for linear mixed models applied to confounding of fixed effects by random effects. *Journal of Statistical Planning and Inference*, **200**, 47-62.

149. Searle, S. R., and Gruber, M. H. J. (2017). *Linear models* (Second edition). Wiley.

150. Seltman, H., J. (2018). *Experimental Design and Analysis.* https://www.stat.cmu.edu/ hseltman/309/Book/Book.pdf.

151. Shah, K. R. (1970). On the Loss of Information in Combined Inter- and Intra-Block Estimation. *Journal of the American Statistical Association*, **65**(332), 1562-1564.

152. Shaw, R. G., and Mitchell-Olds, T. (1993). Anova for Unbalanced Data: An Overview. *Ecology*, **74**(6), 1638-1645.

153. Smith, C. L., and Edwards, L. J. (2017). A test of separate hypotheses for comparing linear mixed models with non nested fixed effects. *Communications in Statistics - Theory and Methods*, **46**(11), 5487-5500.

154. Stroup, W. W. (2016). *Generalized Linear Mixed Models: Modern Concepts, Methods and Applications.* CRC Press.

155. Stroup, W. W., and Kachman, S. D. (1994). Generalized Linear Models - An Overview. *Conference on Applied Statistics in Agriculture.*

156. Stroup, W. W., Milliken, G. A., Claassen, E. A., and Wolfinger, R. D. (2018). *SAS for mixed models: Introduction and basic applications.* SAS.

157. Stroup, W. W., and Mulitze, D. K. (1991). Nearest Neighbor Adjusted Best Linear Unbiased Prediction. *The American Statistician*, **45**(3), 194-200.

158. Sullivan, L. M. (2008). Repeated Measures. Circulation, **117**(9), 1238-1243.

159. Sunwoo, H. (1996). Simple algorithms about Kronecker products in the linear model. *Linear Algebra and Its Applications*, **237/238**, 351-358.

160. Swallow, W. H., and Searle, S. R. (1978). Minimum Variance Quadratic Unbiased Estimation (MIVQUE) of Variance Components. *Technometrics*, **20**(3), 265-272.

161. Takemura, A. (1983). Tensor Analysis of ANOVA Decomposition. *Journal of the American Statistical Association*, **78**(384), 894-900.

162. Tan, S. L., and Nott, D. J. (2014). Variational Approximation for Mixtures of Linear Mixed Models. *Journal of Computational and Graphical Statistics*, **23**(2), 564-585.

163. Thiese, M. S., Ronna, B., and Ott, U. (2016). P value interpretations and considerations. *Journal of Thoracic Disease*, **8**(9), E928-E931.

164. Trinca, L. A., and Gilmour, S. G. (2017). Split-Plot and Multi-Stratum Designs for Statistical Inference. *Technometrics*, **59**(4), 446-457.

165. UCLA: Statistical Consulting Group. (n.d.). *Statistical Methods and Data Analytics: Repeated measures analysis using SAS.* Accessed on February 6, 2022, from https://stats.oarc.ucla.edu/sa
repeatedmeasures/

166. van Raden, P. M., and Jung, Y. C. (1988). A General Purpose Approximation to Restricted Maximum Likelihood: The Tilde-Hat Approach. *Journal of Dairy Science*, **71**(1), 187-194.

167. Vargas, M., Glaz, B., Alvarado, G., Pietragalla, J., Morgounov, A., Zelenskiy, Y., and Crossa, J. (2015). Analysis and Interpretation of Interactions in Agricultural Research. Agronomy Journal, **107**(2), 748-762.

168. Wang, K., and Dai, C. (2014). A Mixed-Effect Model for Analyzing Experiments with Multistage Processes. Q*uality Technology and Quantitative Management*, **11**(4), 491-511.

169. Wang, Y., Liu, Y., Li, X., Wang, C., Wang, M., and Song, Z. (2020). GORFLM: Globally Optimal Robust Fitting for Linear Model. *Signal Processing: Image Communication*, **84**, 115834.

170. Wasserstein, R. L., Schirm, A. L., and Lazar, N. A. (2019). Moving to a World Beyond "$p < 0.05$." *The American Statistician*, **73**(sup1), 1-19.

171. Welch, B. L. (1947). The Generalization of Student's' Problem when Several Different Population Variances are Involved. *Biometrika*, **34**(1/2), 28.

172. Wellek, S. (2010). *Testing statistical hypotheses of equivalence and noninferiority* (2nd ed). CRC Press.

173. Westfall, J., Judd, C. M., and Kenny, D. A. (2015). Replicating Studies in Which Samples of Participants Respond to Samples of Stimuli. *Perspectives on Psychological Science*, **10**(3), 390-399.

174. Westfall, P. H., and SAS Institute (Eds.). (1999). *Multiple comparisons and multiple tests: Using the SAS system.* SAS Institute.

175. Wilk, M. B., and Kempthorne, O. (1955). Fixed, Mixed, and Random Models. *Journal of the American Statistical Association*, **50**(272), 1144-1167.

176. Winter, B. (2013). *Linear models and linear mixed effects models in R with linguistic applications.*
https://Arxiv.Org/Abs/1308.5499, arXiv:1308.5499.

177. Wu, M., Zhao, J., Wang, T., and Zhao, Y. (2017). The ANOVA-type inference in linear mixed model with skew-normal error. *Journal of Systems Science and Complexity*, **30**(3), 710-720.

178. Xu, L., Yang, F., Chen, R., and Yu, S. (2015). A Parametric Bootstrap Test for Two-Way ANOVA Model Without Interaction Under Heteroscedasticity. *Communications in Statistics - Simulation and Computation*, **44**(5), 1264-1272.

179. Yang, R.-C. (2010). Towards understanding and use of mixed model analysis of agricultural experiments. *Canadian Journal of Plant Science*, **90**, 602-627.

180. Yavarizadeh, B., Rasekh, A., and Babadi, B. (2020). Estimation of parameters in linear mixed measurement error models with stochastic linear restrictions. *Communications in Statistics - Theory and Methods*, **49**(23), 5853-5865.

181. Ye, R. D., and Wang, T. H. (2015). Inferences in linear mixed models with skew-normal random effects. *Acta Mathematica Sinica, English Series*, **31**(4), 576-594.

182. Zare, K., Rasekh, A., and Rasekhi, A. A. (2012). Estimation of variance components in linear mixed measurement error models. *Statistical Papers*, **53**(4), 849-863.

183. Zhang, H., and Ding, F. (2013). On the Kronecker Products and Their Applications. *Journal of Applied Mathematics*, 2013, 1-8.

184. Zhong, X.-P., Fung, W.-K., and Wei, B.-C. (2002). Estimation in linear models with random effects and errors-in-variables. *Annals of the Institute of Statistical Mathematics*, **54**(3), 595-606.

185. Zucker, D. M., Lieberman, O., and Manor, O. (2000). Improved small sample inference in the mixed linear model: Bartlett correction and adjusted likelihood. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **62**(4), 827-838.

APPENDICES

Appendix A: Alternative Approach for Estimating $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{u}}$

An easier alternative approach for obtaining the estimates $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{u}}$ involves assuming $\mathbf{u}$ as fixed in (2.3) and apply the least squares method to solve for $\boldsymbol{\beta}$ and $\mathbf{u}$.

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\epsilon}$$
$$\hat{\mathbf{y}} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}$$

$$\begin{aligned}
\boldsymbol{\epsilon}'\boldsymbol{\epsilon} &= (\mathbf{y} - \hat{\mathbf{y}})^T(\mathbf{y} - \hat{\mathbf{y}}) \\
\boldsymbol{\epsilon}'\boldsymbol{\epsilon} &= (\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u}) \\
&= \mathbf{y}^T\mathbf{y} - \mathbf{y}^T\mathbf{X}\boldsymbol{\beta} - \mathbf{y}^T\mathbf{Z}\mathbf{u} - (\mathbf{X}\boldsymbol{\beta})^T\mathbf{y} + (\mathbf{X}\boldsymbol{\beta})^T(\mathbf{X}\boldsymbol{\beta}) \\
&+ (\mathbf{X}\boldsymbol{\beta})^T\mathbf{Z}\mathbf{u} - (\mathbf{Z}\mathbf{u})^T\mathbf{y} + (\mathbf{Z}\mathbf{u})^T\mathbf{X}\boldsymbol{\beta} + (\mathbf{Z}\mathbf{u})^T(\mathbf{Z}\mathbf{u}) \\
\frac{\partial}{\partial\boldsymbol{\beta}}(\boldsymbol{\epsilon}'\boldsymbol{\epsilon}) &= -2\mathbf{y}^T\mathbf{X} + 2\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} + 2\mathbf{X}^T\mathbf{Z}\mathbf{u} = 0 \\
&\implies \mathbf{X}^T\mathbf{X}\boldsymbol{\beta} + \mathbf{X}^T\mathbf{Z}\mathbf{u} = \mathbf{X}^T\mathbf{y}
\end{aligned}$$

$$\begin{aligned}
\frac{\partial}{\partial\mathbf{u}}(\boldsymbol{\epsilon}'\boldsymbol{\epsilon}) &= -2\mathbf{y}^T\mathbf{Z} + 2(\mathbf{X}\boldsymbol{\beta})^T\mathbf{Z} + 2\mathbf{Z}^T\mathbf{Z}\mathbf{u} = 0 \\
&\implies \mathbf{Z}^T\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}^T\mathbf{Z}\mathbf{u} = \mathbf{Z}^T\mathbf{y}
\end{aligned}$$

Adding $\mathbf{G}^{-1}$ to the lower sub-matrix of coefficients results in the following simultaneous equations (Henderson, 1953):

$$\begin{bmatrix} \mathbf{X}^T\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}^T\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}^T\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}^T\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1} \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}^* \\ \mathbf{u}^* \end{bmatrix} = \begin{bmatrix} \mathbf{X}^T\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}^T\mathbf{R}^{-1}\mathbf{y} \end{bmatrix} \tag{8.1}$$

We need to show that the BLUE ($\hat{\boldsymbol{\beta}}$) and BLUP ($\hat{\mathbf{u}}$) in (2.4) and (2.6) are equal to $\boldsymbol{\beta}^*$ and $\mathbf{u}^*$ in (8.1), respectively. To prove that $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}^*$ in (2.4) and (8.1), we solve for $\mathbf{u}^*$ in the second equation of (8.1) and substitute it into the first one.

$$\begin{aligned}
\mathbf{Z}^T\mathbf{R}^{-1}\mathbf{X}\boldsymbol{\beta}^* + (\mathbf{Z}^T\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1})\mathbf{u}^* &= \mathbf{Z}^T\mathbf{R}^{-1}\mathbf{y} \\
(\mathbf{Z}^T\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1})\mathbf{u}^* &= \mathbf{Z}^T\mathbf{R}^{-1}\mathbf{y} - \mathbf{Z}^T\mathbf{R}^{-1}\mathbf{X}\boldsymbol{\beta}^* \\
\implies \mathbf{u}^* &= (\mathbf{Z}^T\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1})^{-1}\mathbf{Z}^T\mathbf{R}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^*)
\end{aligned}$$

$$\begin{aligned}
&\mathbf{X}^T\mathbf{R}^{-1}\mathbf{X}\boldsymbol{\beta}^* + \mathbf{X}^T\mathbf{R}^{-1}\mathbf{Z}\mathbf{u}^* = \mathbf{X}^T\mathbf{R}^{-1}\mathbf{y}\\
\Longrightarrow\ &\mathbf{X}^T\mathbf{R}^{-1}\mathbf{X}\boldsymbol{\beta}^* + \mathbf{X}^T\mathbf{R}^{-1}\mathbf{Z}(\mathbf{Z}^T\mathbf{R}^{-1}\mathbf{Z}+\mathbf{G}^{-1})^{-1}\mathbf{Z}^T\mathbf{R}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}^*) = \mathbf{X}^T\mathbf{R}^{-1}\mathbf{y}\\
\Longrightarrow\ &\mathbf{X}^T\mathbf{R}^{-1}\mathbf{X}\boldsymbol{\beta}^* - \mathbf{X}^T\mathbf{R}^{-1}\mathbf{Z}(\mathbf{Z}^T\mathbf{R}^{-1}\mathbf{Z}+\mathbf{G}^{-1})^{-1}\mathbf{Z}^T\mathbf{R}^{-1}\mathbf{X}\boldsymbol{\beta}^*\\
&= \mathbf{X}^T\mathbf{R}^{-1}\mathbf{y} - \mathbf{X}^T\mathbf{R}^{-1}\mathbf{Z}(\mathbf{Z}^T\mathbf{R}^{-1}\mathbf{Z}+\mathbf{G}^{-1})^{-1}\mathbf{Z}^T\mathbf{R}^{-1}\mathbf{y}\\
\Longrightarrow\ &[\mathbf{X}^T\mathbf{R}^{-1}\mathbf{X} - \mathbf{X}^T\mathbf{R}^{-1}\mathbf{Z}(\mathbf{Z}^T\mathbf{R}^{-1}\mathbf{Z}+\mathbf{G}^{-1})^{-1}\mathbf{Z}^T\mathbf{R}^{-1}]\mathbf{X}\boldsymbol{\beta}^*\\
&= [\mathbf{X}^T\mathbf{R}^{-1} - \mathbf{X}^T\mathbf{R}^{-1}\mathbf{Z}(\mathbf{Z}^T\mathbf{R}^{-1}\mathbf{Z}+\mathbf{G}^{-1})^{-1}\mathbf{Z}^T\mathbf{R}^{-1}]\mathbf{y}\\
\Longrightarrow\ &\mathbf{X}^T[\mathbf{R}^{-1}\mathbf{X} - \mathbf{R}^{-1}\mathbf{Z}(\mathbf{Z}^T\mathbf{R}^{-1}\mathbf{Z}+\mathbf{G}^{-1})^{-1}\mathbf{Z}^T\mathbf{R}^{-1}]\mathbf{X}\boldsymbol{\beta}^*\\
&= \mathbf{X}^T[\mathbf{R}^{-1} - \mathbf{R}^{-1}\mathbf{Z}(\mathbf{Z}^T\mathbf{R}^{-1}\mathbf{Z}+\mathbf{G}^{-1})^{-1}\mathbf{Z}^T\mathbf{R}^{-1}]\mathbf{y}\\
\Longrightarrow\ &\mathbf{X}^T\mathbf{Q}\mathbf{X}\boldsymbol{\beta}^* = \mathbf{X}^T\mathbf{Q}\mathbf{y}
\end{aligned}$$

where $\mathbf{Q} = \mathbf{R}^{-1} - \mathbf{R}^{-1}\mathbf{Z}(\mathbf{Z}^T\mathbf{R}^{-1}\mathbf{Z}+\mathbf{G}^{-1})^{-1}\mathbf{Z}^T\mathbf{R}^{-1}$. If we let $\mathbf{Q} = \mathbf{V}^{-1}$, then $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}^*$.

We need to complete the proof by showing that $\mathbf{VQ} = \mathbf{I}$.

$$\begin{aligned}
\mathbf{VQ} &= (\mathbf{ZGZ}^T + \mathbf{R})[\mathbf{R}^{-1} - \mathbf{R}^{-1}\mathbf{Z}(\mathbf{Z}^T\mathbf{R}^{-1}\mathbf{Z}+\mathbf{G}^{-1})^{-1}\mathbf{Z}^T\mathbf{R}^{-1}]\\
&= \mathbf{ZGZ}^T\mathbf{R}^{-1} - \mathbf{ZGZ}^T\mathbf{R}^{-1}\mathbf{Z}(\mathbf{Z}^T\mathbf{R}^{-1}\mathbf{Z}+\mathbf{G}^{-1})^{-1}\mathbf{Z}^T\mathbf{R}^{-1} + \mathbf{I} - \mathbf{Z}(\mathbf{Z}^T\mathbf{R}^{-1}\mathbf{Z}+\mathbf{G}^{-1})^{-1}\mathbf{Z}^T\mathbf{R}^{-1}\\
&= \mathbf{I} + \mathbf{ZGZ}^T\mathbf{R}^{-1} - \mathbf{Z}(\mathbf{Z}^T\mathbf{R}^{-1}\mathbf{Z}+\mathbf{G}^{-1})^{-1}\mathbf{Z}^T\mathbf{R}^{-1}(\mathbf{ZGZ}^T\mathbf{R}^{-1} + \mathbf{I})\\
&= \mathbf{I} + \mathbf{ZGZ}^T\mathbf{R}^{-1} - \mathbf{ZG}(\mathbf{Z}^T\mathbf{R}^{-1}\mathbf{Z}+\mathbf{G}^{-1})^{-1}(\mathbf{Z}^T\mathbf{R}^{-1}\mathbf{Z}+\mathbf{G}^{-1})\mathbf{Z}^T\mathbf{R}^{-1}\\
&= \mathbf{I} + \mathbf{ZGZ}^T\mathbf{R}^{-1} - \mathbf{ZGZ}^T\mathbf{R}^{-1}\\
&= \mathbf{I}
\end{aligned}$$

Similarly, to prove that $\hat{\mathbf{u}} = \mathbf{u}^*$ in (2.6) and (8.1), we introduce $\mathbf{I} = \mathbf{VV}^{-1}$, where $\mathbf{V} =$

$\mathbf{ZGZ}^T + \mathbf{R}$, and $\mathbf{GZ}^{-1}\mathbf{GZ} = \mathbf{I}$ in the expression of $\mathbf{u}^*$.

$$\begin{aligned}
\mathbf{u}^* &= (\mathbf{Z}^T\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1})^{-1}\mathbf{Z}^T\mathbf{R}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^*) \\
&= (\mathbf{Z}^T\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1})^{-1}\mathbf{Z}^T\mathbf{R}^{-1}\mathbf{V}\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^*) \\
&= (\mathbf{Z}^T\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1})^{-1}\mathbf{Z}^T\mathbf{R}^{-1}(\mathbf{ZGZ}^T + \mathbf{R})\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^*) \\
&= (\mathbf{Z}^T\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1})^{-1}(\mathbf{Z}^T\mathbf{R}^{-1}\mathbf{ZGZ}^T + \mathbf{Z}^T)\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^*) \\
&= (\mathbf{Z}^T\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1})^{-1}(\mathbf{Z}^T\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1})\mathbf{GZ}^T\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^*) \\
&= \mathbf{GZ}^T\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^*) \\
&= \hat{\mathbf{u}}
\end{aligned}$$

Appendix B: R Simulation Code for an FRF Model in CRD

```
library(purrr)
library(broom)
library(broom.mixed)
library(dplyr)
library(ggplot2)
library(lmerTest)
library(tidyverse)

library(haven)
wealthFRR <- read_sav("FRR.sav")
head(wealthFRR)

wealthFRR$soilmgt <- as.factor(wealthFRR$soilmgt)
wealthFRR$loccow <- as.factor(wealthFRR$loccow)
wealthFRR$farmsize <- as.factor(wealthFRR$farmsize)

model4FRR<-lmer(wealth ~ farmsize+(1|soilmgt)+(1|loccow)+(1|farmsize:soilmgt)+
(1|farmsize:loccow)+(1|soilmgt:loccow)+(1|farmsize:soilmgt:loccow),
data=wealthFRR, REML = TRUE)
summary(model4FRR)

farmsize<-wealthFRR$farmsize
soilmgt<-wealthFRR$soilmgt
loccow<-wealthFRR$loccow

set.seed(874)
nfarmsize=3
nsoilmgt=6
nloccow=4
e0=2.00946
e11=0.02723
e12=0.71104
sdsoil=0.2659
```

```
sdloc=0.1989
sd=1.082

(soilmgteff=rnorm(nsoilmgt, 0, sdsoil))
( w = rep(soilmgteff, times=c(207,77,121,72, 74,69)) )
( x = rep(soilmgteff, times=c(88,76,76,74,75,75)) )
( y = rep(soilmgteff, times=c(93,79,75,73,68,78)) )
( z = rep(soilmgteff, times=c(75,76,72,72,76,72)) )
( epssoil = c(w,x,y,z))

(loccoweff=rnorm(nloccow, 0, sdloc))
( epsloc = rep(loccoweff, c(620,464,466,443)) )
(epsilon4=rnorm(n = 1993, mean = 0, sd = 1.082))

( simresp4FRR = e0 + e11*(farmsize == "B")+ e12*(farmsize == "F")+epssoil+ epsloc+epsilon4)
dat4=data.frame(farmsize,soilmgt,loccow,simresp4FRR)
dat4

testmod4=lmer(simresp4FRR ~ farmsize+(1|soilmgt)+(1|loccow)+(1|farmsize:soilmgt)+
(1|farmsize:loccow)+(1|soilmgt:loccow)+(1|farmsize:soilmgt:loccow),
data=dat4, REML = TRUE)
summary(testmod4)

wealthfun4 = function(nfarmsize=3,nsoilmgt=6,nloccow=4, e0=2.00946,e11=0.02723,e12=0.71104,
sdsoil=0.2659,sdloc=0.1989,
sd=1.082) {
soilmgteff=rnorm(nsoilmgt, 0, sdsoil)
w = rep(soilmgteff, times=c(207,77,121,72,74,69))
x = rep(soilmgteff, times=c(88,76,76,74,75,75))
y = rep(soilmgteff, times=c(93,79,75,73,68,78))
z = rep(soilmgteff, times=c(75,76,72,72,76,72))
epssoil = c(w,x,y,z)
loccoweff=rnorm(nloccow, 0, sdloc)
epsloc = rep(loccoweff, c(620,464,466,443))
epsilon4=rnorm(n = 1993, mean = 0, sd = 1.082)
simresp4FRR = e0 + e11*(farmsize == "B")+ e12*(farmsize == "F")+epssoil+ epsloc+epsilon4
dat4=data.frame(farmsize,soilmgt,loccow,simresp4FRR)
testmod4=lmer(simresp4FRR~farmsize+(1|soilmgt)+(1|loccow)+(1|farmsize:soilmgt)+
(1|farmsize:loccow)+(1|soilmgt:loccow)+(1|farmsize:soilmgt:loccow),
data=dat4, REML = TRUE)
testmod4
}

set.seed(874)
wealthfun4()

sims = replicate(10000, wealthfun4(), simplify = FALSE )
sims[[10000]]

library(broom.mixed)
tidy(testmod4)

summary(testmod4)

sims % > %
map_df(tidy) % > %
filter(term=="farmsizeB" ) % > %
ggplot( aes(x = estimate) ) +
geom_density(fill = "green", alpha = .5) +
geom_vline( xintercept =-0.006558)

sims % > %
map_df(tidy) % > %
filter(term=="farmsizeF" ) % > %
ggplot( aes(x = estimate) ) +
geom_density(fill = "green", alpha = .5) +
geom_vline( xintercept = 0.731674)

sims % > %
```

```
map_dbl( summary(.x)$sigma) % > %
data.frame(sigma = .) % > %
ggplot( aes(x = sigma) ) +
geom_density(fill = "blue", alpha = .5) +
geom_vline(xintercept = 1.09660)
```

```
sims % > %
map_dbl(∼summary(.x)$sigma) % > %
. < 1.09660 % > %
mean()
```

```
sims % > %
map_df(tidy) % > %
filter(term == "farmsizeF") % > %
pull(p.value) % > %
. < 0.05 % > %
mean()
```

```
sims % > %
map_df(tidy) % > %
filter(term == "farmsizeB") % > %
pull(p.value) % > %
. < 0.05 % > %
mean()
```

```
suppressPackageStartupMessages(library(dplyr) )
library(ggplot2)
```

```
soilmgt_sims = c(5, 20, 30) % > %
set_names() % > %
map(∼replicate(1000, wealthfun4(nsoilmgt= 6) ) )
```

```
soilmgt_vars = soilmgt_sims % > %
modify_depth(2, ∼tidy(.x, effects = "ran_pars", scales = "vcov") ) % > %
map_dfr(bind_rows, .id = "soilmgt_num") %>%
filter(group == "soilmgt")
head(soilmgt_vars)
```

```
ggplot(soilmgt_vars, aes(x = estimate) ) +
geom_density(fill = "blue", alpha = .25) +
facet_wrap(~soilmgt_num) +
geom_vline(xintercept = 0.01507)
```

```
soilmgt_vars = mutate(soilmgt_vars, soilmgt_num = forcats::fct_inorder(soilmgt_num) )
```

```
add_prefix = function(string) {
paste("Number soilmgt:", string, sep = " ")
}
```

```
groupmed = soilmgt_vars % > %
group_by(soilmgt_num) % > %
summarise(mvar = median(estimate) )
```

```
ggplot(soilmgt_vars, aes(x = estimate) ) +
geom_density(fill = "blue", alpha = .25) +
facet_wrap(~soilmgt_num, labeller = as_labeller(add_prefix) ) +
geom_vline(aes(xintercept = 0.01507, linetype = "True variance"), size = .5 ) +
geom_vline(data = groupmed, aes(xintercept = mvar, linetype = "Median variance"), size =
.5) +
theme_bw(base_size = 14) +
scale_linetype_manual(name = "", values = c(2, 1) ) +
theme(legend.position = "bottom",
legend.key.width = unit(.1, "cm") ) +
labs(x = "Estimated Variance", y = NULL)
```

```
soilmgt_vars % > %
group_by(soilmgt_num) % > %
summarise_at("estimate",
list(min = min, mean = mean, med = median, max = max) )
```

```
loccow_sims = c(5, 20, 30) % > %
set_names() % > %
map(~replicate(1000, wealthfun4(nloccow= 4) ) )

loccow_vars = loccow_sims % > %
modify_depth(2, ~tidy(.x, effects = "ran_pars", scales = "vcov") ) % > %
map_dfr(bind_rows, .id = "loccow_num") % > %
filter(group == "loccow")
head(loccow_vars)

ggplot(loccow_vars, aes(x = estimate) ) +
geom_density(fill = "blue", alpha = .25) +
facet_wrap(~loccow_num) +
geom_vline(xintercept = 0.00918)

loccow_vars = mutate(loccow_vars, loccow_num = forcats::fct_inorder(loccow_num) )

add_prefix = function(string) {
paste("Number loccow:", string, sep = " ")
}

groupmed = loccow_vars % > %
group_by(loccow_num) % > %
summarise(mvar = median(estimate) )

ggplot(loccow_vars, aes(x = estimate) ) +
geom_density(fill = "blue", alpha = .25) +
facet_wrap(~loccow_num, labeller = as_labeller(add_prefix) ) +
geom_vline(aes(xintercept = 0.00918, linetype = "True variance"), size = .5 ) +
geom_vline(data = groupmed, aes(xintercept = mvar, linetype = "Median variance"),
size = .5) +
theme_bw(base_size = 14) +
scale_linetype_manual(name = "", values = c(2, 1) ) +
theme(legend.position = "bottom",
legend.key.width = unit(.1, "cm") ) +
labs(x = "Estimated Variance", y = NULL)

loccow_vars % > %

group_by(loccow_num) % > %

summarise_at("estimate",

list(min = min, mean = mean, med = median, max = max) )
```

Appendix C: SAS Code for Split-split Plot FFR model

```
FILENAME REFFILE ’/folders/myfolders/PhDThesis/Grain − Yield.sim2FFR.sav’;
PROC IMPORT DATAFILE=REFFILE
DBMS=SAV
OUT=WORK.Grain − Yield.sim2FFR;
RUN;


proc MIXED data=Grain − Yield.sim2FFR;
class Replication Nitrogen Management Variety;
model Yield= Nitrogen Management Management*Nitrogen;
random Replication Variety Variety*Nitrogen Variety*Management Nitrogen*Management*Variety
Replication*Nitrogen Replication*Nitrogen*Management;
lsmeans Nitrogen Management Management*Nitrogen;


estimate ’BLUE - Nitrogen N3 "broad"’ intercept 1 Nitrogen 1 0;
estimate ’BLUE - Nitrogen N5 "broad"’ intercept 1 Nitrogen 0 1;
estimate ’BLUE - Nitrogen diff "broad"’ Nitrogen 1 -1;


estimate ’Nitrogen N3 BLUP "narrow"’ intercept 2 Nitrogen 2 0 | Replication 1 1 1 Variety 1
1 Nitrogen*Variety 1 1 0 0 /divisor=2;
estimate ’Nitrogen N5 BLUP "narrow"’ intercept 2 Nitrogen 0 2 | Replication 1 1 1 Variety 1
1 Nitrogen*Variety 0 0 1 1 /divisor=2;
estimate ’BLUE - Nitrogen diff "narrow"’ Nitrogen 2 -2 | Nitrogen*Variety 1 1 -1 -1 /divisor=2;


estimate ’Nitrogen N3 BLUP "interm"’ intercept 2 Nitrogen 2 0 | Replication 1 1 1 Variety 1 1
/divisor=2;
estimate ’Nitrogen N5 BLUP "interm"’ intercept 2 Nitrogen 0 2 | Replication 1 1 1 Variety 1 1
/divisor=2;
estimate ’BLUE - Nitrogen diff "interm"’ Nitrogen 2 -2 /divisor=2;
estimate ’BLUE - Management M3 "broad"’ intercept 2 Management 2 0 /divisor=2;
estimate ’BLUE - Management M4 "broad"’ intercept 2 Management 0 2 /divisor=2;
estimate ’BLUE - Management diff "broad"’ Management 2 -2 /divisor=2;
estimate ’BLUP - Management M3 "interm"’ intercept 2 Management 2 0 | Replication 1 1 1 Variety 1 1 /divisor=2;
estimate ’BLUP - Management M4 "interm"’ intercept 2 Management 2 0 | Replication 1 1 1 Variety 1 1 /divisor=2;
estimate ’BLUP - Management diff "interm"’ Management 2 -2 /divisor=2;

estimate ’BLUE - Management M3 "narrow"’ intercept 2 Management 2 0 | Replication 1 1 1 Variety 1 1 Management*Variety 1 1 0 0 /divisor=2;
estimate ’BLUE - Management M4 "narrow"’ intercept 2 Management 0 2 | Replication 1 1 1 Variety 1 1 Management*Variety 0 0 1 1 /divisor=2;
estimate ’BLUE - Management diff "narrow"’ Management 2 -2 | Management*Variety 1 1 -1 -1 /divisor=2;

estimate ’BLUP - Variety V1 "broad"’ intercept 2 | Variety 2 0 /divisor=2;
estimate ’BLUP - Variety V3 "broad"’ intercept 2 | Variety 0 2 /divisor=2;
estimate ’BLUP - Variety diff "broad"’ intercept 0 | Variety 2 -2 /divisor=2;

estimate ’BLUP - Variety V1 "narrow"’ intercept 1 Nitrogen 1 0 Management 1 0 Nitrogen*Management 1 0 0 0 | Replication 1 1 1 Variety 1 0 Nitrogen*Variety 1 0 0 0 Management*Variety 1 0 0 0 Nitrogen*Management*Variety 1 0 0 0 0 0 0 0;
estimate ’BLUP - Variety V3 "narrow"’ intercept 1 Nitrogen 0 1 Management 0 1 Nitrogen*Management 0 0 0 1 | Replication 1 1 1 Variety 0 1 Nitrogen*Variety 0 0 0 1 Management*Variety 0 0 0 1
Nitrogen*Management*Variety 0 0 0 0 0 0 0 1;
estimate ’BLUP-Variety diff "narrow"’ Nitrogen 1 -1 Management 1 -1
```

```
Nitrogen*Management 1 0 0 -1| Variety 1 -1 Nitrogen*Variety 1 0 0 -1
Management*Variety 1 0 0 -1 Nitrogen*Management*Variety 1 0 0 0 0 0 0 -1;
run;
```

Appendix D: SAS Code for Repeated-Measures

```
/* Data Scrapping and Fitting FFR Model in Proc Mixed */
FILENAME REFFILE ’/home/u35581214/LDH Leakage Data.sav’;
PROC IMPORT DATAFILE=REFFILE
DBMS=SAV
OUT=LDH;
run;


/* View Repeated Measures Data in Multivariate Form */
Proc print data=LDH;
run;


/* Set Repeated Measures Data to Univariate form */
Data LDH-mult(keep=CCI4 CHCI3 Flask Time4 Time5 Time6)
LDH-univ(keep=CCI4 CHCI3 Flask Time Leakage);
set LDH;
output LDH-mult;
Leakage=Time4;Time=1; output LDH-univ;
Leakage=Time5;Time=2; output LDH-univ;
Leakage=Time6;Time=3; output LDH-univ;
run;


/* View Data in Univariate and Multivariate Form */
Proc print data=LDH-univ;
run;
Proc print data = LDH-mult;
run;


/* Subset or partition FRF from LHD univariate original Data */
```

```
Data FRF;
set LDH-univ ;
if (CCI4 =2.5 AND CHCI3 =0) then output ;
if (CCI4 =2.5 AND CHCI3 =0) then output ;
if (CCI4 =2.5 AND CHCI3 =0) then output ;
if (CCI4 =2.5 AND CHCI3 =5) then output ;
if (CCI4 =2.5 AND CHCI3 =5) then output ;
if (CCI4 =2.5 AND CHCI3 =5) then output ;
if (CCI4 =5 AND CHCI3 =0) then output ;
if (CCI4 =5 AND CHCI3 =0) then output ;
if (CCI4 =5 AND CHCI3 =0) then output ;
if (CCI4 =5 AND CHCI3 =5) then output ;
if (CCI4 =5 AND CHCI3 =5) then output ;
if (CCI4 =5 AND CHCI3 =5) then output ;
run;
Proc print data=FRF;
run;


/* Plot differences in Leakages contributed by predictors*/
proc means noprint data=FRF nway;
var Leakage;
class CCI4 CHCI3 Flask Time;
output out=avgFRF mean=avgLeakage;
run;
proc print data=avgFRF;
run;
/* new data set called avg created*/
/* Plot differences in Leakage by predictor CHCI3 and Time */
Proc gplot data=avgFRF;
plot avgLeakage*Time=CHCI3 / haxis=0 to 8 by 1 hminor=0 vminor=0;
```

```
symbol1 v=star c=blue i=join l=1;
symbol2 v=plus c=red i=join l=2;
title "Percentage leakage per time per CHCI3";
run; Quit;
```

/* **Partitioning FRF multivariate data for covariance analysis** */

```
Data FRF-mult;
set LDH-mult ;
if (CCI4 =2.5 AND CHCI3 =0 ) then output ;
if (CCI4 =2.5 AND CHCI3 =0) then output ;
if (CCI4 =2.5 AND CHCI3 =0) then output ;
if (CCI4 =2.5 AND CHCI3 =5) then output ;
if (CCI4 =2.5 AND CHCI3 =5) then output ;
if (CCI4 =2.5 AND CHCI3 =5) then output ;
if (CCI4 =5 AND CHCI3 =0 ) then output ;
if (CCI4 =5 AND CHCI3 =0 ) then output ;
if (CCI4 =5 AND CHCI3 =0 ) then output ;
if (CCI4 =5 AND CHCI3 =5 ) then output ;
if (CCI4 =5 AND CHCI3 =5 ) then output ;
if (CCI4 =5 AND CHCI3 =5 ) then output ;
run;
Proc print data=FRF-mult;
run;
```

/* **Sphericity test H0: Sphericity holds** */

```
proc glm data=FRF-mult;
class CCI4 CHCI3 ;
model Time4 Time5 Time6 =CCI4 CHCI3/ nouni;
repeated Time 3/ printe;
run;
```

```
/* Normality Q-Q plots */
ods graphics on;
proc mixed data=FRF plots=influenceestplot;
class CCI4 CHCI3 Time Flask ;
model Leakage=CCI4 Time CCI4*Time / residual;
random CHCI3 CCI4*CHCI3 CHCI3*Time CCI4*CHCI3*Time;
repeated / subject=Flask(CCI4*CHCI3) type=cs r;
run;
ods graphics off;


/* Checking Covariance Structure */
proc corr data=FRF-mult cov;
var Time4 Time5 Time6;
run;


/* Fitting the model using Proc MIXED Procedure */
proc mixed data=FRF method=reml cl ic covtest;
class CCI4 CHCI3 Time Flask ;
model Leakage=CCI4 Time CCI4*Time /s;
random CHCI3 CCI4*CHCI3 CHCI3*Time CCI4*CHCI3*Time ;
repeated / subject=Flask(CCI4*CHCI3) type=cs r;
run;


proc mixed data=FRF method=reml cl ic covtest;
class CCI4 CHCI3 Time Flask ;
model Leakage=CCI4 Time CCI4*Time ;
random CHCI3 CCI4*CHCI3 CHCI3*Time CCI4*CHCI3*Time /s;
repeated / subject=Flask(CCI4*CHCI3) type=arh(1) r;
```

```
lsmeans CCI4 / pdiff cl adjust=tukey;
run;


proc mixed data=FRF method=reml cl ic covtest;
class CCI4 CHCI3 Time Flask ;
model Leakage=CCI4 Time CCI4*Time ;
random CHCI3 CCI4*CHCI3 CHCI3*Time CCI4*CHCI3*Time /s;
repeated / subject=Flask(CCI4*CHCI3) type=ar(1) r;
lsmeans CCI4 / pdiff cl adjust=tukey;
run;
```

/* **Repeated Measures in Split-plot Design for RFF Model** */

```
Proc mixed data=RFF covtest;
class CCI4 CHCI3 Flask Time;
model Leakage = CHCI3 Time Time*CHCI3 /ddfm=satterthwaite;
random Flask CCI4 CHCI3*CCI4 Time*CCI4 Time*CHCI3*CCI4 Flask(CCI4 CHCI3);
title 'Split-plot Design: RFF';
run;
```

/* **Scrapping Data for the Combined Model FB** */

```
Data FB-univ;
set LDH-univ ;
if (CHCI3 =10) then output ;
if (CHCI3 =25) then output ;
run;
```

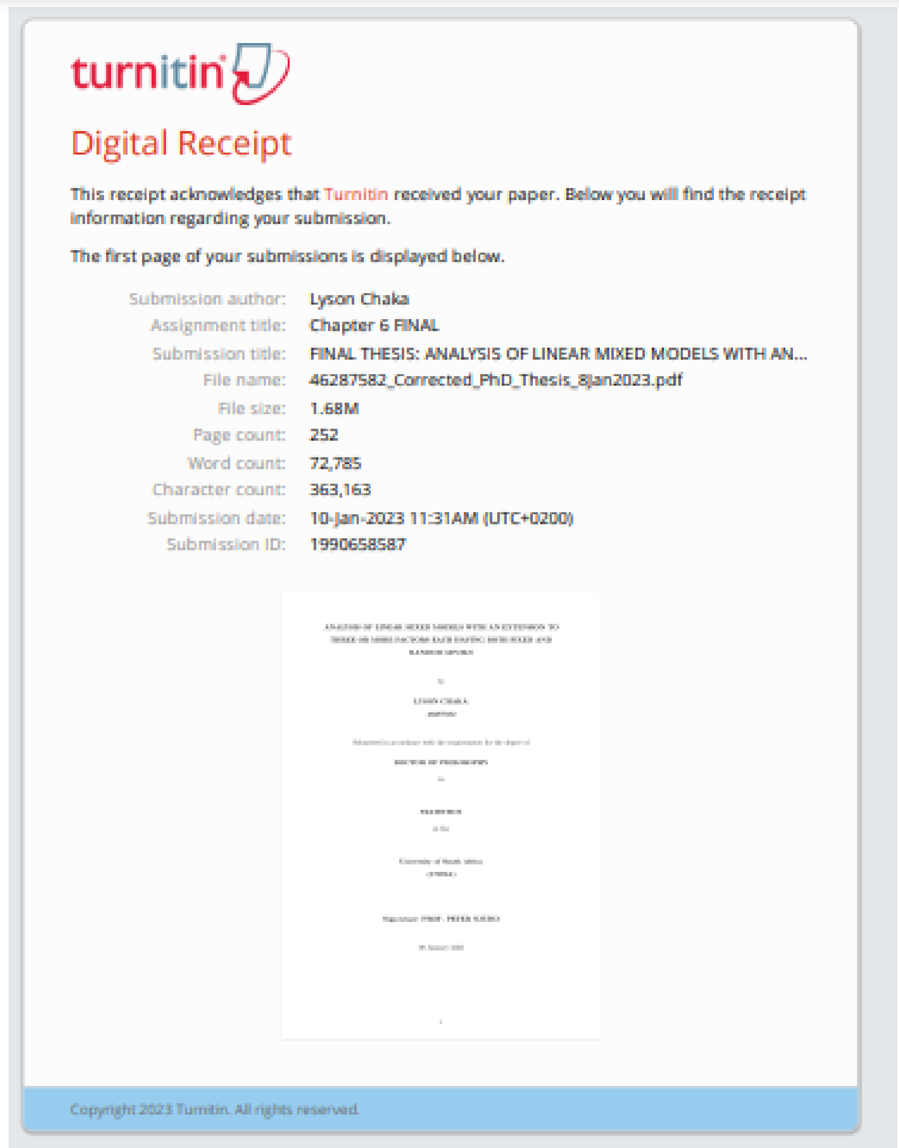
/* **Scrapping Data for the FA x FB Combined Model** */

```
Data FAFB-univ;
set LDH-univ ;
if (CCI4=2.5 and CHCI3 =10) then output ;
if (CCI4=2.5 and CHCI3 =25) then output ;
if (CCI4=5 and CHCI3 =10) then output ;
if (CCI4=5 and CHCI3 =25) then output ;
run;
Proc print data=FB-univ;
run;


/* Fitting the combined model FB (for narrow inferential scope) */
proc mixed data=FB-univ method=reml cl ic covtest;
class CCI4 CHCI3 Time Flask ;
model Leakage= CCI4 CHCI3 CCI4*CHCI3 Time CCI4*Time CHCI3*Time CCI4*CHCI3*Time
/s;
repeated / subject=Flask(CCI4*CHCI3) type=arh(1) r;
run;


/* Fitting the combined model FB (for broad inferential scope) */
proc mixed data=FB-univ method=reml cl ic covtest;
class CCI4 CHCI3 Time Flask ;
model Leakage= CHCI3 Time CHCI3*Time /s;
random CCI4 CCI4*CHCI3 CCI4*Time CCI4*CHCI3*Time;
repeated / subject=Flask(CCI4*CHCI3) type=ar(1) r;
run;
/************************************************************************/
```

```
/** GLIMMIX Syntax for Combined Model (FA) **/
data comb;
set LDH-univ;
if CCI4 in (2.5 , 5) then group=’fixed’; else group=’random’;
run;


proc print data=comb;
run;


data combined;
set comb;
if group=’fixed’ then dummy=’1’;else dummy=’0’;
run;


proc print data=combined;
run;

proc glimmix data=combined;
class CCI4 CHCI3 Flask Time group dummy;
model Leakage=group group*CCI4 /ddfm=satterthwaite;
random CHCI3*Time*dummy;
random _residual_/group=group;
random Time /subject=Flask type=arh(1) ;
lsmeans group group*CCI4/pdiff cl alpha=0.025 /* adjustment for 2 tests, at alpha=0.05*/;
run;
/*************************************************************************/
```

```
/** GLIMMIX Syntax for Combined Model (FAxFB) **/
data combAB;
set LDH-univ;
if (CCI4 in (2.5, 5.0) and CHCI3 in (10, 25)) then group=’fixed’;else group=’random’;
run;


proc print data=combAB;
run;


data combinedAB;
set combAB;
if group=’fixed’ then dummy=’1’;else dummy=’0’;
run;


proc print data=combinedAB;
run;


proc glimmix data=combinedAB;
class CCI4 CHCI3 Flask Time group dummy;
model Leakage=group group*CCI4 group*CHCI3 /ddfm=satterthwaite;
random Time*dummy;
random _residual_/group=group;
random Time /subject=Flask type=ar(1) ;
lsmeans group group*CCI4 group*CHCI3/pdiff cl alpha=0.0125 /*adjustment for 4 tests, at
alpha=0.05*/;
run;


/****************************************************************************/
```

Appendix E: Plagiarism Digital Report



Digital Receipt

This receipt acknowledges that Turnitin received your paper. Below you will find the receipt information regarding your submission.

The first page of your submissions is displayed below.

Submission author:	Lyson Chaka
Assignment title:	Chapter 6 FINAL
Submission title:	FINAL THESIS: ANALYSIS OF LINEAR MIXED MODELS WITH AN...
File name:	46287582_Corrected_PhD_Thesis_8Jan2023.pdf
File size:	1.68M
Page count:	252
Word count:	72,785
Character count:	363,163
Submission date:	10-Jan-2023 11:31AM (UTC+0200)
Submission ID:	1990658587

ANALYSIS OF LINEAR MIXED MODELS WITH AN EXTENSION TO THREE OR MORE FACTORS EACH HAVING BOTH FIXED AND RANDOM EFFECTS

by

LYSON CHAKA
[illegible]

Submitted in accordance with the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

STATISTICS

at the

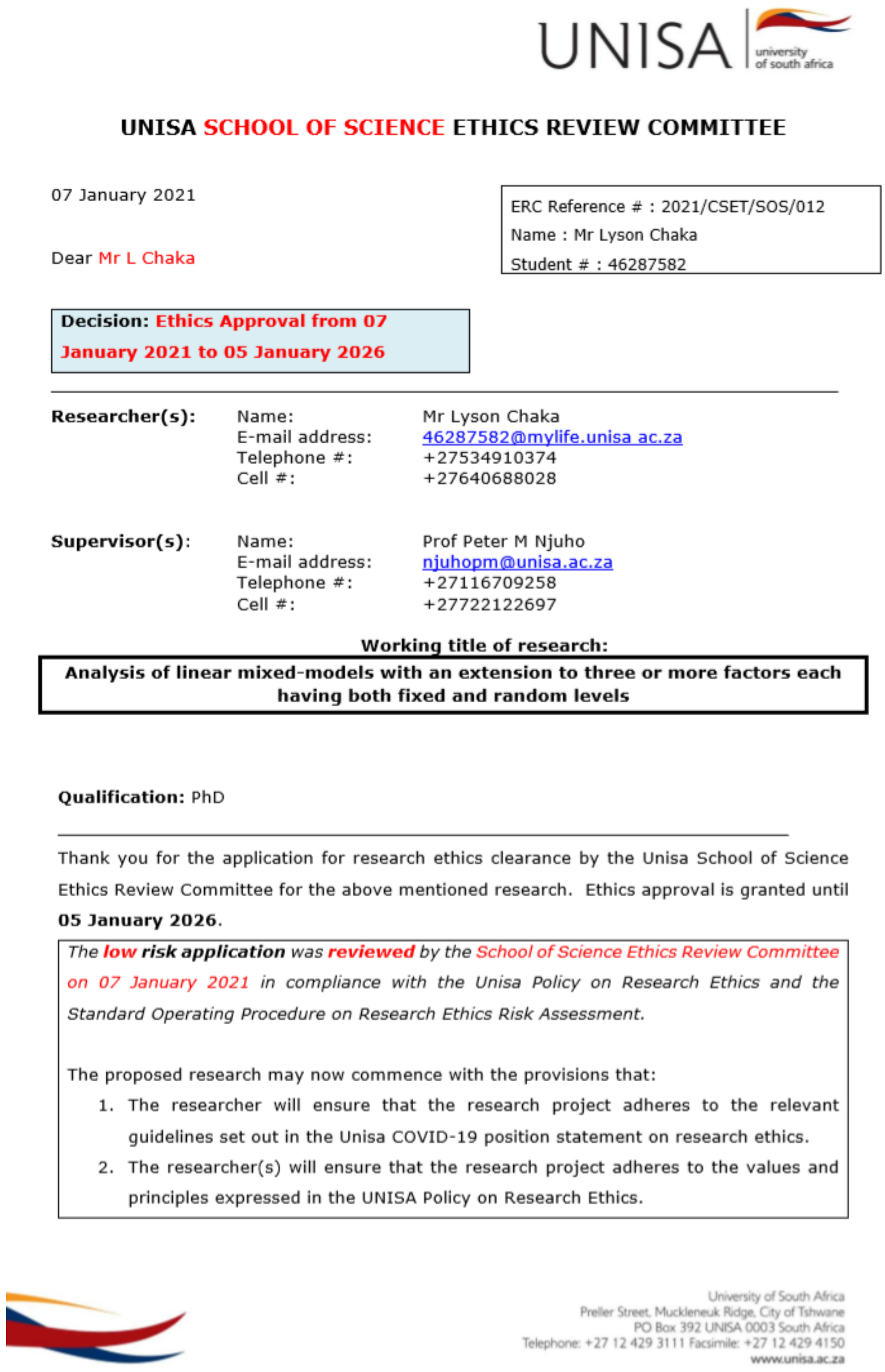
University of South Africa
(UNISA)

Supervisor: PROF. PETER NJUHO

[illegible]

i

Appendix F: Ethical Clearance Approval



UNISA SCHOOL OF SCIENCE ETHICS REVIEW COMMITTEE

07 January 2021

ERC Reference # : 2021/CSET/SOS/012
Name : Mr Lyson Chaka
Student # : 46287582

Dear Mr L Chaka

Decision: Ethics Approval from 07 January 2021 to 05 January 2026

Researcher(s):	Name:	Mr Lyson Chaka
	E-mail address:	46287582@mylife.unisa.ac.za
	Telephone #:	+27534910374
	Cell #:	+27640688028
Supervisor(s):	Name:	Prof Peter M Njuho
	E-mail address:	njuhopm@unisa.ac.za
	Telephone #:	+27116709258
	Cell #:	+27722122697

Working title of research:

Analysis of linear mixed-models with an extension to three or more factors each having both fixed and random levels

Qualification: PhD

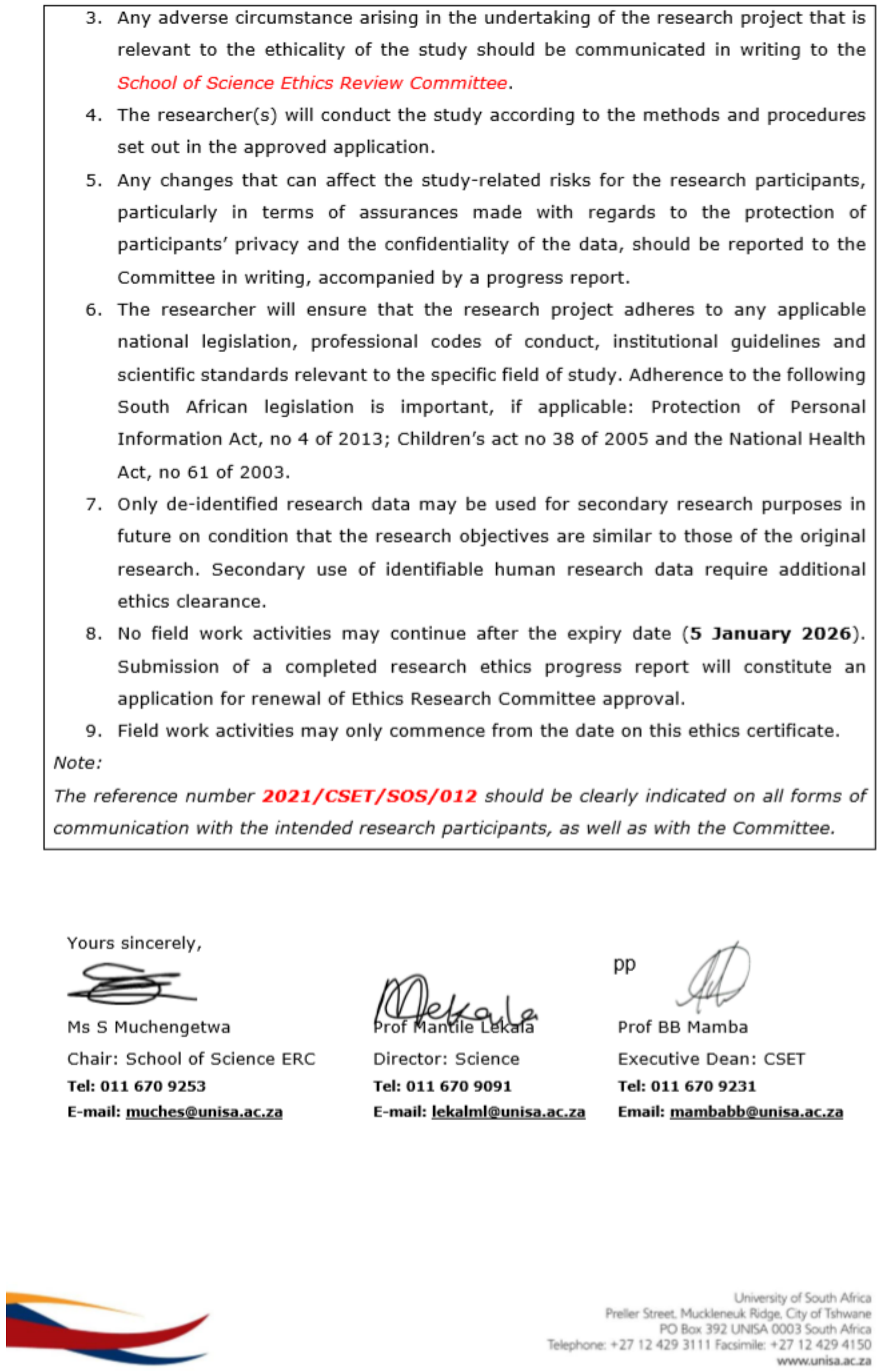
Thank you for the application for research ethics clearance by the Unisa School of Science Ethics Review Committee for the above mentioned research. Ethics approval is granted until **05 January 2026**.

*The **low risk application** was **reviewed** by the School of Science Ethics Review Committee on 07 January 2021 in compliance with the Unisa Policy on Research Ethics and the Standard Operating Procedure on Research Ethics Risk Assessment.*

The proposed research may now commence with the provisions that:

1. The researcher will ensure that the research project adheres to the relevant guidelines set out in the Unisa COVID-19 position statement on research ethics.
2. The researcher(s) will ensure that the research project adheres to the values and principles expressed in the UNISA Policy on Research Ethics.

University of South Africa
Preller Street, Muckleneuk Ridge, City of Tshwane
PO Box 392 UNISA 0003 South Africa
Telephone: +27 12 429 3111 Facsimile: +27 12 429 4150
www.unisa.ac.za



3. Any adverse circumstance arising in the undertaking of the research project that is relevant to the ethicality of the study should be communicated in writing to the *School of Science Ethics Review Committee*.
4. The researcher(s) will conduct the study according to the methods and procedures set out in the approved application.
5. Any changes that can affect the study-related risks for the research participants, particularly in terms of assurances made with regards to the protection of participants' privacy and the confidentiality of the data, should be reported to the Committee in writing, accompanied by a progress report.
6. The researcher will ensure that the research project adheres to any applicable national legislation, professional codes of conduct, institutional guidelines and scientific standards relevant to the specific field of study. Adherence to the following South African legislation is important, if applicable: Protection of Personal Information Act, no 4 of 2013; Children's act no 38 of 2005 and the National Health Act, no 61 of 2003.
7. Only de-identified research data may be used for secondary research purposes in future on condition that the research objectives are similar to those of the original research. Secondary use of identifiable human research data require additional ethics clearance.
8. No field work activities may continue after the expiry date (**5 January 2026**). Submission of a completed research ethics progress report will constitute an application for renewal of Ethics Research Committee approval.
9. Field work activities may only commence from the date on this ethics certificate.

Note:

The reference number ***2021/CSET/SOS/012*** *should be clearly indicated on all forms of communication with the intended research participants, as well as with the Committee.*

Yours sincerely,

Ms S Muchengetwa
Chair: School of Science ERC
Tel: 011 670 9253
E-mail: muches@unisa.ac.za

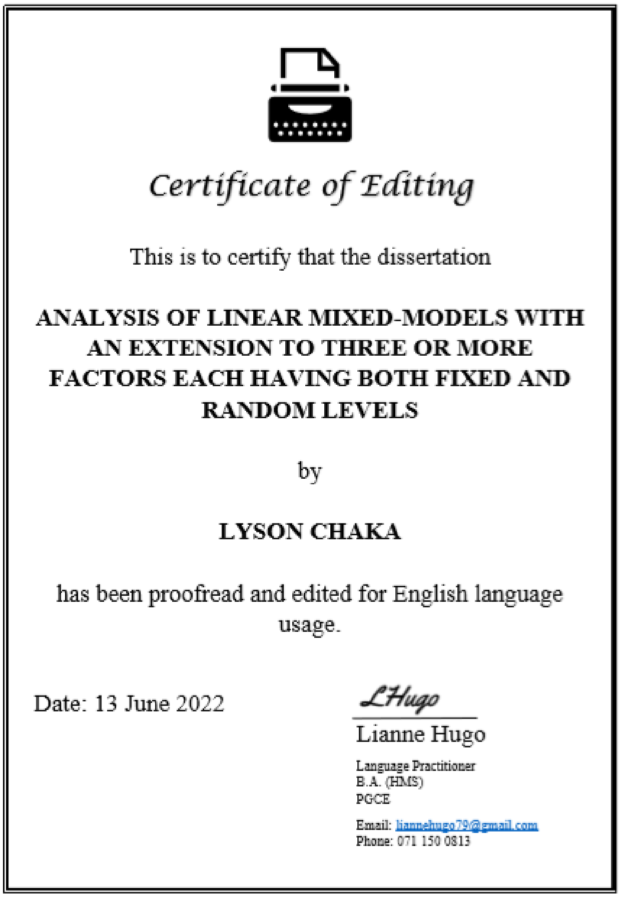
Prof Mantile Lekala
Director: Science
Tel: 011 670 9091
E-mail: lekalml@unisa.ac.za

pp

Prof BB Mamba
Executive Dean: CSET
Tel: 011 670 9231
Email: mambabb@unisa.ac.za

University of South Africa
Preller Street, Muckleneuk Ridge, City of Tshwane
PO Box 392 UNISA 0003 South Africa
Telephone: +27 12 429 3111 Facsimile: +27 12 429 4150
www.unisa.ac.za

Appendix G: Language Editing Certificate



Certificate of Editing

This is to certify that the dissertation

ANALYSIS OF LINEAR MIXED-MODELS WITH AN EXTENSION TO THREE OR MORE FACTORS EACH HAVING BOTH FIXED AND RANDOM LEVELS

by

LYSON CHAKA

has been proofread and edited for English language usage.

Date: 13 June 2022

LHugo

Lianne Hugo

Language Practitioner
B.A. (HMS)
PGCE

Email: liannehugo79@gmail.com
Phone: 071 150 0813