UNIVERSITY OF KWAZULU-NATAL

MODELING THE FACTORS AFFECTING CEREAL CROP YIELDS IN THE AMHARA NATIONAL REGIONAL STATE OF ETHIOPIA

2010

YUNUS HUSSIEN

MODELING THE FACTORS AFFECTING CEREAL CROP YIELDS IN THE AMHARA NATIONAL REGIONAL STATE OF ETHIOPIA

By

YUNUS HUSSIEN

Submitted in fulfilment of the academic

requirements for the degree of

MASTER OF SCIENCE

in

Biometry

in the

School of Statistics and Actuarial Science

University of KwaZulu-Natal

Pietermaritzburg

2010

Dedication

To my father, Ato Hussien Mohammed and my mother, W/ro Merema Mohammed.

Declaration

The research work described in this thesis was carried out in the School of Statistics and Actuarial Science, at the University of KwaZulu-Natal, Pietermaritzburg, under the supervision of Dr. Shaun Ramroop and Prof. Temesgen Zewotir.

I, Mr.Yunus Hussien, declare that this thesis is my own unaided work. It has not been submitted in any form for any degree or diploma to any other university. Where the work of others has been used, it is duly acknowledged.

May, 2010.

_____ _____

Mr. Yunus Hussien Date

_____ _____

Dr. Shaun Ramroop Date

_____ _____

Prof. Temesgen Zewotir Date

Acknowledgements

First of all, I praise the Almighty Allah (God) for all what I have ever achieved. Further, I wish to express my indebted gratitude and sincere thanks to my supervisor Dr. Shaun Ramroop for his encouragement, advice and positive criticism throughout the entire period of my research. My sincere indebtedness is also due to my co-supervisor Prof. Temesgen Zewotir for his support, guidance and advice while conducting the research. I am also respectful to Prof. Mwambi for his invaluable suggestions and comments at different stages of writing the thesis.

I want to extend my appreciation for the help and pieces of advice that I received from my colleagues, specially Dawit Getinet, and to all my friends for their cooperation and encouragement.

I hardly find any word to express my thankfulness to my parents, my eldest sister Sadiya Hussien, my brother Omar Hussien, and all my beautiful sisters, nephews and nieces who tolerated my negligence from domestic duties with patience.

I am grateful to my organization C.S.A and the management body for all their support during the study period. Besides, the financial support from UNDP, which enabled me to accomplish the study, is gratefully acknowledged.

I am truly thankful to all the staff members of the Department of Statistics and Actuarial Science, University of KwaZulu-Natal, for offering me their sincere help when and where I needed.

Abstract

The agriculture sector in Amhara National Regional State is characterised by producing cereal crops which occupy the largest percentage (84.3%) of the total crop area cultivated in the region. As a result, it is imperative to investigate which factors influence the yields of cereal crops particularly in relation to the five major types of cereals in the study region namely barley, maize, sorghum, teff and wheat. Therefore, in this thesis, using data collected by the Central Statistical Agency of Ethiopia, various statistical methods such as multiple regression analysis were applied to investigate the factors which influence the mean yields of the major cereal crops. Moreover, a mixed model analysis was implemented to assess the effects associated with the sampling units (enumeration areas), and a cluster analysis to classify the region into similar groups of zones.

The multiple regression results indicate that all the studied cereals mean yields are affected by zone, fertilizer type and crop damage effects. In addition to this, barley is affected by extension programme; maize crop by seed type, irrigation, and protection of soil erosion; sorghum and teff crops are additionally affected by crop prevention method, extension programme, protection of soil erosion, and gender of the household head; and wheat crop by crop prevention methods, extension programme and gender of the household head. The results from the mixed model analysis were entirely different from the regression results due to the observed dependencies of the cereals mean yields on the sampling unit. Based on the hierarchical cluster analysis, five groups of classes (clusters) were identified which seem to be in agreement with the geographical neighbouring positions of the locations and the similarity of the type of crops produced.

Table of Contents

Dedication.....**i**

Declaration.....**ii**

Acknowledgements.....**iii**

Abstract.....**iv**

Chapter 1 Background and Objectives of the Study.....**1**

1.1 Introduction.....1

1.2 The Study Area.....3

1.3 Objective and Significance of the Study.....6

Chapter 2 The Data Set and Preliminary Analysis.....**7**

2.1 The Data Set.....7

2.1.1 Source of the Data.....7

2.1.2 Sampling Frame, Design and Coverage.....7

2.1.3 Data Collection.....8

2.1.4 Variables of Interest.....8

2.2 Preliminary Analysis of the Data.....10

2.3 Transformation of the Data.....18

2.4 Summary on the Preliminary Analysis of the Data.....22

Chapter 3 Theory and Application of Multiple Regression Analysis.....**23**

3.1 Introduction.....23

3.2 Multiple Linear Regression Models.....24

3.2.1 Estimating Parameters of the Model.....25

3.2.2 Model Selection and Inferences.....27

3.2.3Model Diagnostics.....32

3.3 Application of the Multiple Linear Regression Model to Crop Yield Data.....33

3.3.1 Fitting a Model for Transformed Yields of the Data.....35

3.4 Discussions and Conclusions on Results of the Multiple Regression.....45

Chapter 4 Theory and Application of Mixed Model.....50

4.1 Introduction.....50

4.2 The General Linear Mixed Model.....51

4.3 Estimation of the Model Parameters.....53

4.3.1 Maximum Likelihood Estimation.....54

4.3.2 Restricted Maximum Likelihood Estimation.....55

4.4 The Covariance Structure.....56

4.5 Interpretation of Parameter Estimates (Predictions).....57

4.6 Model Selection and Diagnostics.....58

4.7 Application of the Mixed Models to the Data.....60

4.8 Discussions and Conclusions on Results of the Mixed Models 62

Chapter 5 Review of Cluster Analysis and its Application.....68

5.1 Introduction.....68

5.2 Hierarchical Clustering.....68

5.3 Discussions and Conclusions on Results of Clustering the Data.....71

Chapter 6 Summary and Conclusions.....75

References.....80

Appendix A: Additional Tables.....85

List of Figures

Fig. 1.1 Map of Ethiopia highlighting the Amhara National Regional State.....4

Fig. 1.2 Map of the study region (Amhara National Regional State).....5

Figure 2.1 Frequency distributions of the original yield data by crop types.....13

Figure 2.2 Box plot showing the variability of mean yield by crop types.....14

Figure 2.3 The relationship between yield and seed type, fertilizer type, crop prevention measures,and crop damage by crop type.....15

Figue 2.4 The relationship between yield and extension, prevention of soil erosion, irrigation, and gender of head of the household by crop type.....16

Figure 2.5 Frequency distributions and the Normal Q – Q Plots for transformed cereals yield by crop type.....21

Figure 3.3.1 Frequency distribution and p-p plots of the fitted models residuals by crop type.....39

Figure 3.3.2 Scatter plots and Cook's distance plots for the fitted models by crop type.....41

Figure 5.1 Dendrogram of zones obtained from cluster analysis of complete linkage method.....74

List of Tables

Table 2.1 Summary of Description of the Categorical Variables.....9

Table 2.2 Mean Yield in Quintal per Hectare by Crop Type.....10

Table 2.3 Summary of Frequency Percentages for Use of Agricultural Inputs and Practices on Cereal Crop Farms.....11

Table 2.4 Tests of Normality Results for the Original and Transformed Yield Data by Crop Types.....19

Table 2.5 Skewness and Kurtosis Results for Selecting Transformation Methods..20

Table 3.1 Regression Analysis of Variance (ANOVA) Table.....30

Table 3.3.1 Regression Analysis of Variance Table for Models of Transformed Cereals Yields by Crop Type.....37

Table 3.3.2 Estimates of the Regression Parameter Coefficients for the Fitted Models by Crop Type (Standard Deviations in Parentheses).....43

Table 3.3.3 The 95% Confidence Interval for Parameter Estimates of the Fitted Models by Crop Types.....44

Table 4.1 Unconditional Models Covariance Parameter Estimates by Crop Type.....63

Table 4.2 Conditional Models Covariance Parameter Estimates by Crop Type.....63

Table 4.3 Type 3 Tests of Fixed Effects for Transformed Barley Data.....65

Table 4.4 Type 3 Tests of Fixed Effects for Transformed Maize Data.....65

Table 4.5 Type 3 Tests of Fixed Effects for Transformed Sorghum Data.....65

Table 4.6 Type 3 Tests of Fixed Effects for Transformed Teff Data.....66

Table 4.7 Type 3 Tests of Fixed Effects for Transformed Wheat Data.....66

Table 5.1 Cluster History of the Complete Linkage Cluster Analysis.....72

Table 6.1 Summary for the Type 3 Tests of Significance for Fixed Effects by Crop Type.....76

Table A.1 Model Summary for Transformed Barley Data.....85

Table A.2 Model Summary for Transformed Maize Data.....85

Table A.3 Model Summary for Transformed Sorghum Data 86

Table A.4 Model Summary for Transformed Teff Data 86

Table A.5 Model Summary for Transformed Wheat Data 87

Table A.6 Model Summary for Transformed Barley Data without Zone Effect.....87

Table A.7 Model Summary for Transformed Maize Data without Zone Effect.....88

Table A.8 Model Summary for Transformed Sorghum Data without Zone Effect.88

Table A.9 Model Summary for Transformed Teff Data without Zone Effect.....89

Table A.10 Model Summary for Transformed Wheat Data without Zone Effect.....89

Table A.11 Type 3 Tests of Fixed Effects for Transformed Barley Data (no random factor, EAs).....90

Table A.12 Type 3 Tests of Fixed Effects for Transformed Maize Data (no random factor, EAs).....90

Table A.13 Type 3 Tests of Fixed Effects for Transformed SorghumData (no factor, EAs).....90

Table A.14 Type 3 Tests of Fixed Effects for Transformed Teff Data (no random factor, EAs).....91

Table A.15 Type 3 Tests of Fixed Effects for Transformed Wheat Data (no random factor, EAs).....91

Chapter 1

Background and Objectives of the Study

1.1 Introduction

Ethiopia is located between 3^0-15^0 N latitude and 33^0-48^0 E longitude of the equator. The country covers a land area of about 1.12 million km^2 in the east of Africa. Ethiopia is administratively sub-divided into nine regional states and two city administrations. The population of Ethiopia, according to the 2007 Population and Housing Census preliminary report by the Central Statistical Agency (CSA), was estimated at 73,918,505 people. Of these, 37,296,657 (50.5%) were males and 36,621,848 (49.5%) females. About 84 percent of the total population in the country were found to live in rural areas while the remaining 16 percent lived in urban areas (CSA, 2008).

Ethiopia has different types of climate ranging from semi-arid desert in the lowlands to humid and warm (temperate) in the southwest. The mean annual rainfall distribution has a maximum of over 2000mm and a minimum of less than 300mm over the South-eastern and North-eastern lowlands. The mean annual temperature ranges from a minimum of 15 ^{0}C over the highlands to a maximum of over 25 ^{0}C in the lowlands (NMSA, 2001).

Agriculture is the most important production sector of the country's economy. It provides about 85% of the total employment for the population and contributes about 50% to the country's gross domestic product (GDP). It supplies around 70% of the raw material requirement of agro-based domestic industries (MEDaC, 1999). It is also the

major source of food for the nation and hence the prime contributing sector to food security. In addition, agriculture is expected to play a key role to speed up the overall socio-economic development of the country.

Though agriculture is the backbone of the country's economy, it is dominated by small-scale farmers who have been implementing low input with traditional farming technologies. For this reason the government of Ethiopia has been introducing agricultural extension services based on its strategy for "Agricultural Development-led Industrialisation" starting from the early 1990s to address the use of fertilizer, improved seeds, pesticides, irrigation, and other inputs which are expected to play a major role in increasing crop production. Thus far there have been improvements in the use of modern agricultural inputs by subsistence farmers but the country's agricultural sector is still suffering from the problem of low productivity, shortage of productive farm land, and persistent rural poverty (Samuel, 2006). Bakhsh et al. (2005), studied the factors affecting cotton yield by applying multiple regression method, and their results show that land preparation, irrigation, seed rate, plant protection measures, fertilizer nutrients, and the number of schooling years of respondents were important variables in the production process. It is hence the objective of this study to identify and assess which of the above-mentioned inputs, among other factors, are influencing cereal crop production in the study region, the Amhara National Regional State.

Cereals are the major food crops both in terms of the areas they are planted in and the volumes of production. They are produced in larger quantities when compared with other crops because they are the principal staple crops. Of the total grain crop area cultivated during the 2006/2007 main agricultural season, 79.98% was covered under

cereal crops, 76.05% of which was covered by barley, maize, sorghum, teff[1] and wheat. With respect to the production contribution, these crops made-up 82.55% of the total cereal production in the region (CSA, 2007). For this reason, it is vital to focus on these five main crop types to investigate the factors that affect cereal production or yield. Various statistical methods such as regression analysis are employed to investigate the effects of these factors on crop yield. Moreover, to investigate the effects of the random probability sampling units (i.e. the enumeration areas), a general linear mixed model is applied to the data.

1.2 The Study Area

The Amhara National Regional State (ANRS) which occupies much of north western and north central part (see on Fig 1.1) of Ethiopia, is located between 9°20' and 14°20' North latitude and 36° 20' and 40° 20' East longitude.

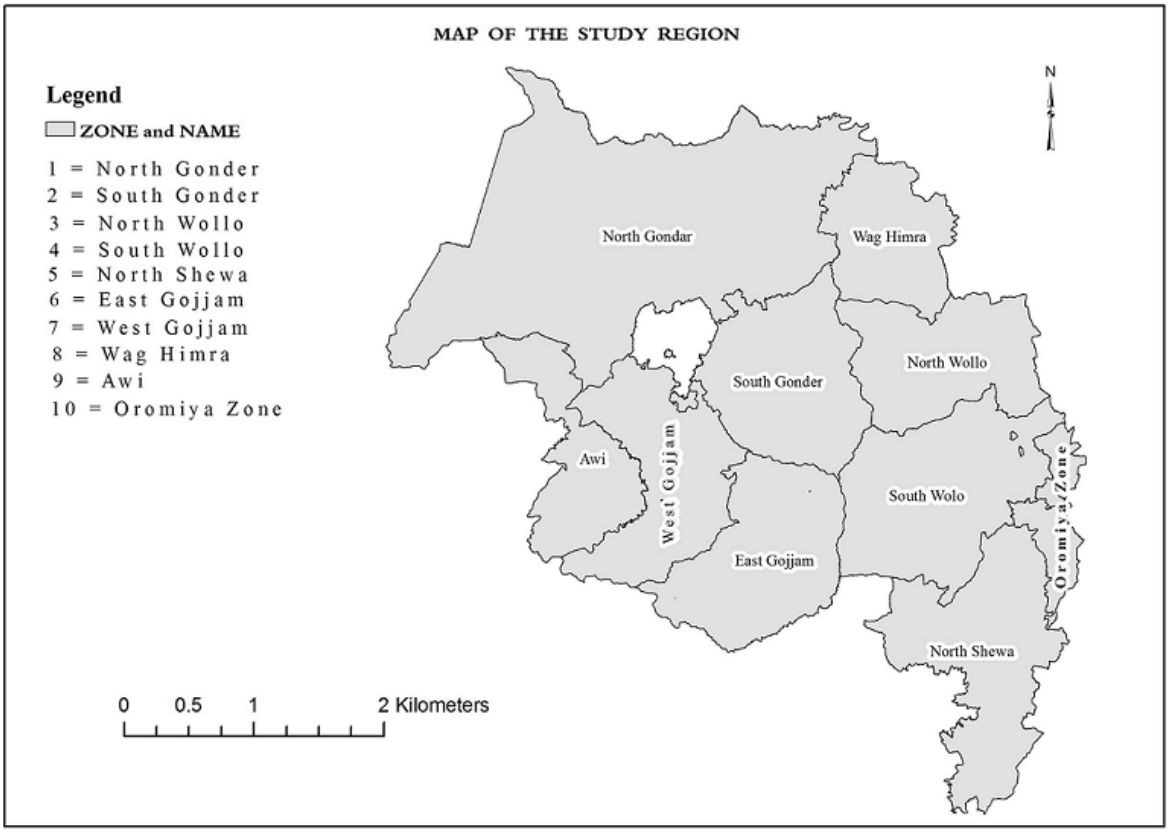
The region is administratively divided into 11 zones[2], 140 districts (locally called 'weredas') and about 3429 localities (called 'kebeles') which are the smallest administrative settings. Based on the 2008 census results reported by the CSA, the region has a total of 17,214,056 people of whom 8,636,875 were men and 8,577,181 women; with an estimated area of 159,173.66 square kilometers and a population density area of 108.15 people per square kilometer (CSA, 2008).

[1] Teff , *Eragrostis tef.* (Zucc.) Trotter, is one of the most important cereals which is endemic to Ethiopia (Alemu, 2005).

[2] Zones are the higher administrative settings dividing the regional state into 11 sceneries.

Figure 1.1 Map of Ethiopia highlighting the Amhara National Regional State

Of the total population 88% depend on agriculture for their livelihood and it accounts for about 55.8% of the GDP of the regional state. Crop production and animal husbandry are the major agricultural activities in the region. With regard to crop production of all crop types, cereals, pulses, oil crops, fibre crops, fruits and vegetables are grown in different parts of the region and cereals account for the highest percentage (84.3%) of cultivated area and 85% of the total production (CSA, 2007).



Source: CSA Department of Cartography, 2007.

Figure 1.2 Map of the study region (Amhara National Regional State)

The 11administrative Zones include one special zone, "Bahir-Dar", which is the capital city of the regional state. However, the agricultural sample survey (2006/07) of the CSA was undertaken by merging the region's capital city within the West-Gojjam zone (Zone7) , as shown on the map (Fig. 1.2), and making a total number of 10 zones in the region. These 10 administrative zones include: North Gondar (Zone1), South Gondar (Zone2), North Wollo (Zone3), South Wollo (Zone4), North Shewa (Zone5), East Gojam (Zone6), West Gojam (Zone7), Wag Himra (Zone8), Awi (Zone9), and Oromia Zone (Zone10). These 10 zones were considered for analysis of the agricultural data in this thesis.

1.3 Objective and Significance of the Study

The main objective of this study is to identify important factors affecting the main cereal crop production in the Amhara National Regional State. The other objective is to group the zones by their cereal production type using cluster analysis. Such a classification is useful to make relevant decisions with regard to:

i) distributing agricultural inputs such as improved seeds, chemical fertilizer, pesticides and insecticides in the region,

ii) marketing a particular cereal crop produced in different parts of the region, and

iii) identifying the zones in the region that are grouped into classes which have similar set-ups with regard to a particular cereal production.

The identification of major input factors affecting cereal crop production are necessary for the assessment , evaluation, and formulation of programmes and policies being put in place to overcome the primary obstacles in the agricultural sector and to identify avenues for future research. The results of this work will also contribute to the literature based on the impact of main agricultural inputs on cereal crop yield in the region.

Chapter 2

The Data Set and Preliminary Analysis

2.1 The Data Set

2.1.1 Source of the Data

The data used in this study were drawn from the main season agricultural sample survey (2006/2007) results conducted by the Central Statistical Agency of Ethiopia in the Amhara National Regional State.

2.1.2 Sampling Frame, Design and Coverage

The CSA (2007) report on area and production of crops indicated that the sampling frame was obtained from the lists of the 2001/02 Ethiopian agricultural sample enumeration. A stratified two-stage cluster sample design was used to select the enumeration areas (EAs) as the primary sampling units, and the agricultural households as secondary sampling units. Enumeration areas from each stratum were systematically selected using a probability proportional to size sampling technique, and from the new list of households within each sample of EAs, a random sample of 20 agricultural households were systematically selected (CSA, 2007).

The survey took place in the rural parts of the 10 administrative zones in the Amhara National Regional State. A total of 8,800 households from 440 selected EAs were planned to be included in the study, however 8,768 households (99.63%) and 439 EAs (99.77%) were successfully covered.

2.1.3 Data Collection

The agricultural data for the year 2006/07 was collected from randomly selected rural agricultural households on cereals, pulses, oilseeds, vegetables, root crops and fruit crops by interviewing and undertaking physical measurements on their fields (CSA, 2007).

2.1.4 Variables of Interest

In our study a number of variables which we assumed to have a potential effect on the main cereal crop production are selected from the 2006/2007 agricultural sample survey data collected by CSA. The variables to be considered are:

Dependent variable: Cereal crop yield

Independent variables: 1) Seed type, 2) Fertilizer type, 3) Extension programme, 4) Type of crop prevention, 5) Crop damage, 6) Protection of soil erosion, 7) Crop irrigation, 8) Gender of the household head, and 9) Zone (the administrative area settings in which the crops are cultivated).

The categorical variables with their corresponding codes and descriptions are summarized as shown in Table 2.1.

Table 2.1 Summary of Description of the Categorical Variables

Factors	Variable Names	Category code	Description of Variable
Seed type	Seedtype1	1	Improved
	Seedtype2	2	Non-improved
Fertilizer type	Fertliz0	0	No fertilizer used
	Fertliz1	1	Natural fertilizer
	Fertliz2	2	Chemical fertilizer
	Fertliz3	3	Both natural and chemical
Extension programme	Ext1	1	Included in the program
	Ext2	2	Not included in the program
Crop prevention methods	Cropprev0	0	No prevention
	Cropprev1	1	Non-chemical prevention
	Cropprev2	2	Chemical-type of prevention
	Cropprev3	3	Both chemical and non-chemical
Crop damage	Damage1	1	Yes
	Damage2	2	No
Crop irrigation	Irrg1	1	Yes
	Irrg2	2	No
Soil erosion protection	Serrop1	1	Yes
	Serrop2	2	No
Gender of the household head	HHsex1	1	Male
	HHsex2	2	Female
Zone	Zone1	01	Zone1 (N.Gondar)
	Zone2	02	Zone2 (S.Gondar)
	Zone3	03	Zone3 (N.Wello)
	Zone4	04	Zone4 (S.Wello)
	Zone5	05	Zone5 (N.Shewa)
	Zone6	06	Zone6 (E.Gojam)
	Zone7	07	Zone7 (W.Gojam)
	Zone8	08	Zone8 (Wag-Hemra)
	Zone9	09	Zone9 (Awi)
	Zone10	10	Zone10 (Oromia Zone)

2.2 Preliminary Analysis of the Data

In order to gain some understanding of the data, an exploratory data analysis was carried out and is presented in this section. The data was checked for the amount of mean yield in quintal (100kg) per hectare for each type of crop considered in this study and the result is presented in Table 2.2. The computed mean yield in quintal per hectare for barley, maize, sorghum, teff and wheat was found to be 12.9 qt/ha, 19.4 qt/ha, 16.3 qt/ha, 10.7 qt/ha, and 14.7 qt/ha with a standard deviation of 5.24, 8.87, 6.13, 4.38, and 5.82 respectively. From these sample means and standard deviations of the cereals, it can be seen that there is a large variation in the yields of the cereals across households in view of the fact that the standard deviations from their respective mean values are quite large. The observed variation in the yields of the cereals across households can be linked to differences in input usage on the farms and other additional factors.

Table 2.2 Mean Yield in Quintal per Hectare by Crop Type

Region	Crop Type	Mean Yield in Quintal per Hectare (Standard deviation)
Amhara	Barley	12.9 (5.24)
	Maize	19.4 (8.87)
	Sorghum	16.3 (6.13)
	Teff	10.7 (4.38)
	Wheat	14.7 (5.82)

The summary of farm holders' frequency percentage on their use of inputs and practices by crop types is displayed in Table 2.3.

Table 2.3 Summary of Frequency Percentages for Use of Agricultural Inputs and Practices on Cereal Crop Farms

Factors	Labels	Crop Types					Average for all crop types (%)
		Barley	Maize	Sorghum	Teff	Wheat	
Seedtype	Improved	.0%	13.1%	.0%	.3%	1.5%	2.98%
	Non-Improved	100.0%	86.9%	100.0%	99.7%	98.5%	97.02%
Fertliz	No fertilizer	67.8%	26.3%	80.4%	50.9%	48.9%	54.86%
	Natural Fertilizer	24.9%	47.2%	18.2%	9.9%	14.6%	22.96%
	Chemical Fertilizer	6.9%	21.7%	1.4%	38.2%	34.8%	20.60%
	Both Natural & Chem	.4%	4.8%	.1%	1.1%	1.7%	1.62%
Cropprev	No Prevention	8.6%	8.5%	10.1%	2.8%	8.7%	7.74%
	Chem Prev	.8%	.3%	1.5%	1.4%	2.9%	1.38%
	Non-Chem Prev	89.7%	90.1%	86.4%	91.2%	82.7%	88.02%
	Both Chem & Non-Chem	.8%	1.1%	1.9%	4.6%	5.8%	2.84%
Damage	Yes	39.8%	42.4%	46.3%	27.2%	30.1%	37.16%
	No	60.2%	57.6%	53.7%	72.8%	69.9%	62.84%
Ext	Included	3.9%	22.6%	1.6%	24.4%	19.8%	14.46%
	Not Included	96.1%	77.4%	98.4%	75.6%	80.2%	85.54%
Irrg	Yes	.9%	2.0%	.5%	.3%	.7%	0.88%
	No	99.1%	98.0%	99.5%	99.7%	99.3%	99.12%
Serrop	Yes	83.7%	75.8%	76.4%	85.9%	87.5%	81.86%
	No	16.3%	24.2%	23.6%	14.1%	12.5%	18.14%
HHsex	Male	89.4%	86.1%	89.9%	90.2%	89.6%	89.04%
	Female	10.6%	13.9%	10.1%	9.8%	10.4%	10.96%

With regard to seed type, 97.02% of the farmers applied local seed varieties (non-improved seed types) and only 2.98% of them used improved seed types. In referring to this factor by crop type, none of the sorghum and barley farm holders used improved

seed types on their farms. The highest users of improved seed types were seen on maize farms (13.1%); and there were only 1.5% and 0.3% users on wheat and teff farms respectively.

It is observed that about 54.86% of the main cereal crop producers did not use any type of fertilizer on their farms. Among these sorghum farm holders were the largest non-users (80.4%) followed by barley (67.8%), teff (50.9%), wheat (48.9%) and maize (26.3%). Chemical fertilizers were applied to approximately 20.6% of the crops, while its largest application was seen by teff farm holders and the smallest proportion was by sorghum farm holders. On average 37.16% of the cereal farm holders reported that there was crop damage during the production season, whereas 62.84% reported no crop damage. The highest damages of 46.3%, 42.4%, and 39.8% were reported for sorghum, maize and barley farms, in that order. In addition to the above tabular examination of the data, it was also assessed via visual representations through the use of histograms and box plots in order to check for normality of distribution and to verify the presence of outliers and extreme cases. The frequency distribution plots of the cereal yields by crop type are displayed on Fig. 2.1. The plots clearly show that all the cereal crops, namely barley, maize, sorghum, teff, and wheat, have a distribution pattern slightly skewed to the right.

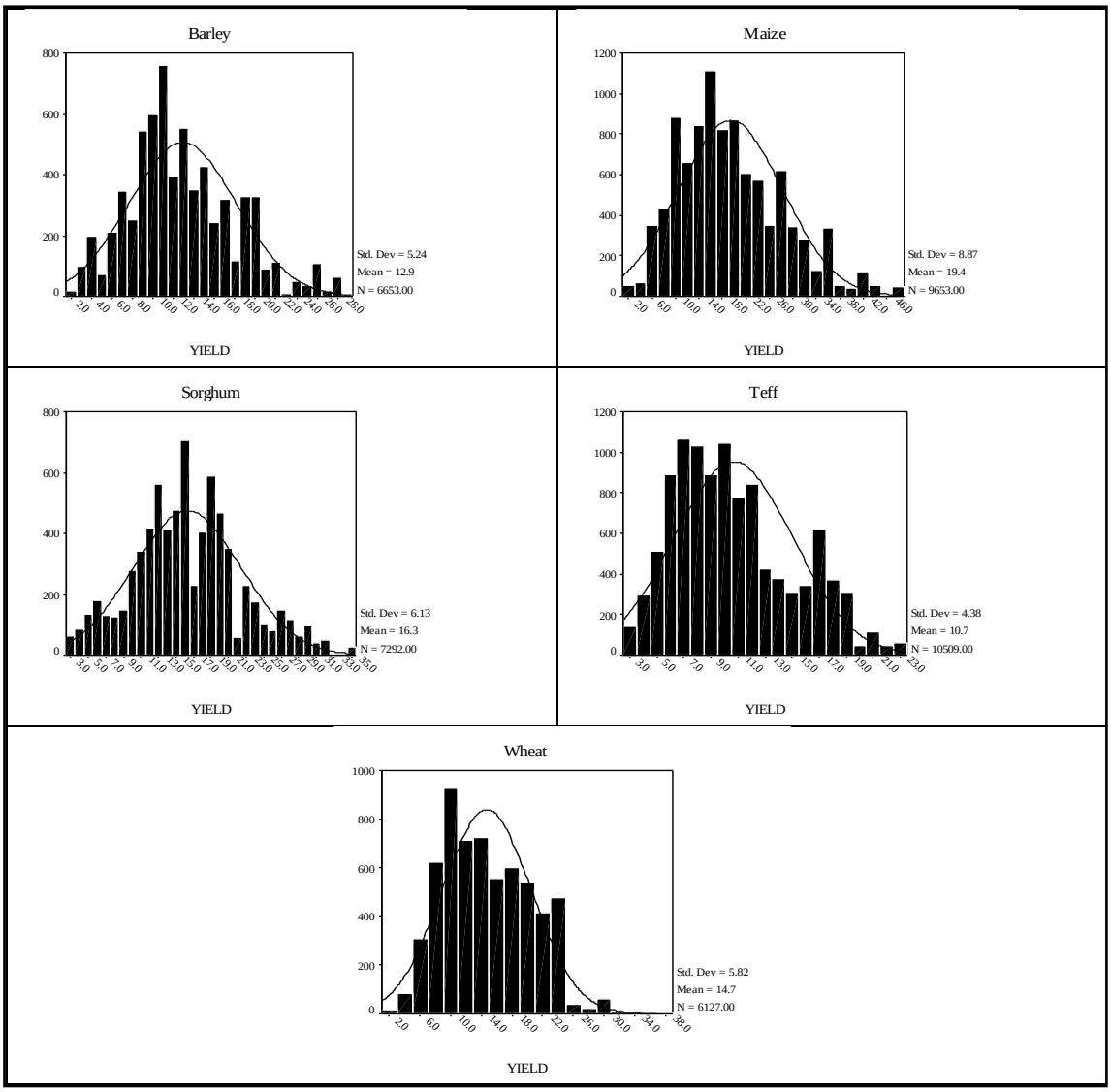


Figure 2.1 Frequency distributions of the original yield data by crop types

Furthermore, the variability in the mean yields of the cereals and their distribution with respect to the crop types as well as that of the independent factors were examined through the use of box plots of mean yields vs crop types, and mean yields vs individual independent factors. The plot in Fig. 2.2 indicates that the variation in the mean yields per hectare differs by crop type and is the highest for maize and the lowest for teff yields. There are few outliers on the high side of the mean yields, and for all cereal types the distribution is found to be slightly skewed towards higher values.

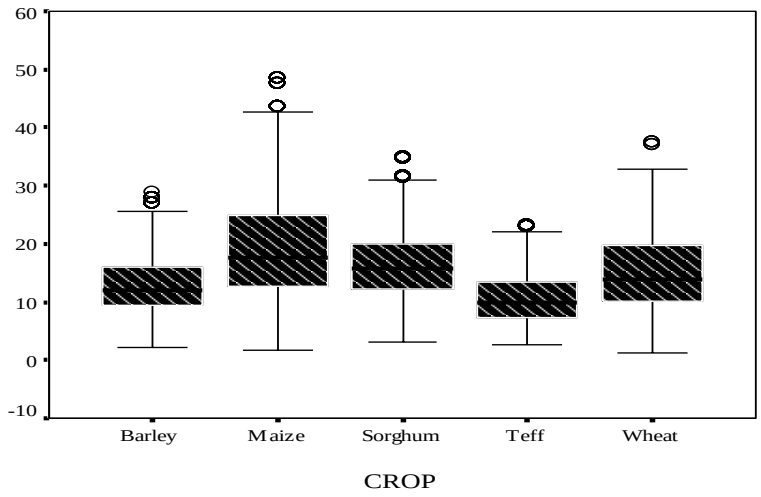


Figure 2.2 Box plot showing the variability of mean yield by crop types

The box plots displayed in Fig. 2.3 and Fig.2.4 (in groups), which indicate the variability pattern in the mean yields of cereals associated with respect to each of the independent factors included in this study, were summarised as follows:

Seed type: The box plot for seed type shows that the variability of the response variable for applying improved and non-improved seed types is approximately similar on maize farms. However, the non (or very few) application of improved seeds on barley, teff and wheat farms showed lower variability on yields. On the other hand, the median response for improved seed types is higher for maize and wheat farms as compared to that of the non-improved seed types. It was not possible to construct a plot for sorghum due to fact that improved seed types are applied on very few farms.

Crop prevention methods: The response variability for chemical type crop prevention measures was the highest for maize and the least for sorghum crops. But with respect to the median response values, there were a maximum for sorghum and a minimum for teff crops.

Fertilizer type: The variability of the cereals responses from the use of both natural and

chemical fertilizers was higher for barley, maize, sorghum, and teff crop farm than wheat. For wheat crops, the highest response variability was observed from the use of chemical fertilizer. It was revealed that for all types of crop farms on which no fertilizer was used and those on which natural fertilizer was applied, found to vary approximately equally in their median responses. There were also potential outliers and extreme cases in each of the crop types for at least two or more levels of fertilizer use.

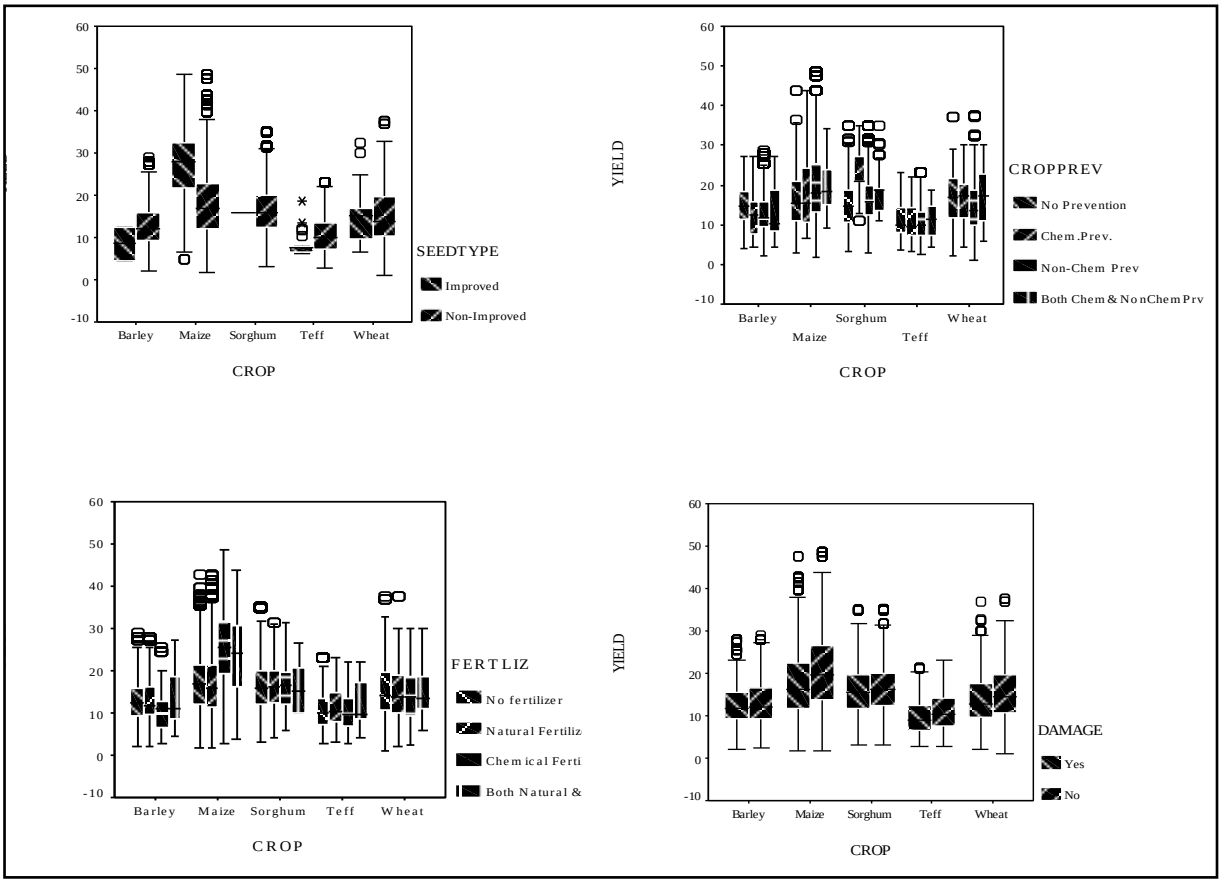


Figure 2.3 The relationship between yield and seed type, fertilizer type, crop prevention measures, and crop damage by crop type

Soil erosion protection: With regard to the soil erosion prevention factor, the variability on the response due to protecting soil erosion was higher for barley, teff and

wheat and lower for maize and sorghum crop yields (see the box-plot "SERRO" in Fig. 2.4).

Crop damage: The response variability for all levels of crop damage factors for barley and sorghum seem to be equal, whereas it is higher for maize, teff, and wheat farms with no crop damage. There are also outliers and extreme values at each level of the independent factor in the data.

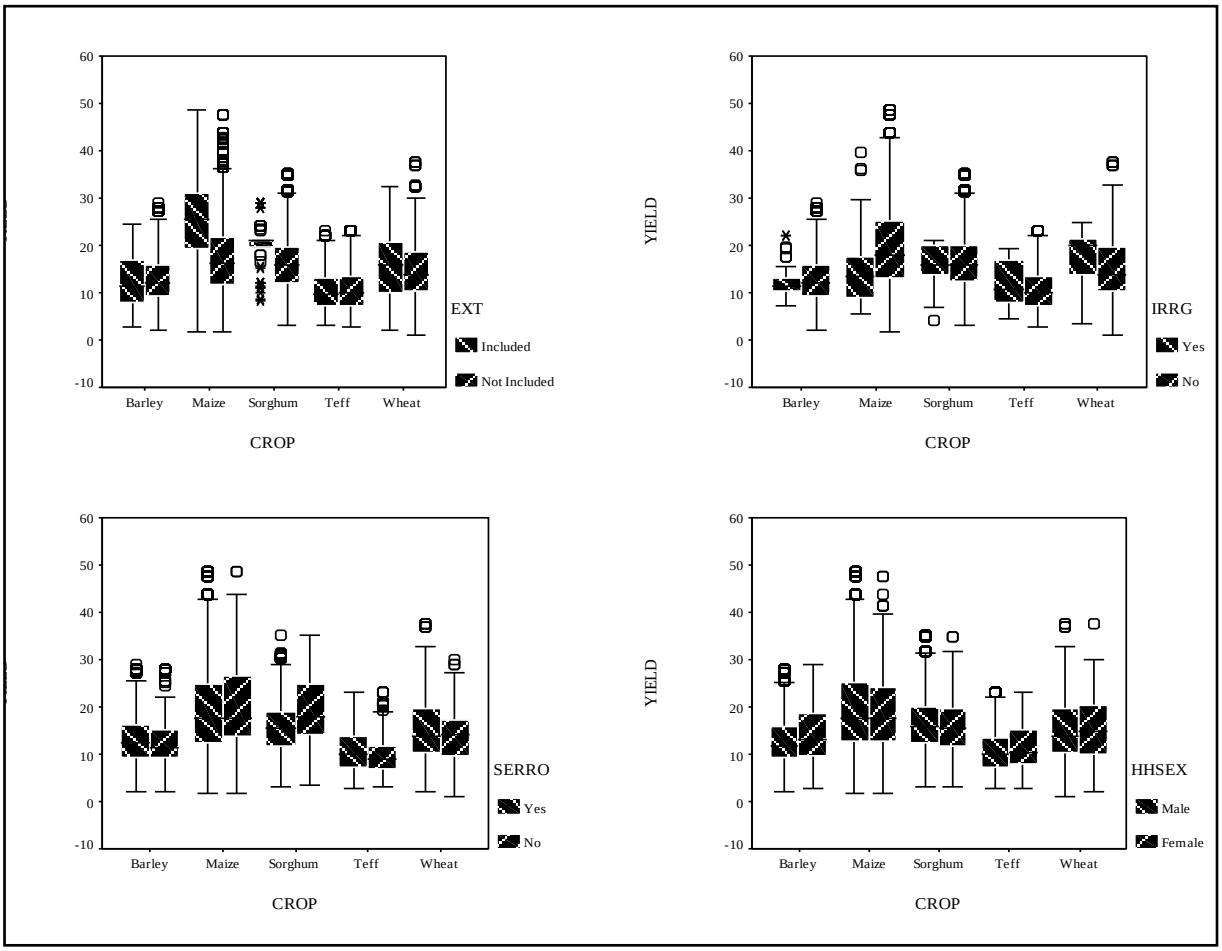


Figure 2.4 The relationship between yield and extension, prevention of soil erosion, irrigation, and gender of head of the household by crop type

Irrigation: unlike for teff and wheat crops, where their median response from irrigated farms was higher than the non-irrigated farms, for maize crops the median response from non-irrigated farms was higher than that of the irrigated farms. On the other hand, the median responses for irrigated and non-irrigated farms for sorghum and barley crops appeared to have equal values.

Extension programme: The effect of agricultural extension programmes on cereal yields variability and its distribution showed that barley and teff crop types had equal median response values for both included and non-included factor levels of extension programmes whereas it was higher for maize, sorghum and wheat included in the programmes. In terms of the response variability, it was higher for barley, maize and wheat farms included in the extension programmes than in the non-included farms.

Gender of head of the household: The response variability for gender was relatively high for barley, sorghum and teff crops, whereas it had very little effect on the variability of the response for maize and wheat crops.

From the exploratory data analysis it was observed that the distributions of original yield data for all cereal crop types were slightly skewed to the right. In light of the observed violation of the normality assumptions, it was clear that an appropriate transformation technique should be selected and applied on the cereals yield data in order to fulfil the requirement for further statistical analysis of the data. Thus, in the next section the transformation techniques applied in this paper are discussed.

2.3 Transformation of the Data

Data transformations are commonly used tools that can serve many functions including improving normality of a distribution and equalizing variance to meet assumptions and improve effect sizes, thus constituting important aspects of data cleaning and preparing for our statistical analyses. Hence, the evaluated transformation techniques include the square root, the logarithmic and the box-cox methods. These are all members of a class of transformations called power transformations. Power transformations are transformations that raise numbers to an exponent (power).

Box and Cox (1964) proposed a parametric power transformation technique in order to reduce non-normality and heteroscedasticity. The original form of the box-cox transformation takes the following form:

$$y_i^{\lambda} = \begin{cases} (y_i^{\lambda} - 1)/\lambda; & \lambda \neq 0 \\ \log y_i\,; & \lambda = 0 \end{cases}$$

where λ stands for transformation parameter estimate.

The three transformation techniques applied on each crop type data is displayed in Table 2.4. Furthermore, before the selection was made, the Kolmogorov-Smirnov test statistic (see Table 2.4) was carried out to test the normality of the data. However, for all tested transformation techniques, it rejects the null hypothesis which states that the data come from the normal distribution. This could be due to the high possibility of the test statistic to reject the null hypothesis as the sample size becomes larger and larger (SAS, 2004). Therefore, it is suggested that examining of other statistics, such as skewness and kurtosis measures and the plots, for instance histograms and Q-Q plots,

are important to make a final assessment of normality tests for large data sets (SAS, 2004).

Table 2.4 Tests of Normality Results for the Original and Transformed Yield Data by Crop Types

Crop Type	Kolmogorov - Smirnov Tests of Normality	Original	Transformation Techniques Tested		
			Square root	Log	Box-Cox
Barley	Statistic	.081	.040	.072	.040
	Df	6653	6653	6653	6653
	Sig.	.000	.000	.000	.000
Maize	Statistic	.082	.037	.059	.037
	Df	9653	9653	9653	9653
	Sig.	.000	.000	.000	.000
Sorghum	Statistic	.048	.041	.094	.040
	Df	7292	7292	7292	7292
	Sig.	.000	.000	.000	.000
Teff	Statistic	.082	.054	.059	.053
	Df	10509	10509	10509	10509
	Sig.	.000	.000	.000	.000
Wheat	Statistic	.089	.065	.072	.065
	Df	6127	6127	6127	6127
	Sig.	.000	.000	.000	.000

As a result, the selection of an appropriate transformation method for our data was made by comparing the improvements in the skewness and kurtosis statistic values of the candidate transformation techniques, as shown in Table 2.5.

Table 2.5 Skewness and Kurtosis Results for Selecting Transformation Methods

Crop Type	Transformations	Skewness		Kurtosis		Selected Transformation Method
		Statistic	Std. Error	Statistic	Std. Error	
Barley	None	.528	.030	.042	.060	
	Square root	-.049	.030	-.110	.060	Square root
	Log	-.749	.030	.892	.060	
	Box-Cox (λ = 0.52)	-.049	.030	-.111	.060	
Maize	None	.647	.025	.056	.050	
	Square root	.084	.025	-.332	.050	Square root
	Log	-.645	.025	.787	.050	
	Box-Cox (λ = 0.50)	.084	.025	-.333	.050	
Sorghum	None	.301	.029	-.036	.057	
	Square root	-.089	.029	.066	.057	Square root
	Log	-.995	.029	1.296	.057	
	Box-Cox (λ = 0.70)	.097	.029	-.095	.057	
Teff	None	.577	.024	-.424	.048	
	Square root	.097	.024	-.632	.048	Square root
	Log	-.243	.024	-.464	.048	
	Box-Cox (λ = 0.42)	-.099	.024	-.630	.048	
Wheat	None	.372	.031	-.517	.063	
	Square root	-.041	.031	-.616	.063	Square root
	Log	-.577	.031	.326	.063	
	Box-Cox (λ = 0.65)	-.041	.031	-.616	.063	

In particular, the frequency distribution plots (Histograms) and the Normal Q-Q plots of the transformed response variable were examined for each data set. Thus, based on the results on different trials on transforming the cereals yield, a square root transformation method was selected as the best technique to apply on the original yield data of the cereals.

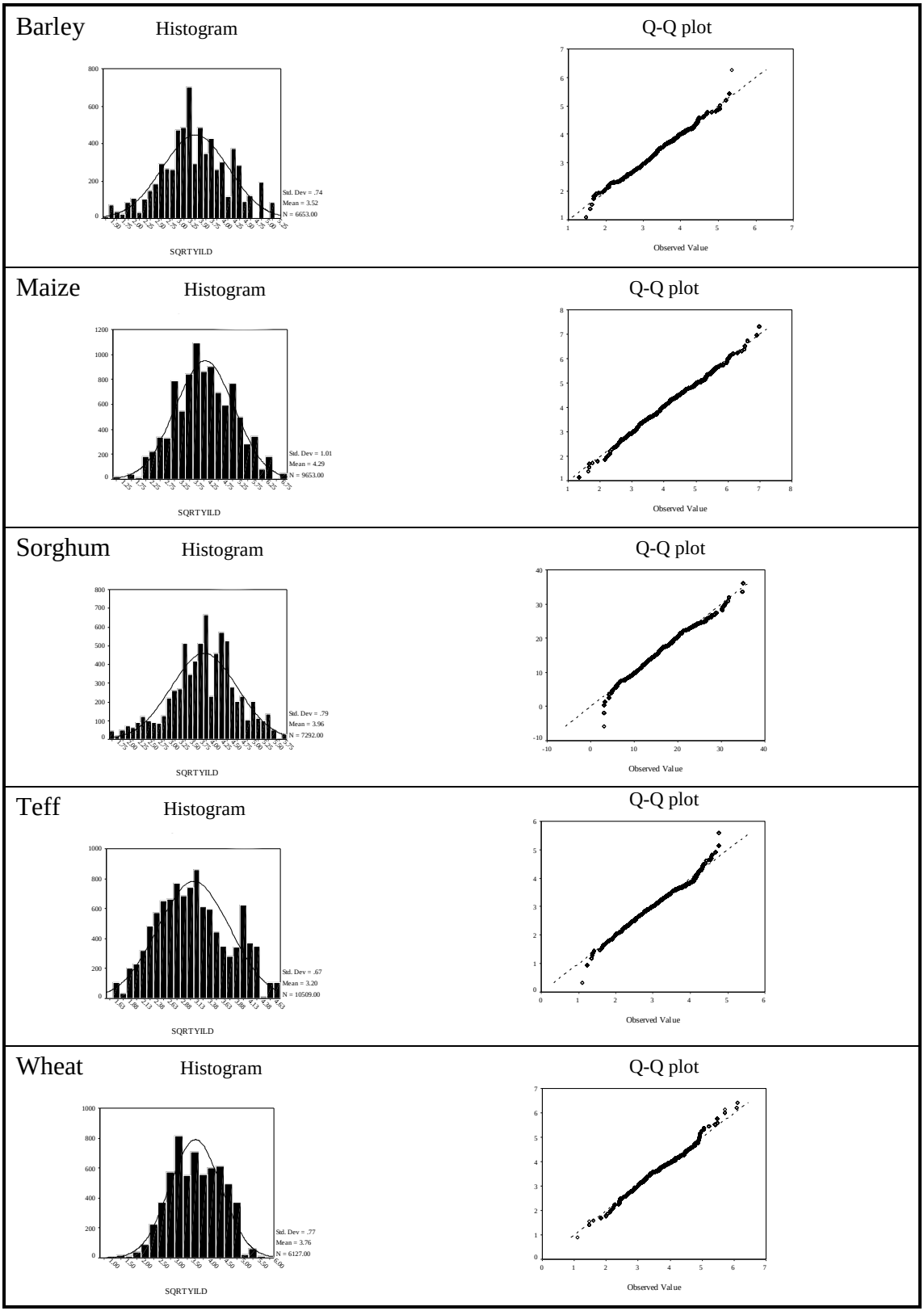


Figure 2.5 Frequency distributions and the Nnormal Q – Q plots for transformed cereal yields by crop type.

The histograms with the corresponding Q-Q plots in Fig 2.5 for the selected transformed yield data (square root of yield data) by crop types clearly shows that the application of this transformation resulted in an improvement of the normality of the data. Thus, the square root transformation of the yield data for each cereal type will be used in the analysis section of the subsequent chapters.

2.4 Summary of the Preliminary Analysis of the Data

Based on the results of the preliminary analysis, the data for all crop types seem to follow a slightly skewed distribution to the right. Therefore, for further statistical analysis to be employed, a square root transformation method was selected and applied to the cereals yield data. The transformation consequently revealed that the observed violation of the data in the assumption of normal distribution had been improved.

The next chapter reviews the theory and application of a multiple regression analysis to the transformed yield data for each particular crop type. Furthermore, model diagnostics and interpretation of the results obtained from the fitted models are discussed.

Chapter 3

Theory and Application of Multiple Regression Analysis

3.1 Introduction

Regression analysis is a statistical tool for the investigation of relationships between a continuous dependent variable and of one or more independent variables. This analysis allows us to understand which variables influence the response, and to predict a value of one variable for a given value of another. The independent variables used in regression can be either continuous or categorical. Independent categorical variables with more than two levels must be converted into variables that only have two levels, called dummy variables, before they are to be used in the regression analysis (Rawlings, 1988; Weisberg, 1985).

In simple linear regression we study the relationship between a response variable y and a single explanatory variable x whilst in multiple regression we study the relationship between y and a number of p explanatory variables ($x_1, x_2, \ldots, x_p$). This enables us to estimate models of greater complexity and investigate the relationship of each explanatory variable to the dependent variable while controlling for the effects of the other variables in the model. An understanding of these statistical techniques is therefore essential to identify the effects of agricultural factors on crop production. It is worthwhile noting that the data in this thesis cannot be analysed through a multivariate regression method of analysis since for every dependent crop yield variable there is a

different set of values of the agricultural factor levels. Thus, a brief discussion of multiple linear regression models which is to be used for analysing our data is presented in the subsequent sections of this chapter.

3.2 Multiple Linear Regression Models

A multiple linear regression model involves the dependent variable, Y_i specified as a linear combination of independent variables (x_{i1}, x_{i2}, $x_{i,p-1}$) and the population parameters, ($\beta_0, \beta_1, \beta_2, \ldots, \beta_{p-1}$). The regression equation takes the form:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots\ldots + \beta_{p-1} X_{i,p-1} + \varepsilon_i \qquad (3.1)$$

where: Y_i is the dependent variable, the $\beta_1, \beta_2, \ldots, \beta_p$, are the regression coefficients. A multiple regression coefficient β tells us about the average amount the dependent variable increases when the independent variable increases by one unit keeping the other variables constant. In the case of dummy variables, it tells about the difference in the expected value of the dependent variable between the conditions described by the 0 value of the variable and the condition described by the 1 value of the variable. The variable β_o is the intercept, representing the amount of Y when all the independent variables are zero and ε_i represents the random error term of the regression equation.

The sign of a regression coefficient is interpreted as the direction of the relationship between the dependent and an independent variable, that means if a coefficient β is positive (negative), then the relationship of the variable with the dependent variable is positive (negative). Moreover, if the β coefficient is equal to 0 then it implies that, there is no linear relationship between the variables.

In regression analysis we need to have a set of assumptions that are required to validate model estimation and hence to make inferences from a sample to a population. The most common assumptions used in this analysis about the error terms are:-

• The mean of the probability distribution of the error term is zero; i.e. $E[\ \varepsilon_i\ ]= 0$,

• The probability distribution of error terms ε_i's are assumed to have a constant variance σ^2,

• The probability distribution of the error term is assumed to be distributed as a normal distribution and,

• The errors terms are uncorrelated to each other.

3.2.1 Estimating the Model Parameters

The sample regression model is formulated by modifying the regression model in equation 3.1 as follows:

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \hat{\beta}_2 X_{i2} + \ldots\ldots + \hat{\beta}_{p-1} X_{i,p-1} + e_i \qquad (3.2)$$

where: Y_i's are the observed values, i = 1, 2, . . ., n,

e_i is the residual (an estimate of the error term ε_i) related with the i^{th} observation.

The fitted model is then given by:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \hat{\beta}_2 X_{i2} + \ldots\ldots + \hat{\beta}_{p-1} X_{i,p-1} \qquad (3.3)$$

The sample regression model equation 3.2 can also be represented in matrix form as follows:

$$\mathbf{Y} = \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{e} \qquad (3.4)$$

where:

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \qquad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1,p-1} \\ 1 & x_{21} & x_{22} & \cdots & x_{2,p-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{n,p-1} \end{bmatrix},$$

$$\widehat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_{p-1} \end{bmatrix}, \qquad \text{and} \qquad \mathbf{e} = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}.$$

And the fitted values are represented by:

$$\widehat{\mathbf{Y}} = \mathbf{X}\widehat{\boldsymbol{\beta}}$$

where $\widehat{\mathbf{Y}}$ is an (n × 1) vector of fitted values and,

$\widehat{\boldsymbol{\beta}}$ is a (p × 1) vector of the estimated value of population parameter $\boldsymbol{\beta}$.

To estimate the unknown population parameter ($\boldsymbol{\beta}$'s), a method of Ordinary Least Squares (OLS) is employed to obtain $\widehat{\boldsymbol{\beta}}$ that minimises the sum of the squares of the residuals, S (Berk, 2004).

$$S = \sum e_i^2,$$

or

$$S = \sum (Y_i - \widehat{Y}_i)^2,$$

or

$$\boldsymbol{S} = (\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}})'(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}).$$

Then, the minimum value of **S** (Berk, 2004) occurs at

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}.$$

The unbiased estimator of the variance, which represents the estimation of the random error variance, σ^2, (Berk, 2004) is computed as:

$$\widehat{\sigma}^2 = \sum \frac{e_i{}^2}{n-p}.$$

And hence, the estimate of variance covariance matrix for $\widehat{\boldsymbol{\beta}}$ (Berk, 2004) is given by:

$$\text{Var}(\widehat{\boldsymbol{\beta}}) = \widehat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}.$$

3.2.2 Model Selection and Inferences

The investigator usually wishes to reduce the number of explanatory variables to be used in the final model. There are several reasons for this. Primarily, a regression model with numerous explanatory variables may be difficult to maintain. In addition to this, regression models with a limited number of explanatory variables are easier to work with and understand. Furthermore, the presence of many highly inter-correlated explanatory variables may substantially increase the sampling variation of the regression coefficients and it could adversely affect the descriptive abilities of the model. Likewise, elimination of key explanatory variables can seriously damage the explanatory power of the model and lead to biased estimates of regression coefficients, mean responses, and predictions of new observations, as well as biased estimates of error variance. Thus, the choice of an appropriate model with a few explanatory variables for final consideration needs to be done with great care. Basically, there are three approaches (Bowerman, 1986) for selecting explanatory variables in dealing with the best regression model:

i. Forward Selection

The forward selection method starts by choosing the independent variable which explains the most variation in the dependent variable and continues to include variables by their order of significance until no variables significantly explain the variation in the outcome variable.

ii. Backward Selection

This method starts with all the variables in the model, and excludes the non-significant variable until we are left with only significant variables.

iii. Stepwise selection

This method involves the combination of the above two selection methods in which case variables entered are checked at each step for removal, and at the same time, variables excluded will be checked for re-entry in the removal method.

The next stage, after selecting the regression model to be employed in this study, is to check in detail whether the selected model fits our data well. The diagnostic checks are useful for identifying influential or outlying observations, multicolinearity and the like. Besides, a variety of residual plots and analyses can be employed to identify any lack of fit, outliers, and influential observations in the data.

- **Goodness of fit tests**

The goodness of fit of the model can be measured by the coefficients of multiple determinations denoted by R^2 which measures the proportion of the variability explained by the model, and its significance is then tested by the F-test, which actually means testing the significance of the regression model as a whole.

The coefficient of multiple determinations (R^2) is given by:

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

where: SSR, SSE and SST denote the regression, error, and total sum of squares respectively.

Although R^2 measures the proportion of variation explained by the model, the high value of R^2, however, does not necessarily imply that the model is adequate. This is due to the fact that an increase in the number of independent variables included in the model ultimately results in a higher value of R^2. Therefore, to discourage the unnecessary inclusion of explanatory variables, the adjusted coefficient of multiple determinations denoted by R^2_{adj} is used instead to test the adequacy of the model. And it is calculated as:

$$R^2_{adj} = 1 - \left(\frac{n-1}{n-p}\right)\frac{SSE}{SST}.$$

- **Analysis of variance (ANOVA) Table**

The analysis of a variance table is used as a tool for testing the possibility of using the regression models. The Analysis of Variance table for linear regression analysis is presented below:

Table 3.1 Regression Analysis of Variance (ANOVA) Table

Source of Variation	SS	df	MS
Regression	$\mathbf{SSR} = \widehat{\boldsymbol{\beta}}' \mathbf{X}' \mathbf{Y} - \left(\frac{\mathbf{1}}{\mathbf{n}}\right) \mathbf{Y}' \mathbf{J}\, \mathbf{Y}$	p - 1	$\mathbf{MSR} = \frac{\mathbf{SSR}}{\mathbf{p-1}}$
Error	$\mathbf{SSE} = \mathbf{e}' \mathbf{e} = \mathbf{Y}'\mathbf{Y} - \widehat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{Y}$	n - p	$\mathbf{MSE} = \frac{\mathbf{SSE}}{\mathbf{n-p}}$
Total	$\mathbf{SST} = \mathbf{Y}'\mathbf{Y} - \left(\frac{\mathbf{1}}{\mathbf{n}}\right) \mathbf{Y}' \mathbf{J}\, \mathbf{Y}$	n – 1	

We test a hypothesis to see whether the dependent variable (Y) and the independent variables $X_1, X_2, \ldots, X_{p-1}$ have significant relation. Thus, the null and alternative hypothesis is presented as:

$$H_o : \beta_1 = \beta_2 \ldots = \beta_{p-1} = 0$$

$$H_1 : \text{Not all } \beta_j \text{ are zeros,}$$

where $j = 1, 2, \ldots, p-1$ (i.e. at least one coefficient is different from zero).

To test the above hypothesis we use the statistic:

$$\mathbf{F}_{cal} = \frac{\mathbf{MSR}}{\mathbf{MSE}}.$$

We will compare $\mathbf{F}_{cal}$ with $\mathbf{F}$ at a specific level of significance, α, with (p-1) and (n-p) degrees of freedom.

If $\mathbf{F}_{cal} \leq \mathbf{F}_{1-\alpha,p-1,n-p}$, then we do not reject H_o and conclude that the independent variables do not contribute significantly to the dependent variable. On the other hand, if $\mathbf{F}_{cal} > \mathbf{F}_{1-\alpha,p-1,n-p}$, we reject Ho in favour of H_1, we conclude that at least one independent variable has significant relation with the dependent variable. If we conclude the latter,

we have to test the coefficients individually to identify which variable is significantly linearly related to the dependent variable.

To test that the regression coefficients equal to zero, the null hypothesis H_o and the alternative hypothesis H_1 are constructed as:

$$H_o : \beta_j = 0$$

$$H_1 : \beta_j \neq 0.$$

The test statistic t, with (n–p) degrees of freedom is calculated as:

$$t = \frac{\hat{\beta}_j}{s(\hat{\beta}_j)}$$

where, $\hat{\beta}_j$ is the estimated value of β_j and $s(\hat{\beta}_j)$ is the standard error of $\hat{\beta}_j$.

If $|t| \leq t_{\alpha/2\ (n-p)}$, we do not reject H_o and conclude that there is no evidence to reject that $\hat{\beta}_j$ is not significantly different from zero, otherwise if $|t| > t_{\alpha/2\ (n-p)}$, reject H_o and conclude that $\hat{\beta}_j$ is significantly different from zero.

The hypotheses on β can also be tested using a $(1-\alpha)100\%$ confidence interval which is constructed from the estimated $\hat{\beta}_j$ for each regression coefficients $(\beta_j's)$.

We have: $$\frac{\hat{\beta}_j - \beta_j}{s(\hat{\beta}_j)} \sim t\,(n-p)$$

$$\Rightarrow -t_{1-\alpha/2\,,(n-p)} \leq \frac{\hat{\beta}_j - \beta_j}{s(\hat{\beta}_j)} \leq t_{1-\alpha/2\ (n-p)},$$

$$\Rightarrow \hat{\beta}_j - t_{1-\alpha/2\,,(n-p)}\, s(\hat{\beta}_j) \leq \beta_j \leq t_{1-\alpha/2\ (n-p)}\, s(\hat{\beta}_j).$$

Hence, the $(1-\alpha)100\%$ confidence interval for β_j is:

$$\hat{\beta}_j \pm t_{1-\alpha/2,\,(n-p)}\, s(\hat{\beta}_j), \qquad \text{where } j = 1, 2, \ldots, p.$$

3.2.3 Model Diagnostics

- **Outliers**

An outlier can be defined as a data point which is far away from the rest of the data. If outliers are occurred due to errors in recording, they can be rejected automatically. Otherwise they should be carefully investigated since they could represent new information (Draper, 1966). Therefore, it is proposed that some procedures should be employed in dealing with severe outliers as follows:

i) remove the observations from the data set and repeat the regression to see whether the fit or the sign of one of the coefficients change or not - if there is no sign change, we can conclude that the outliers do not affect the results and they can be removed from the data;

ii) if there is a change in the fit or in the sign of one of the coefficients, further care should be taken in deciding either to drop them or to keep them in the data and during interpretation of the estimates (Berk, 2004).

- **Influential observations**

Influential observation is an observation that causes the least square point estimates to be substantially different from what they would be if the observation was removed from the data (Bowerman, 1986). An observation could be an outlier but it does not necessarily mean that all outliers are influential. There are a number of measures to

identify observations which significantly influence the estimates of the model parameters; however in this study Cook's distance measure will be used.

- **Cook's Distance (C_i)**

Cook's Distance is defined as the standardised difference between $\hat{\beta}_{(i)}$ the vector of estimate obtained by omitting the i^{th} observation, and $\hat{\beta}$ the vector of parameter estimate obtained using all the data. It is an important diagnostic measure in making decisions about observations that influence the fitted model (Berk, 2004). Cook's distance is formulated as:

$$C_i = \frac{(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)})(\text{var}(\hat{\boldsymbol{\beta}}))^{-1}(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)})}{k}$$

where, var$(\hat{\boldsymbol{\beta}})$ is the variance covariance matrix of parameter estimate of $\hat{\beta}_{(i)}$, and k is the number of explanatory variables in the model.

A large C_i implies that the i^{th} observation has an influence on the set of parameter estimates. With regard to the decision criteria, an observation with C_i value in excess of 1 is commonly taken as an influential observation (Berk, 2004).

3.3 Application of the Multiple Regression Model to Crop Yield Data

The multiple linear regression model discussed in Section 3.2 was fitted to each of the five types of crops yield data and the relationship between the explanatory variables and the yield is specified and analysed by the SPSS (Statistical Package for Social Scientists) statistical software through the use of the following linear relationship:

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1(\text{Seedty2}) + \hat{\beta}_2(\text{Fertliz1}) + \hat{\beta}_3(\text{Fertliz2}) + \hat{\beta}_4(\text{Fertliz3}) + \hat{\beta}_5(\text{Cropp1}) + \hat{\beta}_6(\text{Cropp2}) + \hat{\beta}_7(\text{Cropp3}) + \hat{\beta}_8(\text{Damage1}) + \hat{\beta}_9(\text{Ext1}) + \hat{\beta}_{10}(\text{Irrg1}) + \hat{\beta}_{11}(\text{Serro1}) + \hat{\beta}_{12}(\text{HHsex2}) + \hat{\beta}_{13}(\text{Zone1}) + \hat{\beta}_{14}(\text{Zone2}) + \hat{\beta}_{15}(\text{Zone3}) + \hat{\beta}_{16}(\text{Zone4}) + \hat{\beta}_{17}(\text{Zone5}) + \hat{\beta}_{18}(\text{Zone6}) + \hat{\beta}_{19}(\text{Zone7}) + \hat{\beta}_{20}(\text{Zone8}) + \hat{\beta}_{21}(\text{Zone9}) + e_i \quad (3.5)$$

where: Y_i = Crop yield in quintal per hectare,

Seedty2 = Dummy variable for local seed (1: if local seed, 0: if improved seed),

Fertliz1 = Dummy variable for chemical fertilizer (1: if chemical, 0: otherwise),

Fertliz2 = Dummy variable for non-chemical fertilizer (1: if non-chemical, 0: otherwise),

Fertliz3 = Dummy variable for both chemical and non-chemical fertilizer use (1: if both types used together, 0: otherwise),

Cropp1 = Dummy variable for use of chemical for crop prevention (1: if chemical, 0: otherwise),

Cropp2, = Dummy variable for use of non-chemical type of crop prevention (1: if non-chemical, 0: otherwise),

Cropp3 = Dummy variable for use of both chemical and non-chemical type of crop prevention (1: if both types used together, 0: otherwise),

Damage1 = Dummy variable for crop damage (1: if there is crop damage, 0: if no crop damage),

Ext1 = Dummy variable for farms under extension programme (1: if included in the programme, 0: if not),

Irrg1 = Dummy variable for farm irrigation (1: if irrigated, 0: if not irrigated),

Serrop1 = Dummy variable for soil erosion protection (1: if it exists, 0: if not),

HHsex2 = Dummy variable for gender of the head of the household (1: if female, 0: if male),

Zone1, Zone2, . . . , Zone9 are dummy variables for ZONE coded with '1' if the crop is within the particular administrative zone or '0' otherwise, and

e_i = the error term; $\hat{\beta}_0$ is the constant; and $\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \ldots, \hat{\beta}_{21}$ are the coefficients of the independent variables.

On the basis of the data available in this study, the mean crop yield is hypothesised to be affected by the zone (location) in which the crop is cultivated, seed type, fertilizer use, crop prevention method, crop damage, agricultural extension programme, irrigation, protection of soil erosion, and gender of the head of the household. At the start of the analysis, all of the categorical dummy variables stated in equation (3.5) were considered for inclusion in the model. Then a stepwise selection method was employed to identify and retain only the dummy variables which significantly explain the model as compared to their respective reference categories at 5% level of significance, and to exclude those dummy factor levels which do not. In the sections that follow the regression models that are fitted to the transformed yield data by crop type and their results are discussed.

3.3.1 Fitting a Model for Transformed Yields of the Data

Recall that in the exploratory analysis section of Chapter 2, the distributions of all the studied cereal yields were slightly skewed to the right. Thus, the linear stepwise regression procedure was applied to the transformed data (square root of yield) to estimate the regression coefficients for the particular crop types. The stepwise regression results for the fitted cereals models which considered the transformed yield

as dependent variable and all the dummies explained in the model equation (3.5) as independents, are displayed in the Tables 3.3.1, 3.3.2 and 3.3.3.

Table 3.3.1, the regression analysis of variance (ANOVA) table by crop type, showed that the overall model for barley, maize, sorghum, teff and wheat crops were significant with overall F ratio values of $F_{(11, 6567)} = 172.765$, $F_{(15, 9598)} = 307.915$, $F_{(6, 7252)} = 107.345$, $F_{(16, 10437} = 335.323$ and $F_{(14, 6079)} = 243.056$ respectively at $\alpha = 0.05$ significance level. Therefore, it is concluded that at least one of the regression coefficients significantly contributed to explain the variability in the transformed yield of their respective crop types. The table also showed that the R^2 values of 22.4%, 32.5%, 25.2%, 34%, and 35.9% for barley, maize, sorghum, teff and wheat crops respectively. which indicates the percentage of variation explained by the independent factor levels (dummies) included in the particular models. In addition, the results on the amount of variability contributed by each of the included dummy factor levels, i.e. the change in R^2 and the corresponding significance measures of the changes in F values are presented in the Appendix A section from Tables A.1 to A.5.

Table 3.3.1 **Regression Analysis of Variance Table for Models of Transformed Cereals Yields by Crop Type**

Crop Type		Sum of Squares	Df	Mean Square	F	Sig.
Barley	Regression	814.890	11	74.081	172.765	.000
	Residual	2815.899	6567	.429		
	Total	3630.788	6578			
	$R^2 = 0.224$, Adjusted $R^2 = 0.223$					
Maize	Regression	3200.430	15	213.362	307.915	.000
	Residual	6650.688	9598	.693		
	Total	9851.118	9613			
	$R^2 = 0.325$, Adjusted $R^2 = 0.324$					
Sorghum	Regression	368.866	6	61.478	107.345	.000
	Residual	4153.301	7252	.573		
	Total	4522.167	7258			
	$R^2 = 0.252$, Adjusted $R^2 = 0.250$					
Teff	Regression	1568.677	16	98.042	335.323	.000
	Residual	3051.590	10437	.292		
	Total	4620.267	10453			
	$R^2 = 0.340$, Adjusted $R^2 = 0.339$					
Wheat	Regression	1304.571	14	93.184	243.056	.000
	Residual	2330.588	6079	.383		
	Total	3635.159	6093			
	$R^2 = 0.359$, Adjusted $R^2 = 0.357$					

After the significance of the overall model had been confirmed, it was necessary to identify which variables were important and significant in explaining the variability in the mean yields of the fitted models. However, before making any inferences about the identified parameter estimates it is also imperative to undertake model diagnostics to check whether the regression assumptions are not violated. Therefore, in the sections

that follow, the validity of the basic assumptions of normality and homoscedasticity of the residuals for the fitted models are assessed.

- **Model Checking and Diagnostics**

The fitted models were assessed using graphic tools such as the frequency distribution plots of the residuals, p-p plots, scatter plots and Cook's distance plots to investigate whether the regression assumptions were not violated. With regard to our fitted models, Figure 3.3.1 was presented to look at the frequency distributions (histograms) of the transformed yields and the normal p-p plots by crop type. The p-p plot is a graphical technique used for assessing whether or not a data set is normally distributed. If the distribution is normal, the points on the normal p-p plot fall reasonably close to a straight line. Therefore, as shown in the figure, the normal p-p plot for transformed yields of the studied cereal types indicated a pattern of clustering of points close to a straight line. Thus, it is concluded that the assumption of normality of the residuals for the fitted models under all types of the crops were not violated.

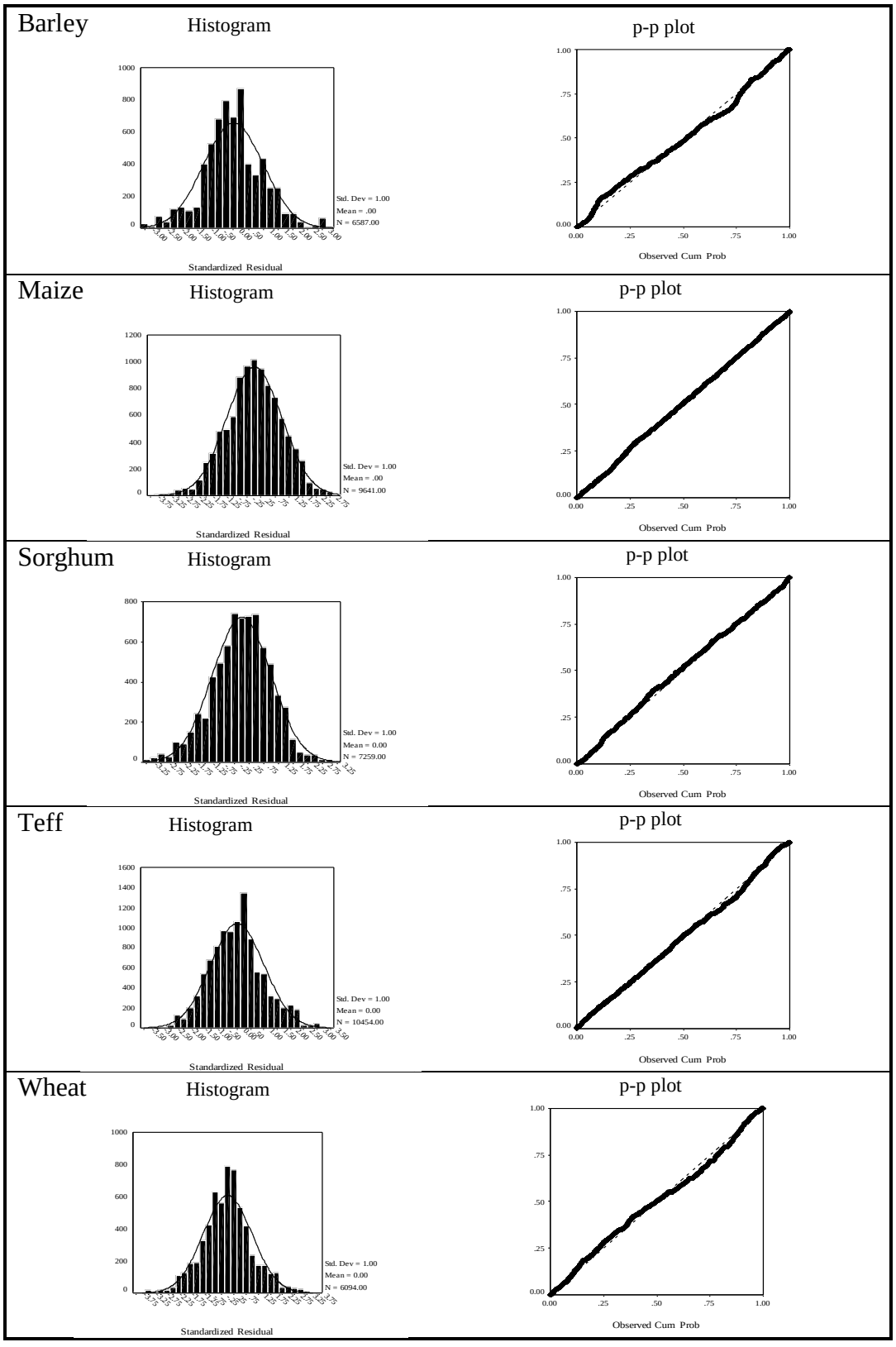


Figure 3.3.1 Frequency distribution and p-p plots of the fitted models residuals by crop type

Moreover, scatter plots and Cook's distance plots were used as further model diagnostic techniques of the analysis. The scatter plots are used to observe any change in the spread or dispersion of the plotted points and thus to check whether the assumptions of constant variance were not violated; and Cook's distance (C_i) is used to measure the influence of an observation and how much the regression coefficients are changed by deleting the particular observation in question. It is suggested that an observation with a $C_i > 1$ may deserve closer inspection, and if the model is correct, then the expected C_i is <1 (i.e there are no influential cases that should be dropped).

For our models of the cereals in this section, the scatter plot of the standardised residuals versus the standardised predicted values and Cook's distance plots were plotted and presented in Figure 3.3.2 by crop type to examine the assumption of homoscedasticity of the error variance as well as to check for the existence of influential observations in the data. From the figure, it was observed that there was no evidence of a specific pattern revealed in any of the scatter plots shown by crop types. This confirms that the assumption of homoscedasticity was valid after transforming the yields of the cereals. Moreover, the figure also showed that Cook's distance value of each of the observations under the corresponding fitted models by crop types were less than one in their magnitude. This suggests that none of the observations were influential on the parameter estimates in the fitted models of their respective cereal types. Therefore the observed outliers were retained in the data.

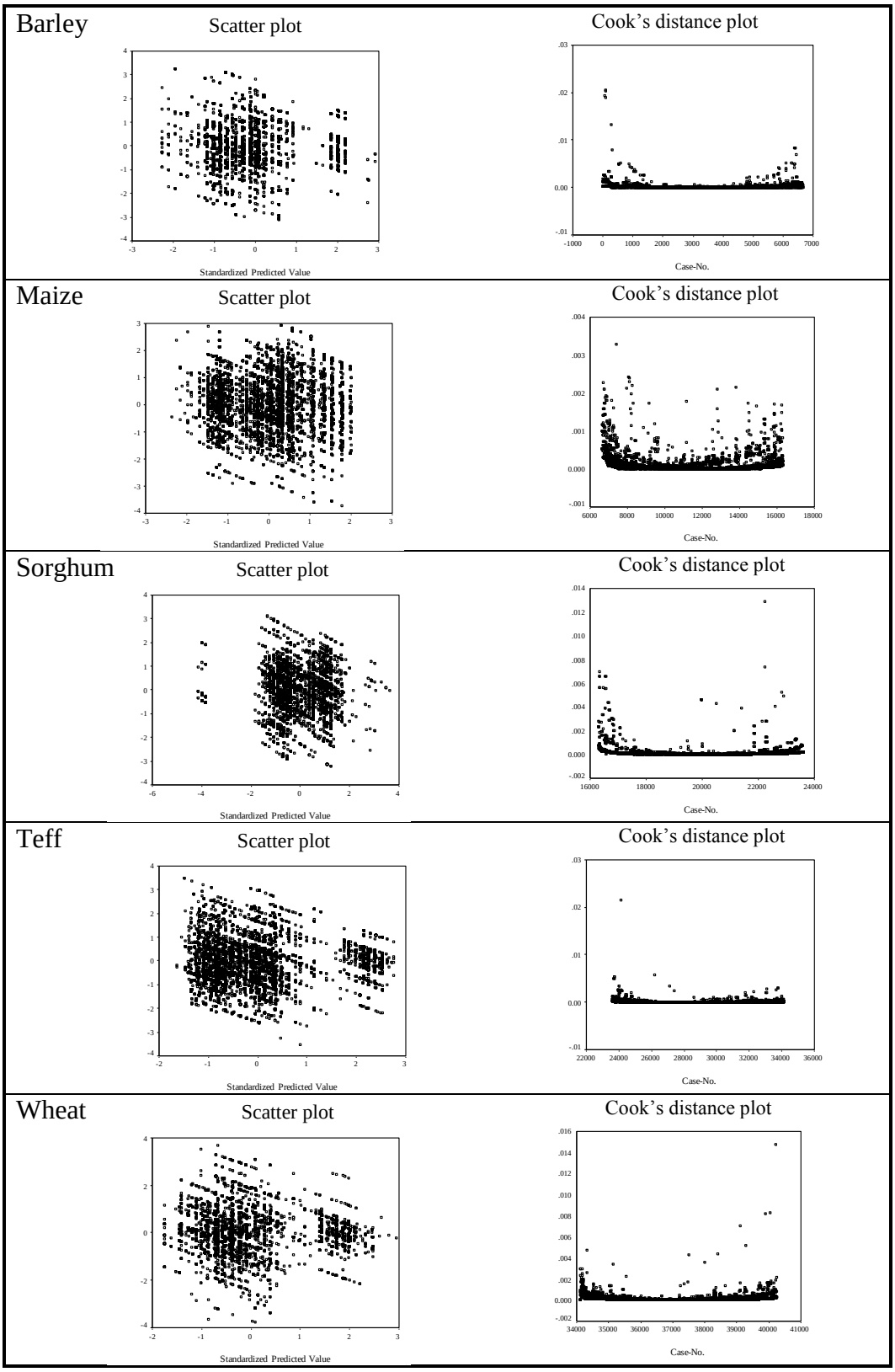


Figure 3.3.2 Scatter plots and Cook's distance plots for the fitted models by crop type

Table 3.3.2 presents the magnitude and significance of the estimated regression coefficients of the fitted models by crop type. These coefficients are useful for constructing the regression equations and also for making direct interpretations. For Barley crop type the t-value of all the estimated regression coefficients, except for Zone2, were found to be significant at $\alpha=0.05$ level of significance, and hence Zone2 will not be included in the prediction model. Likewise, the different sets of predictors under each of the other four crop types were assessed and found to affect the mean yields of their respective crop types significantly. Therefore, it is evident that only these dummy variables, which have been found to contribute significantly to the variability of their respective mean yields, could be used to formulate the regression equations.

The $(1-\alpha)$ 100% confidence interval, the interval in which the true parameter lies, could also be used as an additional means for testing the significance of the regression parameter estimates. Thus, the 95% confidence interval of the coefficients estimates under all the cereal types is presented in Table 3.3.3. It is shown that the coefficients of all the predictor dummy variables, except Zone2 under barley crop type, do not include zero values in their estimated 95% confidence intervals; this confirms that those coefficients had a significant effect on the variability of the transformed yield of the cereals as compared to the effects of their respective reference category levels.

Table 3.3.2 Estimates of the Regression Parameter Coefficients for the Fitted Models by Crop Type (Standard Deviations in Parentheses)

Effects	Barley		Maize		Sorghum		Teff		Wheat	
	Estimates	Sig.	Estimates	Sig.	Estimates	Sig.	Estimates	Sig.	Estimates	Sig.
(Constant)	3.205 (.026)	.000	4.025 (.046)	.000	3.758 (.036)	.000	3.551 (.037)	.000	3.700 (.031)	.000
Seedtype2 (Reference = Seedtype1)	...	...	-.256 (.032)	.000	...	...	...	...	...	...
Fertliz1 (Reference = Fertliz0)	.067 (.019)	.000	.055 (.021)	.010	.173 (.022)	.000	.113 (.019)	.000	.074 (.024)	.002
Fertliz2	-.383 (.039)	.000	.437 (.031)	.000	...	...	.036 (.016)	.026	.157 (.021)	.000
Fertliz3	...	...	.286 (.045)	.000	...	...	...	...	...	...
Ext1 (Reference = Ext2)	.257 (.051)	.000	...	...	.239 (.076)	.002	.095 (.017)	.000	.218 (.025)	.000
Cropprev1 (Reference = Cropprev0)	...	...	...	...	.811 (.081)	.000	...	...	...	...
Cropprev2	...	...	...	...	.106 (.027)	.000	.090 (.019)	.000	.110 (.026)	.000
Cropprev3	...	...	...	...	.331 (.064)	.000	...	...	.159 (.042)	.000
Damage1 (Reference = Damage2)	-.055 (.017)	.001	-.155 (.018)	.000	-.060 (.016)	.000	-.153 (.012)	.000	-.143 (.018)	.000
Irrg1 (Reference = Irrg2)	...	...	.452 (.060)	.000	...	...	...	...	...	...
Serrop1 (Reference = Serrop2)	...	...	-.120 (.022)	.000	-.410 (.021)	.000	.062 (.016)	.000	...	...
HHsex2 (Reference = HHsex1)	...	...	...	...	-.065 (.027)	.016	.052 (.018)	.004	.058 (.026)	.025
Zone1 (Reference = Zone10)	.511 (.035)	.000	.699 (.033)	.000	.577 (.025)	.000	-.585 (.035)	.000	.347 (.036)	.000
Zone2	-.055 (.033)	.095	.115 (.035)	.001	-.129 (.039)	.001	-.640 (.033)	.000	-.342 (.026)	.000
Zone3	.311 (.032)	.000	...	...	.156 (.029)	.000	-.326 (.033)	.000	-.096 (.029)	.001
Zone4	.197 (.033)	.000	.478 (.040)	.000	.207 (.028)	.000	-.201 (.032)	.000	...	...
Zone5	1.008 (.032)	.000	.831 (.042)	.000	.763 (.029)	.000	.532 (.033)	.000	.972 (.023)	.000
Zone6	.107 (.035)	.002	.528 (.037)	.000	.729 (.047)	.000	-.341 (.034)	.000	...	...
Zone7	...	...	.971 (.033)	.000	.429 (.121)	.000	-.662 (.035)	.000	.236 (.033)	.000
Zone8	.331 (.036)	.000	-.124 (.045)	.005	...	...	-.320 (.036)	.000	.130 (.039)	.001
Zone9	...	...	.984 (.035)	.000	-1.151 (.075)	.000	-.717 (.035)	.000	-.234 (.053)	.000

Table 3.3.3 The 95% Confidence Interval for Parameter Estimates of the Fitted Models by Crop Types

Effects	95% Confidence Intervals									
	Barley		Maize		Sorghum		Teff		Wheat	
	Lower bound	Upper bound	Lower bound	Upper bound	Lower bound	Upper bound	Lower bound	Upper bound	Lower bound	Upper bound
(Constant)	3.155	3.256	3.936	4.115	3.687	3.828	3.478	3.625	3.640	3.760
Seedtype2 (Reference = Seedtype1)	...	...	-.319	-.192	...	...	...	...	...	...
Fertliz1 (Reference = Fertliz0)	.029	.105	.013	.097	.129	.217	.076	.149	.027	.121
Fertliz2	-.460	-.307	.375	.498	...	...	.004	.067	.115	.199
Fertliz3	...	...	.197	.374	...	...	...	...	...	...
Ext1 (Reference = Ext2)	.156	.358	...	...	.090	.387	.061	.129	.170	.267
Cropprev1 (Reference = Cropprev0)	...	...	...	...	.652	.969	...	...	...	...
Cropprev2	...	...	...	...	.053	.159	.052	.128	.060	.160
Cropprev3	...	...	...	...	.207	.456	...	...	.077	.241
Damage1 (Reference = Damage2)	-.088	-.022	-.189	-.120	-.092	-.028	-.176	-.129	-.177	-.109
Irrg1 (Reference = Irrg2)	...	...	.334	.571	...	...	...	...	...	...
Serrop1 (Reference = Serrop2)	...	...	-.163	-.078	-.451	-.369	.031	.094	...	...
HHsex2 (Reference = HHsex1)	...	...	...	...	-.117	-.012	.017	.087	.007	.109
Zone1 (Reference = Zone10)	.443	.579	.633	.765	.528	.626	-.654	-.517	.276	.418
Zone2	-.120	.010	.047	.184	-.204	-.053	-.705	-.576	-.394	-.291
Zone3	.248	.375	...	...	.101	.212	-.391	-.261	-.153	-.039
Zone4	.133	.262	.399	.557	.153	.262	-.264	-.139	...	...
Zone5	.946	1.070	.748	.913	.707	.819	.467	.596	.927	1.017
Zone6	.039	.176	.457	.600	.636	.821	-.407	-.274	...	...
Zone7	...	...	.907	1.035	.193	.666	-.730	-.594	.172	.301
Zone8	.260	.401	-.212	-.037	...	...	-.391	-.249	.054	.206
Zone9	...	...	.915	1.052	-1.297	-1.004	-.787	-.648	-.337	-.130

The regression equation to explain the variability in the transformed mean yields of barley crop can be built into a final model as:

$$\widehat{Y_B^*} = 3.205 + (1.008)\text{Zone5} + (0.511)\text{Zone1} + (-0.383)\text{Fertliz2} + (0.311)\text{Zone3} + (0.331)\text{Zone8} + (0.257)\text{Ext1} + (0.197)\text{Zone4} + (-0.055)\text{Damage1}(0.067)\text{Fertliz1} + (0.107)\text{Zone6} \quad (3.6)$$

where: $\widehat{Y_B^*}$ is the transformed mean yield for Barley.

From the above prediction equation we can observe that the dummy variables Zone1, Zone3, Zone4, Zone5, Zone6, Zone8, Ext1, and Fertliz1which are included in the selected model, had positive significant effects whereas Fertliz2, and Damage1 had negative significant effects on the transformed mean yield for barley as compared to their respective reference category levels. The rest of the other dummy variables were found to have no significant effects on the transformed mean yield of barley as compared to their respective reference categories. Thus, they were not included in the prediction model for studying the variability in the mean yields of barley crops.

Similarly, the regression equations for the rest of the other four cereal types could be formulated and interpreted in the same manner as it has been done for Barley crop; but to avoid the repetitive nature of the interpretation, we only presented brief discussions of the regression results in Section 3.4 that follows.

3.4 Discussions and Conclusions on Results of the Multiple Regression

The ANOVA tables for all the five types of cereals based on the multiple regression analysis results show that the p-value was significantly smaller than the significance level $\alpha = 0.05$, which confirms the overall relevance of the fitted models to each crop data. Moreover, the R^2 value which indicates the percentage of variations in the mean yield that is explained by the independent variables included in the models for barley, maize, sorghum, teff and wheat crops are 22.4%, 32.5%, 25.2%, 34%, and 35.9% respectively. The multiple regression estimates of the effects of independent factors on cereal crop yields are summarized as follows:

Seed Type: Seed type is one of the factors considered to have an effect on the mean yields of cereal crops production. Obviously, farmers who apply improved seed types on their farms are expected to get more yields as compared to those who used non-improved seed types. Thus, the results for the studied cereal crop types indicated that the use of non-improved seed type significantly influenced the mean yield of maize crop to decrease by 0.256 units as compared to that of using improved seed type at 5% significance level. However, there was no significant difference between the use of non-improved and improved seed type effects on the mean yields for barley, sorghum, teff, and wheat crops. This could be due to the low level use of improved seed type on farms of these crop types in the study region (See Table 2.3).

Irrigation: Results presented in Table 2.3 in the preliminary analysis section clearly showed that irrigation was applied on less than 1% (0.88%) of the total cultivated cereal crop farms in the region. As a result, the effects of irrigation on the mean yields from barley, sorghum, teff, and wheat farms were not statistically significant at the 5% level. The reason for the non-significance of the effects of irrigation on these cereal crops mean yields could be associated with the inconsistent application of irrigation practices on farms in the region. Whilst, for maize crop the estimated coefficient for the irrigation factor level (Irrg1= 0.452) was statistically significant and positive. This positive effect could be interpreted as the effects of irrigating maize farms increases the mean yields by 0.452 units as compared to that of non-irrigating the farms.

Crop prevention measures: Crop prevention measures, according to the definition given in the CSA reports (2007), include weeding, hoeing and application of pesticide to control pest and disease on cereal crops. The results of the regression analysis for

transformed sorghum mean yield reveals that crop protection measures had statistically significant positive coefficients for non-chemical, chemical, and both chemical and non-chemical categories as compared to the effects of the reference category. This result indicated that the application of chemical, non-chemical, or both chemical and non-chemical crop prevention methods have significant influence on the transformed mean yields of sorghum. Besides, non-chemical prevention methods applied on teff and wheat crop farms resulted in a positive significant difference effect on their mean yields. On the other hand, for barley and maize crops, all dummies of the factor (i.e. crop prevention measures) have shown no significant difference effects on their transformed mean yields.

Fertilizer use: The application of either of the natural or chemical, or both combinations of natural and chemical types of fertilizers to cereal crop farms were expected to considerably increase mean yields of the crops as compared to non use of any of the fertilizer types. Therefore, the dummies for the use of fertilizer types were included in the models to investigate their influence on the mean yields of the studied crop types. Consequently, the SPSS stepwise regression analysis results showed that the estimated coefficients of Fertliz1 (a category representing "Natural fertilizer type") was positive and statistically significant in all the particular models of the crops, at $\alpha = 0.05$ level of significance. The results indicated that the use of fertilizer had contributed towards an increase in the transformed mean yields of the cereal crops as compared to no use of fertilizer. Furthermore, the higher values of the coefficients of chemical fertilizer level for maize and wheat indicate that the transformed mean yields of these crops could be more enhanced by applying chemical fertilizers as compared to the reference level (no use of fertilizer).

Extension Programme: There were positive effects of extension programmes on the transformed mean yields of all cereal crops except Maize. This could be due to the fact that farmers have applied improved agricultural inputs and better management practices on farms which were taught to them in the extension programme. This result highlights the need to bring more number of farmers (crop farms) into the extension programmes for increased productivity of the cereal crops.

Crop Damage: The frequency of occurrence of crop damages such as crop diseases, frost, flood, pests, weeds, etc on farm fields could greatly influence the gain in the transformed mean yields of cereal crops. The stepwise regression analysis results for each type of cereal crop types revealed that the "Crop Damage" factor labelled as "Damage1", representing the incidence of crop damage, had a significantly decreasing effect on the transformed mean yields of all the cereal crop types as compared to the effects with no crop damage (Damage2) level. This piece of evidence is reflected by the estimated negative coefficients of the factor level "Damage1" included in Table 3.3.2 for barley, maize, sorghum, teff, and wheat crop types respectively.

Prevention of Soil Erosion: Terracing, planting trees, ploughing along the contour are among the methods used for preventing the soil on farm fields from severe erosion. The results for the dummy variables of this factor indicated that for teff crops the protection measures has resulted in a positive significant difference effect on the mean yields of the teff crops, whereas its effect was negative on maize and sorghum cereals mean yields as compared to the effects with no protection measures for soil erosion on the respective cereals mean yields. The unexpected negative effects could implicate the

severity of the soil erosion problems in those particular areas producing maize and sorghum crops irrespective of the efforts made to protect the soil.

Gender of head of the household: Accounting for gender differences is important in view of the fact that adoption and use of new agricultural technologies and inputs could be affected by who owns and controls the crop farms. Thus, the gender factor dummy variable 'HHsex2', which stands for female headed households, was included in the regression model. The results from stepwise regression indicated that the female headed households had a positive significant difference effects on the mean yields of Teff and Wheat crops as compared to male headed households. Whereas, this effect of gender on the Sorghum mean yields was negative and significant at 5% level. It was observed from Table 3.3.2 that the prediction model equations for maize and barley could not include the gender factor since it fails to significantly affect the transformed mean yields of the crops as compared to its reference category (i.e. male headed households).

Zones: To account for the effects due to differences among the administrative zones in terms of, for instance, topographic and climatic variability, we included the zone dummy variables into each crop type regression models. Although it is not possible to clearly point out what is being controlled, the results from the regression analysis indicated that the inclusion of the zone dummies had resulted in a highly significant difference effects on the mean yields of all cereal crops. Moreover, a much better R^2 value of the fitted models than the R^2 values for models without the zone effects (See in Appendix Tables A.6 to A.10) were obtained.

Chapter 4

Theory and Application of Mixed Model

4.1 Introduction

Though the primary interest in this thesis is identifying the factors (fixed effects) which affect cereal crop yield, it is worthwhile to see if there is any effect associated with the sampling units (random effects) to improve our model efficiency. The ordinary regression model discussed in Chapter 3 may not be appropriate for this type of analysis, since it does not allow including the random sampling units effect (random effect) into the model. This drawback of the linear regression model could be overcome if we fit a linear mixed model to the data.

Fixed effects are effects which can be used only if our interest is in the effect of the levels of the factors used in the study, whereas the effect is random if the levels in the study are randomly selected and our interest is in the effects of the population of the levels of a factor or factors. Thus, a mixed model is a model which is capable of handling both the fixed and random effects simultaneously. Moreover, the capability of mixed models to deal with unbalanced data, and the possibility of predicting random effects through the Best Linear Unbiased Prediction (BLUP) methods are among the major features that make the mixed model more beneficial (Duchateau et al., 1998:18; cited in Ramroop, 2002). Consequently, in the following section, the theory of general linear mixed model is reviewed and the results of the fitted models to the transformed yield data are discussed.

4.2 The General Linear Mixed Model

A mixed effects model is a model which includes both fixed and random factors in one model. A factor is said to be fixed if all the levels of the factor are selected by a researcher to identify the effects of levels on the response variable of interest. The purpose of the fixed factors is to compare the effects of the levels on the response variable. Whereas, a factor is random if the effects associated with the levels of the factor can be viewed as being like a random sample from a population of effects. The purpose of random factors is to draw conclusions about variation in the population of random effects.

A linear mixed model can be formulated by generalizing the ordinary regression model which was represented in equation (3.1). Now, the linear mixed model takes a broad view of the regression model in the following way:

$$\mathbf{Y} = \mathbf{X\beta} + \mathbf{ZU} + \boldsymbol{\varepsilon} \qquad (4.1)$$

Where, **Y** represents the (n×1) vector of observed responses,

β is an unknown (p×1) vector of fixed-effects parameters,

X is known design matrix of dimension (n×p),

U is an unknown vector of random-effects parameters of dimension (q×1),

Z is known design matrix of dimension (n×q) and

ε is an unknown random error (n×1) vector of residual components.

The residuals (**ε)** are assumed to be independent and normally distributed with mean vector zero and covariance matrix $\sigma^2 I_n$; where I_n is an n×n identity matrix (Verbeke and Molenberghs, 2000).

The vectors $\mathbf{Y}$, $\boldsymbol{\beta}$ and $\boldsymbol{\varepsilon}$ have the same interpretation as in the linear regression case. Since the model (4.1) contains both fixed and random effects, it is referred to as a mixed effects model. The assumptions required of the linear mixed effects model are (Laird and Ware, 1982):

$$\mathbf{U} \sim \mathrm{N}(0, \mathbf{G}),$$

$$\boldsymbol{\varepsilon} \sim \mathrm{N}(0, \mathbf{R}), \text{and } \mathbf{U} \text{ and } \boldsymbol{\varepsilon} \text{ are independent}$$

where $\mathbf{G}$ denotes the variance-covariance matrix associated with the random effects and $\mathbf{R}$ denotes the variance-covariance matrix of the residuals. Note that the residuals are no longer required to be independent or homogeneous as it is assumed in the linear regression. Within the structure of the mixed effect model, the residuals (and random effects) can have correlated and heterogeneous variances. However, it is still required to assume that both residuals and random effects are normally distributed. In addition, we assume that the random effects are independent of the residuals. In matrix form, it can be summarized as:

$$\mathrm{E}\begin{bmatrix}\mathbf{U}\\ \boldsymbol{\varepsilon}\end{bmatrix} = \begin{bmatrix}0\\ 0\end{bmatrix} \text{ and } \quad \mathrm{Var}\begin{bmatrix}\mathbf{U}\\ \boldsymbol{\varepsilon}\end{bmatrix} = \begin{bmatrix}\mathbf{G} & 0\\ 0 & \mathbf{R}\end{bmatrix}.$$

The variance of $\mathbf{Y}$ is denoted by Var($\mathbf{Y}$) or simply $\mathbf{V}$, and it can be shown that

$$\mathbf{V} = \mathbf{Z}\mathbf{G}\mathbf{Z}' + \mathbf{R}. \qquad (4.2)$$

The linear mixed model implies the marginal model (Verbeke and Molenberghs, 2000),

$$\mathbf{Y} \sim \mathbf{N}(\mathbf{X}\boldsymbol{\beta}, \mathbf{Z}\mathbf{G}\mathbf{Z}' + \mathbf{R}) \implies \mathbf{Y} \sim \mathbf{N}(\mathbf{X}\boldsymbol{\beta}, \mathbf{V}).$$

It should be noted that if $\mathbf{R} = \boldsymbol{\sigma}_{\boldsymbol{\varepsilon}}^{\mathbf{2}}\,\mathbf{I}$ and $\mathbf{Z} = \mathbf{0}$ then the linear mixed effects model has identical structure to that of the standard linear model (SAS, 2004).

4.3 Estimation of the Model Parameters

The procedures for estimating parameters in the linear mixed effects model appears to be more complex than the standard regression model in that the mixed model requires us not only to obtain estimates of the unknown fixed parameters **β**, but also to obtain predictors of the unknowns in the random parameters **R**, and **U**. This complexity of estimating the parameters is attributable mainly to the dependence of the fixed effects on the estimates of the covariance parameters. Thus, before trying to estimate the parameters, there should be an appropriate means for adjusting the covariance structure of the data (SPSS, 2005).

There are several methods available, such as the ANOVA method, Hederson's methods I, II, & III, the maximum likelihood (ML), restricted maximum likelihood (REML), and the minimum norm quadratic unbiased estimation (MINQUE) in estimating the unknown parameters in the mixed model. Although many authors (Harville, 1977; Robinson, 1991 and Searle, Casella and McCulloch, 1992 cited in Zewotir and Galpin, 2005) do not seem to reach to consensus on any of the methods as the best way of estimating parameters, the ML, REML, and MINQUE are considered as standard estimating methods for the linear mixed models (Zewotir and Galpin, 2005).

In the SAS user's guide (2004), it is recommended that the ML and REML methods are the best approaches used in many situations for estimating the parameters in mixed models (Hartley and Rao 1967; Patterson and Thompson 1971; Harville 1974; Laird

and Ware 1982; Jennrich and Schluchter 1986). Also, Ramroop (2002) suggested, based on results from simulated data, that ML and REML methods work well in estimating the variance components of the linear mixed model.

In light of the fact that the applications for estimating parameters are computationally intensive, it is necessary to make use of statistical packages. Thus, the SAS Proc Mixed procedure, which is the most efficient procedure to analyze mixed effects models and which also implements the ML and REML techniques for estimating (predicting) the parameters, is used to analyze the crop yield data in this thesis. The next sections discuss these two parameter estimation (prediction) approaches.

4.3.1 Maximum Likelihood Estimation

The maximum likelihood approach makes inference based on estimators obtained from maximizing the log-likelihood function (Verbeke and Molenberghs, 2000):

$$l_{\mathrm{MLE}}(\boldsymbol{\theta}) = \frac{-\mathrm{n}}{2}\log(2\pi) - \frac{1}{2}\log|\mathbf{V}| - \frac{1}{2}(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})'\mathbf{V}^{-1}(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}) \qquad (4.3)$$

with respect to $\boldsymbol{\theta}$, where $\boldsymbol{\theta} = (\boldsymbol{\beta}', \boldsymbol{\alpha}')'$ and $\boldsymbol{\alpha}$ denotes the vector of all variance and covariance parameters (i.e. variance components) contained in equation (4.2). It can be shown that the maximum likelihood estimators (MLE) for the fixed effects parameters, $\boldsymbol{\beta}$, and random effects parameters, $\mathbf{U}$, obtained from maximizing (4.3) conditional on $\boldsymbol{\alpha}$ (i.e. $\mathbf{V}$) respectively are given by (Laird and Ware, 1982):

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{Y}$$

and

$$\widehat{\mathbf{U}} = \mathbf{G}\mathbf{Z}'\mathbf{V}^{-1}\left(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right)$$

which are dependent on **V**, the matrix of unknown variance components. Since all estimators (predictors) are dependent on **V** an iterative procedure is used to obtain solutions for all unknowns, based on some initial values of **V**. The iterative procedure refines the solution with successive iterations until a likelihood convergence criterion has been satisfied. Popular iterative procedures include the Newton-Raphson algorithm, and Fisher-Scoring method. The detailed discussion of these procedures is given in Littell et al. (2006) and Lindstrom and Bates (1988).

4.3.2 Restricted Maximum Likelihood Estimation

The MLE of the variance components does not take into account of the loss of degrees of freedom resulting from estimating the fixed effects. The Restricted Maximum Likelihood technique for estimating the fixed effects and the variance components does not suffer from this defect. The REML technique differs from MLE in that REML maximizes the portion of the likelihood function which is invariant to the fixed effects in the linear mixed model. It has been shown (Harville, 1974) that the log-likelihood function for the REML is given by:

$$l_{\text{REML}}(\boldsymbol{\theta}) = \frac{-(\text{n}-\text{p})}{2}\log(2\pi) - \frac{1}{2}\log|\mathbf{V}| - \frac{1}{2}(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})'\mathbf{V}^{-1}(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}) - \frac{1}{2}\log\left|\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right|. \quad (4.4)$$

The REML technique can be used to obtain estimates of the fixed effects parameter **β**, and the covariance component **α** by maximizing the REML log likelihood in equation (4.4) with respect to **θ** and **β**.

Comparatively, the REML estimation has better bias characteristics (Diggle, 1988), handles high correlations more effectively, and is less sensitive to outliers than ML, but cannot be used for model comparison of fixed effects. ML estimation ignores the degrees of freedom used up by fixed effects in mixed models, leading to underestimation of variance components. However, ML may nonetheless be preferred when comparing two models with different parameterizations of the same effect (for example, simple variable vs. quadratically transformed version of the variable), because ML is invariant to different parameterizations of a fixed effect but REML will treat different parameterizations as different models and compute different likelihood ratios.

4.4 The Covariance Structure

In mixed models, we have to specify the type of covariance structure to be assumed for random effects; so that it can be used as a starting point to work out the estimation for parameters in the ML or REML iterative procedures. The default covariance structure for random effects is termed the 'variance components' structure which assumes the variances of the random effects be independent of the random errors and their sum equals the variance of the dependent variable. This variance-covariance matrix is the basis for estimating between-groups effects. Few of the other possible covariance structure types for assumptions include Compound Symmetry (CS), Unstructured (UN), and Autoregressive (AR).

The Goodness of fit statistics such as Akaike Information Criterion (AIC) or Schwarz's Bayesian Criterion (BIC) are commonly used to select the best covariance structure type. This is achieved by comparing the AIC (or BIC) values for candidate models under different covariance structures and selecting the one with the lowest value on AIC

(or BIC) as the best model. Besides, it is also possible to use a likelihood ratio test based on the difference between a model with a given covariance structure assumption and another model under a different assumption (Verbeke and Molenbergh, 2000).

4.5 Interpretation of Parameter Estimates (Predictions)

The estimated fixed effects parameters are interpreted the same way as the regression coefficients interpreted in the ordinary regression analysis in Chapter 3 of this thesis. Whereas, the covariance estimates for random effects and random error (residual) are obtained from the SAS output labeled "Covariance Parameter Estimates" and hence, checking for significance of their estimates is possible since the table includes the standard errors and the test statistic corresponding to the estimates.

If the predicted variance component of the random factor has significant effect, we conclude that the dependent variable varies by the random effect. This further means that a fixed-effects analysis of dependent variables ignoring the random effect would violate the assumption of independence of observations since observations vary depending on random effects.

- **Intra-class correlation (ICC)**

For variance components models the intra-class correlation coefficient is calculated as the random effect variance divided by total variance. This indicates the percentage value of the variability accounted for by the random effects of the total variability. In equation form, the intra-class correlation coefficient can be formulated as:

$$\mathrm{ICC} = \frac{\boldsymbol{\sigma_i^2}}{\boldsymbol{\sigma_i^2 + \sigma_\varepsilon^2}}$$

where $\boldsymbol{\sigma_i^2}$ represents the variance component associated with the i^{th} random effect,

$\boldsymbol{\sigma_\varepsilon^2}$ is the variance component associated with the random error.

Note: To make comparisons between models fitted with only random effects and models fitted with both the random and fixed effects will help in drawing appropriate conclusions regarding the random as well as the fixed factors included in the models.

- Unconditional model: is a model only with a random variable and which excludes all the fixed factors to look at the variability in the response variable.
- Conditional model: is a model which includes a random variable and one or more fixed factors. If there is any difference in the fit between the conditional model and the corresponding unconditional model, then its significance can be evaluated by using the likelihood ratio tests and conclusions about the covariance components as well as the fixed factors in the model can be made accordingly.

4.6 Model Selection and Diagnostics

In mixed model analysis the selection of variables (fixed effects, random effects and covariance parameters) that enter the model can be conducted based on the Likelihood Ratio Tests (LRT). It should be noted that the likelihood ratio test for fixed effects assumes ML estimation, while for random effects and covariance components either the ML or REML estimation methods could be assumed.

There are a number of statistical measures for goodness of fit tests which includes

-2×Restricted Log Likelihood (-2RLL), Akaike Information Criterion (AIC), and Schwarz's Bayesian Criterion (BIC). The Akaike's Information Criterion (AIC) (Akaike, 1974) is defined as:

$$\mathrm{AIC} = -2l(\hat{\theta}) + 2\mathrm{p}$$

where $l(\hat{\theta})$ is the maximized log likelihood or the residual log likelihood, and p is the number of parameters in the model (SAS, 2004). It can be used to compare models with the same fixed effects but different variance structures; the model having the smallest AIC is deemed best. The Schwarz's Bayesian Criterion (BIC) (Schwarz, 1978) is computed as:

$$\mathrm{BIC} = -2l\left(\hat{\theta}\right) + \mathrm{plog}\mathrm{N}^{*}$$

where: N^{*} stands for the total number of observations (N) for ML estimation and $(\mathrm{N} - \mathrm{p})$ for REML estimation method. Again, we prefer models with smallest BIC, but note that BIC penalizes models with a greater number of covariance parameters more than AIC does, and the two criteria may not agree as to which covariance model is best (SAS, 2004).

If one covariance model is a sub model of another, it is possible to carry out a likelihood ratio test for the significance of the more general model by computing -2 times the difference between their log likelihoods (SAS, 2004). This test is used to determine whether it is necessary to model the covariance structure of the data at all. The "Chi-Square" value is -2 times the log likelihood from the null model minus -2 times the log likelihood from the fitted model, where the null model is the one with only the fixed

effects (i.e. $\chi^2 = -2\left(l(\hat{\beta}_o) - l(\hat{\beta})\right)$ where $l(\hat{\beta}_o)$ is the maximized log likelihood under H_0 and $l(\hat{\beta})$ is the maximized log likelihood over all β).

- **Influence diagnostics and detecting outliers**

Further diagnosing measures with regard to detecting outliers and identifying influential observations in mixed models are applicable by extending the statistical measures and graphical methods used in the ordinary linear models (for the details see, Zewotir and Galpin, 2005; Schabenberger, 2004). It is indicated that the extended diagnostics tools are analogues of, for example, Cook's distance (Cook, 1977), and likelihood distance (Cook and Weisberg, 1982) which were used to measure influence on the fixed factors (Zewotir and Galpin, 2005).

4.7 Application of the Mixed Model to the Data

The general linear mixed model discussed in this chapter is used to fit the yield data. In the fitted model, transformed cereal crop yields is taken as the response variable; and all the categorical fixed factors explained in Section 2.1.4, (Seed type, Fertilizer type, Extension program, Type of crop prevention, Crop damage, Protection of soil erosion, Crop irrigation, Household head Sex, and Zone), and a random factor (i.e. the Enumeration Area), are included in the model as the explanatory variables. The models used to fit the data are formulated as shown below:

$$\mathbf{Y} = \mathbf{X\beta} + \mathbf{ZU} + \boldsymbol{\varepsilon}$$

where **Y**: is a (40,235 x 1) vector of values resulting from a square root transformation of cereal crops mean yields,

β: is a (22 x 1) vector of the overall mean and the main effects parameters of the fixed effects; which represents: **β** = (β_0, seedtype1, Fertliz0, Fertliz1, Fertliz2, Cropp0, cropp1, Cropp2, Damage1, Ext1, Irrg1, Serrop1, HHsex1, Zone1, Zone2, Zone3, Zone4, Zone5, Zone6, Zone7, Zone8, and Zone9).

X: is the known design matrix of dimension (40,432 x 22),

U: represents a (439 x 1) vector of the random effect parameters (i.e. the 439 randomly selected Enumeration areas (EAs) in the region),

Z: is a matrix of size (40,432 x 439) for the random (EA) effects and,

ε: is the random error of size (40,432 x 1) vector of residual components.

Before summarizing the above general linear model equation in a matrix notation to represent the data, we would look at the Variance Components Model for estimating the variance components, in our case variances of the random variables (EAs) and the random effects (Residuals), as follows:

$$Y_{ijklmnopqrs} = \mu + \alpha_i + \beta_j + \gamma_k + \delta_l + \theta_m + \nu_n + \kappa_o + \chi_p + \tau_q + \eta_r + \varepsilon_{ijklmnopqrs}$$

Where: μ stands for the overall mean

α represents the seed type effect for i = 1; 2,

β represents the type of fertilizer use effect for j = 0; 1; 2; 3,

γ represents crop prevention effect for k = 0; 1; 2; 3,

δ represents crop damage effect for l = 1; 2,

θ represents extension programme for m = 1; 2,

ν represents crop irrigation for n = 1; 2,

κ represents prevention of soil erosion for o = 1; 2,

χ represents head of household sex for p = 1; 2,

τ represents the zone effect for q = 1; 2; . . . ; 10,

η represents the random effects of the random factor (EA) for r = 1; 2; . . . ; 439, and

ε represents the residual components for the specified levels.

The next section is devoted to fit the mixed model discussed in this chapter to our transformed cereals yield data and discuss the results obtained for each crop type.

4.8 Discussions and Conclusions on Results of the Mixed Models

Analyzing the crop yields data using mixed models procedures was implemented initially by fitting the unconditional means models (i.e. models with only the random factors, EAs) to the yield data of the cereal crops; and then the conditional means models (i.e. models including the random factors and all the fixed factors) are fitted. This helps to use the estimated outputs for covariance components as a basis for making comparisons with the subsequently fitted models and to draw appropriate conclusions regarding the effects of random factors as well as the fixed factors included in the models. Thus, the covariance parameter estimates corresponding to the unconditional and conditional mean yields models, by crop type, are presented in Table 4.1 and Table 4.2 respectively. In Table 4.1, it is revealed that the mean yields for all types of the cereal crops have shown significant variability between the random factors (EAs). This can be observed from Table 4.1, for example, that the unconditional model estimated the variance for barley mean yields due to the random effects, EAs, as $\sigma^2_{EA} = 0.5306$; and the random errors as $\sigma^2_{\varepsilon} = 2.013 \text{ x} 10^{-7}$.

Table 4.1 Unconditional Models Covariance Parameter Estimates by Crop Type

Crop Type	Covariance Parameter Estimates					
	Cov Parm	Subject	Estimate	Standard Error	Z Value	Pr Z
Barley	Intercept	EAID	0.5306	0.04171	12.72	<.0001
	Residual		2.013E-7	0	.	.
Maize	Intercept	EAID	0.9624	0.06929	13.89	<.0001
	Residual		2.38E-7	0	.	.
Sorghum	Intercept	EAID	0.6473	0.05518	11.73	<.0001
	Residual		1.575E-7	0	.	.
Teff	Intercept	EAID	0.4468	0.03267	13.68	<.0001
	Residual		1.06E-7	0	.	.
Wheat	Intercept	EAID	0.6903	0.05598	12.33	<.0001
	Residual		5.319E-7	0	.	.

Table 4.2 Conditional Models Covariance Parameter Estimates by Crop Type

Crop Type	Covariance Parameter Estimates					
	Cov Parm	Subject	Estimate	Standard Error	Z Value	Pr Z
Barley	Intercept	EAID	0.4490	0.03583	12.53	<.0001
	Residual		2.039E-7	0	.	.
Maize	Intercept	EAID	0.7681	0.05606	13.70	<.0001
	Residual		2.387E-7	0	.	.
Sorghum	Intercept	EAID	0.5427	0.04704	11.54	<.0001
	Residual		1.584E-7	0	.	.
Teff	Intercept	EAID	0.3408	0.02526	13.49	<.0001
	Residual		1.066E-7	0	.	.
Wheat	Intercept	EAID	0.5415	0.04462	12.14	<.0001
	Residual		5.052E-7	0	.	.

The estimated component of variances for the random factors suggested that the mean crop yield for barley varies considerably between the enumerations areas (EAs) in the region. Likewise, the estimated variance for residuals, which shows the unexplained variance in the mean yields for barley after controlling for the random sampling factor, EAs, was close to zero. This is true for the other four crop types as well, because their estimated residual values are similarly close to zero.

We can also make more interpretations by taking the estimated covariances for the random effects shown in the Tables 4.1 and 4.2 and computing for barley crop as $\frac{0.5306-0.4490}{0.5306} = 0.1538$. This result (0.1538) pointed out that only 15.38% of the portion of explainable variation is explained by the fixed factors included in the model for barley crop yield data. This means that a very small fraction of the EA to EA variation in the barley mean yields was explained by the included fixed factors. Or, in other words, there is still a large fraction of variation (84.62%) of the explainable EA to EA variations for barley mean yields remain unexplained. For maize, sorghum, teff, and wheat crops, the explained portion of variation by the fixed factors of the explainable EA to EA variations were calculated to be 20.19%, 16.16%, 23.72%, and 21.56% respectively. These values can be interpreted in a similar way as it was done with the values for barley mean yield.

- **Results for Type 3 tests of fixed effects**

Tests of fixed effects table, from SAS Proc Mixed Procedure outputs, can be used to test the collective effects of all levels of a categorical variable included in a model. Thus, for our fitted models, the table for [3]Type 3 tests of fixed effects by crop type is given in Tables 4.3, 4.4, 4.5, 4.6, and 4.7 for barley, maize, sorghum, teff, and wheat crops respectively. The results showed that the zone effects on the transformed mean yields of the crops were strongly significant for all cereal crop types; whereas crop prevention methods, gender of head of the household, and extension programme were significantly affecting the transformed mean yields of maize, sorghum, and teff crops respectively.

[3] Type 3 tests examine the significance of each partial effect, that is, the significance of an effect with all the other effects in the model (SAS, 2004).

Table 4.3 Type 3 Tests of Fixed Effects

for Transformed Barley Data

Effect	Num DF	Den DF	F Value	Pr > F
Zone	9	314	7.56	<.0001
Seedtype	1	1	0.00	0.9791
Fertliz	3	291	0.08	0.9696
Cropprev	3	149	0.66	0.5800
Damage	1	233	2.10	0.1489
Irrg	1	15	0.00	0.9994
Ext	1	51	0.00	0.9835
Serrop	1	118	0.00	0.9768
HHsex	1	186	0.26	0.6140

Table 4.4 Type 3 Tests of Fixed Effects

for Transformed Maize Data

Effect	Num DF	Den DF	F Value	Pr > F
Zone	9	375	11.94	<.0001
Seedtype	1	98	0.02	0.8787
Fertliz	3	516	0.69	0.5606
Cropprev	3	201	3.19	0.0248
Damage	1	332	0.03	0.8593
Irrg	1	61	0.03	0.8739
Ext	1	132	2.49	0.1173
Serrop	1	182	0.82	0.3666
HHsex	1	320	0.83	0.3625

Table 4.5 Type 3 Tests of Fixed Effects

for Transformed Sorghum Data

Effect	Num DF	Den DF	F Value	Pr > F
Zone	9	266	6.90	<.0001
Seedtype	1	1	0.00	0.9735
Fertliz	3	187	0.21	0.8880
Cropprev	3	170	0.14	0.9355
Damage	1	224	0.30	0.5860
Irrg	1	13	0.00	0.9914
Ext	1	18	0.13	0.7239
Serrop	1	118	0.14	0.7139
HHsex	1	194	5.09	0.0252

Table 4.6 Type 3 Tests of Fixed Effects for Transformed Teff Data

Effect	Num DF	Den DF	F Value	Pr > F
Zone	9	364	13.49	<.0001
Seedtype	1	11	0.03	0.8557
Fertliz	3	382	0.79	0.4978
Cropprev	3	161	0.48	0.6993
Damage	1	295	1.42	0.2347
Irrg	1	21	0.00	0.9935
Ext	1	119	8.41	0.0045
Serrop	1	152	1.51	0.2204
HHsex	1	249	0.25	0.6145

Table 4.7 Type 3 Tests of Fixed Effects for Transformed Wheat Data

Effect	Num DF	Den DF	F Value	Pr > F
Zone	9	294	10.34	<.0001
Seedtype	1	46	0.25	0.6172
Fertliz	3	317	0.04	0.9891
Cropprev	3	163	0.86	0.4639
Damage	1	217	0.75	0.3878
Irrg	1	17	0.01	0.9244
Ext	1	100	2.36	0.1273
Serrop	1	105	2.28	0.1338
Hhsex	1	167	0.11	0.7441

The factors such as seed type, type of fertilizer used, crop damage, crop irrigation, and protection of soil erosion, included in the model were found to have no significant difference effects between the respective categorical variable levels in terms of the mean yields the cereal crops.

These results of the mixed model analysis parameter estimates were completely different from the results obtained by applying OLS estimating methods of the multiple regression analysis in Chapter 3. This could be due to ignoring the EAs effect in the OLS analysis which resulted in biased parameter estimations; and the capability of mixed models to account for the variability in the mean yields due to the survey random

sampling units (i.e. the EAs). This suggests that the regression results based on the OLS methods for our data would likely be misleading since the assumptions of independence and homoscedasticity are being violated by the observed dependences of the crops mean yields on the Enumeration Areas.

In conclusion, the results implied that we need to use such as mixed models to our data and to include additional explanatory variables to explain the fraction of variations between the EAs which is not explained by the already included fixed factors. Furthermore, the results also justify the need for clustering of cereal crop yields within the enumeration areas (EAs).

Chapter 5

Cluster Analysis and its Application

5.1 Introduction

Cluster analysis is one of the multivariate methods used for displaying the similarities and dissimilarities between pairs of objects or cases in a set. The aim of cluster analysis is to identify the actual groups of objects or cases that are similar to each other but different from objects or cases in the other group (Kaufman, 1990).

The procedures for forming clusters or groups can be classified into two broad clustering methods: Hierarchical and Non-Hierarchical. In the hierarchical procedures, a hierarchy or tree-like structure is constructed to see the relationship among observations or individuals. In the non-hierarchical method a position in the measurement is taken as central place and distance is measured from such central point which is usually called a seed. Since it is not easy to identify the right central point, the non-hierarchical methods are rarely used in clustering (Romesburg, 1984).

5.2 Hierarchical Clustering

As mentioned above, hierarchical clustering method enables us to find successive clusters by using previously established clusters. The clusters could be formed based on either agglomerative or divisive method.

Agglomerative hierarchical clustering begins with every case being a cluster by itself and at successive steps, similar clusters are merged to form larger clusters until all cases

merged in one cluster. Divisive clustering starts with all set of cases/objects in one cluster and end up with each case/object as individual clusters.

- **Steps for Hierarchical Clustering**

In general, grouping of cases through the use of hierarchical cluster analysis methods follows the following three main steps:

i. Choose a statistic that quantifies how far apart or similar two cases are,
ii. Determine which clusters are to be merged at successive steps, and
iii. Decide on the number of clusters needed to represent the data.

- **Measure of Distance Between Clusters**

Distance (similarity) is a measure of how far apart (how similar) two objects are. Distance measures are smaller for objects/cases that are similar, while their similarity measures are large. Although there are many different definitions of distance and similarity, the most accepted agglomerative methods are discussed as under:

Single linkage (nearest neighbour): this method is based on the smallest distance between two cases in the different clusters,

Complete linkage (furthest neighbour): this is based on the distance between the two furthest points in a cluster, and

Average linkage: this is based on the average distance from samples in one cluster to samples in other clusters. All of these methods use some measure of distance between data points as a basis for creating groups. The most frequently used distance measure is the generalized Euclidian distance,

$$d_{ij} = \left\{\sum_{k=1}^{p}\left(x_{ik} - x_{jk}\right)^2\right\}^{1/2}$$

Where d_{ij} is the distance measure, which defines the distance between two clusters for combining clusters at each stage of clustering procedures, x_{ik} is the value of variable X_k for object i and x_{jk} is the value of the same variable for object j (Manly, 2005).

- **Standardizing Variables**

Standardizing of variables is necessary if variables are measured on different scales. This is due to the fact that variables with large values tend to contribute more to the distance measure than variables with small values. Thus, they are standardized in some way so that all of the variables will be equally important in determining the distance measure. In this thesis, however, all the variables considered for clustering are measured on the same measurement scale, and hence, that is not of a problem.

- **Plotting the Distances**

A visual representation of the distance at which clusters are combined can be displayed by a tree-like plot called dendrogram. The distinct clusters produced are then interpreted by observing the grouping and relating them with some practical meaning in terms of the objective in this research which is to find possible groupings of similar zones based on quantitative data, i.e. the area percentage value allocated for each crop types as compared to the total cultivated area of the cereals within each particular zones in the region.

- **Determining the number of clusters**

In order to have some measure to help deciding on the number of clusters, in addition to the visual assessment of the dendrogram, we can consider the statistics values of semi-partial R^2 (SPRSQ) and R-square (RSQ). SPRSQ value represents the decrease in the proportion of variance accounted for by joining the two clusters. RSQ indicates the proportion of variance accounted for by the clusters.

There are no general rules available for assessing whether or not values of the statistics SPRSQ and RSQ are small or large, but the relative changes in the values of the statistics as the number of clusters increase can be useful in determining the number of clusters. A marked decrease or increase for SPRSQ and RSQ respectively may indicate that a satisfactory number of clusters have been reached (SAS).

5.3 Discussions and Conclusions on Results of Clustering the Data

In this study we aimed to group the zones into similar classes which minimizes the variance within the classes and at the same time maximizes the variance between classes. The classes are determined based on the percentage value of the ratio of cultivated area for each type of the cereal crops to the total cultivated area of the crops within the respective zones in the region. The data collected during the year 2007/2008 main agricultural season of the region by CSA was used to calculate the magnitudes (percentage values) for the classification. Hence, the complete linkage hierarchical clustering results presented in the clustering history table (Table 5.1) clearly showed that Zone 4 and Zone 5 were the two closest areas at a distance of about 0.23 units apart. As the distance increases slightly larger, i.e. at about 0.27 units, Zone 3 and Zone 8

joined together and the joining of the other zones continued and finally at a distance of about 2.04 units, it ends with Zone 1 and Zone 10 joining with the other groups of zones in one cluster.

Table 5.1 Cluster History of the Complete Linkage Cluster Analysis

NCL	CLUSTERS JOINED		FREQ	SPRSQ	RSQ	NORM Max
9	**4**	**5**	2	0.0047	.995	0.2273
8	**3**	**8**	2	0.0065	.989	0.2673
7	**2**	**6**	2	0.0086	.980	0.3084
6	**7**	**9**	2	0.0126	.968	0.373
5	**1**	**10**	2	0.0237	.944	0.5113
4	**CL8**	**CL9**	4	0.0380	.906	0.653
3	**CL7**	**CL6**	4	0.1065	.800	0.8729
2	**CL3**	**CL4**	8	0.2770	.523	1.3768
1	**CL5**	**CL2**	10	0.5225	.000	2.0379

This result implied that the most similar group has been determined by Zone 4 and Zone 5. These two zones are neighbours sharing similar geographic position which contributed in the similarity of the proportion of farm lands allocated for cultivating each of the five cereal crop types. In contrast, the cluster determined by Zone 1 and Zone 10 had been distinctively different from the rest of the other clusters of zones. These two zones are characterized by their hot climatic conditions and being major Sorghum producing areas in the region where more than 50% (approximately 54% in Zone 1 and 70% in Zone 10) of the total cultivated cereal farms was allocated to sorghum crops in each Zone.

In Table 5.1 it is shown that the changes in SPRSQ are great when going from 1 to 2, 2 to 3, 3 to 4 and from 4 to 5 clusters. Whereas, the additional decrease from having 6 (or from 5 to 6) clusters is not that large as compared to the decrease from 4 to 5

clusters, indicating that the choice of five clusters would be reasonable. Likewise, the RSQ value also shows that going from 1 to 5 numbers of clusters yields a large gain whereas additional clustering does not produce large increase in RSQ value indicating how well the clusters are separated. Thus, based on these results obtained from SPRSQ and RSQ values and by considering the interpretability of the clusters, we determine that five groups of clusters would be satisfactory.

The identified five groups of clusters resulting from the complete linkage analysis were as follows:

Cluster 1 “Zone 4 and Zone 5”: mainly characterized by cultivating wheat and barley crop types;

Cluster 2 “Zone 3 and Zone 8”: mainly characterized by cultivating barley and sorghum crop types;

Cluster 3 “Zone 2 and Zone 6”: mainly characterized by cultivating teff and wheat crop types;

Cluster 4 “Zone 7 and Zone 9”: mainly characterized by cultivating maize and teff crop types; and

Cluster 5 “Zone 1 and Zone 10”: mainly characterized by cultivating sorghum crops.

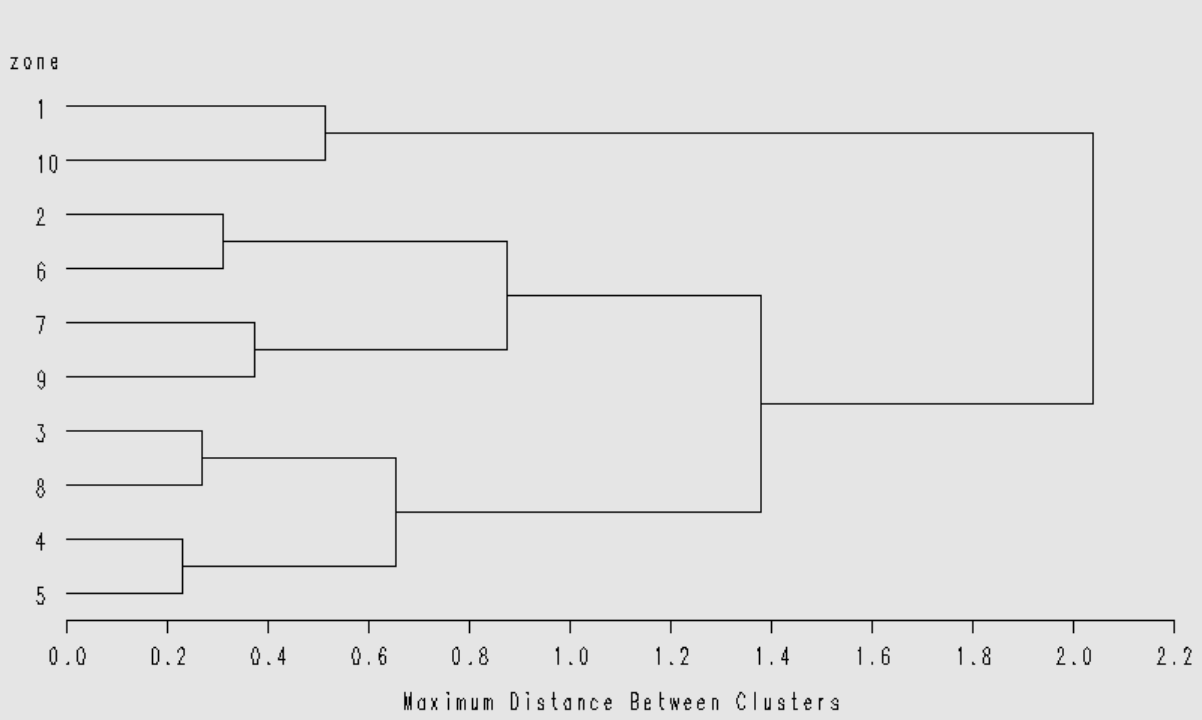


Figure 5.1 Dendrogram of zones obtained from cluster analysis of complete linkage analysis method.

Except for Cluster 5, the groupings of the zones revealed by the dendrogram (Fig. 5.1) were consistent with geographic locations of the zones in the region. There were also close relationship in the cultivated cereal crop types in the zones grouped within the same cluster and differences in crop types between different clusters. Thus the clustering pattern of the zones could also reflect the distribution patterns of different cereal crop types cultivated in the region since the apparent differences in the geographical and environmental conditions revealed the distribution of different cereal crop types cultivated in the region. Furthermore, this could be considered as the main reason for the similarities of the cultivated crops within each of the five groupings of the zones by the cluster analysis. This result brought to light the fact that the yield variability of cereals in the region could also be influenced by differences in the zones geographical and climatic conditions.

Chapter 6

Summary and Conclusions

The primary objective of this study was to identify the factors influencing the yield of main cereal crops using the data collected by CSA in the Amhara Nationals Regional State of Ethiopia. The factors along with some of the agricultural practices presented in the study included: seed type, fertilizer type, crop prevention measures, crop damage, extension program, irrigation, prevention of soil erosion, gender of the head of the household, and the zones. The effects of these factors on the transformed mean yields of barley, maize, sorghum, teff and wheat crops were investigated using stepwise multiple regression analysis by applying the ordinary least squares (OLS) estimation method.

From the stepwise regression, the R^2 values of 0.224, 0.325, 0.252, 0.340, and 0.359 resulted for barley, maize, sorghum, teff, and wheat models respectively; and these values actually account for the percentage of variation in the transformed mean yields of the cereals that is explained by the independent variables included in the models. They translate to 22.4%, 32.5%, 25.2%, 34%, and 35.9% respectively.

In general, the most important factors that are identified to significantly influencing the transformed mean yields of the studied cereals are summarised and presented in Table 6.1. The summaries are made based on the regression results summarised in Section 3.4 and the "Type 3 Tests of Fixed Effects" tables (see the tables by crop type in Appendix Tables A.11-A.15) obtained by fitting a linear mixed model to the transformed data only

for fixed factors (i.e. no random factor effects except the error term). Hence, the combined effects of all levels of a factor which significantly influences mean yields of the particular cereals are identified and presented as shown in the table below.

Table 6.1 Summary for the Type 3 Tests of Significance for Fixed Effects by Crop Type

Crop types	Factors significantly affecting the mean yields of cereals
Barley	Zone, Fertilizer type, Extension programme, and Crop damage.
Maize	Zone, Seed type, Fertilizer type, Crop damage, Irrigation, and Protection of soil erosion.
Sorghum	Zone, Fertilizer type, Crop prevention method, Extension programme, Crop damage, Protection of soil erosion, and Gender of the household head.
Teff	Zone, Fertilizer type, Crop prevention method, Extension programme, Crop damage, Protection of soil erosion, and Gender of the household head.
Wheat	Zone, Fertilizer type, Crop prevention method, Extension programme, Crop damage, and Gender of the household head.

The summarised results of the Type 3 tests of the fixed factors in the above table depicted that the location factor (zone), the type of fertilizers applied on farms and the effects of crop damages (i.e. crop diseases, pests, flood, frost, locust etc.) were the factors in all cereal models which have significantly influenced mean yields of the cereals in the region. Likewise, the effects of extension programmes have shown to have significant effect on the mean yields of all, except maize, cereal crops. There were also no significant differences of effects with respect to seed type and irrigation factors as compared to their respective reference categories on the transformed mean yields of

all cereal crops except maize. The application of crop prevention methods on the cereal farms had significant difference effects between its categorical levels on the mean yields of sorghum, teff and wheat crop types. However, it fails to significantly differ between its levels to influence the transformed mean yields of barley and maize crops. In addition, the transformed mean yields of these three crop types were affected by difference effects between the levels of gender which highlights the importance of accounting for gender differences when dealing with the productivity of cereal crops in the region.

Subsequently, the linear mixed model was applied to the transformed data to improve our model efficiency and to see effects associated with the random effects. The results (see Tables 4.3 - 4.7) show that the difference effects between the levels of the zone variable were strongly significant in influencing the transformed mean yields of all the cereal types. In addition to this, crop prevention methods, gender, and extension programmes were found to have significant difference effects between their levels on the transformed mean yields of maize, sorghum and teff crops respectively. The other factors, such as seed type, the type of fertilizer use, crop damage, crop irrigation, and protection of soil erosion, were found to have no significant difference effects between their levels on the transformed mean yields of the cereals. These results were entirely different from the prior summaries made on the OLS estimates of the multiple regression analysis. This could be as a result of the capability of mixed models to account for the variability in the transformed mean yields due to the random effects (i.e. the EAs). This suggests that the regression results based on the OLS methods for our data must be interpreted with caution in view of the observed dependencies of the transformed mean yields of the crops on the enumeration areas. Furthermore, it

confirms the need to use models which are capable of accounting for the random effects, such as mixed models, instead of applying the ordinary regression models to our data.

Finally, cluster analysis was implemented to identify the similarities and dissimilarities of zones in different groups, using the complete linkage hierarchical clustering methods. The clustering methods were implemented based on the percentage of cultivated area values for the particular cereal crop within the zone. The five identified groups of classes (clusters) resulting from the analysis seems to be in agreement with the type of crops produced in the respective geographical locations. These clusters of zones were identified as: cluster 1 ‘Zone 4 and Zone 5’ (or S.Wello and N.Shewa); cluster 2 ‘Zone 3 and Zone 8’ (or N.Wello and Wag-Hemra); cluster 3 ‘Zone 2 and Zone 6’ (or S.Gondar and E.Gojam); cluster 4 ‘Zone 7 and Zone 9’ (or W.Gojam and Awi); and cluster 5 ‘Zone 1 and Zone 10’ (or N.Gondar and Oromia zone).

In conclusion, the study showed that the yield variability of cereals in the region is strongly influenced by the differences in their locations (zones). This could be reflected mainly by the differences in zonal environmental and geo-climatic conditions. Likewise the significance of the difference effects between the levels of extension programme on transformed mean yields of the cereals points us in the direction for the need to bring more number of farmers (cereal farms) into the extension programmes. This is due to the fact that the possibility of applying improved agricultural inputs and better management practices of the farmers would increase accordingly. The low level use of improved seed types as well as the existing inconsistent application of irrigation practices in the region could be considered among the main reasons for their insignificance of the difference effects between their levels to influence the transformed

mean yields of the cereals (see Table 2.3). This shows the importance to increase application of improved seeds and better irrigation practices on the cereal farms. Moreover, efforts should be made to provide farmers with the best possible means of reducing the prevalent crop damages so as to enhance the productivity of cereal crops in the region.

Cereal crop production in Ethiopia in general and in the Amhara National Regional State in particular, is characterized by its reliance on low input usage and high dependence on rain fed agriculture. In such type of agriculture unsteadiness concerning the yield of cereal crops was mainly the result of variation in weather condition (Alemu, 2005, Jaeger, 1991). Thus, the limitations associated with this study are due to the nature of the available data which fails to include factors regarding weather variability. Therefore, future studies in this area need to incorporate the climatic factors and other relevant additional variables in their data.

References

Akaike, H. (1974). A New Look at the Statistical Model Identification. *IEEE Transaction on Automatic Control*, AC - 19, 716 - 723.

Alemu, Z.G. (2005). Causes of Instability in Cereal Production in Ethiopia. Unpublished Working paper, University of Free State.

Bakhsh, K., Hassan, I., and Maqbool, A. (2005). Factors Affecting Cotton Yield: A Case Study of Sargodha (Pakistan). *Journal of Agriculture and Social Sciences 1813–2235/2005/01–4–332–334.* Retrieved in March 2009 from http://www.ijabjass.org.

Berk, A., Richard (2004). Regression Analysis: A Constructive Critique. Advanced Quantum Techniques in the Social Science Series. SAGE Publication Inc., No.11. Thousand Oaks.

Bowerman, L.B., O' Connell, T.R., and Dickey, A.D. (1986). Linear Statistical Methods: An Applied Approach. Duxbury Press, Boston.

Box, G. E. P., & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society, B, 26*(211-234).

CSA (Central Statistical Agency, Ethiopia) (2007). Agricultural Sample Survey for 2006/07, Report on Area and Production for Major Crops, Statistical Bulletin 388 Vol. I. Addis Ababa.

CSA (Central Statistical Agency, Ethiopia) (2008). Summary and Statistical Report of the 2007 Population and Housing Census. Population and Housing Census Commission / Central Statistical Agency, Addis Ababa, Ethiopia. Retrieved in February 2009 from www.csa.et.gov.

Cook, R. D. (1977). Detection of Influential Observation in Linear Regression. *Technometrics 19: 15 - 18.*

Cook, R. D. and S. Weisberg (1982). Residuals and Influence in Regression. Chapman and Hall, New York.

Diggle, P. J. (1988). An Approach to the Analysis of Repeated Measurements. *Biometrics 44, 959-971*.

Draper, N.R. and Smith, H. (1966). Applied Regression Analysis. John Wiley & Sons, Inc., New York.

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

Hartley, H. O. and J. N. K. Rao (1967). Maximum Likelihood Estimation for the Mixed Analysis of Variance Model. *Biometrics 23: 93 - 108.*

Harville, D. A. (1974). Bayesian Inference for Variance Components Using Only Error Contrasts. *Biometrica 61: 383 - 385*.

Harville, D. A. (1977). Maximum Likelihood Approaches to Variance Component Estimation and to Related Problems (with discussion). *Journal of the American Statistical Association* **72**, 320-340.

Jaeger, W. (1991). The Impact of Policy in African Agriculture, an Empirical Investigation. *WB Working Paper No. 147.*

Jennrich, R. I., and Schluchter, M.D. (1986). Unbalanced Repeated-Measures Models with Structured Covariance Matrices. *Biometrics, 38, 967–974.*

Kaufman, Leonard and Rousseeuw (spelling?) Peter J. (1990). Finding Groups in Data: An Introduction to Cluster Analysis. Wiley-Interscience, New York.

Laird, N. M. and J. H. Ware (1982). Random-Effects Models for Longitudinal Data. *Biometrics 38: 963-974.*

Lindstrom, M. J., and Bates, D. M. (1988). Newton-Raphson and EM Algorithms for Linear Mixed-Effects Models for Repeated-Measures Data. *Journal of the American Statistical Association, 83, 1014–1022.*

Littell, R. C., G. A. Milliken, Stroup, W.W., Wolfinger, R. D. and Schabenberger, O. (2006). SAS Systems for Mixed Models, 2nd Edition. Cary NC: SAS Institute Inc.

Manly, F.J., Bryan (2005).Multivariate Statistical Methods: A primer, Third Edition. Western EcoSystems Technology, Inc. Laramie, Wyoming, USA.

MEDaC (Ministry of Economic Development and Cooperation) (1999). Survey of the Ethiopian economy, Review of Post Reform Developments (1992/93-1997/98). Addis Ababa, Ethiopia.

NMSA (National Meteorological Services Agency, Ethiopia) (2001). Initial National Communication of Ethiopia to the United Nations Framework Convention on Climate Change (UNFCCC). Addis Ababa, Ethiopia.

Patterson, H.D. and Thompson, R. (1971). Recovery of Inter-block Information when Block Sizes are Unequal. *Biometrica 58, 545 – 554.*

Ramroop, S. (2002). An Approach to Estimating the Variance Components to Unbalanced Cluster Sampled Survey Data and Simulated Data. MSc Thesis, UNISA.

Rawlings, J.O. (1988). Applied Regression Analysis: A Research Tool, Wadsworth and Brooks/Cole Advanced Books and Software. Pacific Grove, CA.

Robinson, G. K. (1991). That BLUP is a Good Thing: The Estimation of Random Effects (with discussion). *Statistical Science* **6**, 15-51.

Romesburg Charles (1984). Cluster Analysis for Researchers. Lifetime Learning, Belmont, California.

Samuel Gebreselassie (2006). Intensification of Smallholder Agriculture in Ethiopia: Options and Scenarios. Policy Paper Prepared for the Future Agricultures Consortium and Presented at Institute Development Studies, University of Sussex.

SAS Institute Inc. (2004). SAS/STAT 9.1 User's Guide. Cary, NC: SAS Institute Inc.

Schabenberger, O. (2004). Mixed Model Influence Diagnostics. Cary, NC SAS Institute Inc. Retrieved on 02/11/2009 from http://suport.sas.com/rnd/app/papers/abstracts/mixeddiag.html.

Schwarz, G. E. (1978). Estimating the Dimension of a Model. *Annals of Statistics Volume 6 Number 2: 461-464.*

Searle, S. R., Casella, G. and McCulloch, C. E. (1992). *Variance Components.* Wiley

SPSS, Inc. (2005). Linear Mixed-Effects Modelling in SPSS. SPSS white paper. Retrieved in November 2009 from http://www.spss.com/home_page/wp127.htm.

Verbeke, G., and Molenberghs, G. (2000). Linear Mixed Models for Longitudinal Data. Springer.

Weisberg, S. (1985). Applied Linear Regression, Second Edition. John Wiley and Sons, Inc. New York.

Zewotir, T. and J. S. Galpin (2005). Influence diagnostics for linear mixed models. *Journal of Data Science 3: 153-177.*

Appendix A: Additional Tables

Appendix Tables

Table A.1 Model Summary for Transformed Barley Data

Step-wisely included factors	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
				R Square Change	F Change	df1	df2	Sig. F Change
Zone5	.155	.155	.68305	.155	1205.012	1	6577	.000
Zone1	.177	.176	.67425	.022	173.842	1	6576	.000
FERTLIZ2	.192	.192	.66781	.016	128.522	1	6575	.000
Zone2	.206	.206	.66221	.014	112.614	1	6574	.000
Zone3	.210	.210	.66050	.004	35.104	1	6573	.000
Zone8	.216	.215	.65833	.005	44.388	1	6572	.000
EXT1	.218	.217	.65733	.003	21.054	1	6571	.000
Zone4	.221	.220	.65619	.003	23.814	1	6570	.000
DAMAGE1	.222	.221	.65566	.001	11.525	1	6569	.001
FERTLIZ1	.223	.222	.65524	.001	9.457	1	6568	.002
Zone6	.224	.223	.65482	.001	9.395	1	6567	.002

Table A.2 Model Summary for Transformed Maize Data

Step-wisely included factors	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
				R Square Change	F Change	df1	df2	Sig. F Change
Fertliz2	.150	.150	.93321	.150	1699.617	1	9612	.000
Zone7	.180	.180	.91680	.030	348.128	1	9611	.000
Zone9	.226	.226	.89068	.046	573.172	1	9610	.000
Zone1	.250	.250	.87690	.024	305.246	1	9609	.000
Zone5	.272	.271	.86418	.022	285.962	1	9608	.000
Zone6	.289	.288	.85401	.017	231.175	1	9607	.000
Zone4	.299	.299	.84781	.010	142.074	1	9606	.000
Seedty2	.306	.306	.84340	.007	101.671	1	9605	.000
Damage1	.312	.311	.84034	.005	71.184	1	9604	.000
Fertliz3	.316	.315	.83792	.004	56.505	1	9603	.000
Irrg1	.320	.319	.83545	.004	57.961	1	9602	.000
Serro1	.322	.321	.83394	.003	35.631	1	9601	.000
Zone2	.324	.323	.83301	.002	22.578	1	9600	.000
Zone8	.324	.323	.83267	.001	8.863	1	9599	.003
Fertliz1	.325	.324	.83242	.001	6.686	1	9598	.010

Table A.3 Model summary for Transformed Sorghum Data

Step-wisely included factors	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
				R Square Change	F Change	df1	df2	Sig. F Change
Zone1	.070	.070	.76108	.070	550.038	1	7257	.000
Zone5	.142	.141	.73142	.071	601.461	1	7256	.000
Zone9	.177	.177	.71628	.035	310.985	1	7255	.000
Zone6	.195	.195	.70842	.018	162.878	1	7254	.000
Cropp1	.213	.212	.70064	.018	163.000	1	7253	.000
Serrop1	.227	.226	.69437	.014	132.516	1	7252	.000
Fertliz1	.233	.232	.69169	.006	57.425	1	7251	.000
Zone4	.238	.237	.68949	.005	47.267	1	7250	.000
Zone3	.242	.241	.68746	.005	43.869	1	7249	.000
Zone7	.244	.243	.68676	.002	15.812	1	7248	.000
Cropp3	.246	.244	.68610	.002	14.925	1	7247	.000
Cropp2	.247	.246	.68533	.002	17.244	1	7246	.000
Damage1	.249	.247	.68477	.001	13.008	1	7245	.000
Zone2	.250	.248	.68430	.001	10.985	1	7244	.001
Ext1	.251	.249	.68385	.001	10.382	1	7243	.001
HHsex2	.252	.250	.68363	.001	5.817	1	7242	.016

Table A.4 Model Summary for Transformed Teff Data

Step-wisely included factors	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
				R Square Change	F Change	df1	df2	Sig. F Change
Zone5	.238	.238	.58049	.238	3259.328	1	10452	.000
Zone4	.259	.258	.57249	.021	295.007	1	10451	.000
Zone6	.272	.272	.56733	.013	192.120	1	10450	.000
Zone3	.284	.283	.56285	.012	167.879	1	10449	.000
Damage1	.292	.291	.55965	.008	120.974	1	10448	.000
Zone8	.300	.299	.55649	.008	119.851	1	10447	.000
Fertliz1	.305	.305	.55437	.005	81.151	1	10446	.000
Zone9	.308	.307	.55333	.003	40.237	1	10445	.000
Zone2	.310	.310	.55245	.002	34.514	1	10444	.000
Zone7	.316	.315	.55026	.006	84.136	1	10443	.000
Zone1	.335	.334	.54258	.019	298.618	1	10442	.000
Ext1	.337	.336	.54180	.002	31.183	1	10441	.000
Cropp2	.338	.337	.54135	.001	18.386	1	10440	.000
Serrop1	.339	.338	.54103	.001	13.486	1	10439	.000
HHsex2	.339	.338	.54083	.001	8.767	1	10438	.003
Fertliz2	.340	.339	.54072	.001	4.934	1	10437	.026

Table A.5 Model Summary for Transformed Wheat Data

Step-wisely included factors	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
				R Square Change	F Change	df1	df2	Sig. F Change
Zone5	.278	.278	.65651	.278	2342.105	1	6092	.000
Zone2	.307	.307	.64291	.030	261.572	1	6091	.000
Zone7	.318	.317	.63821	.010	90.918	1	6090	.000
Zone1	.329	.328	.63306	.011	100.598	1	6089	.000
Cropp2	.335	.334	.63020	.006	56.420	1	6088	.000
Damage1	.341	.341	.62723	.006	58.729	1	6087	.000
Ext1	.344	.343	.62602	.003	24.554	1	6086	.000
Fertliz2	.351	.350	.62270	.007	66.126	1	6085	.000
Zone9	.353	.352	.62182	.002	18.166	1	6084	.000
Zone3	.355	.354	.62094	.002	18.271	1	6083	.000
Cropp3	.356	.355	.62027	.001	14.022	1	6082	.000
Zone8	.357	.356	.61983	.001	9.765	1	6081	.002
Fertliz1	.358	.357	.61938	.001	9.742	1	6080	.002
HHsex2	.359	.357	.61918	.001	4.998	1	6079	.025

Table A.6 Model Summary for Transformed Barley Data without Zone Effect

Model	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
				R Square Change	F Change	df1	df2	Sig. F Change
1	.013	.013	.73822	.013	85.443	1	6577	.000
2	.020	.020	.73557	.007	48.387	1	6576	.000
3	.022	.022	.73478	.002	15.087	1	6575	.000
4	.024	.023	.73435	.001	8.750	1	6574	.003
5	.025	.024	.73403	.001	6.811	1	6573	.009
6	.025	.025	.73377	.001	5.537	1	6572	.019

Table A.7 Model Summary for Transformed Maize Data without Zone Effect

Model	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
				R Square Change	F Change	df1	df2	Sig. F Change
1	.150	.150	.93321	.150	1699.617	1	9612	.000
2	.175	.175	.91950	.025	289.813	1	9611	.000
3	.188	.187	.91254	.013	148.117	1	9610	.000
4	.193	.193	.90936	.006	68.459	1	9609	.000
5	.197	.197	.90730	.004	44.668	1	9608	.000
6	.200	.200	.90561	.003	36.866	1	9607	.000
7	.201	.201	.90493	.001	15.392	1	9606	.000
8	.203	.202	.90430	.001	14.332	1	9605	.000
9	.204	.203	.90380	.001	11.663	1	9604	.001
10	.204	.203	.90348	.001	7.901	1	9603	.005

Table A.8 Model Summary for Transformed Sorghum Data without Zone Effect

Model	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
				R Square Change	F Change	df1	df2	Sig. F Change
1	.050	.050	.76923	.050	385.383	1	7257	.000
2	.066	.066	.76278	.016	124.421	1	7256	.000
3	.073	.073	.76013	.007	51.608	1	7255	.000
4	.076	.075	.75914	.003	19.973	1	7254	.000
5	.078	.077	.75831	.002	16.803	1	7253	.000
6	.082	.081	.75678	.004	30.457	1	7252	.000

Table A.9 Model Summary for Transformed Teff Data without Zone Effect

Model	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
				R Square Change	F Change	df1	df2	Sig. F Change
1	.019	.018	.65868	.019	197.099	1	10452	.000
2	.023	.023	.65718	.005	49.042	1	10451	.000
3	.028	.028	.65553	.005	53.460	1	10450	.000
4	.031	.031	.65443	.003	36.097	1	10449	.000
5	.035	.034	.65335	.003	35.784	1	10448	.000
6	.037	.036	.65277	.002	19.342	1	10447	.000
7	.037	.037	.65255	.001	8.313	1	10446	.004
8	.038	.037	.65242	.000	5.069	1	10445	.024
9	.038	.037	.65230	.000	4.823	1	10444	.028

Table A.10 Model Summary for Transformed Wheat Data without Zone Effect

Model	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
				R Square Change	F Change	df1	df2	Sig. F Change
1	.029	.029	.76119	.029	181.951	1	6092	.000
2	.044	.043	.75547	.015	93.471	1	6091	.000
3	.046	.046	.75455	.002	15.869	1	6090	.000
4	.051	.051	.75254	.005	33.593	1	6089	.000
5	.053	.052	.75192	.002	11.062	1	6088	.001
6	.055	.054	.75131	.002	10.885	1	6087	.001
7	.056	.055	.75084	.001	8.587	1	6086	.003
8	.057	.056	.75052	.001	6.312	1	6085	.012

Table A.11 Type 3 Tests of Fixed Effects for Transformed Barley Data (no random factor, EAs)

Effect	Num DF	Den DF	F Value	Pr > F
ZONE	9	6557	187.96	<.0001
SEEDTYPE	1	6557	1.18	0.2782
FERTLIZ	3	6557	38.22	<.0001
CROPPREV	3	6557	1.46	0.2235
DAMAGE	1	6557	11.14	0.0009
IRRG	1	6557	2.05	0.1523
EXT	1	6557	25.38	<.0001
SERROP	1	6557	0.70	0.4035
HHSEX	1	6557	2.75	0.0975

Table A.12 Type 3 Tests of Fixed Effects for Transformed Maize Data (no random factor, EAs)

Effect	Num DF	Den DF	F Value	Pr > F
ZONE	9	9592	191.07	<.0001
SEEDTYPE	1	9592	48.25	<.0001
FERTLIZ	3	9592	57.00	<.0001
CROPPREV	3	9592	1.46	0.2239
DAMAGE	1	9592	78.70	<.0001
IRRG	1	9592	53.63	<.0001
EXT	1	9592	1.50	0.2214
SERROP	1	9592	32.06	<.0001
HHSEX	1	9592	1.72	0.1894

Table A.13 Type 3 Tests of Fixed Effects for Transformed Sorghum Data (no random factor, EAs)

Effect	Num DF	Den DF	F Value	Pr > F
ZONE	9	7237	182.76	<.0001
SEEDTYPE	1	7237	0.04	0.8352
FERTLIZ	3	7237	20.88	<.0001
CROPPREV	3	7237	35.83	<.0001
DAMAGE	1	7237	14.60	0.0001
IRRG	1	7237	0.42	0.5182
EXT	1	7237	13.27	0.0003
SERROP	1	7237	168.69	<.0001
HHSEX	1	7237	5.81	0.0160

Table A.14 Type 3 Tests of Fixed Effects for Transformed Teff Data (no random factor, EAs)

Effect	Num DF	Den DF	F Value	Pr > F
ZONE	9	1E4	530.41	<.0001
SEEDTYPE	1	1E4	0.66	0.4182
FERTLIZ	3	1E4	15.23	<.0001
CROPPREV	3	1E4	8.90	<.0001
DAMAGE	1	1E4	162.19	<.0001
IRRG	1	1E4	3.14	0.0766
EXT	1	1E4	27.13	<.0001
SERROP	1	1E4	14.11	0.0002
HHSEX	1	1E4	8.71	0.0032

Table A.15 Type 3 Tests of Fixed Effects for Transformed Wheat Data (no random factor, EAs)

Effect	Num DF	Den DF	F Value	Pr > F
ZONE	9	6072	321.32	<.0001
SEEDTYPE	1	6072	2.02	0.1549
FERTLIZ	3	6072	26.40	<.0001
CROPPREV	3	6072	23.88	<.0001
DAMAGE	1	6072	66.56	<.0001
IRRG	1	6072	0.61	0.4360
EXT	1	6072	61.08	<.0001
SERROP	1	6072	0.05	0.8275
HHSEX	1	6072	5.11	0.0238