Statistical Modelling of Availability of Major Food Cereals in Lesotho: Application of Regression Models and Diagnostics

By

Makhala Bernice Khoeli

Submitted in fulfillment of the academic
requirements for the degree of

DOCTOR OF PHILOSOPHY
in
Statistics

in the
School of Mathematics, Statistics and Computer Science
University of KwaZulu-Natal
Pietermariztburg
2012

Dedication

To my parents, my lovely son Katleho, my brother and sisters

Declaration

The research work described in this thesis was carried out in the School of Mathematics, Statistics and Computer Science, University of KwaZulu-Natal, Pietermaritzburg, under the supervision of Prof. Henry G. Mwambi.

The thesis presents original work by the author and has not been submitted in any form for any degree or diploma to any University. Where use has been made of the work of others it is duly acknowledged in the text.

November, 2012.

Student: _____ _____

Makhala B. Khoeli Date

Supervisor: _____ _____

Prof. Henry G. Mwambi Date

Acknowledgments

Firstly, I wish to thank God for the strength He gave me to be able to complete this work, despite my ailing health. I would like to express my sincerest gratitude to my supervisor, Professor Henry Mwambi, for his guidance, encouragement, and his patience in helping me complete this Ph.D. His humility and strength of purpose have urged me on even when it was very difficult to go further. Thank you Professor, may all who come after me harness the benefit of your academic guidance and strength.

My deepest gratitude also goes to my family and friends for their support even when I hardly had time for them. My son, Katleho, has been a pillar of strength for bearing with me and also encouraging me even when he also needed the parental support during his studies, thank you son. Great many thanks to Ms Ts'episo Thabane for her support and putting up with me when my studies prevented me from executing some of our chores and obligations. I cannot forget the solid shoulder of strength provided by Mr Kedisitse Waynes Lofafa, throughout my tenure of studies and during the dark days of my ill-health.

I would like to thank the National University of Lesotho for granting me study leave to pursue a Ph.D degree and the University of KwaZulu Natal for this opportunity to study for a Ph.D degree in Statistics. To these two universities I owe a lot in terms of both personal and academic development. I would also like to thank the Lesotho Government for sponsoring my studies and the Association of African Universities (AAU) for giving me a small grant that enabled me to complete my thesis.

Lastly, I wish to thank colleagues and all people whose contribution to the success of this Ph.D was not in vain, may you continue to be pillars of strength to those in need.

Abstract

Oftentimes, application of regression models to analyse cereals data is limited to estimating and predicting crop production or yield. The general approach has been to fit the model without much consideration of the problems that accompany application of regression models to real life data, such as collinearity, models not fitting the data correctly and violation of assumptions. These problems may interfere with applicability and usefulness of the models, and compromise validity of results if they are not corrected when fitting the model. We applied regression models and diagnostics on national and household data to model availability of main cereals in Lesotho, namely, maize, sorghum and wheat. The application includes the linear regression model, regression and collinear diagnostics, Box-Cox transformation, ridge regression, quantile regression, logistic regression and its extensions with multiple nominal and ordinal responses.

The Linear model with first-order autoregressive process AR(1) was used to determine factors that affected availability of cereals at the national level. Case deletion diagnostics were used to identify extreme observations with influence on different quantities of the fitted regression model, such as estimated parameters, predicted values, and covariance matrix of the estimates. Collinearity diagnostics detected the presence of more than one collinear relationship coexisting in the data set. They also determined variables involved in each relationship, and assessed potential negative impact of collinearity on estimated parameters. Ridge regression remedied collinearity problems by controlling inflation and instability of estimates. The Box-Cox transformation corrected non-constant variance, longer and heavier tails of the distribution of data. These increased applicability and usefulness of the linear models in modeling availability of cereals.

Quantile regression, as a robust regression, was applied to the household data as an alternative to classical regression. Classical regression estimates from ordinary least squares method are sensitive to distributions with longer and heavier tails than the normal distribution, as well as to outliers. Quantile regression estimates appear to be more efficient than least squares estimates for a wide range of error term distribution. We studied availability of cereals further by categorizing households according to availability of different cereals, and applied the logistic regression model and its extensions. Logistic regression was applied to model availability and non-availability of cereals. Multinomial logistic regression was applied to model availability with nominal multiple categories. Ordinal logistic regression was applied to model availability with ordinal categories and this made full use of available information. The three variants of logistic regression model gave results that are in agreement, which are also in agreement with the results from the linear regression model and quantile regression model.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Background

Food security is an important topic for all countries in the world (Jones, 1982). The topic consists of three main components, namely, food availability, access and utilization, referred to as the basic elements of USAID's food security framework by Gervais (2004). Availability of cereals, in a country, can be defined as physical existence of food cereals measured through food cereals supplies. National food availability derives from a combination of domestic food stocks, domestic food production, commercial food imports, and food aid. Thus food security is a function of production, stocks, commercial imports, food aid and their underlying factors (Riely et al., 1999). These factors are normally used by Food and Agriculture Organization (FAO) in compiling food balance sheets used to inform countries on expected food deficit or surpluses, food import requirements as well as food aid requirements.

Investigations about availability of food cereals, as a component of food security, are important and need to be undertaken at both the national and household levels. This need emanates from an observation by (FAO, 1997) that food availability at the national level does not guarantee food access at the household level. In others words it is possible that food can be available at the national level while households and/or individuals do not have access to such food. This is because adequate national food supplies is a necessary but not sufficient condition for ensuring food security (Cohen, 2005). Households have access to food if they have the entitlement to produce from own land, purchases, and other means such as exchanges or food received as gifts (Maxwell and Frankenberger, 1992). Entitlement to food originates from their resources, assets, employment opportunities at their disposal, own production, prices of commodities and services, and social security benefits (Sen, 1981). If the combination of factors that constitute entitlement is not adequate to enable individuals or households to obtain the minimum food required, such individuals or households become food insecure. Household entitlement and access to food cereals are used to

measure availability of food cereals to households in this research.

Major crops that are grown in Lesotho are maize, sorghum and wheat, each occupying approximately 60%, 20% and 10% of the area planted (FAO and WFP, 2005). Like for most communities in Southern and East Africa, maize is the most important staple food for Basotho, and it constitutes 80% of the diet of people in the rural areas of Lesotho (FAO and WFP, 2007). The majority of farmers in the country are subsistence farmers with low productivity and average yields of less than 1 tonne per hectare (FAO and WFP, 2007). Cereal shortages in Lesotho are generally offset by commercial imports and food aid, and this makes the country highly dependent on cereal imports to meet domestic food requirements. The importance of these factors in contributing to availability of food, cereals included, justifies the need to study and understand availability of cereals in their context.

Understanding of availability of food cereals at both the national and household levels requires application of statistical methods that are appropriate for analysis of the current data. The commonly applied statistical method is regression analysis, which is often limited to estimation and prediction of crop production or yield. Currently crop yield estimation is based on classical regression analysis and trend analysis (Prasad et al., 2005). The general approach has been to regress a time dependent response variable, crop production or yield, on predictor variables that include time and some meteorological variables (Jones, 1982). This is usually done without much consideration of problems that may come with the application of these models to real life data. The problems that are common include collinearity, models not fitting the data correctly, and violation of model and estimation procedures assumptions. Violation of assumptions may be due to factors such as the presence of extreme values in the data. Realizing that such problems exist in real life, modelling of real life data should take into consideration their existence and appropriate statistical methods be used to study and correct them when fitting the regression models. Otherwise they may interfere with the applicability and usefulness of the models.

Generally, regression models may differ based on the form of the regression function, which could be linear or nonlinear. They may also differ based on the shape of the probability distribution of the response variable, which could be symmetric or skewed, and based on other ways. Lewis et al. (1998), Balaghi et al. (2008), Ren et al. (2008), Jones (1982), Schillinger et al. (2008) and Bella et al. (1996) used classical regression to model crop yield. Prasad et al. (2005) used piecewise linear regression, Cutts and Hassan (2003) used two stage least squares, and Hansen and Indeje (2004) used nonlinear regression. Evenson and Mwabu (1998) applied quantile regression to model crop production.

Oftentimes real life data, specifically the data from observational studies, consist of variables that are highly correlated, resulting with collinearity or the presence of near linear dependencies among

predictor variables. This implies that there is redundancy in the set of variables because the same information is being given by more than one variable and thus variables have the same predictive power of the response variable. Collinearity has effects that can affect the efficiency of linear models in estimating and predicting the response variable. Belsley (1991) noted that collinearity is a data problem not a statistical problem. Thus data need to be checked using collinearity diagnostics when fitting the regression model and remedial measures be used to remedy collinearity problems that may exist in the data.

Real life data may contain some observations that are far too extreme when they are compared with other observations on the response or predictor variables. In such cases extreme observations are called outlying observations or high leverage points, respectively. The existence of outlying observations is a concern for ordinary least squares (OLS) estimates as they are very sensitive to outlying observations. On the other hand, one of the important uses of classical regression models of making inferences about parameter estimates requires the distributional assumption about error terms. The standard distributional assumption about error terms is that they are normally distributed. This assumption is required for the applicability and optimality of OLS estimation procedure. However, real life data do not always satisfy the assumption. Furthermore, the classical regression model is based on the assumption that the error terms have a constant variance. In this case the regression model is referred to as the location shift or location model, where predictor variables are assumed to affect only the location of the conditional distribution of the response variable, not the scale or its distributional shape (Koenker and Hallock, 2000). The estimated model of the conditional mean function is appropriate for such situations. However, there is more that can be established by the application of regression to real life situations than what can be learned from the location shift model alone.

Diagnosis of the regression model and how the model agrees with a particular set of data is critical for developing statistical models and sound inferences (Schabenberger and Pierce, 2002). Regression diagnostics can be used to check results from the fitted model and see if assumptions of the model are not violated, the data contain extreme observations and the extent of their influence on different quantities of the fitted model. The existence of outlying observations and asymmetric or symmetric distributions but rather with heavier and longer tails than that of the normal distributions need to be taken care off when fitting the regression models. One of the corrective measures could be to transform the data. Box and Cox (1964) suggested the transformation that automatically identifies an appropriate transformation from the family of power transformations.

Robust regression procedures such as quantile regression are normally preferred from the OLS procedures due to their robustness to violation of some assumptions of the regression models, and the presence of outlying observations in the data set. Quantile regression of Koenker (2005) is a robust and flexible estimation approach that estimates various quantile functions at different parts

of the distribution of the response variable. The majority of the application of regression in studying relationships of variables focus on estimating rates of change in the mean response given a set of predictor variables. This may give an incomplete view of the entire distribution of the response variable, and can lead to wrong conclusions, particularly when the assumptions of homoscedasticity and normality do not hold, and when there are outliers in the data. Predictor variables can influence the distribution of the response variable in a number of other ways. They can influence the variability and result in heteroscedasticity, one tail of the distribution can be elongated while the other is compressed, and the distribution can even have more than one mode. A comprehensive investigation of these effects can be done through the use of quantile regression because unlike the conditional quantiles, the conditional mean model is not able to detect the changes in shape and scale of the distributions that deviate from the normal distributions, such as right or left skewed distributions. Mosteller and Tukey (1977) noted that the mean is not always sufficient by itself and most of the regression analysis resulted with an incomplete picture of how variables are related.

The use of classical regression and quantile regression models to model availability of food cereals becomes inappropriate when households are categorized according to availability of specific cereals. The categorization results with a categorical variable that indicates what cereal was available to which households, and thus invalidating the use of classical regression and quantile regression models, where the response variable is a continuous variable. The categorical response variable takes three different forms not measured in a ratio scale, and as a result does not satisfy the distributional assumptions required for the classical regression model. The three forms are binary response, multiple nominal categories and ordinal categories. In this case generalized linear models (GLMs), of which logistic regression with its three variants is a special case, can be used as an alternative to the linear regression and quantile regression models.

1.2 Objectives

The overall objective is to model availability of food cereals at both the national and household levels, in consideration of problems that accompany the application of regression models to real life data. Specific objectives are:

1. to detect the presence of more than one collinear relationship that exist simultaneously in the current data, and identify variables involved in each of such relationships using condition index and variance-decomposition proportions, respectively.

2. to remedy collinearity problems and control the instability of parameter estimates using ridge regression.

3. to determine the existence of outlying observations and leverage points in the current data, detect if they are influential and establish the extent of their influence on different quantities

of the fitted model.

4. to establish if the distributional assumptions about the error terms are violated, and check the extend to which they are violated.

5. To correct violation of the distributional assumptions that error terms have a normal distribution and a constant variance using the Box-Cox transformation

6. to deal with asymmetric distributions of data by fitting the quantile regression model at the different parts of the distribution of the response variable.

7. to study availability of cereals further using the generalized linear models, in particular logistic regression model in its three different forms

The approach used to achieve the above outlined objectives is to identify problems that may compromise the validity of the results obtained from fitting the regression models to the real life data. This is done by applying statistical tools that have long been established as relevant to the problems but their application in the analysis of availability of food cereals is limited.

1.3 Organization of the Thesis

This thesis started by describing the composition of the two sets of data used in Chapter 2. The two sets are of the national and household data about availability of cereals. The chapter further presents the descriptive analysis of the data, pointing to relationships that may exist between variables, exploring distributions of data and suggesting further and in-depth analysis of the data. Chapter 3 gives a review of classical regression models and diagnostics, where the general linear model and its alternatives specialized in dealing with different situations and problems that emerge in the application of linear models are discussed. Chapter 4 presents the application of the classical regression models and diagnostics on the national data. The application starts by fitting the model, and using collinearity and regression diagnostics to identify problems in the data, which may interfere with the usefulness and efficiency of the models. Furthermore, it applies Box-Cox transformation to correct violation of distributional assumptions about the error terms, and ridge regression to remedy collinearity problems and control instability in estimated regression coefficients.

Chapter 5 presents the application of the classical regression models and diagnostics to the household data. The regression model with categorical predictor variables is applied and problems that accompany this application are identified and dealt with using appropriate remedial measures. Chapter 6 presents a review of robust regression with an emphasis on quantile regression model, which estimates functional relationships for all parts of the distribution of the response variable.

In Chapter 7, the quantile regression model is applied to the household data, where the model is fitted at the $25th$ quantile, median and $75th$ quantile. Chapter 8 gives a review of three variants of logistic regression, namely, simple logistic regression with binary response, multinomial regression with nominal categories, and ordinal logistic regression. Chapter 9 presents the application of the variants of logistic regression on household data. Lastly, chapter 10 presents the discussion and conclusions together with suggestions for future research work.

Chapter 2

Data Description and Exploratory Analysis

2.1 Introduction

The data used in this research are observational in nature rather than being experimental because they were not compiled under a controlled experiment. They consist of two sets, where the first set is of the national data with variables that are normally used to measure availability or supplies of cereals at the national level, and related variables. The second set is of household data collected from households in some villages in Lesotho. The data were collected by the researcher through a small survey that investigated about households entitlement to the three main cereals in Lesotho, namely, maize, sorghum and wheat. This chapter focuses on describing and exploring data in the two sets of data.

2.2 National Data

The national data are secondary data about food cereals and related variables, compiled from different organizations in Lesotho. These are aggregated data that reflect annual figures for the entire country. The variables in the data include domestic production in tonnes, commercial imports in tonnes, food aid in tonnes, and average price per tonne for maize, sorghum and wheat, in South African Rands. The data on production of the cereals include cultivated area in hectares, and crop yield per hectare in kilograms for each of the cereals. Cultivated area was categorized into planted area, harvested area, and area under crop failure. Harvested area is part of planted area which was not affected by crop failure, and household harvested some crop from it. Crop failure occurs when planted crop gets destroyed by either frost, floods, animals, pests, diseases, or other things that may cause damage to crops. Other variables that are part of the national data are the average

amount of rainfall in milliliters, population size in millions, and time in years.

The data on production of cereals were collected by Bureau of Statistics (BoS) Lesotho through agriculture production surveys (APS) conducted annually, and agriculture censuses conducted every 10 years. The agriculture production survey was conducted by selecting a random sample of farming households in the rural areas throughout the country. These households were interviewed about their farming activities and monitored from the beginning of the agricultural year when they start planting until the end of the agricultural year when they harvest. The Lesotho agricultural year starts in August and ends in July of the following year. Agricultural censuses are generally conducted in the same manner as agriculture production surveys, except that a bigger sample size is selected from both the rural and urban areas.

The data on commercial imports and food aid are compiled by the Early Warning Unit under the Lesotho Disaster Management Authority (DMA). In compiling these data the unit uses information from different government's departments, non-governmental organizations (NGOs), and the World Food Programme (WFP). Prices of cereals are compiled by the Department of Marketing of the Ministry of Trade and Industry, Cooperatives and Marketing (MTICM). The rainfall data are collected and compiled by the Lesotho Meteorological Services (LMS) from a number of weather stations throughout the country. Population size is obtained through population censuses conducted every ten years by the Bureau of Statistics Lesotho, and mid-year population projections for the years that are in between the years in which population censuses were conducted, also made by the Bureau of Statistics Lesotho.

Some variables that are considered to be important in studying the national food cereals demand and supplies are not included in this research due to the scantiness and incompleteness of the data. Such variables are cereals domestic stocks, domestic requirements, surplus or deficit, and consumption per capita. The national data on these variables are available but with insufficient observations as there are gaps of data for some years. The scantiness and incompleteness of the data can be attributed to, among other things, the fact that the data were compiled by different organizations that operate independently with little or no coordination, and varying levels of appreciation of the importance of the data and uses to which the data are put, among the organizations that compile the data. The scantiness of data, specifically of the agricultural data is not a problem to Lesotho only as Cutts and Hassan (2003) noted that agricultural data are scanty in most of the Southern Africa countries.

2.3 Household Data

The household data are primary data collected from a sample of 296 households. The data about households entitlement and access of the three main cereals were collected in a survey conducted in Lesotho by the researcher in 2008. A structured questionnaire was designed and administered by the researcher to a sample of 296 households through interviews. The pilot survey was conducted before the main survey to test the questionnaire. Due to financial and time constraints, the survey was conducted in few villages from three districts out of the ten districts of Lesotho. The three districts are Berea, Maseru and Mafeteng. Households that were included in the survey were randomly selected from easily accessible villages of the three districts. The sampling technique used to select a sample of households is systematic sampling, where a pattern of selecting every fifth household in a village was followed. Thus conclusions made from the results of this survey should be confined to the villages from where the households were selected, and not generalized to the three districts. This is because the interviewed sample of households is not representative of the districts where the villages are located.

The data include information on demographic and socio-economic characteristics of household heads, household entitlement to the three main cereals that constitute sources of the cereals and coping strategies. The characteristics include household size, age, sex, education level and occupation of household head. Households entitlement and copying strategies include, household monthly income in South African Rands, the amount of a given cereal produced from own land, purchases of cereals, exchanges of cereals, food aid from government and NGOs, and gifts of cereals, all measured in kilograms. The data are largely based on memory of respondents since households do not normally keep records of their activities concerning procurement, access, and consumption of food.

The variable name used to measure household entitlement and access to food cereals is availability of a given cereal to the household. Availability of a cereal, for example maize, consists of a combination of different sources that households had for acquiring and accessing such food cereal. Respondents in the households were asked to provide information about sources of food cereals for their households and the amount acquired during the observational period. The survey was conducted in December 2008 and the observational period was from the beginning of the last harvest season, which was June 2008, until in December 2008 when the survey was conducted. Normally the harvest season in Lesotho starts in June of every year. Thus availability of cereals in this context is the total amount of a given cereal, in kilograms, acquired or accessed by a household during the observational period.

2.4 Exploratory Data Analysis

The application of regression models and diagnostics to both the national and household data was preceded by an exploratory analysis of the data. The data on maize, sorghum and wheat were explored to reveal the underlying structure of the data for a better understanding of variables in each set, and to establish relationships that may exist between the variables, prior to modelling the data. The establishment of relationships in a set of data is the initial step of modelling the data, which can lead to some variables being classified as response variables and others as predictor variables. Generally, subsequent steps of statistical modeling can be done when some patterns in the data are thought to exist and have been identified (Everitt and Dunn, 1983). The relationships between variables can be simple or complex depending on questions being answered in a research, and the nature of the data. Summary statistics, scatter plots, box plots and correlation matrices were used to explore the data.

The national data were compiled for the period 1973/1974 to 2006/2007, and this period is considered to be the period under investigation in this research. However, data on some variables such as price of cereals and population size were not available for the entire period, meaning that there were data gaps in some years within the specified period. Each of the national data set on maize, sorghum and wheat was sub-divided into three subsets, based on years in which data on some variables were available. The composition of the three subsets vary in terms of the number of variables and observations. The subsets are not mutually exclusive since they overlap and hence have some observations in common. The subdivision of the data was used to find out if different time intervals make a difference in studying the national availability of the three main food cereals in Lesotho.

All the 296 households included in the survey had entitlement or access to maize, while some had access to sorghum and/or wheat. Having a larger number of households in the sample with access to maize than for sorghum and wheat is not surprising because as it was mentioned in Chapter one, maize is the only staple food in Lesotho consumed by almost all people in the country. The villages from which the survey was conducted were grouped into four locations, on the basis of their geographical locations in the country. The locations are Mafeteng, Maseru lowland, Maseru foothill and Berea. The majority of households that were included in the study, 55.07%, were located in Maseru Lowland, followed by Mafeteng, Maseru foothill and Berea with 20.17%, 14.19%, and 10.47%, respectively. The demographic and socio-economic characteristics of household heads, who are regarded as household representatives in this research, were used in studying households entitlement and access to maize and sorghum. Wheat was excluded in the analysis because few households had access to it during the observational period.

2.4.1 Exploration of National Maize Data

The first subset of maize data contains the data for the entire period under investigation, 1973/74 to 2006/2007, and consists of 9 variables and 33 observations, corresponding to the number of years in the observational period. The variables include; time in years, the amount of rainfall, area planted to maize, planted area affected by crop failure, area harvested to maize, yield of maize per hectare, maize production, maize commercial imports, and maize received as food aid. This subset excludes two variables, price of maize and population size, because the variables have data gaps for some of the years within the observational period. The second subset with 31 observations of the years 1976/1977 to 2006/2007, consists of all variables in the first subset plus population size as an additional variable. This variable does not have data for the first three years, 1973/74 to 1975/1976, meaning that it has data starting from 1976/1977 and thus it could not be included in the first subset of data. The third subset with 29 observations for the period 1973/1974 to 2001/2002 is similarly made up of all variables in the first subset but with price of maize per tonne as an additional variable. This variable has missing data for the period 2002/2003 to 2006/2007 and thus it could not be part of either the initial subset or the second subset.

Table 2.1 presents summary statistics of the variables in the subset that contains data for 1973/1974 to 2001/2002. The lowest maize production that Lesotho experienced in this period was 48918 tonnes, with the maximum of 277685 tonnes. The lowest amount of maize imported for commercial purposes was 60100 tonnes with the maximum of 186000 tonnes. The minimum amount of maize received by the country as food aid was zero tonnes with the maximum of 49900 tonnes. The minimum of zero indicates that there was a time within this period when the country did not receive any maize in the form of food aid. The years in which the country did not receive maize in the form of food aid are 2000/2001 and 2001/2002. There is no literature that gives specific reasons why the country did not receive maize as food aid in these two years. On average the country produced 116860.32 tonnes of maize, imported 111216.43 tonnes for commercial purposes, and received 13110.71 tonnes as food aid during the years 1973/1974 to 2001/2002. These statistics show that on average, domestic production of maize contributed a lager proportion of the Lesotho maize requirements in that period.

Food aid varied greatly, when it is compared with the other two supplies of maize, domestic production and commercial imports. This is shown by the highest coefficient of variation (CV) of 85.54. The lowest price of maize per tonne was R49.00 per tonne in 1973/1974 while the maximum price was R739.00 in 2001/2002. The big difference between the minimum and maximum price shows how price of maize increased progressively over the years. The average cost of maize for the period was R345.78.

The smallest area that was planted and harvested to maize in Lesotho during the period under

Table 2.1: Summary Statistics for 1973 - 2002 Maize Data

	Summary Statistics				
Variables	Minimum	Maximum	Mean	Std deviation	CV
Production	48918	277685	116860.32	48748.19	41.71
Rainfall	465	1043	730.79	128.71	17.61
Planted area	55676	208905	143124.29	36387.84	25.42
Harvested area	76954	177503	122272.39	28744.18	23.51
Failed area	4768	126076	24423.32	24687.45	101.08
Yield/hectare	359	1632	821.68	299.11	36.40
Imports	60100	186000	111216.43	32109.29	28.87
Food aid	0	49900	13110.71	11215.51	85.54
Price/Ton	49	739	345.78	235.25	68.03

investigation were 55676 and 76954 hectares, respectively. The country had the highest area under crop failure of 126076 hectares in 1997/1998, which constituted 60% of area planted to maize in that year. The highest yield of maize per hectare attained in the years 1973/1974 to 2001/2002 was 1632 kilograms per hectare, with the minimum and average of 359 and 821.68 kilograms per hectare, respectively. The minimum, maximum and average amount of rainfall were 465, 1043, and 730.79 milliliters, respectively. The planted area affected by crop failure has the highest CV of 101.08, meaning that it varied greatly when it is compared with all other variables in the table. The large variability of crop failure is probably caused by erratic weather conditions that are becoming frequent lately. The amount of rainfall was less variable than all other variables since it has the lowest CV of 17.61.

Table 2.2 presents all variables that appear in the scatter matrices for maize, sorghum and wheat data. Some of the variables in the table do not appear in some scatter matrices, depending on the subset of data represented in a matrix. Figure 2.1 shows a scatter-plot matrix that presents scatter plots of all pairs of variables in the 1973/1974 to 2001/2002 maize data, except yield of maize. Yield is omitted from the matrix because it almost measures the same quantity as maize production and thus its inclusion will cause redundancy in the data.

Scatter plots of maize production and other variables on the first row and the first column show that maize production is linearly related with time in years, the amount of rainfall, area planted, area harvested and price of maize. The plot of maize production and the amount of rainfall shows an outstanding observation. This observation is identified as the 1975/1976 case with the highest amount of rainfall of 1042.52 milliliters, and a relatively low maize production of 49128 tonnes. Scatter plots of time on the second row shows that time has linear relationships with area planted, area harvested, commercial imports and price of maize. The plot of price of maize and time, at a glance, shows a strong positive correlation between the two variables, which is almost perfect. The plots indicate that almost all variables which have a linear relationship with area planted

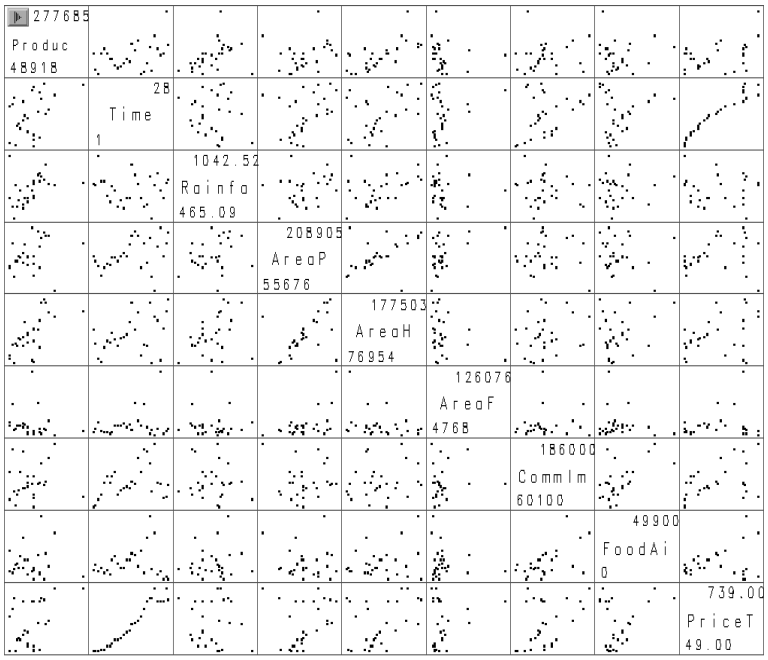

Figure 2.1: Scatter Plot Matrix for 1973 - 2002 Maize Data

are similarly related with area harvested. The variables are maize production, time and amount of rainfall. The reason for this could be that the two variables measure almost the same quantity since area harvested is part of area planted not affected by crop failure. Thus one of the two variables, area planted, was dropped to avoid redundancy when relationships between variables were studied further.

The correlation matrix in Table 2.3 presents Pearson correlation coefficients between all pairs of variables. Pearson correlation coefficient is a quantitative measure of the strength of linear relationship between two variables. Thus the correlation coefficients quantify the strength of linear relationships identified from the scatter-plot matrix in Figure 2.1. Lack of correlation between a pair of variables does not necessarily mean that there is no relationship between the two variables, it only shows lack of a linear relationship. The correlation coefficients of maize production and each of time, the amount of rainfall, area harvested and price of maize are 0.42, 0.39, 0.68 and 0.43, respectively. All the coefficients are positive, indicating positive relationships that maize production increased with an increase in each of the variables. The strong positive relationship between price of maize and time identified from Figure 2.1 is confirmed by a very high positive correlation coefficient of 0.97. When the suspected extreme observation is excluded from the data, the correlation coefficient of production and rainfall increases from 0.39 to 0.62.

Summary statistics of the data for the periods 1973/74 to 2006/2007 and 1976/1977 to 2006/2007

Table 2.2: Variables in the Scatter Matrices

Variable	Label
Production	Production of a given cereal
Lag-Prod	Production of a given cereal in the past immediate year
Time	Time in years
Population	Population of people in Lesotho
Rainfall	Amount of rainfall
AreaP	Area planted to a given cereal
AreaH	Area harvested to a given cereal
AreaF	Area planted to a given cereal affected by crop failure
CommImports	Commercial imports of a given cereal
FoodAid	A given cereal received as food aid
Price	Price of a given cereal per tonne

Table 2.3: Correlation Matrix for 1973 - 2002 Maize Data

	Production	Time	Rainfall	PArea	HArea	FArea	Imports	Food aid	Price/Ton
Production	1.0000	0.4167	0.3934	0.2811	0.6799	-0.1567	0.1016	0.0314	0.4295
Time	0.4167	1.0000	-0.1156	0.3747	0.4103	0.2296	0.3587	0.1027	0.9665
Rainfall	0.3934	-0.1155	1.0000	0.2632	0.3884	-0.0177	-0.1135	0.1113	-0.0966
PArea	0.2811	0.3747	0.2632	1.0000	0.4963	0.5356	-0.0234	-0.1113	0.2471
HArea	0.6799	0.4103	0.3884	0.4963	1.0000	-0.2834	0.1990	0.2469	0.3823
FArea	-0.1567	0.2296	-0.0177	0.5356	-0.2834	1.0000	0.0834	-0.1453	0.1699
Imports	0.1016	0.3587	-0.1134	-0.02335	0.1989	0.0833	1.0000	0.5676	0.4372
Food aid	0.0314	0.1028	0.1113	-0.1113	0.2469	-0.1454	0.5676	1.0000	0.24328
Price/Ton	0.4295	0.9665	-0.0967	0.2471	0.3823	0.1699	0.4372	0.2433	1.0000

are not presented because they do not differ much with that of the subset presented in Table 2.1. Scatter-plot matrices in Figures A.1 and A.2 of Appendix A show linear relationships that exist among some pairs of variables in the two data subsets, respectively. The plots show that in both subsets, maize production is linearly related with only three variables, which are the amount of rainfall, area planted and area harvested to maize. The plots also show two linear relationships of area planted and area harvested, and the amount of rainfall and area harvested. The positive relationships of maize production and each of the amount of rainfall and area harvested, for the data of 1973/74 to 2006/2007, are quantified by the correlation coefficients of 0.39 and 0.61, respectively (Table B.1 in Appendix B). In the case of the 1976/1977 to 2006/2007 data, the correlation coefficients for maize production and each of the amount of rainfall and area harvested are 0.60 and 0.58, respectively (Table B.2 in Appendix B). The correlation coefficients of the amount of rainfall and area harvested for the two data subsets are 0.40 and 0.60, respectively. This indicates that the relationship between the two variables is stronger for the 1976/1977 to 2006/2007 data, where the 1975/1976 observation that appeared outstanding in terms of the amount of rainfall is excluded.

Skewness, kurtosis and box plots were used to explore the distribution of a variable that was identified as a potential response variable when fitting regression models in the subsequent chapters. The variable is production of each of maize, sorghum and wheat. The values of skewness and kurtosis of maize production are 1.26 and 3.06, respectively. The two quantities provide summary information about the shape of the distribution of the data. The value of skewness of 1.26 shows that the distribution deviates slightly from the normal distribution and it is skewed to the right. The value of kurtosis of 3.06 shows that the distribution is a little bid leptokurtic, meaning that it has a slightly higher peak and heavier tails than the normal distribution. The tails that are heavier than that of the normal distribution imply high probability of extreme observations of maize production. The values of skewness and kurtosis for the 1973/74 to 2006/2007 and 1976/1977 to 2006/2007 data in Table 2.8 show similar pattern as that of the 1973/74 to 2001/2002 data, though the 1976/1977 to 2006/2007 data has the highest values.

The box plot of 1973/1974 to 2001/2002 maize data in Figure 2.2, at a glance shows that maize production is approximately normal. This is indicated by the line inside the box showing the median, which is almost equidistant to the lower and upper edges of the box. The Box plot of 1973/1974 to 2006/2007 maize that in Figure M.1 shows a slight positive skewness where the line in the box is not equidistant to the lower and upper edges of the box but also not far from the middle of the box. In this case, the line is slightly towards the lower edge of the box. In the case of the 1976/1977 to 2006/2007 maize data, the line is towards the lower edge of the box showing a positively skewed distribution where the majority of the observations fell in the lower side of the distribution of maize production (Figure M.2). This implies that the majority of the years in the observational period had lower production of maize.

In general, the observations from the plots agree with what was observed from the values of the skewness and kurtosis in Table 2.8. The 1973/1974 to 2001/2002 has the lowest values of skewness and kurtosis showing a relatively slight deviation from normality, followed by the values from the 1973/1974 to 2006/2007 maize data. The 1976/1977 to 2006/2007 maize data is the worst in terms of maize production deviating from normality, shown by both the box plot and values of skewness and kurtosis. All box plots for maize data at different time intervals show one extreme observation of maize production.

2.4.2 Exploration of National Sorghum Data

The sorghum national data were similarly divided into three subsets, which are not mutually exclusive. The first subset of 1973/1974 to 2006/2007 has the same composition of variables as that of maize for the same period. The 1973/1974 to 1997/1998 data consists of all variables in the first subset plus price of sorghum per tonne as an additional variable. The 1976/1977 to 2006/2007 data consists of all variables in the first subset plus population size as an additional variable. Table 2.4

presents summary statistics for the 1976/1977 to 2006/2007 data. Unlike maize supplies, which included food aid, sorghum supplies consisted of domestic production, and commercial imports only. This shows that Lesotho did not receive any sorghum in the form of food aid within the period under investigation.

Table 2.4: Summary Statistics for 1976 - 2007 Sorghum Data

	Summary Statistics				
Variables	Minimum	Maximum	Mean	Std deviation	CV
Production	6887	85775	35421.67	19916.76	56.23
Rainfall	465	969	718.16	121.92	16.98
Area Planted	11047	81594	48267.03	18535.06	38.41
Harvested area	8579	76355	43166.83	17748.43	41.16
Failed area	0	19153	5100.20	4416.22	86.59
Yield/hectare	329	1383	736.03	305.48	41.50
Imports	0	16900	1676.67	3228.82	192.57

The lowest sorghum production experienced in Lesotho in the period, 1976/1977 and 2006/2007, was 6887 tonnes, with the highest production of 85775 tonnes. Sorghum imports had the minimum of zero, indicating that there were years in which the country relied entirely on domestic production to meet its sorghum needs. On average the country produced and imported 35421.67 and 1676.67 tonnes of sorghum, respectively in that period. The highest CV of 192.57 for commercial imports indicates that the amount of sorghum imported by the country within that period varied greatly, when it is compared with other variables (Table 2.4). The minimum of zero for area under crop

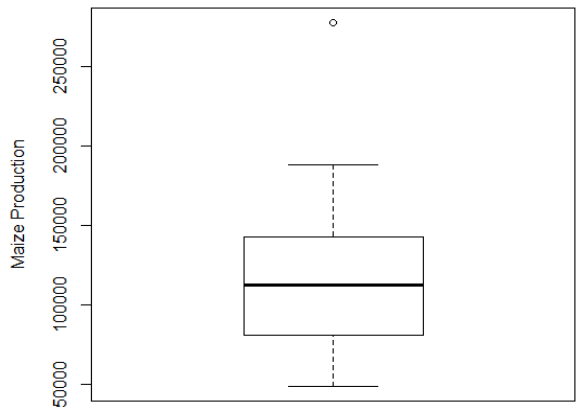

Figure 2.2: Box Plot for 1973 - 2002 Maize Data

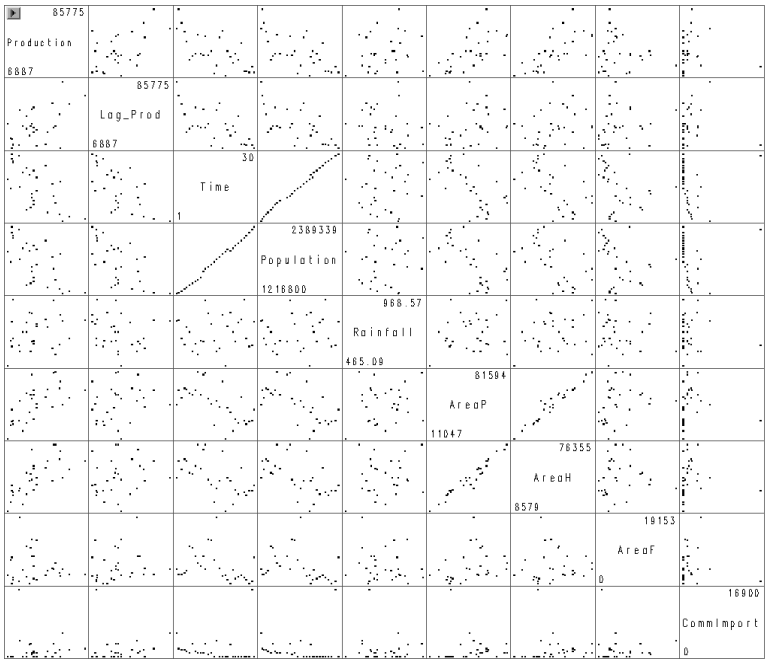

Figure 2.3: Scatter Plot Matrix for 1976 - 2007 Sorghum Data

failure is an indication that the country had good years when sorghum was not affected by factors that cause crop failure. In those years area planted to sorghum turned out to be equivalent to area harvested to sorghum. The lowest and highest yield of sorghum attained during that period were 329 and 1383 kilograms per hectare, respectively. The average yield of sorghum in the same period was 736.03 kilograms per hectare.

Scatter plots on the first row of Figure 2.3 show that sorghum production has a positive linear relationships with production of sorghum in the past immediate year or at time $t-1$, and area harvested, whereas it has a negative relationship with time and population size. Like in the maize data, every variable related with area planted is also related with area harvested. In this case, the variables are sorghum production, time and population size. Almost all the four variables that are linearly related with sorghum production have pairwise relationships among themselves. In particular time shows, at a glance, a negative relationship with production of sorghum in the current year, production of sorghum in the past immediate year and area planted to sorghum show a positive relationship.

The correlation coefficients of sorghum production and each of production of sorghum in the immediate past year, time in years, population size and harvested area are 0.53, -0.67, -0.66 and 0.71, respectively (Table 2.5). These coefficients show that in the years 1976/1977 to 2006/2007, if sorghum production in the immediate past was high, the current sorghum production was also

Table 2.5: Correlation Matrix for 1976 - 2007 Sorghum Data

	Production	Production-t	Time	Population	Rainfall	PArea	HArea	FArea	Imports
Production	1.0000	0.5272	-0.6707	-0.6613	0.2118	0.6531	0.7140	-0.1285	0.0338
Production-t	0.5272	1.0000	-0.6692	-0.6540	-0.2052	0.2411	0.2127	0.1570	0.0642
Time	-0.6707	-0.6692	1.0000	0.9975	0.0941	-0.6245	-0.5733	-0.3169	-0.0039
Population	-0.6613	-0.6540	0.9975	1.0000	0.1011	-0.6271	-0.5785	-0.3072	0.0043
Rainfall	0.2118	-0.2052	0.0941	0.1011	1.0000	0.1597	0.2067	-0.1605	-0.0968
PArea	0.6531	0.2411	-0.6245	-0.6272	0.1597	1.0000	0.9713	0.2935	0.0069
HArea	0.7140	0.2127	-0.5733	-0.5785	0.2067	0.9713	1.0000	0.0577	0.0132
FArea	-0.1285	0.1570	-0.3169	-0.3072	-0.1605	0.2935	0.0577	1.0000	-0.0241
Imports	0.0338	0.0642	-0.0040	0.0043	-0.0968	0.0069	0.0132	-0.0241	1.0000

high, and sorghum production increased with an increase in area harvested, while it declined with time and an increase in population size. The correlation coefficient that measures a relationship between time and area harvested to sorghum is -0.57. This shows that area harvested to sorghum declined with time. Time and population size have a positive perfect relationship shown by the correlation coefficient of 0.9975, showing that population size increased with time.

Scatter plots for 1973/1974 to 1997/1998 sorghum data show that production of sorghum has negative linear relationships with time, area under crop failure, and price of sorghum per tonne (Figure A.3 in Appendix A). In addition, sorghum production has positive linear relationships with area harvested to sorghum. Like it is the case in the 1973/1974 to 2006/2007 data, time is linearly related with almost all other variables that are related with production of sorghum, plus imported sorghum. The strength of the identified relationships is measured by correlation coefficients in Table B.3 of Appendix B. The Correlation coefficients between sorghum production and each of time, area harvested, area under crop failure, and price of sorghum are -0.52, 0.70, -0.45 and -0.41, respectively. The correlation coefficients between time and area harvested, imported sorghum and price of sorghum are -0.52, -0.63 and 0.90, respectively. In the case of 1973/1974 to 2006/2007 sorghum data, the pattern of the relationships between sorghum production and production of sorghum in the past immediate year, time and area harvested, in Figure A.4 of Appendix A, is similar to the pattern in the 1973/1974 to 2006/2007 data. The strength of these relationships is shown by correlation coefficients in Table B.4 of appendix B.

The values of skewness and kurtosis of sorghum production for the period, 1976/1977 to 2006/2007, are 0.60 and -0.14, respectively. The value of skewness of 0.60 shows that the distribution deviates a little bid from the normal distribution and it is skewed to the right. The value of kurtosis of -0.14 shows that the distribution is a little bid platykurtic, meaning that it is a little bid flatter than the normal distribution. The values of skewness and kurtosis for the 1973/74 to 2006/2007 and 1973/1974 to 1997/1998 data in Table 2.8 show similar pattern as that of the 1976/1977 to

2006/2007 data.

The box plot of 1976/1977 to 2006/2007 sorghum data in Figure 2.4, at a glance, shows that the line inside the box is not equidistant to the lower and upper edges of the box, but it is towards the lower edge of the box. The implication is that sorghum production deviates from the normal distribution and is skewed to the right. This observation is confirmed by the positive value of skewness in Table 2.8, though it is not too far from zero. The box plots of 1973/1974 to 2006/2007 and 1973/1974 to 1997/1998 sorghum data in Figures M.3 and M.4 portray the same pattern, though the deviation for the normal distribution is more pronounced in the case of the 1973/1974 to 1997/1998 data. This is the subset of data with relatively the lowest value of skewness but the highest negative value of kurtosis (Table 2.8).

2.4.3 Exploration of National Wheat Data

The national wheat data were similarly divided into three subsets. The subsets are for the 1973/1974 to 2006/2007, 1973/1974 to 2001/2002, and 1976/1977 to 2006/2007 data, with the same composition of variables as that of maize and sorghum data. Summary statistics in Table 2.6 show that in the period, 1976/1977 to 2006/2007, imported wheat contributed a larger part of the national wheat requirements than domestic wheat production. This is shown by a minimum of 6844 tonnes, a maximum of 61381 tonnes, and an average of 26258.48 tonnes for wheat production, which are relatively low when they are compared with their counterparts for imported wheat. The minimum, maximum and average of imported wheat are 21500, 77000 and 43610.34 tonnes, respectively. Food

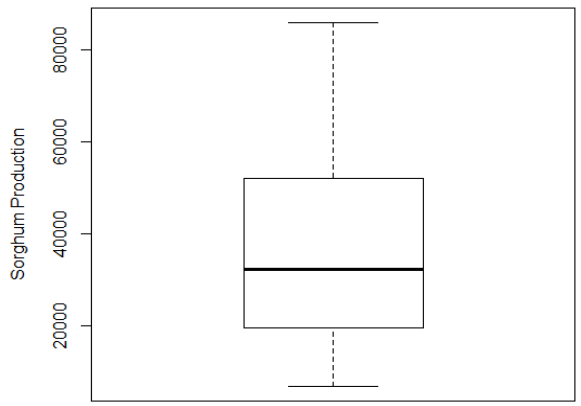

Figure 2.4: Box Plot for 1976 - 2007 Sorghum Data

aid was one of the supplies of wheat until 1994/1995 as the country did not receive wheat in the form of food aid beyond that period. Imported wheat was less variable than domestic production and food aid since it has the lowest CV of 33.14.

Table 2.6: Summary Statistics for 1973 - 2002 Wheat Data

	Summary Statistics					
Variables	Minimum	Maximum	Mean	Std deviation	CV	Skewness
Production	6844	61381	26258.48	16080.21	61.24	0.89
Rainfall	465	1043	738.99	133.88	18.12	0.24
Planted area	2900	82100	34202.00	16660.21	48.71	0.91
Harvested area	11088	76600	29936.52	14377.87	48.03	1.50
Failed area	94	21562	5186.17	4618.43	89.05	2.19
Yield/hectare	215	1925	770.90	363.06	47.10	1.24
Imports	12700	77000	43610.34	14454.37	33.14	0.82
Food aid	0	42900	9527.59	11416.96	119.83	0.66
Price/Ton	61	1095	525.77	366.23	69.66	0.35

Wheat had the minimum yield per hectare of 172 kilograms and the maximum yield per hectare of 1925 kilograms. Area planted to wheat affected by crop failure with a CV of 119.83 was more variable than area planted and harvested. The big variability in area under crop failure was probably due to erratic weather conditions, which are more prevalent lately.

The first row and column of Figure 2.5, at a glance, show that wheat production is linearly related with wheat production in the past immediate year, time in years, the amount of rainfall, area harvested to wheat, wheat received by the country as food aid, and price of wheat per tonne. Correlation coefficients between wheat production and wheat production at time $t-1$, time in years, area harvested, wheat received as food aid and price of wheat are 0.68, -0.47, 0.68, -0.45 and -0.37, respectively (Table 2.7). The strong positive relationship between time and price of sorghum is shown by a correlation coefficient of 0.93.

The scatter plot matrix in Figure A.5 of Appendix A and correlation matrix in Table B.5 of Appendix B, show relationships that exist in the 1973/1974 to 2006/2007 wheat data. Wheat production at time t is linearly related with wheat production in the past immediate year, time, the amount of rainfall, and area harvested, with the correlation coefficients of 0.72, -0.56, 0.36 and 0.72, respectively. In the case of the 1976/1977 to 2006/2007 wheat data, wheat production is linearly related with wheat production in the past immediate year, time, population size, and area harvested (Figure A.6 and Table B.6 in Appendices A and B). The correlation coefficients that measure the strength of the identified relationships are 0.62, -0.45, -0,45 and 0.60, respectively.

The box plot of 1973/1974 to 2001/2002 wheat data in Figure 2.6, at a glance, shows that the line

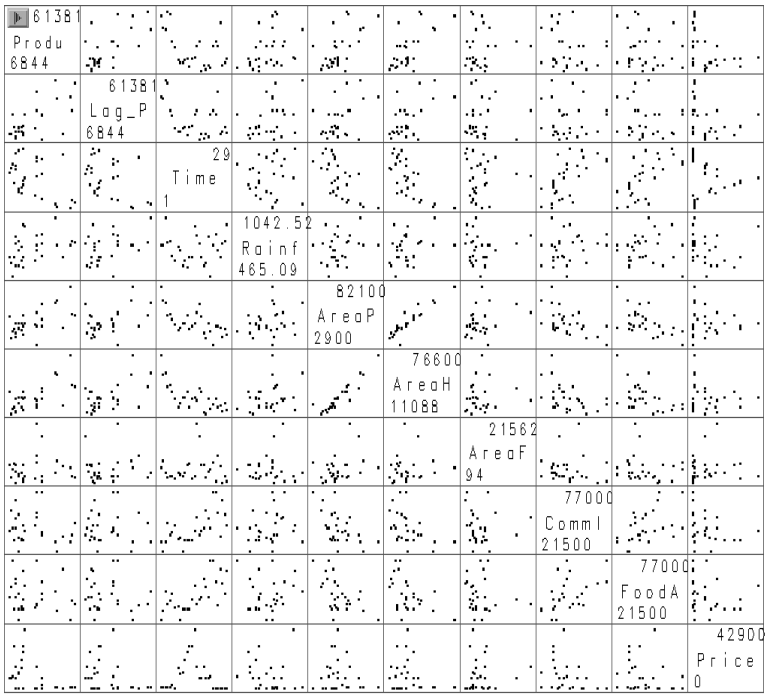

Figure 2.5: Scatter Plot Matrix for 1973 - 2002 Wheat Data

Table 2.7: Correlation Matrix for 1973 - 2002 Wheat data

	Production	Production-t	Time	Rainfall	PArea	HArea	FArea	Imports	Food aid	Price/Ton
Production	1.0000	0.6761	-0.4659	0.3548	0.6033	0.6790	0.1053	-0.1304	-0.4447	-0.3654
Production-t	0.6761	1.0000	-0.4187	0.4096	0.4650	0.5565	0.0674	0.0156	-0.2845	-0.2856
Time	-0.4659	-0.4187	1.0000	0.0008	-0.6627	-0.7127	0.0224	0.6492	-0.1211	0.93042
Rainfall	0.3548	0.4096	0.0008	1.0000	0.2461	0.2991	0.0055	0.2129	-0.1758	-0.0661
PArea	0.6033	0.4650	-0.6627	0.2461	1.0000	0.9239	0.3431	-0.5307	-0.0711	-0.5923
HArea	0.6790	0.5565	-0.7127	0.2991	0.9239	1.0000	0.1084	-0.4468	-0.1686	-0.6119
FArea	0.1053	0.0674	0.0224	0.0055	0.3431	0.1084	1.0000	-0.0466	0.0959	0.0299
Imports	-0.1304	0.0156	0.6492	0.2129	-0.5307	-0.4468	-0.0466	1.0000	-0.3563	0.7194
Food aid	-0.4447	-0.2845	-0.1211	-0.1758	-0.0711	-0.1686	0.0959	-0.3563	1.0000	-0.2244
Price/Ton	-0.3654	-0.2856	0.9304	-0.0661	-0.5923	-0.61187	0.0299	0.7194	-0.2244	1.0000

inside the box is not equidistant to the lower and upper edges of the box, but it is towards the lower edge of the box. The box plot and value of skewness (0.89) in Table 2.8 are in agreement that wheat production deviates from the normal distribution and is skewed to the right. The Box plot of 1973/1974 to 2006/2007 and 1976/1977 to 2006/2007 wheat data in Figures M.5 and M.6 of Appendix M show the same pattern as the one observed in Figure 2.6. Further, Figure M.6 shows that the 1976/1977 to 2006/2007 wheat data has two extreme values. Looking at the values of skewness and kurtosis together with the box plots to determine the deviation of the distribution of

Table 2.8: Values of Skewness and Kurtosis for National Data

Cereal	Subset of Data	Skewness	Kurtosis
Maize	1973/1974 to 2006/2007	1.47	3.85
	1973/1974 to 2001/2002	1.26	3.06
	1976/1977 to 2006/2007	1.54	3.98
Sorghum	1973/1974 to 2006/2007	0.70	-0.04
	1973/1974 to 1997/1998	0.54	-0.20
	1976/1977 to 2006/2007	0.60	-0.14
Wheat	1973/1974 to 2006/2007	0.99	-0.01
	1973/1974 to 2001/2002	0.89	-0.31
	1976/1977 to 2006/2007	1.39	1.59

a given variable from the normal distribution, it looks like the quantity in which the values should be far from zero varies from one set of data to another. This idea comes with an observation that for subsets of maize data, the values are relatively higher than zero but box plots show slight deviations from the normal distribution. Yet in the case of subsets of sorghum and wheat data where the values are not too far from zero the box plots portray a noticeable deviation from normality.

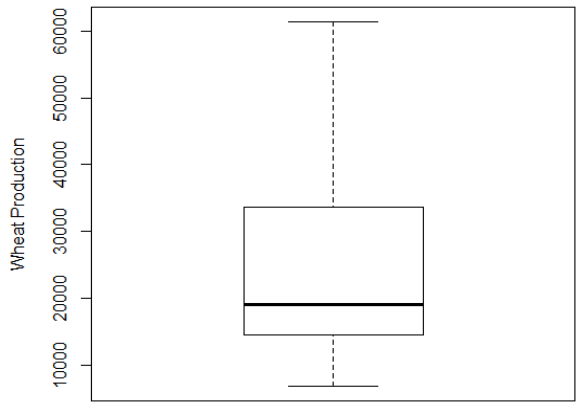

Figure 2.6: Box Plot for 1973 - 2002 Wheat Data

In general, the values of skewness and kurtosis for maize data are higher than zero and show that maize production for all the three periods are skewed to the right and have leptokurtic distributions that are peaked (Table 2.8). In the case of sorghum data, the values are close to zero and show distributions of sorghum production that are do not deviate much from the normal distribution. In

the case of wheat data, the subset with the values that show a deviation from normality in wheat production is of 1976/1977 to 2006/2007.

2.4.4 Exploration of Household Maize Data

Characteristics of households heads, the number of households and their proportions are presented in Table 2.9. More than half (55.4%) of these households were headed by males while 44.5% were headed by females. The biggest proportion of household heads (60.14%) attained primary education only, followed by 24% who did not have any form of formal education, and only 3.72% managed to attain education beyond high school, such as diploma and university degrees. Unemployment was prevalent among heads of the studied households since 43.24% of household heads were unemployed and only 17.91% constitute salary earners. The rest of the households were headed by either subsistence farmers, casual workers or pensioners.

Table 2.9: Characteristics of Heads of Households

Characteristic	Category	Households Heads	Percent
Gender	Female	132	44.59
	Male	164	55.41
Education Status	No Formal Education	73	24.66
	Primary Education	178	60.14
	High School Education	34	11.49
	Post High School Education	11	3.72
Occupation	Casual Worker	39	13.18
	Pensioner	20	6.76
	Salary Earner	53	17.91
	Subsistence farmer	56	18.92
	Unemployed	128	43.24

Summary statistics in this section are based on the information provided by households for the observational period, which was from the beginning of the last harvest season in June 2008 until in December 2008 when the survey was conducted. The statistics in Table 2.10 show that the smallest household in terms of size had one member while the biggest had fifteen members. The average size of households was five members, and the $25th$, $50th$ and $75th$ quantiles of the distribution of household size are four, five and seven, respectively. This shows that 25%, 50% and 75% of the households had less or equal to four, five and seven members, respectively. The mean and median number of household size are both equal to five, showing that the distribution of household size is not skewed but symmetric. Some of the households were headed by teenagers since the youngest head of a household was aged seventeen years, while some were headed by very old people, aged ninety six years. The minimum monthly income of zero suggests that there were households without any monthly income, while the maximum monthly income of R5000.00 suggests that some

households had monthly income as high as five thousand Rands. The $25th$, $50th$ and $75th$ quantiles of the distribution of households monthly income are 200, 350 and 735, respectively. This means that 25%, 50% and 75% of the households had monthly income less or equal to R200.00, R350.00 and R735.00, respectively.

The minimum availability of maize was as little as 18 kilograms for the entire observational period, indicating that availability of maize for some households was 18 kilograms. This could be due to various reasons, such as households did not have means to acquire more or households had alternative food cereals such as sorghum and wheat. The maximum quantity of maize available to households was 4800 kilograms and the mean value of maize availability was 631.63 kilograms, for the specified observational period. The $25th$, $50th$ and $75th$ quartiles of the distribution of maize availability are 336, 504 and 840, respectively. These quartiles indicate that 25%, 50% and 75% of the households had maize availability less or equal to 336, 504 and 840 kilograms, respectively. Both household monthly income and maize availability have distributions that are skewed to the right because their respective mean values are greater than their respective median values. This implies that the majority of households fell in the lower side of the distributions of monthly income and maize availability, while few households fell in the upper side of the distributions.

Table 2.10: Summary Statistics of Continuous Variables for Maize Household data

	Summary Statistics								
Variable	Minimum	Maximum	Q_1	Median	Q_3	Mean	Std. Deviation	CV	MAD
Household Size	1	15	4	5	7	5	2.57	47.83	2.97
Age	17	96	42	56	69	54.80	16.72	30.51	19.27
Income	0	5000	200	350	735	587.40	670.23	114.10	222.40
Availability	18	4800	336	504	840	631.63	488.58	77.35	302.50

The values of skewness and kurtosis of availability of maize are 3.60 and 24.36, respectively. As indicated earlier, the two quantities provide summary information about the shape of the distribution of the data. The value of skewness of 3.60 shows that the distribution deviates slightly from the normal distribution and it is skewed to the right. The value of kurtosis of 24.36 shows that the distribution is leptokurtic, meaning that it has a higher peak and heavier tails than the normal distribution. The tails that are heavier than that of the normal distribution imply high probability of extreme observations of maize availability.

The box plot in Figure 2.7, at a glance, shows that the distribution of maize availability deviates from the normal distribution. This is indicated by the line inside the box, showing the median value, which is not equidistant to the lower and upper edges of the box. This line is towards the lower edge of the box, showing a concentration of observations in the lower end of the distribution.

The concentration means that the majority of households fell in the lower side of the distribution of maize availability with relatively small quantities of maize available to them, while few households fell in the upper side of the distribution with relatively bigger quantities of maize availability. The box plot also shows several observations that appear beyond the end of the upper whisker. This is an indication of the presence of suspected extreme observations.

The variability within each of the variables in Table 2.10 is measured using the standard deviation, coefficient of variation (CV) and median absolute deviation (MAD). The CV of 114.10 for monthly income shows that monthly income was more variable than other variables, followed by availability of maize with the CV of 77.35. Age of household heads with CV of 30.51 was relatively less variable. The MAD is the robust measure of variability calculated as the median of absolute deviations between observations and their median. It is robust in the sense that it is minimally affected by a small fraction of extreme observations. Households maize availability and monthly income have large MAD of 302.50 and 222.40, respectively. This is an indication that the two variables were highly variable around their median.

Tables 2.11, 2.12, and 2.13 present summary statistics of households monthly income and maize availability by households location, education level and occupation of household head, respectively. Since there are suspected extreme values in both household income and maize availability, the median is an appropriate measure of location for both cases. Unlike the mean, the median is robust and not affected by extreme values. Households that resided in Berea were better off in terms of

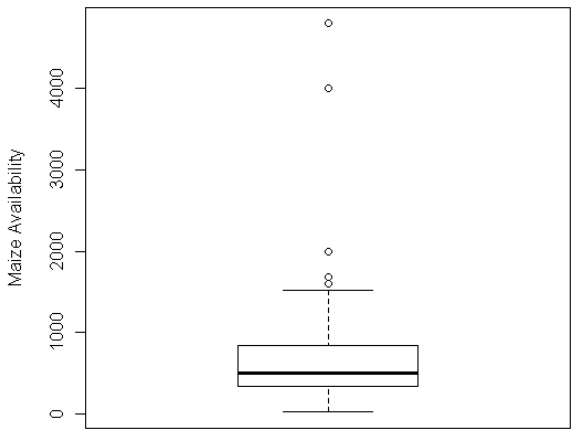

Figure 2.7: Box Plot of Household Maize Availability

Table 2.11: Summary Statistics by Location of Households For Maize Data

		Summary Statistics							
Location	Variable	Minimum	Maximum	Q_1	Median	Q_3	Mean	Std. Deviation	CV
Mafeteng	Income	100	1500	200	250	500	372.92	293.77	78.76
	Availability	43	1680	336	504	840	621.07	396.94	63.91
Maseru FootHill	Income	0	2000	130	238	300	311.07	343.00	110.26
	Availability	120	2000	420	580	860	686.88	419.12	61.02
Maseru Lowland	Income	0	5000	100	200	500	687.55	716.84	104.26
	Availability	18	4800	321	504	800	607.81	499.50	82.17
Berea	Income	100	4000	300	500	960	850.32	993.51	116.84
	Availability	170	4000	420	540	840	702.48	661.30	94.14

monthly income since they had the highest median of R500.00 (Table 2.11). On the other hand households in Maseru foothill were better off in terms of maize availability as they had the highest median of 580 kilograms. High minimum values as well as high values of the first and third quartiles, of both variables, are another indication that households in Berea and Maseru foothill were better-off, in terms of monthly income and maize availability, respectively. According to the first and third quartiles, 25% of Berea households had less or equal to R300.00 as their monthly income, while 75% had less or equal to R960.00. In the case of Maseru foothill, 25% of households had less or equal to 420.00 kilograms of maize availability, while 75% had less or equal to 860.00 kilograms.

The difference in monthly income within households in Maseru lowland was relatively high. Some households had no monthly income while others had monthly income as high as R5000.00. Some households had maize availability as low as 18 kilograms and others had as high as 4800 kilograms. Generally, income had a relatively high variability shown by high coefficients of variation. Maize availability for households in Berea had the highest variability followed by that of households in Maseru foothills, and Maize availability for households in Mafeteng had the lowest variability. The variability is shown by respective coefficients of variation of 116.84, 110.26 and 78.76.

Households with heads who attained post high school education, such as diploma and university degrees, had higher monthly incomes (Table 2.12). This is shown by high values of the first quartile, median and third quartile of income that correspond to post high school level of education of R1000.00, R1200.00 and R3500.00, respectively. This observation is not surprising because the expectation is that the higher the level of education the higher the earnings in terms of a salary. The median and third quartile for maize availability also shows that these households were better off when they are compared with their counterparts headed by people who had lower education qualifications. Interestingly, households headed by persons without any formal education rank second in terms of their maize availability. The distributions of monthly income and maize availability for households headed by people who attained primary education are characterized by too low

Table 2.12: Summary Statistics by Education of Household Head for Maize Data

		Summary Statistics							
Education Level	Variable	Minimum	Maximum	Q_1	Median	Q_3	Mean	Std. Deviation	CV
No Formal Education	Income	100	2000	260	500	770	389.18	437.88	112.57
	Availability	53	1680	356	560	870	659.08	410.05	62.22
Primary	Income	0	5000	200	400	800	608.90	666.59	109.47
	Availability	75	4800	336	504	800	609.39	476.53	78.20
High School	Income	100	1500	200	500	700	523.53	353.39	67.50
	Availability	18	1512	356	510	700	594.35	401.80	67.60
Post High School	Income	150	4000	1000	1200	3500	1752.27	1338.96	76.41
	Availability	168	4000	336	600	1050	924.73	1079.72	116.76

minimum values and high maximum values, respectively. In general, monthly incomes of households headed by people with no formal education and those who attained primary and high school education were almost similar.

Table 2.13: Summary Statistics by Occupation of Household Head for Maize Data

		Summary Statistics							
Occupation	Variable	Minimum	Maximum	Q_1	Median	Q_3	Mean	Std. Deviation	CV
CasualWorker	Income	100	2000	300	500	800	607.69	445.52	73.31
	Availability	18	1008	178	420	672	430.46	253.03	58.78
Pensioner	Income	200	1500	200	300	400	407.50	341.54	83.81
	Availability	168	4000	323	576	1008	802.00	853.67	106.40
Salary earner	Income	100	5000	500	1000	1800	1299.53	1114.05	85.73
	Availability	150	4800	400	672	1020	828.13	711.74	85.94
Subsistence farmer	Income	50	2000	200	300	500	421.43	359.91	85.40
	Availability	43	1680	433	575	871	692.78	393.31	56.77
Unemployed	Income	0	2000	150	255	500	387.07	364.52	94.18
	Availability	53	1680	336	500	820	558.15	334.19	59.87

Households headed by salary earners were better off as they had the highest monthly income. This is shown by values of the maximum of R5000.00, lower quartile of R500.00, middle quartile of R1000.00 and upper quartile of R1800.00 (Table 2.13). This is not surprising as their heads were employed and earned some income on a monthly basis. Monthly incomes for households headed by people of other occupations than salary earners varied slightly. In the case of maize availability, households headed by salary earners were better off, followed by households headed by pensioners and subsistence farmers, which were slightly different. The coefficients of variation of monthly income and maize availability, under different occupations of household heads, show that the variability of the two variables was not that different. However, maize availability for households headed

by pensioners had high variability, shown by the highest CV of 106.40.

2.4.5 Exploration of Household Sorghum Data

Generally, the minimum, maximum, and median values as well as the first, second and third quartiles of sorghum availability are smaller than that of maize availability (Tables 2.10 and 2.14). For example the median value of sorghum availability is 150 kilograms while that of maize availability is 350 kilograms. This is an indication that households in the sample acquired less sorghum than maize during the observational period. The skewness and kurtosis of sorghum availability are 1.78 and 5.96, respectively. The value of skewness shows that the distribution of sorghum availability is slightly skewed to the right and the value of kurtosis shows that the distribution has a peak that is slightly higher than that of the normal distribution.

Table 2.14: Summary Statistics of Sorghum Availability

Minimum	Maximum	Q_1	Median	Q_3	Mean	Std. Deviation	MAD
3	1050	60	150	250	176.53	150.73	133.50

The box plot in Figures 2.8, at a glance, shows that the distribution of sorghum availability is approximately normal since the line in the box is almost in the middle. In addition, it shows two observations that appear as suspected extreme values.

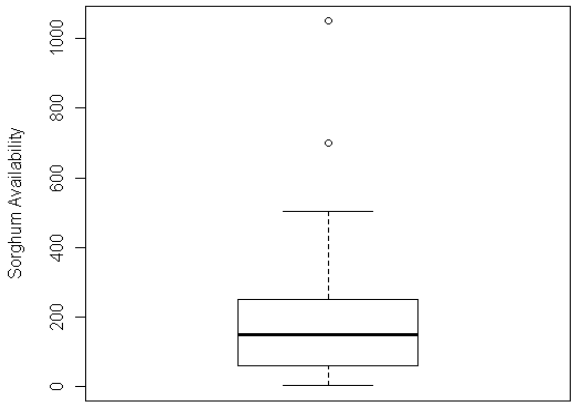

Figure 2.8: Box Plot of Household Sorghum Availability

Tables 2.15, 2.16 and 2.17 show sorghum availability by location of a household, education level and occupation of the head of a household, respectively. Generally, households in Maseru lowlands and Berea were better off in terms of availability of sorghum. This is because they had almost the same sorghum availability, which is higher than that of households in Mafeteng and Maseru foothill. This is shown by higher median value of one hundred and sixty eight (168) kilograms for both Maseru lowlands and Berea. Sorghum availability for households in Berea had the highest variability shown by the highest standard deviation of 215.84.

Table 2.15: Sorghum Availability by Location of Households

	Summary Statistics of Availability						
Location	Minimum	Maximum	Q_1	Median	Q_3	Mean	Std. Deviation
Mafeteng	3	480	36	91	170	126.24	122.47
Maseru Foothill	10	420	50	84	160	126.77	117.55
Maseru Lowland	3	504	84	168	300	196.36	136.34
Berea	17	1050	92	168	278	224.16	215.84

Availability of sorghum for households headed by people who attained post high school education was relatively high when it is compared with that of their counterparts headed by people with lower education qualifications (Table 2.16). This could be an indication that households headed by people who attained diploma and/or university degree had more resources or purchasing power, which enabled them to have more access to sorghum than those headed by people with lower qualifications. The highest standard deviation of 325.36 of sorghum availability for these households show that their sorghum availability was highly variable when it is compare with that of households headed by people with lower qualifications.

Table 2.16: Sorghum Availability by Education of Household Head

	Summary Statistics						
Education Level	Minimum	Maximum	Q_1	Median	Q_3	Mean	Std. Deviation
No Formal Education	3	504	36	95	160	137.14	135.93
Primary	8	700	65	150	240	171.36	128.14
High School	3	500	80	140	316	196.84	157.64
Post High School	50	1050	198	316	460	372.91	325.36

In general, summary statistics in Table 2.17 show that households headed by salary earners had higher availability of sorghum than their counterparts headed by people of other occupations. Similar reasoning used in the case of maize availability that salary earners earn income monthly and

Table 2.17: Sorghum Availability by Occupation of Household Head

	Summary Statistics						
Occupation	Minimum	Maximum	Q_1	Median	Q_3	Mean	Std. Deviation
CasualWorker	60	500	126	200	300	220.51	133.28
Pensioner	10	1050	50	150	310	220.73	267.84
Salary earner	10	500	60	199	336	197.97	145.95
Subsistence farmer	3	700	90	159	232	179.53	139.47
Unemployed	3	504	48	93	210	141.96	124.23

can use their income to acquire maize for their households can be extended to the case of sorghum availability. Households headed by unemployed people had the lowest sorghum availability. This is inline with the general understanding that if households heads are unemployed they do not have enough resources or purchasing power to acquire or access sorghum for their households. Variability of sorghum availability for households headed by pensioners, shown by the standard deviation of 267.84, was the highest.

2.5 Summary

Exploration of the national data showed that on average, domestic production of maize contributed a larger proportion of the Lesotho requirements for maize, in the years 1973 to 2002, than imports. In the case of wheat, commercial imports contributed a larger part of the national requirements than domestic production. The area planted to maize and wheat affected by crop failure varied greatly within the specified period. The big variability could be due to erratic weather conditions that are prevalent lately. Generally the averages of area planted to maize and maize production were the highest when they were compared with their counterparts for sorghum and wheat. This is inline with the known fact that Lesotho produce more maize than sorghum and wheat because maize is the only staple food in the country.

Skewness, kurtosis and box plots were used to explore the distribution of a potential response variable in each subset of data. Deviations from normality were observed and there was an interesting observation concerning the use of skewness, kurtosis and box plots in different sets of data. It looks like the quantity in which the values of skewness and kurtosis should be far from zero to indicate deviation from assumption of normality varies from one set of data to another. This was observed from subsets of the maize data where the values were relatively higher than zero but the box plots showed slight deviations from the normal distribution. Yet in the case of subsets of the sorghum and wheat data where the values were not too far from zero, the box plots portrayed a noticeable deviation from the assumption of normality.

Exploration of household data showed that the distribution of maize availability deviated slightly from the normal distribution and showed outlying observations, while sorghum availability was approximately normally distribution but with two outlying observations. Summary statistics showed that availability of maize was generally higher than availability of sorghum for the studied households. This could be due to the fact that maize is the only staple food in Lesotho and thus every household used available resources to acquire it in big quantities than the other two cereals during the observational period. The exploration of both the national and household data formed the basis for modelling the data by detecting preliminarily relationships that exist in the sets of data and variables that are involved in such variables.

Chapter 3

Regression Modelling and Diagnostics

3.1 Introduction

Regression analysis is the most researched and applied area in statistics, which is used to study relationships among variables. It is a statistical tool that is commonly applied to study the relationship of two or more variables so that the response variable Y can be described and predicted from ($p \geq 1$) predictor variables, normally denoted by $X_1, \ldots, X_p$. The overall analysis includes graphical and analytic methods of exploring relationships between a response variable and predictor variables. The application of regression analysis is common in business, social sciences, behavioral sciences, biological sciences, medicine and many other research areas (Kutner et al., 2005).

Regression models include the simple linear regression model that involves a response variable Y and one predictor variable X, and the multiple regression model that involves a response variable and two or more predictor variables $X_1, \ldots, X_p$, with $p \geq 2$. Typical data associated with a regression analysis problem are generated from a process where the response variable and predictor variables are measured on a sample of say n individuals, years, plots or any other unit of interest, depending on the field of application. The general linear model has alternatives specialized in dealing with different situations and problems that emerge in the application of linear regression models.

3.2 Linear Regression Model

The simplest form of a regression model is called the simple linear regression model and it is used to study bivariate relationships of variables. The model that represents a linear bivariate relationship for n observations can be presented as

$$y_i = \beta_0 + \beta_1 x_i + u_i, \quad i = 1, 2, \ldots, n, \tag{3.1}$$

with the mean response, $E(y_i|x_i) = \beta_0 + \beta_1 x_i$, where y_i is the ith value of the continuous response variable and x_i is the corresponding value of the predictor variable, β_0 and β_1 are unknown model parameters or regression coefficients, and u_i is the error term that gives the random variation in Y not explained by X. The error terms u_i are assumed to be random variables with mean zero and a constant variance σ^2, and to be pairwise independent. The predictor variable is known and assumed to be measured without error. The coefficient β_1 is the slope of the regression line that gives the change in the mean of the probability distribution of Y associated with a unit increase in X. The coefficient β_0 is the Y intercept of the regression line that gives the mean of the probability distribution of Y when $X = 0$. The regression model in equation (3.1) is linear in the parameters.

The simple linear regression can be generalized to the multiple linear regression model used when one predictor variable in the model cannot provide an adequate description of a response variable and the knowledge of more than one predictor variable is necessary to give a better understating and prediction of a response variable, as it is the case in this research. Kutner et al. (2005) noted that in most applications of regression analysis predictions of the response variable based on a model with one predictor variable are not precise to be useful. This happens when a number of variables affect the response variable substantially in a different manner.

The general linear regression model that relates the response variable Y with p predictor variables $X_1, \ldots, X_p$ can be defined as

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots \beta_p x_{ip} + u_i, \tag{3.2}$$

where $\beta_0, \ldots, \beta_p$ are model parameters or partial regression coefficients. The variables $X_1, \ldots, X_p$ are a set of known quantities assumed to be measured without error and thus designated as predictor variables. Like in a simple linear regression model, u_i are error terms assumed to have mean zero, constant variance σ^2, and to be pairwise independent.

Generally the linear multiple regression model can be presented in matrix form as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}, \tag{3.3}$$

where $\mathbf{y}$ is a $n \times 1$ vector of observations on the response variable, $\mathbf{X}$ is a $n \times (p+1)$ matrix of predictor variables or the design matrix with the first column of ones, and $\boldsymbol{\beta}$ is a $(p+1) \times 1$ vector of model parameters including the constant, also called partial regression coefficients. A partial regression coefficient reflects the partial effect of one predictor variable when the rest of the predictor variables included in the model are held constant. The vector $\mathbf{u}$ is a $n \times 1$ vector of independent error terms with mean vector zero, and the covariance matrix, $\sigma^2\mathbf{I}$. The two vectors $\mathbf{y}$ and $\mathbf{u}$ are random because their elements are random variables, and $\mathbf{X}$ is a matrix of known constants. The classical regression analysis assumes that in a regression relationship only the response variable Y is assumed to be measured with error. The ith row of the design matrix is

the vector $\mathbf{x}_i' = (1, x_{i1}, \ldots, x_{ip})$ of observed values of p predictor variables corresponding to the response variable value measured in the ith observational or experimental unit.

3.2.1 General Linear Model

Multiple linear regression models with quantitative predictor variables can be extended to more complex models by including qualitative predictor variables. This extension results in what is commonly known as the general linear model, which consists of a combination of quantitative and qualitative or categorical predictor variables. Categorical variables identify the category to which a studied observation belongs and can be useful in explaining the variability in the response variable. These variables increase flexibility in the application of regression models to real life situations. Dummy variables or indicator variables are used to represent categorical variables in regression models. These variables can take on two values, normally 0 or 1, which signify that an observation belong or does not belong to a category. The number of dummy variables required to represent a categorical variable in a model is one less than the number of categories in the represented variable. In general if a categorical variable has m categories, $m - 1$ dummy variables are sufficient in the model, since the inclusion of all the m categories can result in linear dependency among columns of the design matrix $\mathbf{X}$ (Kutner et al., 2005). The variation in the response variable associated with a particular categorical variable is the total variation among the categories of the variable.

Consider the general linear model with two quantitative variables and one categorical variable with two categories, which can be presented as

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + u_i, \tag{3.4}$$

where x_{i1} and x_{i2} are the ith values of X_1 and X_2 respectively, x_{i3} is the value of the dummy variable representing one of the two categories of the categorical variable, which is 1 if an observation belongs to a category and 0 otherwise. The omitted category is normally referred to as the reference or base category. This is usually the category of the variable of interest to which other categories of the variable are compared. The coefficients, β_1, β_2 and β_3 are model parameters as in model (3.2). However, unlike the other two parameters, β_3 that corresponds to the dummy variable, measures how the mean response varies as a categorical variable changes from one category to another. It represents the differential effect of the categorical predictor variable on the response variable, for fixed values of X_1 and X_2. This is the difference or contrast between a given category and the reference category, showing how much higher or lower the mean response of a given category is than the mean response of the reference category, for fixed values or categories of the other variables.

3.2.2 Multiple Regression Model with First-order Autoregressive Process AR(1)

Sometimes multiple regression model can involve a component of an autoregressive process where some of the predictor variables are past values of the response variable. In such cases the response variable Y is regressed on its past values and separate predictor variables. Consider the linear regression model that include the first-order autoregressive process AR(1) where one autoregressive term Y_{t-1} is included as one of the predictor variables. The response variable at time t, Y_t is regressed on its past immediate values y_{t-1} and some other predictor variables $X_{1t}, \ldots, X_{pt}$ at time t. The regression model with first-order autoregressive process AR(1) can be presented as

$$\mathbf{y}_t = \alpha \mathbf{y}_{t-1} + \mathbf{X}\boldsymbol{\beta} + \mathbf{u}, \tag{3.5}$$

where $\mathbf{y}_t$ is a vector of observations on the response variable at time t, $\mathbf{y}_{t-1}$ is a vector of observations on the response variable at time $t-1$, and $\mathbf{X}$ is the design matrix. The parameters in the model consists of an autoregressive coefficient α associated with the autoregressive term $\mathbf{y}_{t-1}$ and a vector of partial regression coefficients $\boldsymbol{\beta}$. The autoregressive coefficient reflect the dependence of successive observations of the same variable. This is the dependence of the current values of the response variable and their past immediate values. Each of the partial regression coefficients in the vector $\boldsymbol{\beta}$ reflect the partial effect of one predictor variable when the rest of the predictor variables are held constant. The vector $\mathbf{u}$ is as defined under model (3.3), and the ordinary least squares (OLS) procedure can be used to estimate model parameters.

3.3 Estimation of Model Parameters

3.3.1 Ordinary Least Squares Procedure (OLS)

Estimation of model parameters in linear regression model can be done by the OLS estimation procedure. The least squares procedure was first published by Legendre in 1805 and later by Gauss in 1809 (Sen and Srvastava, 1990). The procedure seeks to find the estimates of parameters that minimize the sum of squared deviations of the n observed values $y_1, \ldots, y_n$ from their model predicted values $\hat{y}_i, \ldots, \hat{y}_n$. The sum of squared residuals for the general linear model can be expressed as

$$S(\boldsymbol{\beta}) = \mathbf{u}'\mathbf{u} = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}). \tag{3.6}$$

The estimates are obtained by differentiating the sum of squared residuals with respect to the vector $\boldsymbol{\beta}$. This yields a set of normal equations given by

$$\mathbf{X}'\mathbf{y} = \mathbf{X}'\mathbf{X}\boldsymbol{\beta}. \tag{3.7}$$

This set has a unique solution if and only if the matrix $\mathbf{X}$ is full rank. The full rank implies that the inverse $(\mathbf{X}'\mathbf{X})^{-1}$ exists. The solution to the normal equations, assuming $\mathbf{X}$ is of full rank, is

the unique OLS estimates of $\boldsymbol{\beta}$ given by

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}. \tag{3.8}$$

When the assumptions about the error terms hold, OLS estimators $\hat{\boldsymbol{\beta}}$ have optimal properties of being best linear unbiased estimators (BLUE). The assumptions about the error terms are sometimes referred to as Gauss-Markov (G-M) conditions. Under the G-M conditions the mean and covariance matrix of $\hat{\boldsymbol{\beta}}$ are given by $E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$ and $\text{cov}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$, respectively. Oftentimes σ^2 is not known and is replaced by its estimate S^2, which is the mean square error (MSE), given by

$$S^2 = \frac{S(\boldsymbol{\beta})}{n-p}, \tag{3.9}$$

where $S(\boldsymbol{\beta})$ is as defined in equation (3.6), n is the number of observations and p is the number of model parameters to be estimated.

The least squares estimation procedure of regression coefficients does not require distributional assumption about the error terms to give unbiased estimators with minimum variance. However, for making inferences about the parameter estimates, assumptions about the distribution of the error terms are required. The standard assumption about the distribution of the error terms is that the errors follow a normal distribution. When the Gauss-Markov conditions and normality assumption hold, the regression model is called the normal error regression model. The error terms in this model are independent and identical normally distributed with mean zero and covariance matrix σ^2, that is, $\mathbf{u} \sim iid\ \mathcal{N}(0, \sigma^2\mathbf{I})$. When the distribution of the response variable is skewed, transformation such as the Box-Cox transformation discussed in the subsequent section can be applied to validate the normality assumption. The model specified in terms of the mean response possesses a good feature that it can be extended to other distributions which are not necessarily normal (McCullagh and Nelder, 1989).

When the matrix $\mathbf{X}$ is not full rank, the inverse $(\mathbf{X}'\mathbf{X})^{-1}$ does not exist, and the set of normal equations in equation (3.7) does not have a unique solution. In such situations, the generalized inverse $(\mathbf{X}'\mathbf{X})^{-}$ is used and the resultant parameter estimates are given by

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{y}. \tag{3.10}$$

In this case $\hat{\boldsymbol{\beta}}$ is not an unbiased estimator of $\boldsymbol{\beta}$, and further the solution is not unique because it depends on the g-inverse used.

3.3.2 Generalized Least Squares (GLS)

Generalized least squares (GLS) estimation method is an alternative to ordinary least squares procedure, which works when one or both of the assumptions about the error terms are not satisfied.

Sometimes it happens that the assumption that the variance of the error terms, u_i is constant does not hold. In such cases the covariance matrix of $\mathbf{u}$ is not $\sigma^2\mathbf{I}$, but it is a diagonal matrix with unequal diagonal elements. It can also happen that the assumption that error terms are not correlated is violated, which would mean that the off-diagonal elements of the covariance matrix are not zero, and hence the covariance matrix of error terms is not $\sigma^2\mathbf{I}$, but rather a more general variance-covariance matrix, for example $\mathbf{V}$ multiplied with a variance component σ^2. When one or both of these assumptions are not satisfied the OLS estimates of $\boldsymbol{\beta}$ do not apply and this calls for the transformation of the linear model in such a way that the least squares estimates can be obtained.

Consider the general linear model in equation (3.3), where in this particular case, $E(\mathbf{u}) = \mathbf{0}$, $\text{cov}(\mathbf{u}) = \sigma^2\mathbf{V}$, and $\mathbf{u} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{V})$. Assume the design matrix $\mathbf{X}$ is of full rank, and $\mathbf{V}$ is a known positive definite matrix. In this case the ordinary least squares estimation procedure is not applicable and the response variable Y need to be transformed to some variable Z, which satisfy the assumptions required for the estimation procedure. This can be achieved by multiplying the linear model by the inverse of a nonsingular symmetric matrix $\mathbf{P}$, given by $\mathbf{P}^{-1}$, where $\mathbf{P}'\mathbf{P} = \mathbf{P}\mathbf{P} = \mathbf{P}^2 = \mathbf{V}$. The resultant transformed model can be presented as

$$\mathbf{z} = \mathbf{Q}\boldsymbol{\beta} + \mathbf{f}, \tag{3.11}$$

where $\mathbf{z} = \mathbf{P}^{-1}\mathbf{y}$, $\mathbf{Q} = \mathbf{P}^{-1}\mathbf{X}$, and $\mathbf{f} = \mathbf{P}^{-1}\mathbf{u}$ so that $E(\mathbf{f}) = \mathbf{0}$. Since $\mathbf{f}$ is a random variable with mean zero and is a linear combination of the elements of a normally distributed $\mathbf{u}$, it is also normally distributed with mean zero and covariance matrix $\sigma^2\mathbf{I}$, that is $\mathbf{f} \sim \mathcal{N}(0, \sigma^2\mathbf{I})$.

The ordinary least squares procedure can be applied to obtain the estimates $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ for the transformed regression model in equation (3.11) since the assumptions about the vector of the error terms that, $E(\mathbf{f}) = \mathbf{0}$ and $\text{cov}(\mathbf{f}) = \sigma^2\mathbf{I}$ are satisfied in the transformed model (Montgomery et al., 2012). The estimates of the model parameters in the vector $\boldsymbol{\beta}$ are the unique generalized least squares estimates defined by

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}, \tag{3.12}$$

with the mean $\boldsymbol{\beta}$ and covariance matrix $\sigma^2(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}$. When the Gauss-Markov conditions hold, the generalized least squares estimators $\hat{\boldsymbol{\beta}}$ are best linear unbiased estimators (BLUE) of $\boldsymbol{\beta}$.

3.3.3 Maximum Likelihood Estimation (MLE)

Maximum likelihood estimation (MLE) procedure can also be used to obtain estimates of the parameters for the linear regression model when the functional distribution of the probability distribution of the error term $\mathbf{u}$ is specified. The method of maximum likelihood chooses, as estimates of the

parameters, the values in the parameter space that are most consistent with the sample. MLE requires the specification of the probability of the observations y_i in order to arrive at the estimates of the unknown parameters. In other words the method relies on the distributional assumptions about the data, without loss of generality that assumes the data are generated from a normal distribution.

Generally the density function of the ith observation of the response variable for the normal simple regression model is defined as

$$f(y_i) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{1}{2}\left(\frac{y_i - \beta_0 - \beta_1 x_i}{\sigma}\right)^2\right\}. \tag{3.13}$$

The density function derives from the fact that $E(Y_i|x_i) = \beta_0 + \beta_1 x_i$ and $\text{var}(Y_i) = \sigma^2$.

The likelihood function of n observations $y_1, y_2, \ldots, y_n$, which is the function of unknown parameters, $\boldsymbol{\beta} = (\beta_0 \quad \beta_1)'$, and σ^2, is the product of individual density functions given in equation (3.13). Thus the likelihood function is given by

$$L(\boldsymbol{\beta}, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}}\exp\left\{-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right\}. \tag{3.14}$$

The maximum likelihood estimation method maximizes the likelihood or the log-likelihood function with respect to $\boldsymbol{\beta}$ and σ^2, to obtain the corresponding maximum likelihood estimators $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$. Maximum likelihood estimators are identical to least squares estimators in equation (3.10), by possessing optimal properties of least squares estimators, provided the same data were used for both estimators (Montgomery et al., 2001). In addition, maximum likelihood estimators have other desirable properties that they are consistent, sufficient, and have minimum variance among all linear and nonlinear unbiased estimators. However, the maximum likelihood estimator $\hat{\sigma}^2$ is a biased estimator of the parameter σ^2 since $E(\hat{\sigma}^2) \neq \sigma^2$. The MLE method can be generalized to deal with the case where $\mathbf{y} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \mathbf{V}\sigma^2)$, discussed earlier under generalized least squares procedure. In this case the maximum likelihood estimators turnout to be equal to the ones given in equation (3.12), provided the same data were used.

3.3.4 Box-Cox Transformation

Oftentimes, real life data are not normally distributed, which is a problem concerning the application of linear regression models, since these models require the assumption of normality for making inferences about parameter estimates. An appropriate transformation of data can lead to data that are approximately normally distributed, and increase the applicability and usefulness of linear regression models. The Box and Cox (1964) transformation of the response variable Y automatically identifies a transformation from the family of power transformations. The transformation's aim is to ensure that the normality and constant variance assumptions for linear model are satisfied.

Consider a set of n positive observations $y_1, \ldots, y_n$ on the response variable, Y, then if the ratio of the largest observation to the smallest is of a considerable size, say, 10 or bigger the transformation of of Y can be made (Draper and Smith, 1998). The Box-Cox family of transformation can be expressed as

$$y^{(\lambda)} = \begin{cases} \frac{y^{\lambda}-1}{\lambda}, & \text{if } \lambda \neq 0; \\ \log y, & \text{if } \lambda = 0. \end{cases} \tag{3.15}$$

The transformation is continuous at $\lambda = 0$. For some unknown λ, the transformed observations denoted by $y_i^{(\lambda)}$ are assumed to be independently normally distributed with the mean $\mathbf{X}\boldsymbol{\beta}$ and constant variance σ^2, that is $\mathbf{y}^{(\lambda)} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$, where $\mathbf{y}^{(\lambda)}$ is a vector of transformed observations, $\mathbf{X}$ is a design matrix and $\boldsymbol{\beta}$ is a vector of parameters associated with the transformed observations.

The likelihood of the transformed observations can be presented as

$$\frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} \exp\left\{ -\frac{(\mathbf{y}^{(\lambda)} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y}^{(\lambda)} - \mathbf{X}\boldsymbol{\beta})}{2\sigma^2} \right\} \mathbf{J}(\lambda; \mathbf{y}), \tag{3.16}$$

where $\mathbf{J}(\lambda; \mathbf{y}) = \prod_{i=1}^{n} \left| \frac{dy_i^{(\lambda)}}{dy_i} \right|$ is the Jacobian of the transformation from $\mathbf{y}$ to $\mathbf{y}^{(\lambda)}$.

Box and Cox (1964) suggested two approaches that can be used to make inferences about the parameters in equation (3.16). The two approaches are the maximum likelihood and Bayesian methods. Using the maximum likelihood method, the estimates of the regression parameters $\boldsymbol{\beta}$, σ^2 and the transformation parameter λ are obtained by maximizing the likelihood function of equation (3.16) or its log-likelihood. For the fixed λ, the maximum likelihood estimates of $\boldsymbol{\beta}$ are least squares estimates and the maximum likelihood estimate of σ^2 is given by

$$\hat{\sigma}^2(\lambda) = \frac{\mathbf{y}'^{(\lambda)}(\mathbf{I}_n - \mathbf{G})\mathbf{y}^{(\lambda)}}{n} = \frac{S(\lambda)}{n}, \tag{3.17}$$

where $\mathbf{G} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, $S(\lambda)$ is the residual sum of squares, and the matrix $\mathbf{X}$ is full rank. The $100(1-\alpha)\%$ confidence interval for λ can be derived by inverting the likelihood ratio test of the null hypothesis that $H_0 : \lambda = \lambda_0$.

3.4 Goodness-of-Fit for Linear Regression

It becomes necessary that after the regression model is fitted, some measure is used to detect how good the fit of the model is. The coefficient of determination denoted by R^2 is one of such measures and it is commonly used to measure goodness-of-fit for the classical regression model. This measure is calculated from sum of squares and can be defined as

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}, \tag{3.18}$$

where SS_{res} and SS_{tot} are the residual sum of squares and total sum of squares, respectively. The coefficient of determination is a global measure of goodness-of-fit over the entire distribution of the response variable, which shows the proportion of total variability in the response variable accounted for by the regression model. It assumes values in the range [0, 1], with the value of 0 indicating that predictor variables in the regression model are not useful in predicting or explaining the response variable and thus there is no linear relationship between the predictor and response variables. On the other hand, an R^2 value of 1 indicates that all variability in the response variable is accounted for by the linear regression model or predictor variables in the model.

The application of linear regression model to real life data can sometimes be confronted with problems that occur in the data set. These problems may interfere with applicability and usefulness of the models, if they are not corrected or taken care off when fitting the model. Some of the problems that are common in the real life data sets are collinear variables, models not fitting the data well and assumptions that are violated due to various reasons such as the presence of extreme values in the data set. Collinearity, its diagnostics, and regression model diagnostics are discussed in the subsequent sections.

3.5 Collinearity and Remedies

Most observational studies, in which multiple regression is commonly applied involve predictor variables that are correlated among themselves (Kutner et al., 2005). When predictor variables are highly correlated there exists a condition called multicollinearity or collinearity. This condition where the data are said to be ill-conditioned is sometimes referred to as the presence of near linear dependencies among predictor variables or columns of the design matrix $\mathbf{X}$. Collinearity result in the violation of a classical regression assumption that predictor variables are independent. If there are no linear dependencies among columns of $\mathbf{X}$, predictor variables are said to be orthogonal.

3.5.1 Effects of Collinearity

The presence of collinearity in predictor variables can cause difficulties in answering some of the questions that are addressed in multiple regression analysis. Some of the questions are, on the relative importance of the effects of different predictor variables on the response variable, the size of the effect a given predictor variable has on the response variable, and the decision to drop any predictor variable from the model due to having little or no effect on the response variable (Kutner et al., 2005). When the data are ill-conditioned the matrix $\mathbf{X}'\mathbf{X}$ is not of full rank. This implies that $\mathbf{X}'\mathbf{X}$ is not invertible and the parameter estimates from the data are not unique.

Collinearity does not affect the overall fit of the model, and does not affect inferences about the

mean response, or prediction, provided the inferences are done within the scope of the observations. However, it reduces the efficiency of regression if the purpose of the analysis is to determine the effects that predictor variables have on the response variable. When there is collinearity problem regression coefficients become unstable with large standard errors, they become large in magnitude and have wrong signs that are opposite to the anticipated signs, based on theory or prior knowledge. In addition, the coefficients become very sensitive to small random errors in the response variable and are likely to fluctuate as predictor variables are added or removed from the model.

When two or more predictor variables are highly correlated, it becomes almost impossible to separate their influence on the response variable. The interpretation of regression coefficients that they reflect the change in the mean response variable when a given predictor variable is increased by one unit while the rest of the predictor variables included in the model are held constant, is not applicable. Under severe collinearity, regression coefficients are likely to have large roundoff errors. As it is the case with regression coefficients, the coefficients of partial correlation between the response variable and each predictor variable are likely to be variable from one sample to another.

3.5.2 Diagnosis of Collinearity

There are a number of diagnostics reported in literature for detecting the presence of collinearity among the columns of $\mathbf{X}$ (Draper and Smith, 1998; Kutner et al., 2005; Belsley et al., 1980; Belsley, 1991). Some of the suggested diagnostics are; some estimated regression coefficients having wrong sign that is opposite to the sign anticipated on the basis of theory, intuition or prior knowledge in the subject matter, nonsignificant results of individual tests on the regression coefficients for predictor variables that are considered to be important on the basis of prior knowledge, big changes in the fitted model caused by deleting or adding a predictor variable or an observation, and large correlation coefficients between pairs of predictor variables. According to Kutner et al. (2005) these are informal diagnostic of collinearity, which have some limitations.

Belsley et al. (1980) and Belsley (1991) discussed a number of procedures that have been used to detect collinearity. The discussed collinear diagnostics include examining the correlation matrix of predictor variables, variance inflation factors (VIF), examining the determinant $det(\mathbf{X}'\mathbf{X})$ of the matrix $\mathbf{X}'\mathbf{X}$ or the determinant of the correlation matrix $det(\mathbf{C})$, performing all-subsets of regression on predictor variables, partial correlations of the the data matrix $\mathbf{X}$ and the correlation matrix of the least squares estimates $\hat{\boldsymbol{\beta}}$, and examining the eigenvalues and eigenvectors, or principal components of $\mathbf{X}$ or of the correlation matrix $\mathbf{C}$. The authors outlined shortfalls of these procedures, which rendered them incomplete as collinearity diagnostics, and suggested collinearity diagnostics based on eigenvalues and eigenvectors that take care of the identified shortfalls. This section discusses some of the commonly used procedures.

1. **Correlation Matrix**

   The correlation matrix is a commonly used and easy to compute tool for detecting collinearity between predictor variables, however it has some limitations as a collinearity diagnostics. High correlation coefficient between a pair of predictor variables can indicate collinearity problem, but the absence of high correlation coefficient does not always mean that there is no collinearity problem. The correlation matrix is not able to diagnose collinearity that involves three or more variables when there are no pairwise collinear relationships between the variables. This is due to the fact that three or more variables taken together can be collinear while there are no high pairwise correlations observed among them. The correlation matrix is also not able to show several collinear relationships that coexist in the set of data. Another shortfall of the correlation matrix as a collinearity diagnostic is lack of standard measure of how large should the correlation coefficient be to indicate collinearity.

2. **Variance Inflation Factor (VIF)**

   The variance inflation factor (VIF) is one of the commonly used measures of detecting the presence of collinearity in predictor variables. Variance inflation factors are computed from the correlation matrix $\mathbf{C}$ of the predictor variables. The factors measure the quantity by which the variances of the estimated regression coefficients for correlated variables are inflated as compared to when the predictor variables are not correlated. Assuming that the predictor variables, which are columns of the data matrix $\mathbf{X}$ have been centered and scaled to unit length. The variance inflation factors are the diagonal elements of the inverse $\mathbf{C}^{-1}$ of the correlation matrix $\mathbf{C}$. The variance inflation factor of the jth regression coefficient, VIF_j is defined by

$$VIF_j = \frac{1}{1 - R_j^2}, \qquad 0 \leq R_j^2 \leq 1, \tag{3.19}$$

   where R_j^2 is the coefficient of determination computed based on regressing X_j on other predictor variables.

   The name variance inflation factor was introduced in the 1960s by Marquardt (Belsley, 1991). The high value of VIF_j shows that the value of R_j^2 is close to 1.0. This is an indication of the presence of collinearity, which leads to inflated variances of regression coefficients. When variances of estimated coefficients get inflated they lead to small values of the t-statistics for individual coefficients hence causing insignificance, despite the overall model F-statistic being significant. If predictor variables are orthogonal, meaning that they are not linearly related, R_j^2 will be 0 and VIF_j will be 1.0. A value of variance inflation factor that is greater than 10 is an indication of collinearity problems (Belsley, 1991). Variance inflation factor values that are greater than 30 imply severe collinearity problems. However, values of variance inflation factors should be evaluated in relation to the overall fit of the model of interest (Freund and Wilson, 1998).

3. **Tolerance**

   Sometimes tolerance is used together with the variance inflation factor to detect the presence of near linear relationships among predictor variables. Tolerance measures the amount of variance in the jth predictor variable X_j, which is not explained by other predictor variables. Tolerance can be expressed as the reciprocal of the variance inflation factor defined as

$$\text{Tolerance} = \frac{1}{\text{VIF}_j} = 1 - R_j^2, \tag{3.20}$$

   where R_j^2 is as defined earlier. If there are linear relationships that involve X_j and other predictor variables, R_j^2 will be close to 1.0 and tolerance will be close to 0. This implies that almost all of the variability in X_j is explained by other predictor variables. The variance inflation factor and tolerance are inversely related. Values of tolerance that are less than or equal to 0.10, or equivalently values of variance inflation factor that are greater than or equal to 10 show that there may be problems of near dependencies among predictor variables.

   Like the correlation matrix, the variance inflation factor and tolerance have a number of shortfalls. As it is the case with any measure based on correlation, large values of variance inflation factors and small values of tolerance are a sign of collinearity problems. However, small values of variance inflation factor and large values of tolerance do not necessarily indicate the absence of collinearity problems. The variance inflation factor and tolerance are not able to diagnose several separate collinear relationships that exist simultaneously in the data matrix $\mathbf{X}$. Another shortfall of the variance inflation factor and tolerance is the lack of the well established methods of determining a meaningful cutoff point of large and small values of the two collinear diagnostics.

4. **Eigenvalues and Eigenvectors**

   The eigenvalues and eigenvectors of the matrix $\mathbf{X}'\mathbf{X}$ or of the correlation matrix $\mathbf{C}$ have been used to detect and deal with collinearity. Principal components of the matrix $\mathbf{X}$ can also be used to deal with collinearity problems. Silvey (1969) suggested that small singular values of $\mathbf{X}$ or small eigenvalues of $\mathbf{X}'\mathbf{X}$ indicate the presence of near linear dependency among columns of $\mathbf{X}$, where singular values of $\mathbf{X}$ are the square roots of the eigenvalues of $\mathbf{X}'\mathbf{X}$. The eigenvalues and eigenvectors procedure is superior to the collinearity diagnostics discussed earlier. This procedure has the ability to deal with several collinear relationships that exist at the same time in the data set since there is a small eigenvalue that corresponds to each of such relationships. However the eigenvalues procedure has a shortfall of lacking a clear definition of how small should the eigenvalue be to indicate the presence of near linear dependency.

3.5.3 Singular-value Decomposition (SVD) Method

Belsley et al. (1980) and Belsley (1991) noted that the reported collinearity diagnostics have limitations in detecting near linear dependencies since they are not able to diagnose the presence of more than one collinear relationship that exist simultaneously in the set of data. In addition they fail to determine variables that are involved in each of such relationships, and assess the potential negative impact of collinearity on estimated regression coefficients. The authors applied the numerical analysis and extended the idea of Silvey (1969) to develop two diagnostics that are able to deal with collinearity problems sufficiently. The two diagnostics, which make use of the eigenvalues and eigenvectors, are condition index and variance-decomposition proportion.

Condition index and variance-decomposition proportion are based on the singular-value decomposition (SVD) of the data matrix $\mathbf{X}$, given by $\mathbf{X} = \mathbf{DUV}'$, where $\mathbf{U}'\mathbf{U} = \mathbf{V}'\mathbf{V} = \mathbf{I}_p$. The matrix $\mathbf{D}$ is a $p \times p$ diagonal matrix with nonnegative diagonal elements $\mu_1, \ldots, \mu_p$ referred to as singular values of $\mathbf{X}$, which are positive square roots of the eigenvalues $\lambda_1, \ldots, \lambda_p$ of $\mathbf{X}'\mathbf{X}$. The matrix $\mathbf{V}$ is a $p \times p$ matrix whose columns are eigenvectors of $\mathbf{X}'\mathbf{X}$, and $\mathbf{U}$ is $p \times p$ matrix whose columns are p eigenvectors of $\mathbf{XX}'$ that correspond to its p nonzero eigenvalues.

1. **Condition Index**
   Condition indices are able to detect more than one collinear relationships that coexist among columns of the matrix $\mathbf{X}$. They are developed from the concept of the condition number of the matrix $\mathbf{X}$. The condition number is the ratio of the largest to the smallest singular value defined as

$$\kappa(\mathbf{X}) = \sqrt{\frac{\lambda_{\max}}{\lambda_{\min}}} = \frac{\mu_{\max}}{\mu_{\min}} \geq 1, \tag{3.21}$$

   where $\lambda_{\max}$ and $\lambda_{\min}$ are the largest and smallest eigenvalues, respectively, and $\mu_{\max}$ and $\mu_{\min}$ are the respective largest and smallest singular values. The largest singular value or eigenvalue is used as a measure that determines how small should be a singular value or an eigenvalue that indicates the presence of near linear dependency among columns of the matrix $\mathbf{X}$. A large value of the condition number $\kappa(\mathbf{X})$ suggests high degree of collinearity among predictor variables.

   The concept of condition number was extended to develop the condition index such that the set of condition indices of the matrix $\mathbf{X}$, $\{\eta_k : k = 1, 2, \ldots, p\}$ are given by

$$\eta_k = \sqrt{\frac{\lambda_{\max}}{\lambda_k}} = \frac{\mu_{\max}}{\mu_k}, \qquad k = 1, \ldots, p \tag{3.22}$$

   where $\lambda_{\max}$ and $\mu_{\max}$ are as defined above, and λ_k and μ_k are the kth eigenvalue and singular value, respectively. The condition index $\eta_k \geq 1$ for all values of k, and it is computed for

each singular value. The largest value of the set $\{\eta_k : k = 1, 2, \ldots, p\}$ is equal to the condition number of $\mathbf{X}$, $\kappa(\mathbf{X})$. A singular value μ_k, which is small when it is compared with its yardstick, $\mu_{\max}$, corresponds to a large condition index. A large condition index indicates the presence of near linear dependency among predictor variables that correspond to the smaller singular value.

Belsley et al. (1980) and Belsley (1991) showed empirically that condition indices in the range of 5 to 10 suggest weak collinear relationships whereas condition indices of 30 to 100 suggest moderate to strong collinear relationships. Condition indices have two advantages, as collinearity diagnostics, over other collinearity diagnostics discussed earlier. The first advantage is that the use of condition indices makes it possible to determine when an eigenvalue or a singular value is small. The second advantage is that the occurrence of more than one large condition index indicates the presence of more than one near linear dependencies that coexist in the set of data. This is so because there are as many collinear relationships among the columns of $\mathbf{X}$ as there are large condition indices.

2. **Variance-Decomposition Proportion**

Despite the fact that condition indices are helpful in establishing the existence and number of collinear relationships among predictor variables, they fail to point to variables that are involved in such relationships. Belsley et al. (1980) adapted the work of Silvey (1969) on variance-decomposition and combined it with the idea of condition indices to develop the variance-decomposition proportion to address this limitation. The variance-decomposition proportion establishes predictor variables that are involved in a particular collinear relationship, and measures the extent to which the least squares parameter estimates are adversely affected by the presence of collinearity. This collinear diagnostic is based on the fact that when a predictor variable used in least squares regression is involved in a collinear relationship, the variance of its estimated regression coefficient gets inflated.

Let $\phi_{kj} = \frac{\upsilon_{kj}^2}{\mu_j^2}$ and $\phi_k = \sum_{j=1}^{p} \phi_{kj}$, for $k = 1, \ldots, p$, then (k, j)th variance-decomposition proportion can be defined by

$$\pi_{jk} \equiv \frac{\phi_{kj}}{\phi_k} \qquad k, j = 1, \ldots, p, \tag{3.23}$$

where υ_{kj} are components of the matrix $\mathbf{V}$. The variance-decomposition proportion is derived from the fact that the variance of the estimate of the kth regression parameter $\hat{\beta}_k$, can be expressed as

$$\text{var}(\hat{\beta}_k) = \sigma^2 \sum_{j=1}^{p} \frac{\upsilon_{kj}^2}{\mu_j^2} = \sigma^2 \sum_{j=1}^{p} \frac{\upsilon_{kj}^2}{\lambda_j}, \tag{3.24}$$

where p is the number of predictor variables, μ_j is the jth singular value of $\mathbf{X}$, λ_j is the jth eigenvalue of $\mathbf{X}'\mathbf{X}$, and $\mathbf{V} = (\upsilon_{ij})$. The variance of $\hat{\beta}_k$ follows from the covariance matrix

$\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$ of OLS estimates $\hat{\boldsymbol{\beta}}$, which can be expressed in terms of the SVD of the matrix $\mathbf{X}$ as

$$\text{cov}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1} = \sigma^2\mathbf{V}\mathbf{D}^{-2}\mathbf{V}', \tag{3.25}$$

where σ^2 is the variance of the error term, and the matrices $\mathbf{D}$ and $\mathbf{V}$ are as defined earlier. The $\text{var}(\hat{\beta}_k)$ is the kth diagonal element of the covariance matrix of $\hat{\boldsymbol{\beta}}$ decomposed into the sum of p components, where each component corresponds to only one of the p singular values, μ_j, of the matrix $\mathbf{X}$ or to only one of the p eigenvalues, $\lambda_j = \mu_j^2$, of the matrix $\mathbf{X}'\mathbf{X}$.

The variance-decomposition proportion is the proportion of the variance of $\hat{\beta}_k$ associated with the jth component of the variance-decomposition in equation (3.23). Components of the variance of $\hat{\beta}_k$ associated with collinear relationships among columns of $\mathbf{X}$, or with small singular value μ_j and large condition index η_k are considered to be large when they are compared with other components. Thus large proportions of the variances of at least two regression coefficients associated with one small singular value μ_j, provides evidence that predictor variables that correspond to the regression coefficients whose variances are affected are involved in the identified collinear relationships. Since two or more variables are required to have a near linear dependency, the variances of two or more regression coefficient should be adversely affected to show a linear dependency. The cutoff point for variance proportions that show involvement of a variable in a near linear dependency is 0.5.

3.5.4 Remedies of Collinearity

Once near linear relationships that exist among predictor variables are identified through the use of collinearity diagnostics the next step would be to find ways of dealing with such relationships. Hocking (1996) noted that the method of dealing with near linear dependencies depends on the objective of the analysis at hand and source of collinearity. A number of remedies of collinearity have been reported in literature and for some of them one has to be aware of the existence of near linear dependencies and their consequences and use the best judgement possible. If collinearity is caused by redundancy where several variables measure the same thing then one of the methods, which can be applied is to drop one of the collinear variables. Other collinear remedies are; respecification of the model, use of additional or new data, use of data reduction techniques such as principal component analysis, principal components regression, and the commonly used ridge regression (Kutner et al., 2005). Some of these remedies are not commonly used because of their limitations or because they are not always possible to apply. For example, principal components regression has serious limitations (Hadi and Ling, 1998), and the use of additional or new data is not always possible because it is too costly in terms of time and expense. Ridge regression has been used as a remedial measure of collinearity, in the current research work.

3.5.5 Ridge Regression

One of the methods that are commonly used to remedy collinearity problems in data, and control the instability of parameter estimates is ridge regression. Ridge regression achieves all these by modifying the least squares procedure to allow for a biased estimator of regression coefficients, which is called ridge estimator. Ridge estimation was first introduced by Hoerl in 1962 to control inflation and general instability in least squares estimates (Hoerl and Kennard, 1970a). An estimator with a small bias but with high precision than an unbiased estimator may be preferred because it will have a larger probability of being close to the value of the true parameter. It is worth noting that generally, the precision of an estimator is inversely related to its variance. The smaller the variance of an estimator, the higher the precision of the estimator. The ordinary least squares estimator $\hat{\boldsymbol{\beta}}$ is unbiased but may be imprecise, while the ridge estimator, $\hat{\boldsymbol{\beta}}_R$, may be more precise despite having some bias.

Ridge regression is applied to the standardized regression model, which is the centered and scaled model. According to Kutner et al. (2005), the standardized regression model is given by

$$Y_i^* = \beta_1^* X_1^* + \ldots + \beta_p^* X_p^* + u_i^*, \tag{3.26}$$

where $Y_i^* = \frac{1}{\sqrt{\mathrm{n}-1}}\left(\frac{Y_i - \bar{Y}}{S_Y}\right)$ and $X_{ik}^* = \frac{1}{\sqrt{\mathrm{n}-1}}\left(\frac{X_{ik} - \bar{X}_k}{S_{X(k)}}\right)$. The unbiased least squares estimator for the standardized model is given by

$$\hat{\boldsymbol{\beta}}^* = (\mathbf{X}^{*\prime}\mathbf{X}^*)^{-1}\mathbf{X}^{*\prime}\mathbf{y}, \tag{3.27}$$

where $\mathbf{X}^{*\prime}\mathbf{X}^* = \mathbf{r}_{XX}$ and $\mathbf{X}^{*\prime}\mathbf{y} = \mathbf{r}_{XY}$ are the correlation matrix that consist of correlation coefficients between the predictor variables, and the correlation vector between the response variable and predictor variables, respectively. Both matrices have elements that are between -1 and 1. The use of the standardized regression model helps to control roundoff errors when the matrix $\mathbf{X}^{*\prime}\mathbf{X}^*$ is inverted, and to express regression coefficients in the same units for comparison purposes.

When the data are ill-conditioned due to near linear dependency among columns of the matrix $\mathbf{X}^*$, the standardized least squares estimator becomes unreliable with large variances. In the presence of linear dependencies, normal equations do not have a unique solution because $\mathbf{X}^{*\prime}\mathbf{X}^*$ matrix is singular and cannot be inverted. Ridge regression estimator is used to remedy collinearity problems by introducing a biasing vector δ to the diagonal of $\mathbf{X}^{*\prime}\mathbf{X}^*$. The introduction of δ leads to a non-singular matrix $(\mathbf{X}^{*\prime}\mathbf{X}^* + \delta\mathbf{I})$ and thus reduces the impact of collinearity on parameter estimates. This yields biased estimates of $\boldsymbol{\beta}$ but with smaller variances than ordinary least squares estimates, and narrower confidence intervals of $\boldsymbol{\beta}$. Hoerl and Kennard (1970a); Montgomery et al. (2012) defined the biased ridge estimator as

$$\hat{\boldsymbol{\beta}}_R^* = (\mathbf{X}^{*\prime}\mathbf{X}^* + \delta\mathbf{I})^{-1}\mathbf{X}^{*\prime}\mathbf{y}, \tag{3.28}$$

where δ is a small positive number called the ridge factor, determined by the user. Alternatively the biased ridge estimator can be expressed as a linear transformation of the ordinary least squares estimator $\hat{\boldsymbol{\beta}}^*$ as

$$\hat{\boldsymbol{\beta}}_R^* = (\mathbf{X}^{*\prime}\mathbf{X}^* + \delta\mathbf{I})^{-1}(\mathbf{X}^{*\prime}\mathbf{X}^*)\hat{\boldsymbol{\beta}}^*, \tag{3.29}$$

(Marquardt and Snee, 1975; Montgomery et al., 2012). The covariance matrix of the ridge estimator and the bias in ridge estimator can be presented as $\text{var}(\hat{\boldsymbol{\beta}}_R^*) = \sigma^2(\mathbf{X}^{*\prime}\mathbf{X}^* + \delta\mathbf{I})^{-1}\mathbf{X}^{*\prime}\mathbf{X}^*(\mathbf{X}^{*\prime}\mathbf{X}^* + \delta\mathbf{I})^{-1}$ and $\text{Bias}(\hat{\boldsymbol{\beta}}_R^*) = -\delta(\mathbf{X}^{*\prime}\mathbf{X}^* + \delta\mathbf{I})^{-1}\boldsymbol{\beta}^*$, respectively. Montgomery et al. (2012) defined the mean square error of the ridge estimator, used to measure the precision of regression estimates, by

$$\begin{aligned} MSE(\hat{\boldsymbol{\beta}}_R^*) &= \text{var}(\hat{\boldsymbol{\beta}}_R^*) + \{\text{Bias}(\hat{\boldsymbol{\beta}}_R^*)\}^2 \\ &= \sigma^2 \sum_{j=1}^{p} \frac{\lambda_j}{(\lambda_j + \delta)^2} + \delta^2 \boldsymbol{\beta}^{*\prime}(\mathbf{X}^{*\prime}\mathbf{X}^* + \delta\mathbf{I})^{-2}\boldsymbol{\beta}^*, \end{aligned} \tag{3.30}$$

where $\lambda_1, \ldots, \lambda_p$ are the eigenvalues of $\mathbf{X}^{*\prime}\mathbf{X}^*$. The variance of $\hat{\boldsymbol{\beta}}_R^*$ decreases as the biasing factor increases, but the squared bias in $\hat{\boldsymbol{\beta}}_R^*$ increases as the biasing factor increases. When $\delta = 0$, the ridge regression estimator is similar to the unbiased standardized least squares estimator $\boldsymbol{\beta}^*$. However, when $\delta > 0$, the ridge regression estimator is biased but it is more stable than the ordinary least squares estimator.

Determination of the Biasing Factor

A number of suggestions were made in literature on how to determine the value of the biasing factor δ. Hoerl and Kennard (1970b) recommended the use of ridge trace to determine the value of δ. The ridge trace is a plot of the individual parameter estimates $\hat{\beta}_{jR}^*$ against δ, for values of δ in the interval [0, 1]. The smallest value of δ at which the parameter estimates are stabilized is chosen as the ridge factor. The objective of the ridge trace is to show the sensitivity of estimation of parameters to nonorthogonality of predictor variables.

Hoerl et al. (1975) suggested that a reasonable selection of the biasing factor δ is

$$\delta = \frac{p\hat{\sigma}^2}{\hat{\boldsymbol{\beta}}^{*\prime}\hat{\boldsymbol{\beta}}^*}, \tag{3.31}$$

where $\hat{\boldsymbol{\beta}}^*$ is the least squares estimator under scaled variables, and $\hat{\sigma}^2$ is an estimator of the variance of the error term. The suggested choice of δ gives ridge regression coefficients with smaller mean square error than least squares estimator. Hoerl and Kennard (1976) suggested an iterative method of selecting δ based on the selection method suggested by Hoerl et al. (1975). The selection of the biasing factor based on the iterative method is defined as

$$\delta = \frac{p\hat{\sigma}^2}{\hat{\boldsymbol{\beta}}_{\delta_i}^{*\prime}\hat{\boldsymbol{\beta}}_{\delta_i}^*}, \tag{3.32}$$

where $\hat{\beta}_{\delta_i}$, $i = 0, 1, ...$, is estimated using different values of δ until the termination criteria is satisfied. The iteration process is repeated until the difference between two successive estimates is negligible (Chatterjee and Hadi, 2006). The ridge factor is chosen in such a way that the decrease in the variance term of the mean square square error of $\hat{\boldsymbol{\beta}}_R^*$ is greater than the increase in the squared bias. In that case the mean square error of $\hat{\boldsymbol{\beta}}_R^*$ will be less than the variance of the least squares estimator $\hat{\boldsymbol{\beta}}^*$. Hoerl and Kennard (1970a) showed that there is a positive value of δ, for which the ridge estimator has a smaller mean square error than the variance of the least squares estimator.

Chatterjee and Hadi (2006) noted that ridge regression yields parameter estimates that are more robust to small changes in the data than the least squares procedure. Ridge coefficients estimates have the property of smaller mean square error, which makes them to be closer to the true unknown parameters than the least squares estimates. However, the choice of the biasing factor δ is to some extent subjective since there are several methods of selecting δ that are suggested, of which none is recommended as the best in terms of performance. Like the least squares estimator, a ridge estimator is sensitive to the presence of extreme observations in the data. Hence the need for regression diagnostics that detect extreme observations and their influence on the ridge estimator.

3.6 Regression Model Diagnostics

In regression analysis it is important to scrutinize the results and find out if there are problems that may compromise the validity of the results. This is commonly done through regression diagnostics, which consist of a number of techniques used to check the quality of data, and if the assumptions of the model are satisfied. If the distributional and model assumptions are not satisfied model diagnostics check the extent to which they are violated. Most of the regression diagnostics have three major components, namely, fitted values, $\hat{y}_i$, residuals, e_i and leverages, h_{ii}. Residuals and leverages are diagnostic techniques by themselves, and they also constitute fundamental parts in building almost all diagnostic techniques based on case deletion (Cook and Weisberg, 1982) and (Sen and Srvastava, 1990).

Oftentimes regression analysis is applied on data sets that involve some extreme observations called outliers and high leverage points. The presence of outlying observations can be attributed to a number of factors. Belsley et al. (1980) outlined three sources of outlying observations as, erroneous recording of data either at the data collection stage or at the data capturing stage, observational errors, and correct extreme observations that contain important information. Extreme observations are a concern in ordinary least squares estimation because the procedure is very sensitive to outlying observations. This is mostly attributed to the fact that the procedure is used on the regression based on the mean response. The regression model based on quantiles of the distribution of the

response variable were also studied in the current research. Outlying observations may have large residuals and affect the least squares regression coefficients.

Graphic diagnostics such as scatter plots can be used to identify outlying observations for regression with one or two predictor variables. However, when more than two predictor variables are included in the regression model scatter plots may not show multivariate outliers (Belsley et al., 1980; Hocking, 1996). In such cases residuals can be used to identify observations that are outlying with respect to a response variable, while leverage can be used to identify observations that are extreme with respect to predictor variables. The three components of regression diagnostics, predicted values, residuals and leverage, are discussed in the subsequent sections.

3.6.1 Predicted Values and Leverages

Given the solution to the general linear model in equation (3.10), the vector of n fitted values is defined by $\hat{\mathbf{y}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$. This can also be presented as $\hat{\mathbf{y}} = \mathbf{H}\mathbf{y}$, where $\mathbf{H}$ is the $n \times n$ hat matrix defined as

$$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'. \tag{3.33}$$

The hat matrix is sometimes referred to as the projection matrix. The matrix $\mathbf{H}$ has two important properties that it is symmetric and idempotent. A diagonal element of the hat matrix is the leverage of the ith case denoted by h_{ii}, which is defined as

$$h_{ii} = \mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i \quad i = 1, 2, \ldots, n, \tag{3.34}$$

where $\mathbf{x}_i'$, the ith row of the design matrix $\mathbf{X}$, corresponds to a single observation or case in the data. Leverages are helpful in detecting cases that are extreme with respect to their p predictor variables values in multiple regression analysis. One of their properties is that $0 \leq h_{ii} \leq 1$. The diagonal element h_{ii} measures the distance that indicates how far away an individual observation x_i is from the centroid or the mean, $\bar{x}$, of all n observations in the data. It reflects how extreme is an observation relative to other observations. A large value of h_{ii} shows that the ith observation is far from the center of all observations for a predictor variable, and hence it is extreme in terms of the predictor variable. This observation has a large leverage value and it is called a high leverage point, which has a big influence in determining the fitted value $\hat{y}_i$. According to Belsley et al. (1980), the ith observation is considered as the leverage point when its value of h_{ii} is greater than $2p/n$. A leverage value of zero indicates that the ith observation has no influence on the fitted model.

3.6.2 Residuals

Residuals consist of important information, which can be used to detect inadequacies in the regression model and problems in the data. Thus they are normally used as diagnostic tools to check if

the fitted model deviates from the linear regression model in terms of nonlinearity of the regression function, unequal variance of error terms, correlated error terms, presence of extreme and influential observations, error terms that are not normal, and omission of important variables from the model. Diagnostics plots of residuals against predictor variable or fitted values give information on these departures for the linear model. Some of the plots that are used are scatter plot, box plot, and normal probability plot. Although graphic analysis of residuals is sometimes considered to be an informal diagnostic tool, in most cases it gives adequate information for assessing the correctness of the model (Kutner et al., 2005).

The residuals, $\hat{u}_i = y_i - \hat{y}_i$, are observed errors of the unknown true error terms u_i. Like the true error terms, the residuals have mean zero. However, residuals may have varying variances, where the variance of the ith residual is given by $\text{Var}(\hat{u}_i) = \sigma^2(1 - h_{ii})$. If the error terms are normally distributed the residuals are also normally distributed. Unlike the error terms that are independent, residuals are correlated. The ordinary residuals, $\hat{u}_i$ can be used to detect outlying observations but they are not good indicators of extreme observations, and they cannot be compared because they have different variances (Hocking, 1996). Ordinary residuals can be standardized so that they have a common variance of one and become more effective diagnostic measure. The standardized residuals r_i called internally studentized residuals are defined as $r_i = \frac{\hat{u}_i}{s\sqrt{(1-h_{ii})}}$, where s is the square root of the MSE used to estimate the variance of the error term σ^2, under independence and common variance, and $s\sqrt{(1 - h_{ii})}$ is the estimate of the standard error of the residuals.

An alternative method that is often recommended is to use the leave one observation out method in calculating the standardized residuals. These residuals are called the studentized deleted residual or externally studentized residual, and can be defined as

$$t_i = \frac{\hat{u}_i}{\sqrt{s^2_{(i)}(1 - h_{ii})}}, \tag{3.35}$$

where $s^2_{(i)}$ is the mean square error computed from fitting the regression model when the ith observation is deleted or left out.

Sometimes the study of residuals does not give full information on outlying observations since some of the observations can have relatively small residuals yet they are outliers (Belsley et al., 1980). Cook and Weisberg (1980) noted that in some cases, an individual observation or a group of observations can influence important computations from fitting the linear regression model while they cannot be detected by studying residuals. Such situations call for the use of alternative diagnostics with the ability to identify outlying observations and study them further.

3.6.3 Identification of Influential Cases

When observations that are outlying in terms of their response variable values, predictor variable values or both have been identified, it is important to study them further and establish whether they are influential or not. An observation is considered influential if deleting it from the data causes major changes in the fitted regression model. Sen and Srvastava (1990) noted that a case is considered influential if it has a large residual or it is far away from the centroid of the space of the predictor variables. There are a number of diagnostic techniques based on the deletion of the ith observations, which have been proposed in literature (Cook, 1977; Belsley et al., 1980). These techniques measure the influence of the ith observation on the fitted regression model.

Chatterjee and Hadi (1996) reviewed a number of the proposed influence measures and discussed their relationships. Their relationships emanates from the fact that they are functions of the residuals $\hat{u}_i$, the residual mean square error s^2, and the ith diagonal element of the hat matrix h_{ii}. However, different diagnostic tools are intended to detect influence of the ith case on different quantities of the fitted regression model, such as the estimated coefficients, predicted values, and the covariance matrix of the estimated coefficients. Four measures of influence based on the deletion of a single observation that are commonly applied to detect influence on different quantities of the fitted model are discussed. The measures are, DFBETAS, DFFITS, Cook's Distance, and COVRATIO.

1. **DFBETAS**

   Belsley et al. (1980) proposed DFBETAS as the measure of the influence of the ith case on each regression coefficient $\hat{\beta}_j$, where $j = 1, \ldots, p$. This influence measure is used to study the change in the estimated regression coefficients caused by leaving out the ith observation from the data. If the primary objective of regression analysis is estimation of parameters, DFBETAS is the appropriate measure of the influence of the ith observation. DFBETAS is defined by

   $$(\text{DFBETAS})_{j(i)} = \frac{\hat{\beta}_j - \hat{\beta}_{j(i)}}{\sqrt{s^2_{(i)} c_{jj}}} \qquad j = 1, 2, \ldots, p, \tag{3.36}$$

   where $\hat{\beta}_j$ is the regression coefficient estimated from all n observations, and $\hat{\beta}_{j(i)}$ is the regression coefficient estimated when the ith observation is deleted from the n observations. The term c_{jj} is the jth diagonal element of $(\mathbf{X}'\mathbf{X})^{-1}$, and $s^2_{(i)} c_{jj}$ is an estimate of the variance of $\hat{\beta}_j$ given by $\sigma^2 c_{jj}$. The variance of the error term σ^2 is estimated by $s^2_{(i)}$, the mean square error obtained when the ith case is left out in fitting the regression model.

   The sign of DFBETAS indicates whether including the ith case leads to an increase or a decrease in the estimated regression coefficient. Belsley et al. (1980) noted that large absolute values of DFBETAS show that the ith case has an influence in determining the jth

regression coefficient $\hat{\beta}_j$. The authors suggested a cutoff point of $2/\sqrt{n}$ of DFBETAS for observations that are considered influential and need to be studied further.

2. **DFFITS**

   If the main objective of regression analysis is prediction, the assessment of the influence of the ith observation on predicted values becomes important. Welsch-Kuh's distance by Welsch and Kuh (1977) also referred to as DFFITS by Belsley et al. (1980) is a diagnostic technique used to measure the influence of the ith case on the fitted value $\hat{y}_i$. It measures the change in the ith fitted value in terms of standard error units when the ith observation is excluded (Schabenberger and Pierce, 2002). DFFITS is defined by

$$\text{DFFITS}_i = \frac{\hat{y}_i - \hat{y}_{i(i)}}{s_{(i)}\sqrt{h_{ii}}}, \tag{3.37}$$

   where $\hat{y}_i$ is the fitted value for the ith observation when all the data are used in fitting the regression model, and $\hat{y}_{i(i)}$ is the predicted value for the ith case obtained when the regression model is fitted without the ith case.

   DFFITS can also be expressed as $\text{DFFITS}_i = t_i \left(\frac{h_{ii}}{1-h_{ii}}\right)^{1/2}$. In this case, the DFFITS value for the ith case is the externally studentized residual t_i adjusted by the factor $\left(\frac{h_{ii}}{1-h_{ii}}\right)^{1/2}$. This factor is a function of the leverage value for the same ith case. If the ith case is extreme with respect to the predictor variable X and has a high leverage value, the adjusting factor will be greater than 1 and the value of DFFITS_i will be large in absolute value. A large value of DFFITS_i indicates that the ith observation is influential. Belsley et al. (1980) suggested a cutoff point of $2\sqrt{p/n}$ when using DFFITS to identify observations that are considered influential and need to be investigated further.

3. **Cook's Distance**

   The Cook's distance proposed by Cook (1977) is an aggregate influence measure that measures the influence of the ith case on all n fitted values. The Cook's distance D_i compares each of the n fitted values $\hat{y}_i$ with the corresponding fitted value $\hat{y}_{i(i)}$ obtained when the ith case is deleted in fitting the regression model and can be expressed in vector notation as

$$D_i = \frac{(\hat{\mathbf{y}}_{(i)} - \hat{\mathbf{y}})'(\hat{\mathbf{y}}_{(i)} - \hat{\mathbf{y}})}{ps^2}, \tag{3.38}$$

   where $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$ is the vector of fitted values when all n cases are used for fitting the regression model, and $\hat{\mathbf{y}}_{(i)} = \mathbf{X}\hat{\boldsymbol{\beta}}_{(i)}$ is the vector of the fitted values when the ith case is deleted. The diagnostic D_i is the scaled Euclidean distance moved by the fitted vector when the ith case is deleted from the data (Cook and Weisberg, 1980; Chatterjee and Hadi, 1996).

   In another expression, $D_i = \frac{t_i^2}{p}\frac{h_{ii}}{(1-h_{ii})}$, the Cook's distance depends on the size of the externally studentized residual t_i and leverage value h_{ii}. If t_i, h_{ii} or both of them are large, D_i

will be large, hence the ith case will be influential. Cook (1977) suggested that the values of D_i be compared with the probability $100(1-\alpha)\%$ points of F distribution with p and $n-p$ degrees of freedom. According to Weisberg (2005) if the largest value of D_i is substantially less than 1, the ith observation does not influence $\hat{\boldsymbol{\beta}}$.

4. **COVRATIO**

The covariance matrix of the estimated regression coefficients $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$ is another important aspect of regression analysis, for which the effect of the ith observation needs to be investigated. Belsley et al. (1980) introduced a diagnostic technique based on the deletion of the ith row of the design matrix $\mathbf{X}$ that compares the covariance matrix computed from all the data, $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$, and the covariance matrix computed when the ith observation is deleted from the data, $\sigma^2(\mathbf{X}'_{(i)}\mathbf{X}_{(i)})^{-1}$. This diagnostic technique measures the influence of the ith observation on the covariance matrix and is given by the ratio of the determinants of the two covariance matrices defined by

$$\begin{aligned} \text{COVRATIO} &= \frac{det\left[s^2_{(i)}\left(\mathbf{X}'_{(i)}\mathbf{X}_{(i)}\right)^{-1}\right]}{det\left[s^2\left(\mathbf{X}'\mathbf{X}\right)^{-1}\right]} \\ &= \left(\frac{n-p-t_i^2}{n-p-1}\right)^p \frac{1}{1-h_{ii}}, \end{aligned} \tag{3.39}$$

where s^2 and $s^2_{(i)}$ are mean square errors used to estimate the variance σ^2 computed from all the data and the variance $\sigma^2_{(i)}$ computed when the ith observation is deleted from the data, respectively. COVRATIO measures the effect of the ith observation on the precision of the estimated regression coefficients. According to Belsley et al. (1980) observations that result in COVRATIO values not close to 1 have a potential of being influential and hence they need to be investigated further. Values of COVRATIO that are close to 1 show that the ith observation has a little influence on the precision of the estimates.

Rawlings et al. (1998) noted that when the value of COVRATIO is greater than 1, there is an indication that the ith observation increases the precision of the estimates, whereas when the value of COVRATIO is less than 1 there is an indication that the ith observation decreases precision of the estimates. Belsley et al. (1980) suggested a cut-off point of $1 \pm 3p/n$ for observations that need to be investigated further. That is, if COVRATIO $\geq 1 + 3p/n$ or COVRATIO $\leq 1 - 3p/n$ the ith observation influences the covariance matrix of the estimated regression coefficients.

Once an influential observation and its effect on quantities of the fitted model have been detected, the first step would be to examine if the observation was erroneously recorded and if so the possibility of correcting the error, or if the case is a valid observation that gives important information. In cases where it is a valid observation, robust regression procedures that reduce the influence of

influential observations can be applied, instead of OLS procedure. Quantile regression by Koenker and Bassett (1978), as an alternative robust and flexible estimation approach to classical regression approach is discussed and applied in the subsequent in chapters.

Chapter 4

Linear Regression Models and Diagnostics of National Data

4.1 Introduction

Linear regression models were applied to model the national food cereals availability. Domestic production of each of the three main cereals, maize, sorghum and wheat, as one component of national availability of cereals, was modelled using the general linear regression model. The classical linear regression model was fitted to the national maize data, while the linear regression model with the first-order autoregressive process AR(1) was fitted to the sorghum and wheat data, where production in the past immediate year was identified to be having a relationship with the current production. This model included the autoregressive term, Y_{t-1}, which is production of sorghum or production of wheat in the past immediate year as one of the predictor variable. This variable consists of past values of the response variable, which is the current production of sorghum or current production of wheat in this research. The ordinary least squares estimation procedure was used to estimate regression parameters.

Collinearity diagnostics were used to detect the presence of collinear relationships that exist among predictor variables in the model. The condition index was used to diagnose the presence of more than one collinear relationship that exist simultaneously in the set of data. The variance-decomposition proportion was used to identify variables that are involved in a particular collinear relationship, and assess the extent to which the least squares parameter estimates are adversely affected by collinearity. On the basis of the identified near linear dependencies among columns of the data matrix $\mathbf{X}$, we used ridge regression to control inflation and instability of estimated regression coefficients caused by the dependencies.

We applied linear regression model diagnostics to check if the distributional and model assumptions

are satisfied. Diagnostics based on the deletion of the ith observations were used to identify extreme observations that influence different quantities of the fitted model, such as, the estimated regression coefficients, predicted values, and covariance matrix of the estimates. The model was refitted after excluding each of the identified influential observations, one at the time. The results from fitting the model to the data with and without a particular observation or case, were compared to establish the extent to which the observation was influential. In cases where violation of assumptions such as the non-constant variance of the error term and distributions of error term with longer and heavier tails than the normal distribution were established, we applied the Box-Cox transformation to identify an appropriate transformation from a family of power transformation for correction.

The REG procedure of SAS was used to fit the linear regression model to the national data as well as to perform collinearity and regression diagnostics. In identifying collinear relationships and variables that are involved in such relationships, the procedure uses an approach that follows the one of Belsley et al. (1980). The REG procedure has the INFLUENCE option that produces influence statistics proposed by Belsley et al. (1980) to measure influence of each observation on estimates of model parameters. The diagnostics and box plots were plotted by the R system of graphics, after checking that R gives the same results as the ones produced by SAS when fitting the linear regression models.

4.2 Regression Model of National Maize Data

The linear regression model was fitted to different subsets of maize data. In fitting the model to the 1973/1974 to 2001/2002 data, maize production was regressed on four predictor variables for both cases of with and without suspected extreme observations. The four predictor variables are time, the amount of rainfall, area harvested to maize and price of maize, which appeared in Figure 2.1 and Table 2.1 as having linear relationships with maize production. The suspected extreme cases are the third and twenty seventh observations of the years 1975/1976 and 1999/2000, respectively. The third observation has the highest amount of rainfall of 1042.52 milliliters, and a relatively low maize production of 49128 tonnes. The twenty seventh observation has the highest maize production of 277685 tonnes. This is suspected to be the observation that appeared outstanding in Figures 2.2, M.1 and M.2 of box plots for all the three subsets of maize data, in Chapter 2.

The results from fitting the model with and without the extreme cases show that the overall F test of the null hypothesis that none of the predictor variables is linearly related to maize production, is significant at 5% level of significant. This suggests that at least one of the four predictor variables in the fitted models has a linear relationship with maize production. According to the value of the coefficient of determination R^2 of 0.54 from fitting the model to the data set with all observations, the predictor variables in the regression model explain 54% of the variability in maize production.

Table 4.1: Parameter Estimates for the Full Set of 1973 - 2002 Maize Data

Response Variable: maize production

Variable	DF	Estimate	Std. error	t value	Pr > |t|	Tolerance	VIF
Intercept	1	-62918.00	44599.00	-1.41	0.1717	.	0
Time	1	-1290.13	3378.31	-0.38	0.7061	0.06	16.23
Rainfall	1	82.11	61.85	1.33	0.1974	0.75	1.33
Harvested Area	1	0.86	0.30	2.85	0.0091	0.63	1.58
Price/Ton	1	96.85	115.04	0.84	0.4086	0.06	15.40

The significance of regression coefficients are tested both at the 5% and 10% level of significance for comparison of strength of evidence. According to the results of the t test of the significance of each variable in affecting the mean response of maize production, only area harvested to maize is significant at 5% level of significant (Table 4.1). This shows that among the four predictor variables in the model, area harvested is the only variable which had a significant effect on maize production in the years, 1973/1974 to 2001/2002. The size of the effect of harvested area on maize production is given by its parameter estimate of 0.86. The estimate shows that an increase in area harvested to maize by 1 hectare resulted with an increase of the mean maize production by 0.86 of a tonne during the specified period, when other variables in the model are held constant. The additional variable in the subset of data, price of maize per tonne, is not significant and thus it does not add value in explaining and predicting maize production in the period, 1973/1974 to 2001/2002.

The value of tolerance of 0.06 that corresponds to each of the two predictor variables, time and price of maize is less than 0.10. In addition, the values of VIF, 16.23 and 15.40 that correspond to each of the variables are greater than 10 (Table 4.1). These are signs that there are collinearity problems in the subset of data that might be affecting the two predictor variables. The tolerance value of 0.06 for both time and price shows that only 6% of the variability in each of the variables is independent of other predictor variables in the model. This means that there is at least one collinear relationship in the data set since almost all (94%) of the variability in each of the variables is explained by other predictor variables in the model. The VIF values for time and price of maize, 16.23 and 15.40, show that the variance of the parameter estimates that correspond to these variables are inflated due to collinearity by 16.23 and 15.40, respectively.

The eigenvalues, condition indices, and variance-decomposition proportions for parameters estimates were used to identify the number of near linear dependencies that exist in the data set, and the variables that are involved in each of them (Table 4.2). The largest condition index of 24.08 that corresponds to the smallest eigenvalue of 0.01 indicates the presence of one near linear dependency among the columns of the design matrix $\mathbf{X}$. Variables that are involved in the identified near linear dependency are indicated by high variance-decomposition proportions of their regres-

Table 4.2: Collinear Diagnostics for 1973 - 2002 Maize Data

Number	Eigenvalue	Condition Index	Proportions of Variance			
			Time	Rainfall	Harvested Area	Price/Ton
1	4.62	1.00	0.00063	0.00095	0.00144	0.00082
2	0.33	3.72	0.01004	0.01256	0.00654	0.02261
3	0.03	13.53	0.00010	0.04508	0.86995	0.00763
4	0.01	18.05	0.07307	0.73878	0.03840	0.12941
5	0.01	24.08	0.91618	0.20263	0.08367	0.83953

sion coefficients that are greater or equal to 0.50 and correspond to one small eigenvalue and the highest condition index. High variance proportions of 0.91 and 0.84 correspond to the regression coefficients of time and price of maize, respectively. Thus in the period, 1973/1974 to 2001/2002, two variables, namely, time in years and price of maize in Rands per tonne were highly correlated. These results confirm a strong linear relationship between the two variables portrayed in their scatter plot in Figure 2.1 and its strength shown by high correlation coefficient of 0.97 in Table 2.3.

The strong collinear relationship that exists between time and price of maize has an effect on the standard error of the estimated regression coefficients and the value of the t statistic of price of maize. This is because when time is excluded in fitting the same model, the standard error of the estimate declines from 115.04 to 32.54 and the value of t statistic increases from 0.84 to 1.68, but price of maize remains not having a significant effect on maize production. Thus there is no need to remove the time from the model because though its exclusion when fitting the model increases the value of the t statistic for price of maize, it does not change the status of the variable in terms of it having a significant effect on maize production in the period, 1973/1974 to 2001/2002.

The case deletion regression diagnostics were applied to check if the third and twenty seventh observations, which were identified as suspected extreme cases are influential. The diagnostics that were used to measure the influence of the observations on different quantities of the fitted regression model are externally studentized residual t_i, leverage h_{ii}, DFFITS, Cook's Distance, DFBETAS and COVRATIO. The values of the diagnostics and their cut off points are given in Table 4.3. According to the results, the third observation is not considered as an outlying observation since its absolute value of t_i, 1.95 is less than the cutoff point of 2. However, it is considered as a high leverage point because its value of h_{ii}, 0.47 is high when it is compared with the cutoff point of 0.29. On the other hand, the twenty seventh observation is not a high leverage point, but it is an outlying observation since its value of t_i of 4.13 is two times bigger than the cutoff point. Thus the third observation of the year 1975/1976 is extreme in terms of one of the predictor variables, while the twenty seventh observation of the year 1999/2000 is extreme in terms of the response variable, production of maize.

Table 4.3: Case Deletion Diagnostic for 1973 - 2002 Maize Data

Diagnostic Measure	3rd Case Diagnostics	27th Case Diagnostics	Cutoff Point
RStudent t_i	-1.95	4.13	2
Leverage h_{ii}	0.47	0.16	$2p/n = 0.29$
DFFITS	-1.85	1.78	$2\sqrt{p/n} = 0.75$
Cook's Distance D_i	0.61	0.37	$F_{(0.5,p,n-p)} = 1.16$
COVRATIO	1.07	0.08	$1 \pm 3p/n = 1 \pm 0.43$
DFBETAS			$2/\sqrt{n} = 0.38$
Time	0.03	-0.19	0.38
Rainfall	-1.55	0.27	0.38
Harvested Area	1.11	0.41	0.38
Price/Ton	-0.05	0.50	0.38

The absolute values of DFFITS, 1.85 for the third case and 1.78 for the twenty seventh case, are greater than the cutoff point of 0.75. Thus the two observations have influence on the ith fitted values and fitting the model when each of them is excluded will change the third fitted value $\hat{y}_3$ and the twenty seventh fitted value $\hat{y}_{27}$ by 1.85 and 1.78 standard errors, respectively. The negative sign of the value of DFFITS for the third observation indicates that $\hat{y}_3$ from fitting the model with all the observations is less than $\hat{y}_3$ from fitting the model without the third observation. The positive sign of the value of DFFITS for the twenty seventh observation indicates that $\hat{y}_{27}$ from fitting the model with all the observations is larger than $\hat{y}_{27}$ from fitting the model without the twenty seventh observation.

The value of COVRATIO, 1.07, for the third observation is neither greater than the cutoff point of $1 + 0.43 = 1.43$ nor less than the cutoff point of $1 - 0.43 = 0.57$. This suggests that the exclusion of the third case from the data does not influence the covariance matrix of the estimated coefficients, and thus the observation does not affect the precision of the estimates. The twenty seventh observation has an effect on the precision of the estimates since its COVRATIO of 0.08 is less than the cutoff point of $1 - 0.43$. According to Rawlings et al. (1998) the observation decreases the precision since its COVRATIO is less than 1.

In the case of the influence of the two cases on the jth parameter estimates $\hat{\beta}_j$, the deletion of the third observation influences the estimated coefficients of the amount of rainfall and harvested area. This is so because the respective absolute values of DFBETAS for the two variables of 1.55 and 1.11 are greater than the cutoff point of 0.38. On the other hand, the deletion of the twenty seventh observation affects the estimated coefficient of price of maize because its absolute value of DFBETAS, 0.50, is bigger than the cutoff point. In the case of harvested area the difference between the absolute value of DFBETAS, 0.41, and the cutoff point is minimal. The results from

the case deletion diagnostics suggest that the third and twenty seventh observations are influential since fitting the model after their deletion from the data has an effect on a number of quantities of the fitted model.

Table 4.4: Influence of the 3rd Case on the Fitted Model for 1973 - 2002 Maize Data

Predictor Variable	With the 3rd case $R^2 = 0.54$				Without the 3rd Case $R^2 = 0.58$			
	Estimate	Std. error	t value	Pr $> \|t\|$	Estimate	Std. error	t value	Pr $> \|t\|$
Intercept	-62918.00	44599.00	-1.41	0.1717	-87860.00	44018.00	-2.00	0.0585
Time	-1290.13	3378.31	-0.38	0.7061	-1394.38	3190.53	-0.44	0.6663
Rainfall	82.11	61.85	1.33	0.1974	172.89	74.72	2.31	0.0304
Harvested Area	0.86	0.30	2.85	0.0091	0.54	0.33	1.66	0.1111
Price/Ton	96.85	115.04	0.84	0.4086	102.20	108.67	0.94	0.3572

Tables 4.4 and 4.5 present the comparison of the results from fitting the regression model when all observations in the data set were used and when each of the third and twenty seventh observations were excluded, one at the time. The results helped to assess the influence of the observations on the estimated parameters and their summary statistics. The exclusion of the observations caused a slight increase in the coefficient of determination from 0.54 to 0.58 and 0.56 for the third and twenty seventh observations, respectively. This is an indication that the exclusion of each observation do not have a considerable impact on the goodness-of-fit of the model. However, when the observations were deleted one at a time the size of all parameter estimates changed. Substantial changes were observed from 82.11 to 172.89 in the parameter estimate of the amount of rainfall and from 96.85 to 102.20 in the parameter estimate of price of maize when the third observations is excluded (Table 4.4). Similarly, when the twenty seventh observations was excluded substantial changes were observed for parameter estimates corresponding to the same predictor variables, though in this case the estimates declined (Table 4.5). This is in agreement with the results from DFBETAS in Table 4.3 that the estimated coefficients of amount of rainfall and price of maize are affected by the exclusion of the third and twenty seventh observations from the data, respectively.

Table 4.5: Influence of the 27th Case on the Fitted Model for 1973 - 2002 Maize Data

Predictor Variable	With the 27th case $R^2 = 0.54$				Without the 27th Case $R^2 = 0.56$			
	Estimate	Std. error	t value	Pr $> \|t\|$	Estimate	Std. error	t value	Pr $> \|t\|$
Intercept	-62918.00	44599.00	-1.41	0.1717	-38145.00	34714.00	-1.10	0.2838
Time	-1290.13	3378.31	-0.38	0.7061	-798.18	2593.16	-0.31	0.7611
Rainfall	82.11	61.85	1.33	0.1974	69.25	47.53	1.46	0.1592
Harvested Area	0.86	0.30	2.85	0.0091	0.76	0.23	3.29	0.0034
Price/Ton	96.85	115.04	0.84	0.4086	52.49	88.86	0.59	0.5607

The values of the standard errors of estimated coefficients, t statistics and p-values are affected by the deletion of each of the third and twenty seventh observations (Tables 4.4 and 4.5). When the model was fitted using the complete data set, area harvested to maize is the only variable that had a significant effect on maize production in the period, 1973/1974 to 2001/2002. However, when the third observation was deleted harvested area is not significant any more, instead the amount of rainfall appears as the only variable which had a significant effect on maize production. On the contrary, when the model was fitted without the twenty seventh observation, the significance of the variables in affecting maize production does not change since area harvested remains the only variable with a significant effect on maize production.

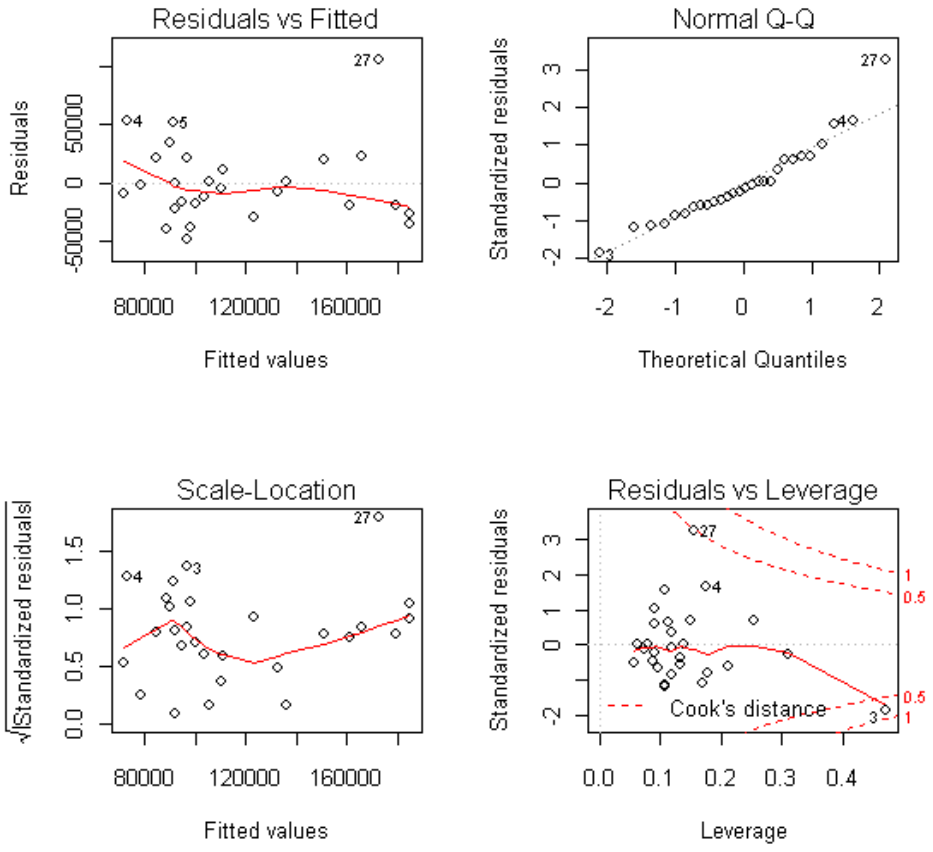

Figure 4.1: Plot of Residuals for 1973 - 2002 Maize Data

Residual plots in Figure 4.1 were used to check the assumptions of constant variance and normality of error terms for the fitted model. The first plot on the top left is a plot of residuals against fitted values. It shows a pattern of residuals that decrease as the fitted values get larger. This is a sign of violation of constant variance assumption or heteroscedasticity. The next plot on the top right is

a normal QQ plot that plots standardized residuals against theoretical normal quantiles. The plot shows that plotted points lie on a straight line with one point on the right end being above the line, another point on the left being above the line, and one outstanding case. This is an indication of a symmetric distribution with tails that deviate slightly from that of normal distribution. The plot also shows evidence of the presence of outliers in the set of data. The bottom left plot is a plot of square root of standardized residuals against fitted values. This plot is similar to the first plot but with positive values only. It also shows a pattern of residuals decreasing with bigger values of fitted values.

The last plot on the bottom right is a plot of standardized residuals against leverage that shows the upper and lower Cook's distance contour. The plot highlights outliers and high leverage points that are influential on the quantities of the fitted model. The twenty seventh observation that stands out in the first three plots with the highest residual is considered influential since it appears in the upper Cook's distance contour. The third observation is considered as a high leverage point since it appears in the lower Cook's distance contour, with the highest leverage. The observations from the plot confirms the findings from the case deletion diagnostics that the third observation is a high leverage point, observation twenty seven is an outlier, and that both observations are influential.

The two observations could not be removed from the data because they are legitimate cases in the data, providing an important observation. The third observation is the case of the year 1975/1976 with the highest amount of rainfall of 1042.52 milliliters and relatively low maize production of 49128 tonnes. Probably the country experienced heavy rains in 1975/1976 which resulted with flooding that adversely affected production of maize. Maize yields are highly sensitive to rainfall deficiency since they are adversely affected by dry spells with little rain as well as flooding caused by heavy rains, depending on the stage of maize crop growth (Singh et al., 1985). The authors further noted that young crop of maize is more prone to damaging effects of flooding, which lead to decreased maize yields. The twenty seventh observation is the case of 1999/2000 showing that maize production was high in that year.

In the case of the 1973/1974 to 2006/2007 and 1976/19777 to 2006/2007 maize data, maize production was regressed on two predictor variables only. The two variables are the amount of rainfall and area harvested, which appear to have a linear relationship with maize production from the scatter plot and correlation matrices in Figures A.1 and A.2, and Tables B.1 and B.2 of Appendices A and B, respectively. The model for the 1973/1974 to 2006/2007 data was fitted with and without each of the third and twenty seventh observations. This was done to assess if the influence of each observation on the fitted model is similar in the two periods, 1973/1974 to 2006/2007 and 1973/1974 to 2001/2002. In the case of the 1976/1977 to 2006/2007 data, the model was fitted using all the observations only, since none of the observations was identified as a suspected extreme observation in this period.

The results from fitting the model to the 1973/1974 to 2006/2007 maize data, with and without the third and twenty seventh observations, are almost similar to the results from fitting the model to the 1973/1974 to 2001/2002 maize data for the two scenarios. They are similar in the sense that like in Table 4.1, only one variable, area harvested, had a significant effect on maize production (Table C.1 in Appendix C). On the Contrary, for the 1976/19777 to 2006/2007 maize data, the amount of rainfall instead of area harvested is the only variable with a significant effect on maize production (Table C.3 in Appendix C). This is an indication that different time intervals at which maize data were compiled make a difference in terms of which variables had significant effects on maize production. However, the additional variable in the 1976/19777 to 2006/2007 subset of data, population size, does not add value in explaining and predicting maize production since it is not linearly related with maize production (Table B.2 in Appendix B) and thus it is not included in the model.

Collinear diagnostics for both the 1973/1974 to 2006/2007 and 1976/19777 to 2006/2007 maize data do not show the presence of any near linear dependency among the two predictor variables in the model (Tables C.1, C.2, C.3, C.4 in Appendix C). The tolerance of 0.84 and 0.64 for the amount of rainfall and harvested area from fitting the model to the two subsets of data show that 84% and 64% of variability in one of the two variables, say area harvested to maize, is not dependent on the amount of rainfall. In addition, the VIF of 1.19 and 1.55 for each of the two variables in each subset of data are less than 10, indicating that the variances of their respective estimated coefficients are not inflated due to collinearity. Thus the values of tolerance and VIF show that there are no collinearity problems in the two subsets of maize data.

The results of condition indices and variance-decomposition proportions are in line with what is shown by the values of tolerance and VIF that there is no near linear dependency in the data. For example, the highest condition index of 14.11 that corresponds to the small eigenvalue of 0.02 does not indicate the presence of collinearity in the 1973/1974 to 2006/2007 data (Table C.2). This is because to have a collinear relationship in a data set, two or more variables should have a variance proportion of 0.50 or more, which correspond to the small eigenvalue and highest condition index (Belsley et al., 1980; Belsley, 1991). This applies to both subsets of data because there is only one regression coefficient with high variance proportion that corresponds to the highest condition index (Tables C.2, C.4).

The results of case deletion diagnostics for the 1973/1974 to 2006/2007 data in Table D.1 of Appendix D are similar to that of the 1973/1974 to 2001/2002 data in Table 4.3. The comparison of the results from fitting the model with and without the third observation in Table D.2 and twenty seventh observation in Table D.3 of Appendix D, yields almost similar results as that of 1973/1974 to 2001/2002 data in Tables 4.4 and 4.5. Thus despite the time interval in which the maize data

were compiled, the influence of the third and twenty seventh observations on different quantities of the fitted model remains the same regarding the parameter estimates and their summary statistics.

The top and bottom left plots in Figure E.1 for the 1973/1974 to 2006/2007 maize data show that residuals are randomly scattered without showing any specific pattern. This is indicative of homoscedasticity where the assumption of constant variance holds. The normal QQ plot on the top right shows that plotted points lie on a straight line with points on the right and left ends being above the line, and one outstanding point. This is an indication of a symmetric distribution but with a rather longer and heavier tail than that of the normal distribution. Furthermore, the plot shows evidence of the presence of an outlier in the set of data. The last plot on the bottom right shows that, unlike in the 1973/1974 to 2001/2002 data, the twenty seventh observation, which stands out in the first three plots, is an extreme but not influential outlier in the 1973/1974 to 2006/2007 data. This is because the point is close to the upper Cook's distance contour but not inside it. On the other hand, the plot shows that, as it is the case in the 1976/1977 to 2001/2002 maize data, the third observation appears in the lower Cook's distance contour and thus it is an influential high leverage point even in this subset of data.

The top and bottom left plots in Figure E.2 for the 1976/1977 to 2006/2007 data show residuals that increase with the fitted values. This suggests that the assumption of constant variance does not hold. The normal QQ plot on the top right shows that plotted points lie on a straight line with an exception of one point on the right end being above the line and one outstanding case. This suggests a symmetric distribution but with tails that deviate slightly from that of the normal distribution. Further, the plot shows evidence of the presence of an outlier in the set of data. The last plot on the bottom right shows that the twenty fourth observation is an extreme but not influential outlier, since it appears close to the upper Cook's distance contour but not inside it. This is the 1998/1999 observation that stands out in residuals plots for all the three subsets of maize data.

4.3 Regression Model of National Sorghum Data

Linear regression models were fitted for all the three subsets of the national sorghum data. Similarly, the models were fitted with and without the identified suspected extreme observations, for subsets of data with such observations. The linear regression model with first-order autoregressive process in which sorghum production in the current year was regressed on time, production of sorghum in the past immediate year, population size and area harvested to sorghum was fitted to the 1976/1977 to 2006/2007 data. This model was applied because one of the predictor variables, production of sorghum in the past immediate year, consists of past values of the response variable. The suspected extreme observation is the eleventh case of the year 1986/1987 with high area har-

vested to sorghum of 70909 hectares and a relatively low sorghum production of 31232 tonnes. The observation deviates from the pattern observed from this subset of data that sorghum production increased with area harvested.

According to the results, the F test is significant at 5%, indicating that at least one of the predictor variables in the model had a significant linear relationship with sorghum production in the period 1976/1977 to 2006/2007, when the eleventh case was included and when it was excluded. The respective values of the coefficient of determination R^2 from fitting the model with and without the eleventh case, show that the regression model explains 67% and 78% of the variability in sorghum production.

Table 4.6: Parameter Estimates for the 1976 - 2007 Sorghum Data

Response Variable: Sorghum production

Variable	DF	Estimate	Std. error	t value	Pr > |t|	Tolerance	VIF
Intercept	1	-95696.00	105604.00	-0.91	0.3735	.	0
Time	1	-3877.89	3875.18	-1.00	0.3266	0.004	225.40
Production-t	1	0.28	0.17	1.71	0.0988	0.480	2.08
Population	1	0.09	0.09	0.93	0.3607	0.005	217.84
Harvested Area	1	0.65	0.16	3.98	0.0005	0.617	1.62

Table 4.6 presents results from fitting the model using all observations in the data subset, including the eleventh observation. Two variables, production in the immediate past year and area harvested to sorghum are significant at 10% and 5% levels of significance, respectively. The smaller value of the p-value [0.0005] for harvested area shows that though both variables were significant in affecting sorghum production, the evidence for harvested area is stronger than that of sorghum production in the immediate past year. The estimate of the autoregressive coefficient for sorghum production in the immediate past year of 0.28 indicates a weak positive serial correlation between sorghum production at time t and sorghum production at time $t-1$. The serial correlation shows that sorghum production in the immediate pas year $t-1$ predicted current sorghum production at time t in a positive direction. This means that if sorghum production at time $t-1$ was high, sorghum production at time t was also high, when other variables are held contant.

The estimated coefficient of area harvested shows that, in the years 1976/1977 to 2006/2007, an increase in area harvested to sorghum by 1 hectare increased the mean sorghum production by 0.65 of a tonne when the rest of the variables in the model are held constant. Despite the strong correlation between population size and sorghum production shown by the correlation coefficient of -0.66 in Table 2.5, population size is not significant in affecting sorghum production. This is probably because the perfect correlation between time and population size with the correlation coefficient of 0.99 conceals the significant effect that population size could have had on sorghum

production. The presence of collinear relationships among predictor variables can lead to regression coefficients with small values of t statistics for predictor variables that are anticipated to be having an effect (Draper and Smith, 1998; Kutner et al., 2005). It is worth noting that the sign of the regression coefficient for population size in Table 4.6 has a positive sign while the correlation coefficient between population size and sorghum production in Table 2.4 has a negative sign. This could be another effect of the perfect correlation observed between time and population size, since when there is collinearity problem, regression coefficients can have wrong signs that are opposite to the anticipated (Kutner et al., 2005).

Table 4.7: Collinear Diagnostics for 1976 - 2007 Sorghum Data

Number	Eigenvalue	Condition Index	Proportions of Variance			
			Time	Production-t	Population	Harvested Area
1	4.37	1.00	0.00004	0.00484	0.000009	0.00394
2	0.48	3.02	0.00097	0.09868	0.000025	0.01848
3	0.14	5.56	0.00023	0.31718	0.000005	0.36366
4	0.02	16.98	0.01692	0.53144	0.000721	0.60926
5	0.00	196.07	0.98182	0.04786	0.999240	0.00467

The values of tolerance of 0.004 for time and 0.005 for population size indicate that almost all the variability in each of the two predictor variables is explained by other predictor variables in the model (Table 4.6). This is a sign of the presence of collinearity problem in the 1976/1977 to 2006/2007 sorghum data. The values of VIF for time and population size show that the variance of the regression coefficient for both variables is inflated due to collinearity by 225.40 and 217.84, respectively.

Table 4.7 shows the presence of two near linear dependencies among the variables in the data. The dependencies are indicated by two high condition indices of 196.07 and 16.98, which are associated with the small eigenvalues of 0.00 and 0.02, respectively. The highest conditional index of 196.07 is indicative of serious collinear problems that involve time and population size. The estimated coefficients of the two variables have high variance decomposition proportions of 0.98 for time and 0.99 for population size. Both proportions correspond to the highest condition index of 196.07 and small eigenvalue of 0.00. The condition index of 196.07 and high variance proportions confirm the perfect correlation between time and population size observed earlier from their scatter plot and correlation coefficient of 0.99 (Figure 2.3 and Table 2.5). The variance of regression coefficients for the two variables are adversely affected by collinearity. The second near dependency that involve two variables, sorghum production in the past and area harvested to sorghum is shown by the second highest condition index of 16.98. The involvement of the two variables in the collinear relationship is indicated by high variance-decomposition proportions of their coefficients of 0.53 for

sorghum production in the past immediate year and 0.61 for harvested area, which correspond to the condition index of 16.98 and the eigenvalue of 0.02.

Table 4.8: Case Deletion Diagnostic for for 1976 - 2007 Sorghum Data

Diagnostic Measure	11th Case Diagnostics	Cutoff Point
RStudent t_i	-3.40	2
leverage h_{ii}	0.21	$2p/n = 0.27$
DFFITS	-1.74	$2\sqrt{p/n} = 0.73$
Cook's Distance D_i	0.42	$F_{(0.5,p,n-p)} = 1.16$
COVRATIO	0.22	$1 \pm 3p/n = 1 \pm 0.40$
DFBETAS		$2/\sqrt{n} = 0.36$
Time	0.38	0.36
Production-t	-0.83	0.36
Population	-0.48	0.36
Harvested Area	-1.35	0.36

The results for the case deletion diagnostics of the eleventh observation are presented in Table 4.8. The absolute value of the externally studentized residual of 3.40 is greater than the cut off point of 2, whereas the value of the leverage h_{ii} of 0.21 is less than the cut off point of 0.27. Thus the eleventh case is considered as an outlier, not a high leverage point, because it is extreme in terms of sorghum production, the response variable. The absolute value of DFFITS, 1.74, is higher than the cut off point of 0.73 and thus the eleventh predicted value is influenced by the deletion of the eleventh case. The exclusion of the eleventh case also has an influence on the covariance matrix of the estimates of $\boldsymbol{\beta}$ since the COVRATIO value of 0.22 is less than the cut off point of $1 - 0.40 = 0.60$. This means that the eleventh observation decreases precision of estimated coefficients. All parameters estimated from the model are influenced by the deletion of the case since their absolute values of DFBETAS are greater than the cut off point of 0.36. However, the outlying case seem to be influencing the parameter estimate for harvested area, with DFBETAS of -1.35, more than any other parameter estimate.

Table 4.9: Influence of the 11th Case on the Fitted Model for 1976 - 2007 Sorghum Data

Predictor Variable	With the 11th case $R^2 = 0.67$				Without the 11th Case $R^2 = 0.78$			
	Estimate	Std. error	t value	Pr $> \|t\|$	Estimate	Std. error	t value	Pr $> \|t\|$
Intercept	-95696.00	105604.00	-0.91	0.3735	-154045.00	90157.00	-1.71	0.1004
Time	-3877.89	3875.18	-1.00	0.3266	-5109.86	3268.11	-1.56	0.1310
Production-t	0.28	0.17	1.71	0.0988	0.40	0.14	2.80	0.0100
Population	0.09	0.09	0.93	0.3607	0.12	0.08	1.58	0.1275
Harvested Area	0.65	0.16	3.98	0.0005	0.83	0.15	5.67	<.0001

The comparison of the results from fitting the model when the eleventh case was included and when it was excluded is given in Table 4.9. The coefficient of determination R^2 increased from 0.67 of when the model was fitted with the observation to 0.78 of when the model was fitted without the observation. Thus the eleventh observation has an influence on the goodness-of-fit of the model since the predictor variables in the model explain a larger proportion of the variability in production of sorghum when the case is included than when it is included. The deletion of the case causes an increase in magnitude of all parameter estimates and their t statistics, and a decrease in standard errors of estimates and p-values. When the case is deleted, the evidence that sorghum production in the current time t and sorghum production in the past immediate time $t-1$ have serial dependence becomes stronger. This is shown by the increase in the autoregressive coefficient from 0.28 to 0.40, and the p value that declines from 0.0988 to 0.0100. The results show that the eleventh observation is an influential case that improves the efficiency of the model in predicting sorghum production of the period 1976/1977 to 2006/2007.

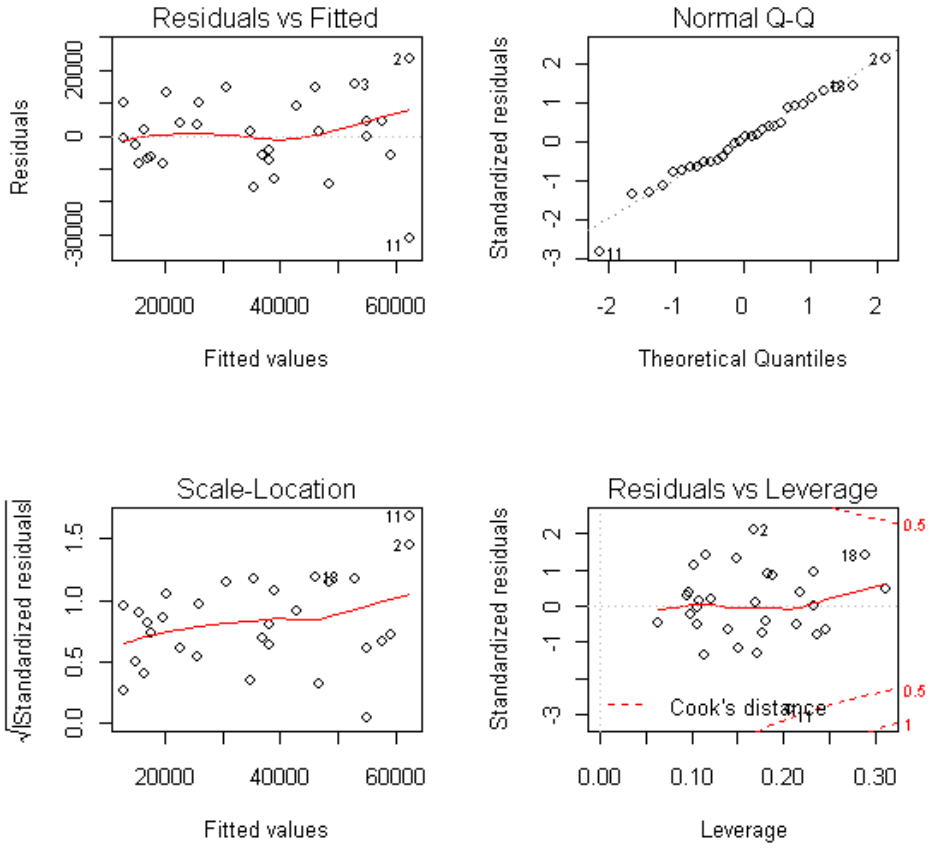

Figure 4.2: Plots of Residuals for 1976 - 2007 Sorghum Data

The plots on the top and bottom left show a random pattern of residuals, indicating that the assumption of constant variance holds in these data (Figure 4.2). The top right plot shows that plotted points almost lie on a straight line with an exception of the eleventh point on the left end, which is far below the line. This is an indication of an approximation of a symmetric distribution but with evidence of the presence of an extreme observation. The last plot on the bottom right shows that the eleventh case is not influential since it does not appear in any of the Cook's distance contours.

Like in the previous subset of sorghum data, the regression model with first-order autoregressive process was fitted for the 1973/1974 to 2006/2007 subset. In this case sorghum production at time t was regressed on three predictor variables used in the model fitted to the previous subset of data. The variables are time, sorghum production in the past immediate year, and area harvested to sorghum (Tables C.5 and C.6 in Appendix C). The influence of the 1986/1987 observation identified as the fourteenth observation in this subset of data and as the eleventh case in the previous subset was studied by fitting the model when it was included and when it was excluded (Table D.5 in Appendix D). The interest here is to find out if the influence of the observation is the same in the two different time periods.

The results of the F test for both cases of fitting the model with and without the suspected extreme case indicate that at least one of the variables in the model had a significant linear relationship with sorghum production during the period under investigation. The regression model fitted when the extreme case is included, explains 66% of the change in sorghum production. The pattern of significant variables in this subset of data is similar to that in the 1976/1977 to 2006/2007 subset (Table C.5). Sorghum production in the past immediate year and area harvested to sorghum are both significant but with varying strength of evidence of significance, [p-value=0.0521] for Sorghum production in the past immediate year and [p-value<0.0001] for harvested area. The interpretation of the estimated coefficients for the two variables in this subsets is similar to that of the previous subset because their magnitudes are almost equal.

All values of tolerance are greater than 0.10 and all values of VIF are less than 10 (Table C.5 in Appendix C). This is an indication that there are no collinear relationships among the predictor variables in the model. On the contrary, the highest condition index of 15.56 associated with the smallest eigenvalue of 0.01 shows that there is one near linear dependency among the predictor variables (Table C.6 in Appendix C). The identified near dependency involves all the three variables in the model. The involvement of the variables is indicated by high variance-decomposition proportions of their regression coefficients, which are associated with the highest condition index of 15.56. The proportions are 0.84, 0.50 and 0.67 for time, sorghum production at time $t-1$ and harvested area, respectively. These results establish that condition index and variance-decomposition proportion are superior to both tolerance and VIF, as collinear diagnostics, since they are able to

detect the presence of a collinear relationship where tolerance and VIF failed to do so. In addition, condition index and variance-decomposition proportion manage to point to more than two variables that are involved in a collinear relationship, which is not the case with tolerance and VIF.

Case deletion diagnostics in Table D.4 of Appendix D show that the pattern of influence of the 1986/1987 observation on different quantities of the fitted model in the period 1973/1974 to 2006/2007 is the same as that in the period 1976/1977 to 2006/2007. The comparison of results from fitting the regression model with and without the observation in Table D.5 of Appendix D show that, like in 1976/1977-2006/2007 data, the deletion of the extreme case improves the efficiency of the model in predicting sorghum production. The results show that the observation is considered influential in both periods. These suggest that the period under which the data were observed and position of the extreme observation in the data set do not make a difference in terms of its influence on different quantities of the fitted model.

In the case of the 1973/1974 to 1997/1998 sorghum data, sorghum production was regressed on time, harvested area, area affected by crop failure and price of sorghum per tonne. The model was fitted when the twenty fourth observation identified as a suspected extreme case was included and when it was deleted. Results are given in Tables C.7 and C.8 in Appendix C. According to the results of the F test, at least one of the predictor variables had a significant linear relationship with sorghum production in both cases. The regression model explains 68% of the variability in sorghum production when all observations are used in fitting the model. The p-values show that there is evidence that area harvested to sorghum and area under crop failure had a significant effect on sorghum production. The parameter estimates of the two variables show that an increase of harvested area by 1 hectare increased the mean sorghum production by 0.50 of a tonne, while the increase of area under crop failure by 1 hectare reduced the mean production by 1.77 tonnes, when the rest of the variables in the model are held constant (Table C.7).

The values of tolerance are greater than 0.10 and the values of VIF are smaller than 10, indicating the absence of collinear relationships among the predictor variables in the model (Table C.7). This is inline with the highest condition index of 13.67 that corresponds to the smallest eigenvalue of 0.02, which is associated with only one high variance-decomposition proportion of the regression coefficient of 0.61 for harvested area (Table C.8). When there is only one high variance-decomposition proportion associated with the highest condition index, there is no collinearity because the variance of two regression coefficients should be affected to indicate the presence of collinearity in the data set (Belsley, 1991).

It is interesting to note a peculiar finding from the results of this subset of data that the highest condition index of 13.67 failed to identify the presence of collinear relationships, while the second highest condition index of 12.33 corresponding to the second smallest eigenvalue of 0.03

shows the presence of one such relationship. This condition index is associated with two high variance-decomposition proportions of the regression coefficients for time and price of sorghum, and hence the existence of a near linear dependency. This finding deviates from the normal pattern of condition indices and variance proportions that show the presence of collinear relationships in the set of data. Normally the highest condition index should be associated with high variance proportions of coefficients of two or more variables to indicate the presence of collinear relationships. The identified linear dependency confirms the strong linear relationship of the two variables observed from Figure A.3 and Table B.3 in Appendices A and B. The peculiar pattern shown by condition index and variance proportion is an empirical finding that needs to be investigated further in the future research to establish the statistical theory behind it.

The absolute values of t_i, 1.24, and h_{ii}, 0.65, suggest that the twenty fourth observation identified as a suspected extreme case is not an outlier but it is a high leverage point (Table D.6). This observation is extreme with respect to one of the predictor variable. The absolute values of DFFITS, COVRATIO and DFBETAS for time and sorghum price are greater than their respective cutoff points. This suggests that the deletion of the observation has an influence on the twenty fourth predicted value, precision of the regression coefficients, and two estimated coefficients of time and price of sorghum, and thus it is considered influential. The results from fitting the model when the observation was included are almost similar to the results from fitting the model without the observation (Table D.7). Though the deletion of the twenty fourth observation affect some quantities of the fitted model, it does not affect the overall efficiency of the model in predicting sorghum production for the years 1973/1974 to 1997/1998 since the change in R^2 is not that much.

The top and bottom left plots in Figure E.3 for the 1973/1974 to 2006/2007 sorghum data suggest that the assumption of constant variance is violated since the residuals are increasing with the values of the predicted values. The normal QQ plot on the top right shows that plotted points are almost on a straight line with points on the right end being slightly below the line. This is an indication of a symmetric distribution but with a rather heavier tail than that of the normal distribution. The last plot on the bottom right shows that the fourteenth observation, which stands out in the first three plots, is an extreme but not influential. In the case of the 1973/1974 to 1998/1999 sorghum data, the top left plot in Figure E.4 shows residuals that are randomly scattered. The random pattern shows that the the assumption of constant variance of error terms holds. The normal QQ plot shows a symmetric distribution but with a rather heavier tails than that of the normal distribution. The bottom right plot shows that though the twenty fourth observation does not appear outstanding in any of the first three plots, it is an influential high leverage point because it appears inside the lower Cook's distance contour.

4.4 Regression Model of National Wheat Data

Linear regression models were fitted for all the three subsets of the national wheat data, as it was the case with the national maize and sorghum data. The models were fitted with and without the observations identified as suspected extreme observations. The regression model with a first-order autoregressive process where wheat production at time t was regressed on four predictor variables was fitted for the three subsets of 1973/1974 to 2001/2002, 1973/1974 to 2006/2007 and 1976/1977 to 2006/2007 data. Three of the predictor variables namely, time in years, production of wheat at time $t-1$, and area harvested appear in the models fitted to each of three subsets of data. The fourth variable varies from one subset to another, where it is price of wheat in the 1973/1974 to 2001/2002 data, the amount of rainfall in the 1973/1974 to 2006/2007 data, and population size in the 1976/1977 to 2006/2007 data.

Table 4.10: Parameter Estimates for 1973 - 2002 Wheat Data Using All Variables

Response Variable: wheat production

Variable	DF	Estimate	Std. error	t value	Pr > |t|	Tolerance	VIF
Intercept	1	-2862.02	11583.00	-0.25	0.8069	.	0
Time	1	129.86	798.64	0.16	0.8722	0.09	10.51
Production-t	1	0.46	0.17	2.67	0.0132	0.64	1.55
Harvested Area	1	0.53	0.23	2.34	0.0278	0.41	2.45
Price/Ton	1	-0.55	16.51	-0.03	0.9737	0.12	8.31

The 1998/1999 observation that occupies the $26th$ and $23rd$ positions in the three subsets of wheat data was identified as a suspected extreme case in all the subsets. When the model was fitted to the data including the suspected extreme case, the results of the F test show that at least one of the predictor variable in the model have a significant linear relationship with wheat production, in all the subsets. The results from fitting the model to each of the subsets are almost similar (Table 4.10, and Tables C.9 and C.11 of Appendix C). They are the same in the sense that only two variables, wheat production at time $t-1$ and harvested area, appear to have had a linear relationship with wheat production at time t, in all the subsets of data. This shows that variables that are additional to each of the two subsets of wheat data, which are price of wheat in the 1973/1974 to 2001/2002 data, and population size in the 1976/1977 to 2006/2007 data, do not add any value in explaining the mean response of production of wheat, because they are not linearly related to production wheat.

The size of the parameter estimates of the wheat production at time $t-1$ and harvested area, differs slightly in all subsets of data and thus the interpretation of the estimates is similar. The estimate of the autoregressive coefficient for wheat production in the immediate past year is positive and is in the interval [0.45, 0.55] for all subsets. This is an indication that the serial correlation

between production of wheat at time t and production of wheat at time $t-1$ does not differ much in the three periods, 1973/1974 to 2001/2002, 1973/1974 to 2006/2007, and 1976/1977 to 2006/2007. The positive sign of the autoregressive coefficient shows that the low or high production of wheat in the past immediate year was followed by a similarly low or high production of wheat in the following year, for all the periods under investigation, when other variables are held constant. The estimated coefficient for harvested area is also positive and is within the interval [0.51, 0.80], which shows that an increase in area harvested to wheat increased wheat production by at least a half of a kilogram for all the three periods, when other variables are held constant.

Table 4.11: Collinear Diagnostics for 1973 - 2002 Wheat Data

Number	Eigenvalue	Condition Index	Proportions of Variance			
			Time	Production-t	Harvested Area	Price/Ton
1	4.14	1.00	0.00109	0.00775	0.00342	0.00175
2	0.69	2.46	0.00785	0.05406	0.02682	0.01625
3	0.12	5.87	0.00048	0.81702	0.15097	0.01279
4	0.04	10.66	0.03259	0.06046	0.56410	0.27994
5	0.01	18.63	0.95800	0.06071	0.25469	0.68927

The tolerance of 0.09 and the VIF of 10.51 associated with the variable, time, is a sign of the presence of near linear dependencies in the 1973/1974 to 2001/2002 wheat data (Table 4.10). The highest condition index of 18.63 associated with the eigenvalue of 0.01 reveals that there is one near linear dependency in the data (Table 4.11). The identified collinearity involves time and price of wheat since their regression coefficients have high variance-decomposition proportions of 0.96 for time and 0.69 for price, which are associated with the highest condition index of 18.63. This finding confirms the strong linear relationship of the two variables observed from their scatter plot in Figure 2.5 and high correlation coefficient of 0.93 in Table 2.7.

The values of tolerance and VIF calculated from the 1973/1974 to 2006/2007 wheat data are greater than 0.10 and less than 10, respectively (Table C.9), indicating that there are no near linear dependencies in this subset of wheat data. On the other hand, the condition index and variance-decomposition proportions of coefficients give contradicting results that show the presence of one collinear relationship (Table C.10). The identified collinear relationship involves time and area harvested because they have high variance-decomposition proportions of 0.82 and 0.71, respectively. These proportions are associated with the second highest condition index of 12.55. It is worth noting the peculiar finding similar to that of the 1973/1974 to 1997/1998 sorghum data that high variance-decomposition proportions of coefficients of two variables are associated with the second highest condition index of 12.55, instead of the highest condition index of 18.13.

In the case of 1976/1977 to 2006/2007, the tolerance of 0.01 for each of time and population size indicates that almost all the variability in the two variables is explained by other predictor variables in the model (Table C.11 in Appendix C). According to the values of VIF, the variance of estimated regression coefficients for time and population size are inflated by 214.41 and 209.28, respectively, due to collinearity. In addition, the highest condition index of 192.58 shows that there is one collinear relationship in the 1976/1977 to 2006/2007 sorghum data (Table C.12 in Appendix C). This relationship involves time and population size since the variance-decomposition proportions of their respective regression coefficients are high and correspond to the highest condition index. The proportions of the variances that are affected by collinearity are 0.99 for time and 0.99 for population size. A very large value of condition index is the sign of severe collinearity between the two variables. The severe collinearity could have concealed the effect of population size on wheat production. This is because though the correlation coefficient between the two variables in Table B.6 of Appendix B, show a negative linear relationship, population size appear to have not been significant in affecting wheat production in the years, 1976/1977 to 2006/2007.

Table 4.12: Case Deletion Diagnostic for 1973 - 2002 Wheat Data

Diagnostic Measure	26th Case Diagnostics	Cutoff Point
RStudent t_i	4.32	2
leverage h_{ii}	0.13	$2p/n = 0.28$
DFFITS	1.69	$2\sqrt{p/n} = 0.74$
Cook's Distance D_i	0.33	$F_{(0.5,p,n-p)} = 1.16$
COVRATIO	0.07	$1 \pm 3p/n = 1 \pm 0.41$
DFBETAS		$2/\sqrt{n} = 0.37$
Time	-0.23	0.37
Production-t	-0.58	0.37
Harvested Area	0.39	0.37
Price/Ton	0.75	0.37

The values of externally studentized residual in the interval [4.32, 4.92], for the three subsets of wheat data, are larger than the cutoff point of 2, while the values of leverage in the interval [0.07, 0.13] are smaller than the cutoff points in the interval [0.24, 0.28] (Table 4.12, and Tables D.8 and D.10 in Appendix D). These values suggest that the 1998/1999 observation is an outlier in all the subsets, but not a high leverage point. Thus the observation is extreme in terms of the response variable, wheat production, but it is not extreme in terms of any of the predictor variables in the model. Since the DFFITS value that corresponds to the observation is greater than the given cutoff point in all the subsets, the observation has an influence on the ith predicted value that correspond to it in each subset of data. The absolute values of DFBETAS show that the observation has influence on parameter estimates of production of wheat in the immediate past year, area harvested,

and price of wheat, in the 1973/1974 to 2001/2002 wheat data (Table 4.12). In the case of the 1973/1974 to 2006/2007 data, the observation has an influence on parameter estimates of time, the amount of rainfall and area harvested (Table D.8). For the 1976/1977 to 2006/2007 data, the observation has an influence on parameter estimates of time and population size (Table D.10).

Table 4.13: Influence of the 26th Case on the Fitted Model for 1973 - 2002 Wheat Data

Predictor Variable	Full Data Set $R^2 = 0.59$				26th Case Deleted $R^2 = 0.75$			
	Estimate	Std. error	t value	Pr $> \|t\|$	Estimate	Std. error	t value	Pr $> \|t\|$
Intercept	-2862.02	11583.00	-0.25	0.8069.00	-1232.21	8791.85	-0.14	0.8898
Time	129.86	798.64	0.16	0.8722	271.40	606.53	0.45	0.6587
Production-t	0.46	0.17	2.67	0.0132	0.54	0.13	4.07	0.0005
Harvested Area	0.53	0.23	2.34	0.0278	0.47	0.17	2.69	0.0132
Price/Ton	-0.55	16.51	-0.03	0.9737	-10.00	12.71	-0.79	0.4394

The comparison of the results from fitting the model with and without the outlying observation is presented in Table 4.13, and Tables D.9 and D.11 in Appendix D. When the case is deleted, the proportion of the variability in production of wheat attributed to the predictor variables in the model changes from 59% to 75% in the 1973/1974 to 2001/2002 data, from 64% to 78% in the 1973/1974 to 2006/2007 data, and from 57% to 75% in the 1976/1977 to 2006/2007 data. The substantial changes in the proportions show that the exclusion of the observation has an effect on the goodness-of-fit of the model used to predict wheat production in all subsets of wheat data.

On the other hand, the significance of individual variables does not change when the 1998/1999 observation is excluded, for all the subsets. The two variables, wheat production in the past immediate year and area harvested remain as the only variables with a significant linear relationship with wheat production. However, the declining p-values in the 1973/1974 to 2001/2002 and 1973/1974 to 2006/2007 wheat data show that the strength of evidence that area harvested had a significant linear relationship with wheat production increased when the outlying case was deleted. In addition, the magnitude of parameter estimates and values of their standard errors changed when the model was fitted without the observation. The influence of the observation on different quantities of the fitted model is similar in all the three periods under investigations.

The top and bottom left plots in Figure 4.3 and Figures E.5 and E.6 of Appendix E show a pattern where residuals increase with the values of the predicted values up to some point of the distribution of data, and then they started to decrease. This is an indication that the variance of error terms is not constant up a certain point of the distribution, for all subsets of wheat data. The plot on the top right of Figure 4.3 and Figure E.5 show that plotted points lie on a straight line with points on the right end being above the line, and one outstanding point. This is an indication of a symmetric distribution but with rather longer and heavier tails than the normal distribution. On the other

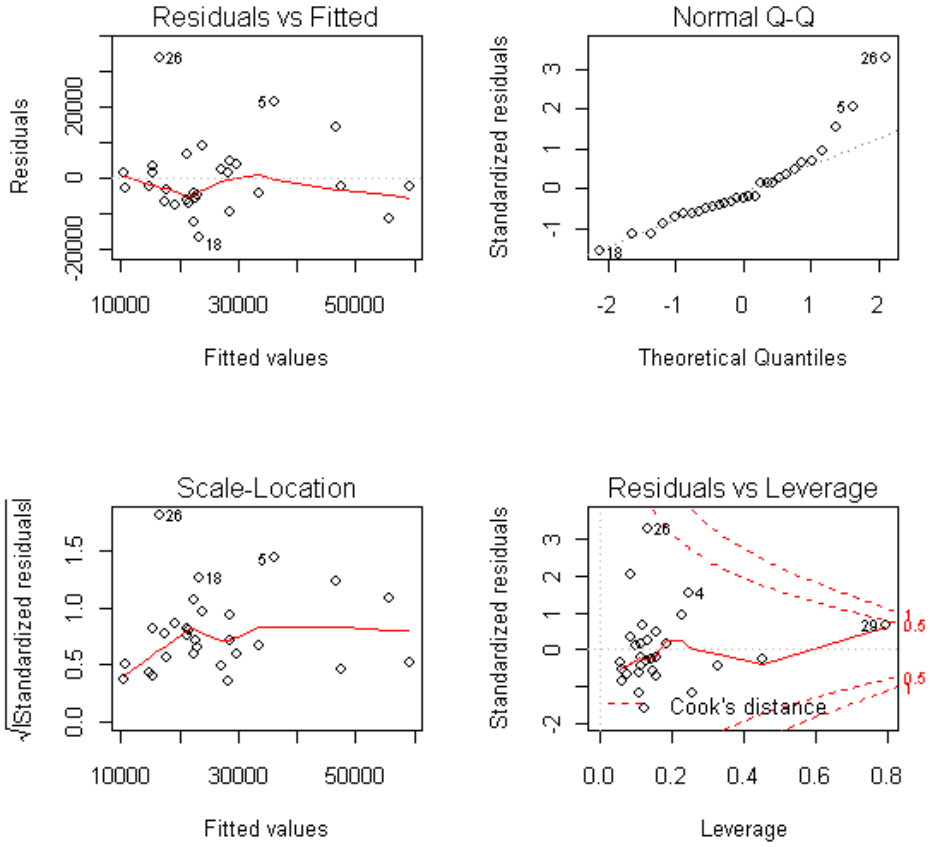

Figure 4.3: Plots of Residuals for 1973 - 2002 Wheat Data

hand, the plot on the top right of Figure E.6 show that almost all the points, except the twenty third and second observations, fall on a straight line. The plots for all the three subsets show evidence of the presence of two outliers in the subsets. The last plot on the bottom right of Figure 4.3 highlights the twenty sixth point, that stand out in the first three plots, as an extreme but not influential outlier. This is because the point is close to the upper Cook's distance contour, but not inside it. Thus the 1998/1999 observation is identified as an outlier, which is not considered influential in all subsets of wheat data.

4.5 Box-Cox Transformation of National Data

The violation of the assumption of constant variance of error terms and the problem of distributions of error terms that are symmetric but with rather longer and heavier tails than the normal distribution were corrected through the transformation of the response variable. These problems were observed from the values of skewness and kurtosis, box plots and diagnostics plots in Chapter 2 and preceding sections of this chapter respectively. In particular, the problems were identified for the 1976/1977 to 2006/2007 maize data, 1973/1974 to 2006/2007 sorghum data and all subsets of wheat data. The TRANSREG procedure was used to perform a Box-Cox transformation of the response variable. The procedure gives an option of specifying a list of transformation parameters and chooses the optimal value of the parameter.

Table 4.14: Values of Skewness and Kurtosis for National Data

		Original Data		Transformed Data	
Cereal	Subset of Data	Skewness	Kurtosis	Skewness	Kurtosis
Maize	1976/1977 to 2006/2007	1.47	3.85	0.19	0.41
Sorghum	1973/1974 to 2006/2007	0.70	-0.04	-0.52	-0.35
Wheat	1973/1974 to 2006/2007	0.99	-0.01	0.43	-0.43
	1973/1974 to 2001/2002	0.89	-0.31	0.06	-0.95
	1976/1977 to 2006/2007	1.39	1.59	-0.69	2.21

The values of skewness and kurtosis for the original data and the transformed data are given in Table 4.14. In the case of maize data for the period 1976/1977 to 2006/2007, both values decreased after the transformation, implying that the transformation of maize production reduced the degree of the deviation from the normal distribution. The absolute value of skewness for sorghum data of 1973/1974 to 2006/2007 decreased a bid from 0.70 to 0.52 but with a change of sign from positive to negative, while the absolute value of kurtosis increased from 0.04 to 0.35 with the sign remaining negative. This implies that when sorghum availability is transformed it becomes negatively skewed but with a flatter curve than the normal curve. In the case of wheat data, the transformation

of wheat production caused a decrease in the values of skewness for all the three subsets of data with the 1973/1974 to 2001/2002 data having the biggest decrease. This decrease shows that the transformation corrected skewness of the distribution of data for all the subsets. The values of kurtosis changed signs for all the three subsets and their absolute values increased.

The Box-Cox family of power transformation was applied to identify the appropriate power transformation used to transform the response variable in each subset of data, so that the applicability and usefulness of linear regression models in modelling cereals availability at the national level is increased.

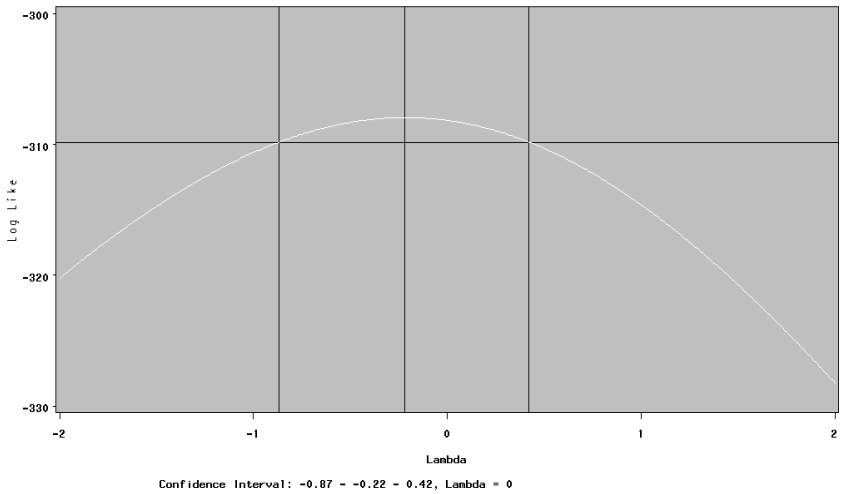

Figure 4.4: Log Likelihood Plot for 1976 - 2007 Maize Data

In the case of the 1976/1977 to 2006/2007 maize data, the estimated optimum value of the transformation parameter, λ obtained by the maximum likelihood method, is -0.22. The plot of the log likelihood against values of λ shows an approximate 95% confidence interval for λ as $[-0.87, 0.42]$ (Figure 4.4). The confidence interval contains the standard power transformation value of zero, which is used as the nearest convenient value of $\hat{\lambda}$. This value suggests a natural logarithm transformation for the response variable, maize production. The results obtained from fitting the model using the transformed maize production are not shown since they do not differ much with the results obtained using the untransformed maize production in Table C.3 of Appendix C. Two variables, the amount of rainfall and area harvested remain significant in affecting maize production. The use of the transformed response variable improved the fit of the model since the value of R^2 increased from 0.43 to 0.51.

The box plot in Figure 4.5 confirms the observation from the values of skewness and kurtosis for the

transformed maize production of 1976/1977 to 2006/2007 data that the suggested transformation corrected the deviation from assumption of normality. This is portrayed by the line in the box being equidistant to the lower and upper edges of the box, rather than being towards the lower edge as it was the case in the box plot for the original maize production in Figure M.2 of Appendix M. Further, the extreme observation shown in the box plot for original data does not appear in the box plot of the transformed data.

The top and bottom left plots in Figure 4.6 do not show any specific pattern of the residuals, the residuals are scattered randomly. This is an indication that the heteroscedasticity observed in Figure E.2 of Appendix E was corrected by using the suggested natural logarithm transformation to transform maize production. The points on the lower end of the QQ plot in Figure E.2, which were above the straight line, are now on the line in Figure 4.6. However, the points on the upper end of the QQ plot have now moved and they are below the straight line. This shows that the distribution of the error term with longer and heavier tails, was corrected partly by the suggested transformation of the response variable.

The estimated optimum value of $\hat{\lambda}$ for the 1973/1974 to 2006/2007 sorghum data is 0.34. Figure 4.7 shows an approximate 95% confidence interval of $[-0.06, 0.74]$. Like in the case of the 1976/1977 to 2006/2007 maize data, the confidence interval contains the value of the standard power transformation as zero. This value is used as the nearest convenient value of $\hat{\lambda}$ and it suggests a natural logarithm transformation for the response variable, sorghum production.

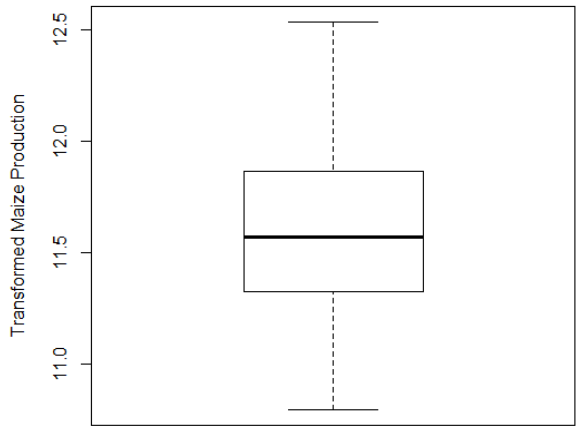

Figure 4.5: Box Plot of the Transformed 1976 - 2007 Maize Data

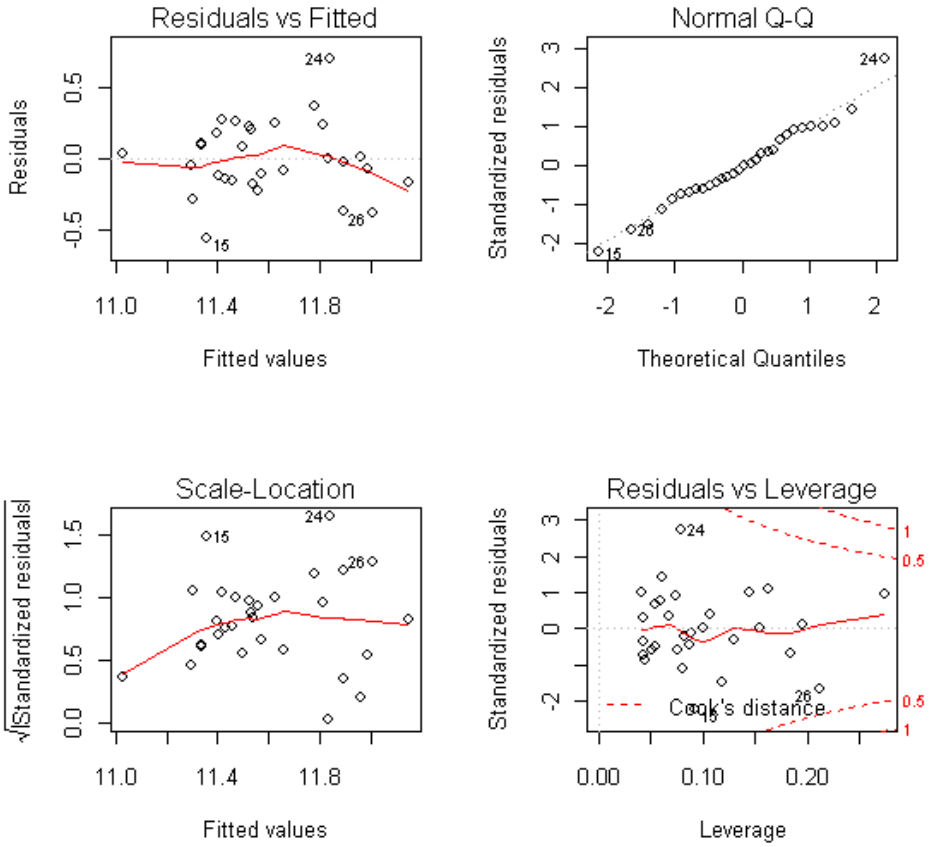

Figure 4.6: Plot of Residuals for Transformed 1976 - 2007 Maize Data

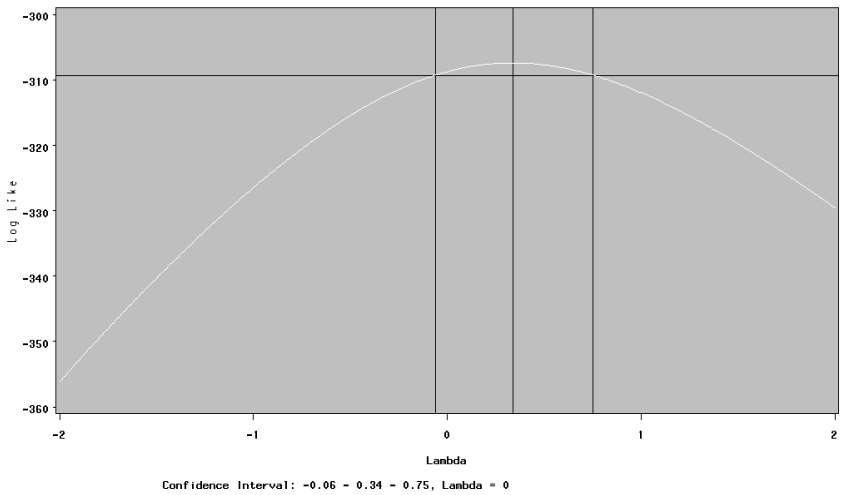

Figure 4.7: Log Likelihood Plot for 1973 - 2007 Sorghum Data

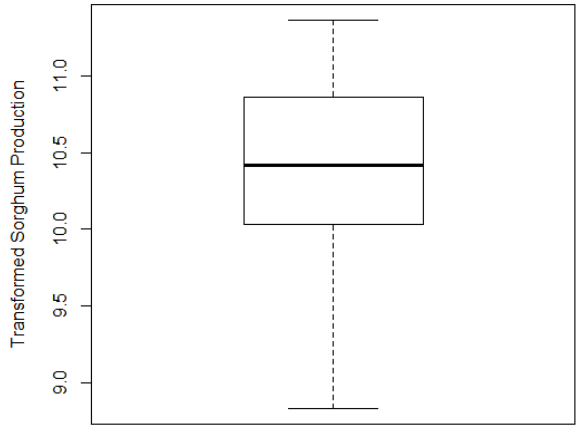

Figure 4.8: Box Plot of the Transformed 1973 - 2007 Sorghum Data

Similarly, the box plot of the transformed sorghum data for the period 1973/1974 to 2006/2007 in Figure 4.8 shows that transforming sorghum production corrected its distribution to being approximately normally distributed. This is shown by the line inside the box being almost in the middle of the box, rather than being towards the lower edge of the box as it was the case in Figure M.3 of Appendix M for the original sorghum production.

The results from using the natural logarithm of sorghum production as the response variable in fitting the model are the same as the results obtained using the untransformed sorghum production in Table C.5 of Appendix C. Two variables, production of sorghum at time t-1 and area harvested, remain significant and the strength of evidence for their significance remains the same. The value of the coefficient of determination increased slightly from 0.66 to 0.69, indicating a small improvement in the fit of the model. The violation of the assumption of constant variance observed from Figure E.3 of Appendix E was corrected because the residuals on the top and bottom left plots in Figure 4.9 do not show any specific pattern, they are scattered randomly. However, the QQ plot on the top right shows that the longer and heavier tails have been corrected but partially.

The estimated optimum value of the transformation parameter for the 1973/1974 to 2006/2007 wheat data is 0.42. Figure 4.10 shows an approximate 95% confidence interval for λ as $[0.1, 0.79]$. The interval contains the standard power transformation value of 0.5, suggesting a square root transformation of the response variable, wheat production. The results from fitting the model

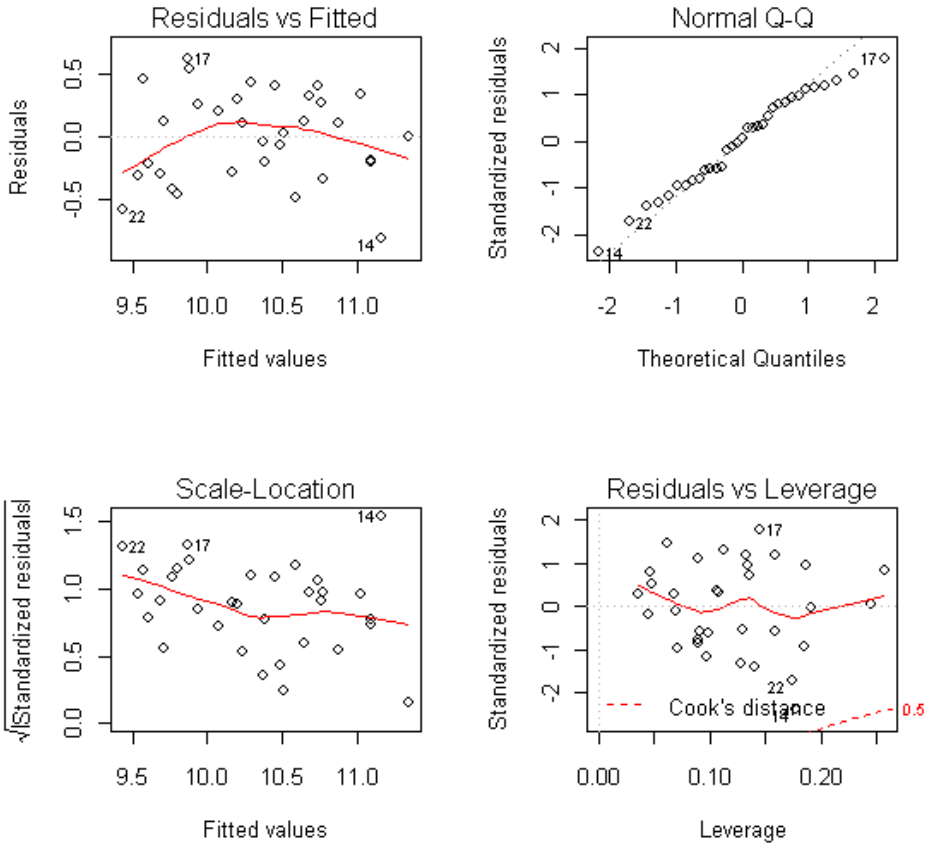

Figure 4.9: Plot of Residuals for Transformed 1973 - 2007 Sorghum Data

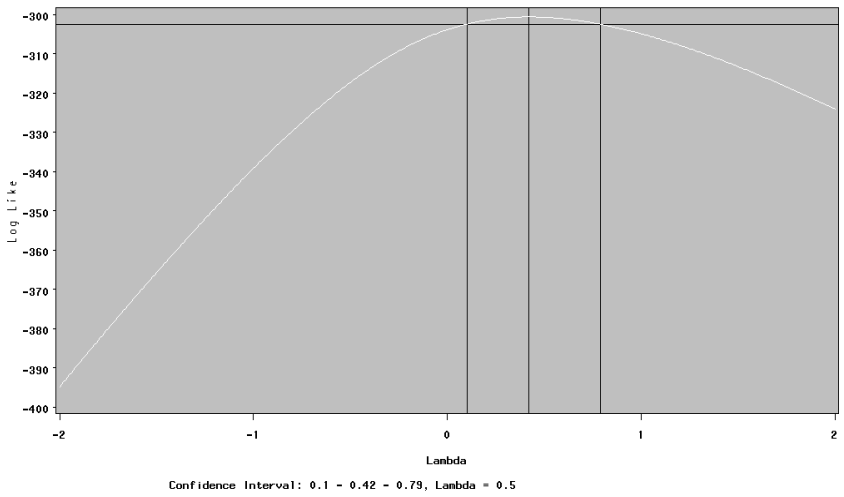

Figure 4.10: Log Likelihood Plot for 1973 - 2007 Wheat Data

using the square root of wheat production are the same as the results obtained from using wheat production in its original form, given in Table C.9 of Appendix C. Two variables, production of wheat at time t-1 and area harvested remain significant and the strength of evidence for their significance remains the same. The value of the coefficient of determination decreased slightly from 0.64 to 0.63. This is an indication that the use of the transformed wheat production as the response variable does not improve the fit of the model.

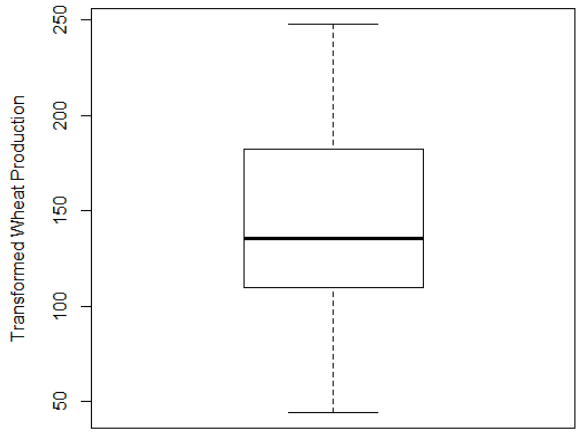

Figure 4.11: Box Plot of the Transformed 1973 - 2007 Wheat Data

In the case of wheat data for 1973/1974 to 2006/2007, the box plot in Figure 4.11 shows that though the line in the middle of the box shifted upwards a bid after the transformation of wheat production, the distribution of the variable remains skewed to the right. The box plot of the transformed wheat production for the 1973/1974 to 2001/2002 data in Figure M.7 of Appendix M shows that the suggested transformation did not correct the deviation of the distribution of the variable from the normal distribution. Probably this is due to the fact that though the value of skewness decreased from 0.89 to 0.06, the absolute value of kurtosis increased from 0.31 to 0.95. The box plot for the 1976/1977 to 2006/2007 wheat data in Figure M.8 of Appendix M shows the same pattern as the one in the previous subset of wheat data in the sense that the suggested transformation did not correct the deviation from the normal distribution.

In the case of 1973/1974 to 2001/2002 and 1976/1977 to 2006/2007 wheat data, the estimated optimum values of the transformation parameter are 0.23 and 0.28, respectively. Figure F.1 in Appendix F for the 1973/1974 to 2001/2002 data and Figure F.2 in Appendix F for the 1976/1977

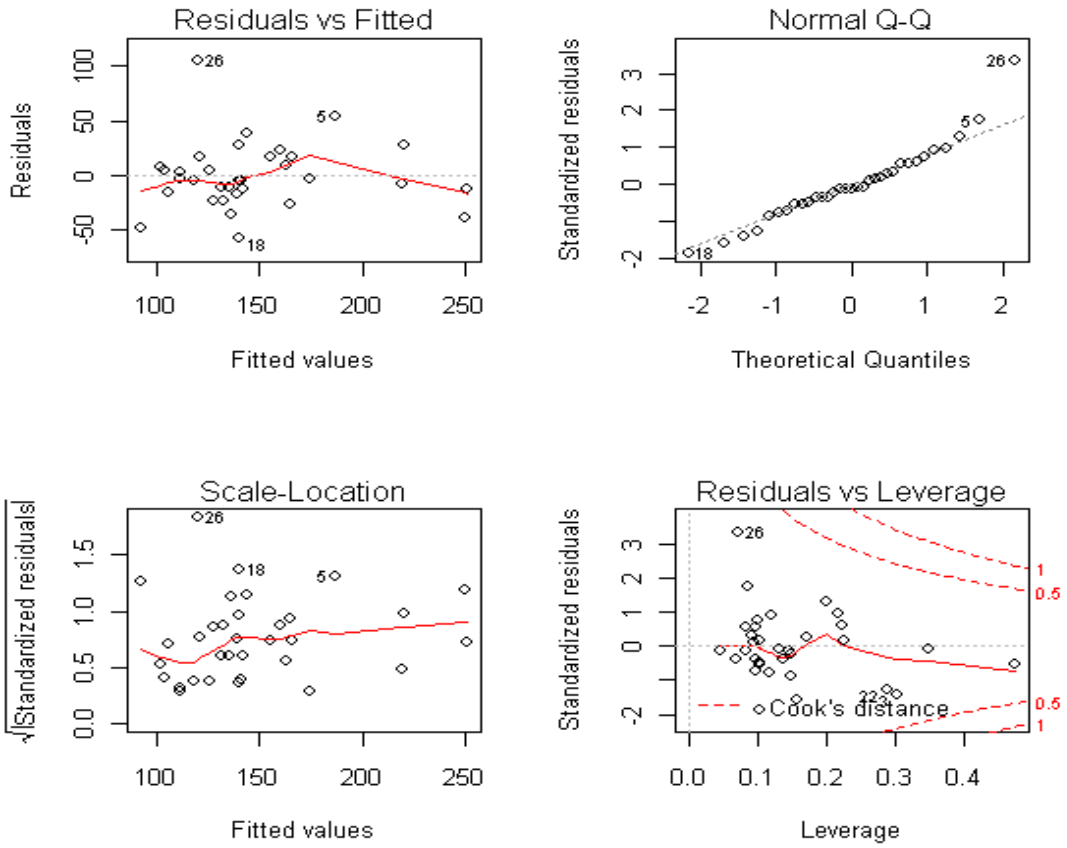

Figure 4.12: Plot of Residuals for Transformed 1973 - 2007 Wheat Data

to 2006/2007 data show the approximate 95% confidence intervals for λ as $[-0.34, 0.79]$ and [-0.04, 0.64], respectively. Both intervals contain the standard power transformation value of zero, suggesting a natural logarithm transformation of wheat production. Like in the case of 1973/1974 to 2006/2007 wheat data, the results from fitting the model to the transformed data in both subsets remain the same as the ones obtained from using wheat production in its original form in Table 4.10 and Table C.11 of Appendix C. Two variables, production of wheat at time t-1 and area harvested remain significant. The strength of evidence for the significance of production of wheat at time t-1 remains the same in both subset of data. However, in the 1973/1974 to 2001/2002 data, the strength of evidence for harvested area has decreased. The value of the coefficient of determination decreased in both cases, showing that the fit of the model gets worse when the transformed Wheat production is used as the response variable.

The top left plot in Figure 4.12 and Figure G.2 in Appendix G of diagnostics plots from fitting the model using the transformed wheat production, show that a pattern of residuals is not very different from that of residuals from fitting the model using untransformed wheat production. This shows that using the square root transformation of wheat production for the 1973/1974 to 2006/2007 data and natural logarithm transformation of wheat production for the 1973/1974 to 2001/2002 data does not correct the observed heteroscedasticity completely. On the other hand, almost all points on the QQ plots of the two subsets of wheat data fall on the straight line, except the outlying observation of 1998/1999 and few observations on the left and upper ends. These show that the transformation have corrected longer and heavier tails of the distributions for the two subsets, but partially. In the case of 1976/1977 to 2006/2007 data, the top left plot in Figure G.3 of Appendix G shows residuals that are randomly scattered. This shows that the violation of non-constant variance was corrected by using the natural logarithm transformation of wheat production.

Though there are a number of cases where the Box-Cox transformation failed to correct the violation of the assumptions of constant variance of the error terms and normality, quantile regression could not be applied to the national data as a robust alternative to the leas squares procedure because these data have a limitation of a relatively small sample size. (Cade, 2003) noted that quantile regression does not give reliable confidence intervals when the sample size is small.

4.6 Application of Ridge Regression in National Data

Ridge regression was applied to control inflation and instability of parameter estimates caused by collinearity or near linear dependencies among columns of the data matrix $\mathbf{X}$. The dependencies were identified from the results of collinearity diagnostics obtained when fitting the regression model to the national maize, sorghum and wheat data. We used ridge trace to select the biasing factor δ for which the ridge regression estimate $\hat{\boldsymbol{\beta}}_R^*$ is closer to the true underlying regression parameters $\boldsymbol{\beta}$, than the least square estimate $\hat{\boldsymbol{\beta}}$. This is the estimate of $\boldsymbol{\beta}$ with a smaller mean square error (MSE) than the one of $\hat{\boldsymbol{\beta}}$. Ridge traces were used to show the sensitivity of parameter estimates to nonorthogonality of predictor variables in the fitted regression model for maize, sorghum and wheat data. These are plots where parameter estimates $\hat{\beta}_{jR}^*$ for different values of δ are plotted against values of δ in the interval [0, 1]. The REG procedure was used to perform the ridge regression analysis because it has a RIDGE option that requests a ridge regression analysis and specifies the values of the biasing factor. Microsoft Excel was used to plot ridge traces using the data generated from the REG procedure when performing the ridge analysis.

The ordinary least squares estimates are parameter estimates at $\delta = 0$, while the ridge regression estimates are at values of δ that are greater than zero. Unlike the least squares estimation procedure with established distributional theory, ridge regression lacks distributional theory (Gunst and Mason, 1980). The lack of distributional theory does not allow inferences such as t test because the t statistic rely on the assumption of normality, and in ridge regression the validity of approximated t statistics is not guaranteed. Thus inferences about ridge parameter estimates were not made.

The ridge trace for the 1973/1974 to 2001/2002 maize data in Figure 4.13 shows that, in general, ordinary least squares estimates at $\delta = 0$ are unstable and overestimated in absolute values. The small increase from $\delta = 0$ to $\delta = 0.05$ causes a rapid decline in absolute values of estimated regression coefficients for both time and price of maize. The parameter estimate for time has a negative value at $\delta = 0$, which moves rapidly to zero and further to positive values with a small increase in the biasing factor. The parameter estimate for price of maize has the second highest positive value at $\delta = 0$, but it moves quickly towards zero when δ increases. The rapid movement indicates the instability involved in the parameter estimates, and severity of collinearity that exist between time and price of maize. Gunst and Mason (1980) and Schabenberger and Pierce (2002) noted that parameter estimates for variables that are involved in a near linear dependency change rapidly as the biasing factor increases from zero, and signs for some estimates may change. The more the ridge trace moves rapidly, the higher the degree of collinearity among the variables. The estimates stabilize at the values of δ in the interval [0.10, 0.40], since it is within this interval where the estimates begin to show small change as the biasing factor increases. Thus ridge estimates at this interval are likely to be closer to the true parameter $\boldsymbol{\beta}$ and more stable. The value of δ within the

interval that gives a biased but more stable estimate of $\boldsymbol{\beta}$ is 0.25.

Table 4.15: Parameter Estimates for 1973 - 2002 Maize Data at $\delta = 0$ and $\delta = 0.25$

Variable	$\delta = 0$			$\delta = 0.25$		
	Estimate	Std. Error	VIF_i	Estimate	Std. Error	VIF_i
Time	-1290.13	3378.31	16.23	564.4	547.26	0.41
Rainfall	82.11	61.85	1.33	80.81	44.79	0.67
Harvested area	0.86	0.30	1.58	0.67	0.21	0.72
PriceTon	96.85	115.04	15.40	34.29	19.58	0.43

Table 4.15 shows parameter estimates, their standard errors and variance inflation factors (VIF) for both ordinary least squares estimation procedure at $\delta = 0$ and ridge regression at $\delta = 0.25$, for the 1973/1974 to 2001/2002 maize data. Ordinary least squares estimates have large absolute values for all variables. They also have large standard errors, which decline rapidly as the value of the biasing factor increases to 0.25. Variance inflation factors associated with time and price of maize declined from 16.23 and 15.40 to 0.41 and 0.43, respectively, when the value of the biasing factor increases from zero to 0.25. The reduction in the variance inflation factors to values that are less than 10 is an indication of the extent to which the use of ridge regression remedies collinearity problems. Ridge regression remedies collinearity by reducing the quantities by which the variances of the estimated regression coefficients for time and price are inflated due to collinearity.

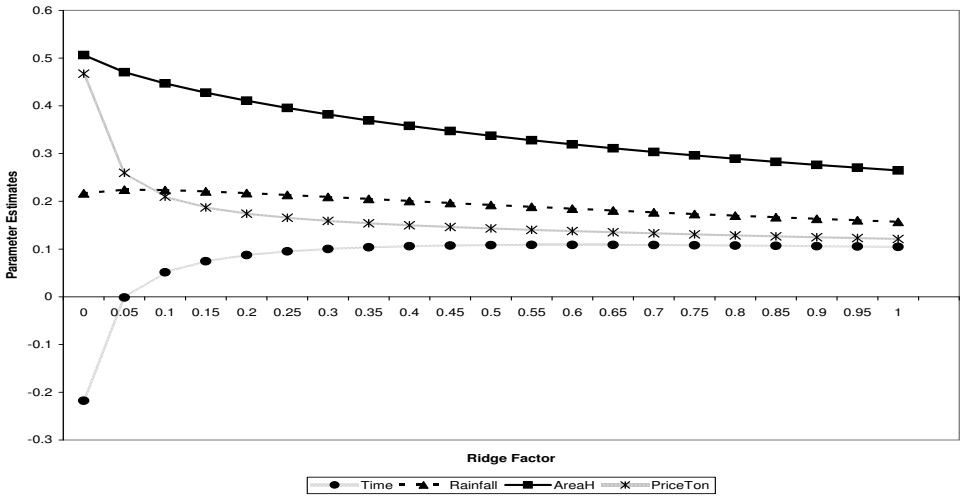

Figure 4.13: Ridge Trace for 1973 - 2002 Maize Data

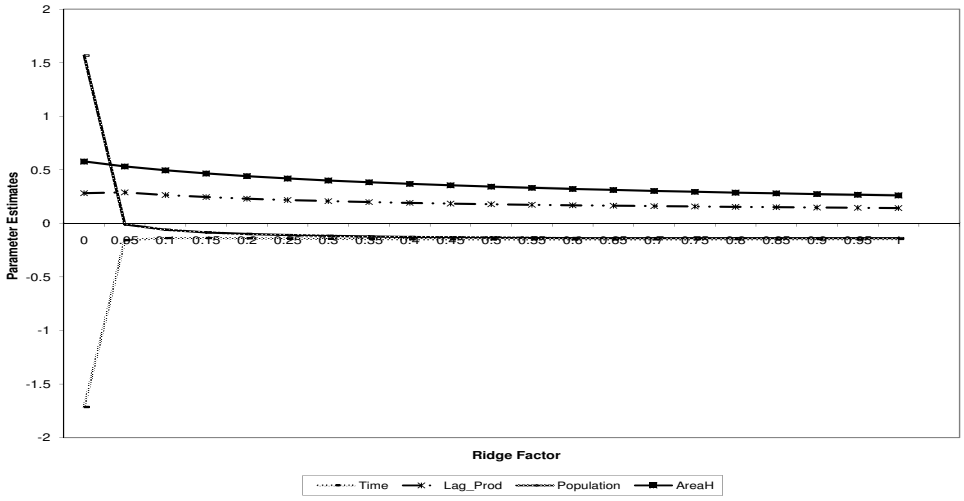

Figure 4.14: Ridge Trace for 1976 - 2007 Sorghum Data

Similarly, least squares estimates for the 1976/1977 to 2006/2007 sorghum data are overestimated and unstable (Figure 4.14). The increase of the ridge factor from 0 to 0.05 yields a rapid decline of absolute values of the estimated coefficients for time and population size. The least squares estimate for population size has the largest positive value. However an increment in the biasing factor from 0 to 0.05 pushes it to zero and further to negative values as the factor increases further. The estimate of time at $\delta = 0$ has a negative value, which changes rapidly towards zero as the biasing factor increases. The estimates for the two variables are the most unstable, indicating that they are strongly correlated. The parameter estimates stabilize at values of δ in the interval [0.05, 0.30]. A biased but more stable estimate $\hat{\boldsymbol{\beta}}_R^*$ is at $\delta = 0.15$.

Table 4.16: Parameter Estimates for 1976 - 2007 Sorghum Data at $\delta = 0$ and $\delta = 0.15$

Variable	$\delta = 0$			$\delta = 0.15$		
	Estimate	Std. Error	VIF_i	Estimate	Std. Error	VIF_i
Time	-3877.89	3875.18	225.4	-307.33	152.55	0.39
Production-t	0.28	0.17	2.08	0.24	0.12	1.02
Population	0.09	0.09	217.84	-0.005	0.004	0.42
Harvested area	0.65	0.16	1.62	0.52	0.12	0.95

The least squares estimates for the 1976/1977 to 2006/2007 sorghum data are large in absolute

values, their standard errors and VIF are large when they are compared to that of ridge estimates at $\delta = 0.15$ (Table 4.16). The use of ridge estimation seem to be remedying the effects of collinearity since it reduces substantially the quantity by which the variance of each of the coefficients for time and population size are inflated. The VIF that corresponds with time is reduced from 225.4 to 0.39, and the VIF that corresponds with population size is reduced from 217.84 to 0.42.

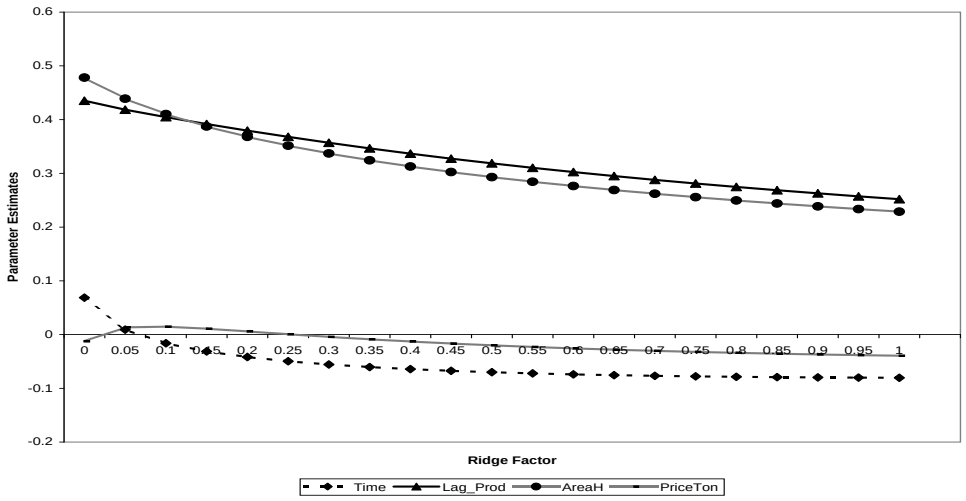

Figure 4.15: Ridge Trace for 1973 - 2002 Wheat Data

Estimated regression coefficients for time, production of sorghum in the immediate past year and area harvested to sorghum, for the 1973/1974 to 2006/2007 sorghum data, show the same pattern of movement as δ increases from zero (Figure H.1). In addition, their signs do not change as the biasing factor increases. The gradual movement of the estimates suggests the presence of moderate collinearity problems that involve the three variables. The ridge trace confirms the results from the condition index of 15.56 and variance-decomposition proportions in Table C.6 in Appendix C, which suggested that the three variables are involved in a moderate collinear relationship.

Estimated coefficient for time increases slightly in absolute value as the biasing factor increases from zero, whereas the estimates for production of sorghum in the past immediate year and area harvested decline slightly (Table I.1 in Appendix I). The standard errors of the estimates decline as δ increases. The estimates stabilize at values of the biasing factor in the interval [0.05, 0.20]. The value of the biasing factor that gives an unbiased but stable estimate of $\boldsymbol{\beta}^*$ is 0.05.

The ridge trace of 1973/1974 to 1997/1998 sorghum data in Figure H.2 in Appendix H shows that an increase in the biasing factor from zero results in an increase and a decrease in absolute value of parameter estimates for price of sorghum and time, respectively. Parameter Estimates for area harvested and area under crop failure for sorghum are stable. These show that time and price are collinear, while area harvested and area under crop failure are not involved in any near linear dependency in the data set. The movement in the estimates of time and price is not as rapid as that of the same variables in maize data in the 1973/1974 to 2001/2002 period, which indicates that collinear relationship that exists between time and price for sorghum data is moderate. The least squares estimates of time and price are large in absolute values, their standard errors are large and decline as δ increases (Table I.2 in Appendix I). The unstable estimates become stable at values of δ that range from 0.05 to 0.25. A stable but bias estimated coefficient $\hat{\boldsymbol{\beta}}_R^*$ is at $\delta = 0.20$.

In the case of 1973/1974 to 2001/2002 wheat data, the least squares estimates are slightly larger in absolute values than the ridge regression estimates (Figure 4.15). The estimates change gradually with the increase in the biasing factor. The estimated coefficient for time has a small positive value at $\delta = 0$, which is driven to zero and further slightly below zero as δ increases. The estimated coefficient for price of wheat has a small negative value at $\delta = 0$, which is driven to zero and further slightly below zero as δ increases. The stabilization of parameter estimates occurs at values of δ that range between 0.15 and 0.35. A more stable estimate of $\boldsymbol{\beta}^*$ is at $\delta = 0.25$.

Table 4.17: Parameter Estimates for 1973 - 2002 Wheat Data at $\delta = 0$ and $\delta = 0.25$

Variable	$\delta = 0$			$\delta = 0.25$		
	Estimate	Std. Error	VIF_i	Estimate	Std. Error	VIF_i
Time	129.86	798.64	10.51	-93.71	182.00	0.52
Production-t	0.46	0.17	1.55	0.39	0.12	0.71
Harvested area	0.53	0.23	2.45	0.39	0.13	0.79
Price/Ton	-0.55	16.51	8.31	0.03	4.42	0.57

The least squares estimates have relatively bigger absolute values, standard errors and VIF than ridge regression estimates, specifically for the variable, time (Table 4.17). Like in maize data, the variables that are affected by collinearity are time and price. The gradual movement of the ridge trace and small decline in absolute values of estimates is an indication that the degree of collinearity between the two variables is not as severe as it is for the 1973/1974 to 2001/2002 maize data.

Like for sorghum data set in the period 1973/1974 to 2006/2007, estimated regression coefficients for all variables in the ridge trace of wheat data set for the same period change gradually as the biasing factor increases from zero (Figure H.3 in Appendix H). The estimate for time moves from a

positive value at $\delta = 0$ to negative values as δ increases. The stabilization of parameter estimates occurs at values of the biasing factor in the interval [0.10, 0.35]. Table I.3 shows a change in the parameter estimate of time from 36.33 at $\delta = 0$ to -114.61 at $\delta = 0.15$. Similarly its standard error changes with big magnitude from 298.94 to 196.69. The sizeable changes in the parameter estimate and its standard error is a sign of severe collinearity that has been remedied by ridge regression.

The pattern of parameter estimates of the 1976/1977 to 2006/2007 wheat data in Figure H.4 in Appendix H is similar to that of sorghum data in the same period in Figure 4.14. The increase of the ridge factor from 0 to 0.05 results with a rapid decline of absolute values of the estimated coefficients for time and population size. The decline of estimates as δ increases shows that the two variables are highly correlated. The estimates stabilize at values of δ in the interval [0.05, 0.45]. The value of the biasing factor that provides a biased but more stable estimate of $\boldsymbol{\beta}^*$ is 0.25. Parameter estimates and their standard errors decline with an increase in the biasing factor (Table I.4 in Appendix I).

4.7 Summary

Fitting the linear regression model to the national maize data show that area harvested was the only factor that affected maize production during the two periods, 1973/1974 to 2001/2002 and 1973/1974 to 2006/2007. However, when the third observation of 1975/1976 identified as an influential observation was left out from the 1973/1974 to 2001/2002 data, the results show that area harvested did not have an effect on maize production, instead the amount of rainfall had a positive effect. During the period 1976/1977 to 2006/2007, both the amount of rainfall and harvested area affected maize production positively, but with the amount of rainfall having a stronger evidence. In the case of sorghum data, area harvested had a positive effect on sorghum production, and production of sorghum in the past immediate year had a positive association with sorghum production in the current year, for the two periods, 1976/1977 to 2006/2007 and 1973/1974 to 2001/2002. Concerning the years, 1973/1974 to 1997/1998, area harvested affected sorghum production positively, while area under crop failure affected sorghum production negatively. With regard to wheat data, the results from all the three subsets are almost similar in the sense that harvested area and wheat production in the past immediate year are the only variables with a relationship with sorghum production in the current year. The relationship is positive for both variables in all the subsets of data. Though the additional variables in the subsets of data on the three cereals did not add value in predicting production of cereals, some subsets of sorghum and wheat data showed interesting patterns of condition index and variance proportions in detecting collinear diagnostics.

The results from collinearity diagnostics indicate that high variance-decomposition proportions that correspond to one small eigenvalue and the highest condition index pointed to collinear re-

lationships, and thus confirmed strong collinearity suggested by high correlation coefficients. The superiority of condition index and variance-decomposition proportion, to both tolerance and variance inflation factors, as collinear diagnostics was not only shown by them pointing to more than two variables involved in a collinear relationship. It was also shown by the diagnostics being able to detect collinear relationships where tolerance and variance inflation factors failed to do so. In the 1973/1974 to 1997/1998 sorghum data and 1973/1974 to 2006/2007 wheat data, condition index and variance-decomposition proportion showed a peculiar pattern that deviated from the normal that the highest condition index indicates the presence of collinear relationship. Instead the second highest condition index pointed to the presence of collinear relationships. The observed peculiar pattern is an empirical finding that needs to be investigated further in the future research to establish the statistical theory behind it. Strong collinear relationship between variables decreased values of the t statistics for variables involved in such relationships. A very large value of condition index signified severe collinearity that concealed the effect that the involved predictor variables could have had on the response variable. Ridge estimates at values of the biasing factor selected by ridge traces have smaller standard errors and variance inflation factors, indicating the extent to which ridge regression remedied collinearity problems and controlled the instability in parameter estimates.

Fitting the model without the ith observation identified by case deletion diagnostics as influential on different quantities of the fitted model showed how quantities such as the goodness-of-fit measure (R-squared), parameter estimates, their standard errors and values of t statistics were affected by the observation. The plot of standardized residuals against leverage that shows the upper and lower Cook's distance contour, highlighted outliers and high leverage points that are influential. The use of Box-Cox transformation to correct violation of assumptions increased the strength of evidence for significance of some variables in predicting the response variable, and improved the goodness of fit of the model in some cases. In the case where the Box-Cox transformation corrected the violation of assumptions partially or failed to correct it completely, the best option could have been to use quantile regression as a robust alternative to the OLS procedure. However, it could not be applied because the national data have a limitation of having a relatively small sample size for quantile regression to give reliable confidence interval. In the next chapter, the linear regression models were used to model availability of cereals at the household level. Regression diagnostics were used to detect problems in the data, which may compromise the quality of the results, and appropriate remedial measures were used where possible.

Chapter 5

General Regression Model and Diagnostics of Household Data

5.1 Introduction

In the previous chapter, we applied linear regression models and diagnostics to model national availability of cereals. The focuss in this chapter is to apply the general linear model and diagnostics to model availability of cereals at the household level, in particular maize and sorghum availability. The models where each of maize availability and sorghum availability was regressed on predictor variables consisting of continuous, discrete and categorical variables were fitted. Continuous and discrete predictor variables in the model are household monthly income and household size, respectively. Categorical variables are location of a household, sex, education level and occupation of the head of a household. Location of a household is divided into four categories depending on the place of location of households included in the study. The categories are Berea, Mafeteng, Maseru lowland and Maseru Foothill. Education level is divided into four categories, namely, no formal education, primary, high school and post high school education. Occupation is divided into five categories, namely, casual worker, salary earner, subsistence farmer, pensioner and unemployed. When fitting the model, one category of each categorical variable served as a reference or base category.

Conclusions made from the analysis of the household data are restricted to the villages from which households were selected, not the districts where the villages were located. The reason being that households within locations included in the study were selected at random from villages that were easily accessible, using systematic sampling where every fifth household was selected and interviewed on the spot. This did not give all households in the districts an equal chance of being included in the study. Thus households in the study are not a true representative of households in the districts.

The model fitted to the household data is the general linear regression model with categorical predictor variables. The GLM procedure of SAS and command lm() of R were used to fit the general linear model as well as to perform regression diagnostics. The two statistical softwares give the same results for the general linear model and influence measures. The diagnostics and box plots were plotted by the R system of graphics. The type III sum of squares that evaluate the effect of each variable after all other variables have been accounted for was used for both cases of household maize availability and sorghum availability. This is the partial sum of squares used when one predictor variable is added one at time and examining its effect on the response variable given that all other predictor variables are already in the model. It shows by how much is the residual sum of squares reduced by adding a particular term to the model that contains all other terms.

In the case of type III sum of squares, there is one degree of freedom sum of squares measuring the contribution of each regression coefficient β_i on the regression sum of squares, given that all the terms not involving β_i were already in the model. Estimated coefficients for categories of categorical predictor variables represent differential effects of the variables on the response variable. These are differences or contrasts between a given category and the respective reference category and thus they are interpreted relative to the reference categories. For example, the estimate that corresponds to the category of Berea, -13.99, is the mean response difference between Berea and the reference category Mafeteng (Table 5.2).

5.2 Modelling of Maize Availability

The overall F test of the null hypothesis that none of the predictor variables is linearly related to maize availability for households is significant at 5% level of significant. The significance of the tests shows that at least one of the predictor variables in the model is linearly related to availability of maize. The value of the coefficient of determination R^2 of 0.30 shows that predictor variables in the model explain 30% of the variability in maize availability.

Three variables, household size, household monthly income and occupation of the head of a household, are significant at 5% level of significance, suggesting that maize availability is linearly related to the three variables (Table 5.1). The linear relation of the three variables with maize availability means that they had a significant effect on household availability of maize.

The effect of household size and household monthly income was measured by their estimated regression coefficients of 87.22 and 0.14, respectively (Table 5.2). The coefficients show that an increase in the size of a household by one member increased the mean maize availability by 87.22 kilograms, and an increase in household monthly income by one Rand increased the mean maize availability

Table 5.1: Sources of Variation and Sum of Squares for Maize Household Data

Response Variable: Maize availability to households

Source of Variation	DF	Type III Sum of Squares	Mean Square	F Value	Pr > F
Household Size	1	12226241.96	12226241.96	70.13	<.0001
Income	1	1726321.14	1726321.14	9.90	0.0018
Sex	1	3922.92	3922.92	0.02	0.8809
Location	3	349506.14	116502.05	0.67	0.5722
Education	3	432105.60	144035.20	0.83	0.4803
Occupation	4	1886103.40	471525.85	2.70	0.0307

Table 5.2: Regression Parameter Estimates for Maize Availability

Response Variable: Maize availability to households

Variable	Category	Estimate	Standard Error	t value	P-value
Intercept		265.20	103.66	2.56	0.01104
Household Size		87.22	10.41	8.37	<.0001
Income		0.14	0.05	3.15	0.00183
Sex	ref=Male				
	Female	8.43	56.22	0.15	0.88087
Location	ref=Mafeteng				
	Berea	-13.99	98.15	-0.14	0.88678
	Maseru Foothill	-61.26	99.59	-0.62	0.53897
	Maseru Lowland	-88.54	67.09	-1.32	0.18798
Education	ref=No educ				
	High School	29.27	94.96	0.31	0.75817
	Primary	-53.35	69.11	-0.77	0.44074
	Post High School	131.98	156.43	0.84	0.39954
Occupation	ref=farmer				
	Casual Worker	-195.97	92.85	-2.11	0.03568
	Pensioner	92.74	112.39	0.83	0.40998
	Salary Earner	-105.42	95.56	-1.10	0.27086
	Unemployed	-159.55	72.27	-2.21	0.02807

by 140 grams, when the rest of the variables in the model are held constant.

Sex of household head is not significant and this suggests that maize availability for a household was not affected by whether the household was headed by a male or a female. All categories of location of a household in Table 5.2 have negative estimated coefficients, indicating that maize availability for households in Berea, Maseru lowland and Maseru foothills was less than maize availability for households in Mafeteng, the reference category. However, large p-values is an indication that the differences are not statistically significant for all categories. Similarly, in the case of education level, maize availability for households that were headed by people with high school, primary or post high school education, was not significantly different from maize availability for households

headed by people with no formal education. The results show that sex of household head, location of a household and education level of household head did not have a significant effect on how much maize was available to households.

On the contrary, two categories of occupation of household head, namely, casual workers and unemployed, are significant at 5% level of significance. This means that maize availability for households headed by casual workers and unemployed heads was significantly different from that of households headed by subsistence farmers. However, maize availability for households headed by pensioners and salary earners was not significantly different from that of households headed by subsistence farmers. The estimated regression coefficients show that maize available to households headed by casual workers and unemployed people was less than that of households headed by subsistence farmers by 195.97 and 159.55 kilograms, respectively. The difference in maize availability was higher for households headed by casual workers than for those headed by unemployed people.

An observation that households headed by casual workers and unemployed people were worse off than households headed by subsistence farmers, in terms of maize availability, is not surprising because subsistence farming is an occupation devoted to agricultural production for subsistence including production of food cereals. On the other hand, it could be that households headed by casual workers and unemployed people did not have enough resources that enabled them to acquire maize for their households through purchases or food production because their heads were not employed. In the case of households headed by salary earners the reason for their maize availability being not significantly different from that of households headed by subsistence farmers may be that they earned income, which enabled them almost the same access to maize as households headed by subsistence farmers who engaged in food production. The reason for lack of significant difference in the category of households headed by pensioners could be that some pensioners engaged in food production after they retired from their respective jobs and thus their availability of maize was almost similar to that of subsistence farmers.

We used residual plots in Figure 5.1 to check the assumptions of constant variance and normality of error terms for the fitted model. The plots on the top and bottom left show a pattern of residuals that increase as the fitted values get larger. This is a sign of violation of constant variance assumption also referred to as heteroscedasticity. The normal QQ plot on the top right shows that plotted points lie on a straight line with points on the right end being above the line, and two outstanding points. This is an indication of a symmetric distribution but with a rather longer and heavier tail than the normal distribution. The plot also shows evidence of the presence of outliers in the set of data.

The last plot on the bottom right highlights outliers and high leverage points that are influential. Points 219 and 281 that stand out in the first three plots with large residuals are considered

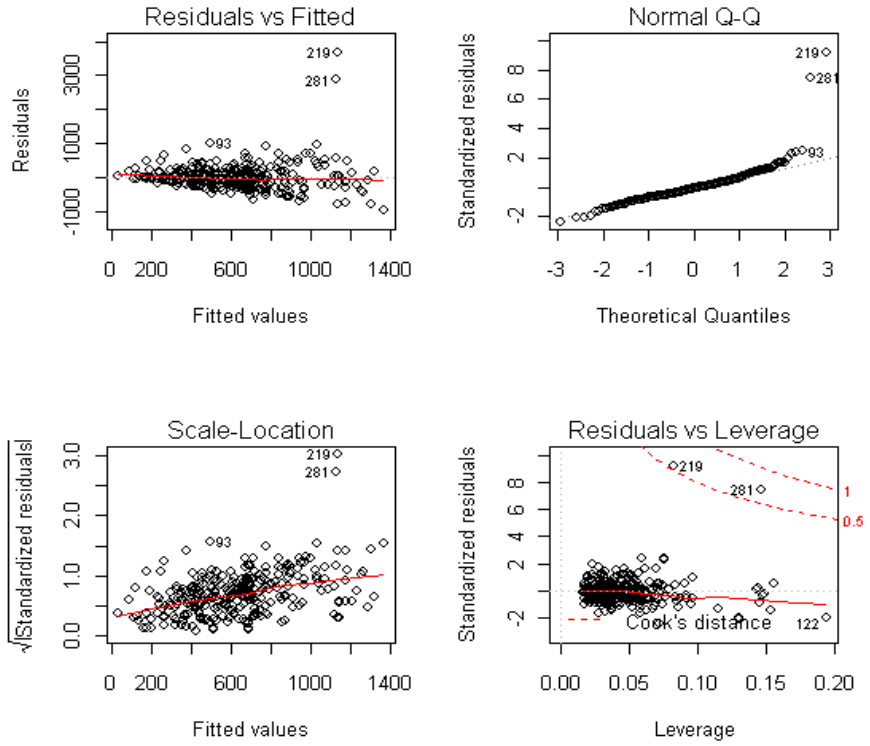

Figure 5.1: Plot of Residuals for Maize Availability

influential since they appear in the upper Cook's distance contour in this plot. These observations represent households 219 and 281 in the sample, with 4800 and 4000 kilograms of maize availability. The two amounts are too big when they are compared with maize availability of the rest of the households in the sample. Thus the two observations are considered influential outliers since they are extreme with respect to the response variable, maize availability, and appear in the upper Cook's distance contour.

Table 5.3: Case Deletion Diagnostic for Maize Availability

Diagnostic Measure	219th Case Diagnostics	281st Case Diagnostics	Cutoff Point
RStudent t_i	10.92	8.28	2
Leverage h_{ii}	0.08	0.15	$2p/n = 0.04$
DFFITS	3.26	3.42	$2\sqrt{p/n} = 0.29$
Cook's Distance D_i	0.54	0.67	1
COVRATIO	0.01	0.06	$1 \pm 3p/n = 1 \pm 0.06$
DFBETAS			$2/\sqrt{n} = 0.12$
Income	2.75	0.14	0.12
Post High School	-1.00	1.89	0.12
Pensioner	0.14	1.51	0.12

The two observations identified as influential were studied further using the case deletion diagnostics to examine their influence on different quantities of he fitted model. According to the diagnostics results in Table 5.3, observations 219 and 281 are considered to be outlying observations since their absolute values of the externally studentized residual t_i are greater than the cutoff point of 2. This is inline with what is reflected in all the plots in Figure 5.1. In addition, both observations are considered as high leverage points because their values of h_{ii} are higher than the cutoff point of 0.04. Though the difference between the value of leverage and the cutoff point for observation 281 is relatively small. This means the observations are extreme with respect to both the response variable and at least one of the predictor variable.

The values of the DFFITS of 3.26 and 3.42 for observations 219 and 281 respectively, are both higher than the cutoff point of 0.29. This shows that the two observations have influence on the $219th$ and $281st$ fitted values, and excluding each of them one at time when fitting the model will change their respective fitted values, $\hat{y}_{219}$ and $\hat{y}_{281}$. The positive sign of the values of the DFFITS means that the $219th$, and $281st$ fitted values from fitting the model with all observations are larger than the two $219th$, and $281st$ fitted values from fitting the model without each of the observations 219 and 281. The values of COVRATIO for both observations are less than the cutoff point of $1 - 0.06 = 0.94$, indicating that the exclusion of each of the two observations affects the covariance

matrix of the estimated regression coefficients. The observations decrease precision of the estimates as their COVRATIO values are less than 1.

The absolute values of DFBETAS that are greater than the cutoff point of 0.12 is an indication that both observations affect parameter estimates of variables that correspond to large DFBETAS. Observation 219 affects the parameter estimates of household monthly income and one category of education level of household heads, post high school education. Observation 281 affects the parameter estimates of the category of heads with post high school education, and the category of heads who were pensioners.

The model was fitted without each of the observations 219 and 281, to examine how different quantities of the fitted model are affected by excluding each observation when fitting the model. When the model was fitted without observation 219, the value of R^2 increased from 0.30 to 0.35. This shows that the deletion of the influential observation improved the goodness-of-fit of the model slightly, since the proportion of the variability in household maize availability explained by predictor variables in the model increased from 30% to 35%. In addition, the deletion of this observation affected the significance of some of the variables in the model in terms of the significant effect they had on maize availability. For example, when the observation was excluded, household monthly income is not significant anymore and one category of education level of household head is significantly different from the reference category of heads with no formal education. The category that becomes significant is of household heads with beyond high school qualifications such as diploma and university degrees. The results suggest that observation 219 is more influential than observation 281 because when the model was fitted without observation 281, R^2 increased from 0.30 to 0.33 and the significance of the variables in the model was not affected.

5.3 Modelling of Sorghum Availability

The overall F test is significant at 5% level of significant, showing that at least one of the predictor variables in the model is linearly related to availability of sorghum. The value of R^2 of 0.19 suggests that predictor variables in the model explain only 19% of the variability of sorghum availability. Three variables, household size, the location of a household, and education level of the head of a household are significant at 5% level of significance (Table 5.4). This means that household size, the location of a household and education level of the head of a household had a significant effect on household sorghum availability.

Though Table 5.4 shows that the location of a household had a significant effect on household sorghum availability, the categories of this variables are not significantly different from the reference category, Berea (Table 5.5). This implies that location of a household had a significant effect on

Table 5.4: Sources of Variation and Type III Sum of Squares for Sorghum Availability

Response Variable: Sorghum availability to households

Source of Variation	DF	Type III Sum of Squares	Mean Square	F Value	Pr > F
Household Size	1	92309.05	92309.05	4.61	0.0332
Income	1	3549.43	3549.43	0.18	0.6741
Sex	1	1283.01	1283.01	0.06	0.8004
Location	3	161740.29	53913.43	2.70	0.0478
Education	3	271789.39	90596.46	4.53	0.0045
Occupation	4	114536.90	28634.22	1.43	0.2259

sorghum availability but the the effect of the other three locations of households was not significantly different from that of Berea. Like it was the case with maize availability, sex of household head did not have an effect on household sorghum availability. However, sorghum availability was affected by education level of household head, where all categories in the table are significantly different from the reference category of post high school education, at 5%. Only one category of occupation of household head was significantly different from the reference category of casual workers at 10%.

Table 5.5: Regression Parameter Estimates for Sorghum Availability

Response Variable: Sorghum availability to households

Variable	Category	Estimate	Standard Error	t value	P-value
Intercept		397.85	79.85	4.98	<.0001
Household Size		10.63	4.95	2.15	0.03318
Income		0.01	0.02	0.42	0.67414
Sex	Ref=Male				
	Female	6.41	25.30	0.25	0.80039
Location	Ref=Berea				
	Mafeteng	-44.62	36.32	-1.23	0.22098
	Maseru Foothill	-78.99	53.27	-1.48	0.14005
	Maseru Lowland	20.69	34.37	0.60	0.54800
Education	Ref=Post COSC				
	High School	-210.94	73.51	-2.87	0.00466
	No Formal Education	-250.68	71.47	-3.51	0.00059
	Primary	-245.50	70.24	-3.50	0.00059
Occupation	Ref=Casual Worker				
	Subsistence Farmer	-23.66	41.29	-0.57	0.56741
	Pensioner	-26.99	53.46	-0.50	0.61437
	Salary Earner	-84.41	46.15	-1.83	0.06925
	Unemployed	-64.23	40.53	-1.59	0.11494

The results in Table 5.5 show that it did not matter whether households were headed by a female or a male their sorghum availability was not significantly different. In addition it did not matter

where households were located their availability of sorghum was not significantly different from that of households residing in Berea. Furthermore, the occupation of household head did not make a significant difference in terms of household sorghum availability. All categories of education of a household head are significant at 5% level of significance, indicating that education level of a head of household played a significant role in determining availability of sorghum to a household. Sorghum availability for households headed by people with high school, primary and no formal education was less than that of households headed by people with post high school education by 210.94, 250.68 and 245.50 kilograms, respectively. The reason could be that household heads with diploma and university degrees qualifications got employed and earned better salaries than those with lower qualifications. Hence they were able to acquire more sorghum for their households, either through purchases or own produce.

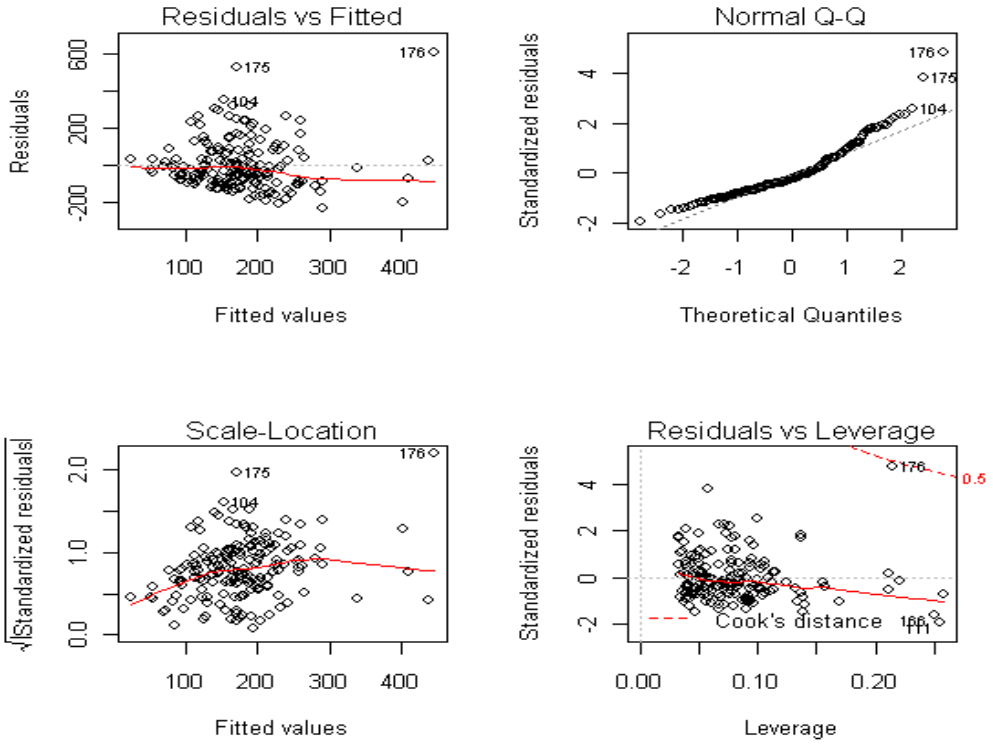

Figure 5.2: Plot of Residuals for Sorghum Availability

The plots on the top and bottom left of Figure 5.2 show a pattern of residuals that increase as the fitted values get bigger, which is a sign of heteroscedasticity. The normal QQ plot shows that most of the plotted points lie on a straight line with points on both right and left ends being above the line, and two outstanding cases. This suggests a symmetric distribution but with rather longer and heavier tails than the normal distribution. In addition the plot shows the presence of two outliers,

which are observations 175 an 176. The plot on the bottom right shows that observation 176 that stand out in the first three plots, is not an influential case since though it is close to the Cook's distance contour in the plot, but does not appear inside it.

5.4 Box-Cox Transformation of Household Data

The Box-Cox family of power transformation was applied to identify the appropriate power transformation used to correct the nonconstant variance problem and longer and heavier tails of the distribution of data. These were identified through skewness, kurtosis, box plots and diagnostic plots for the model fitted to the household maize and sorghum data, respectively. As it was mentioned earlier, the correction of violation of assumptions about the distribution of the error term increases the applicability and usefulness of linear regression models in modeling data. The TRANSREG procedure was used to perform a Box-Cox transformation of the response variable. The procedure gives an option of specifying a list of transformation parameters and chooses the optimal value of the parameter.

In the case of household maize data, the estimated optimum value of the transformation parameter $\hat{\lambda}$ for maize availability is 0.20. The plot of the log likelihood against values of λ shows an approximate 95% confidence interval of λ as $[0.11, 0.30]$ (Figure 5.3). The confidence interval does not contain any of the standard power transformation values that could have been used as the nearest convenient value of $\hat{\lambda}$. Thus the estimated optimum value 0.20 is used to transform the response variable.

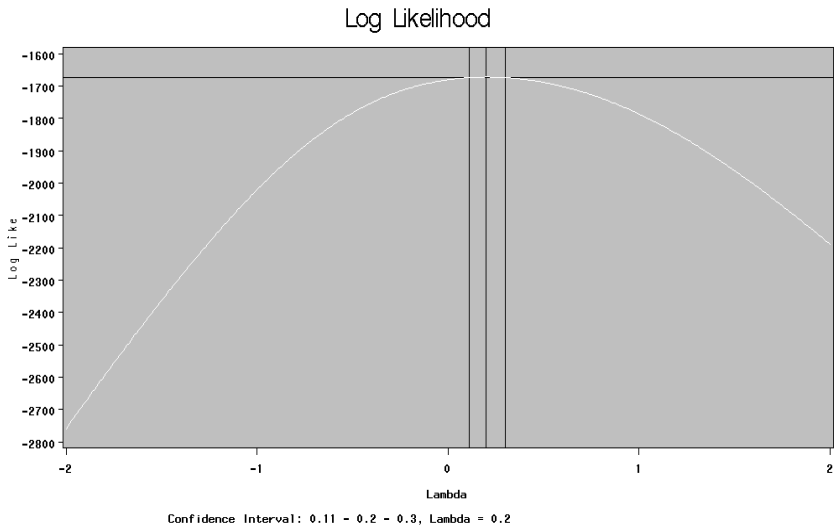

Figure 5.3: Box-Cox Plot for Maize Availability

The values of skewness and kurtosis of the transformed maize availability are 0.03 and 0.99 respectively. They are less than the ones of the maize availability in its original form, where the values are 3.60 for skewness and 24.36 for kurtosis. The values for the transformed data are close to zero, indicating that the transformation corrected the positive skewness and leptokurtic distribution observed in Chapter 2. Thus sorghum availability becomes approximately normally distributed after the transformation. When we compare the box plots of the transformed maize availability in Figure 5.4 and untransformed maize availability in Figure 2.7, we notice that the transformation corrected the deviation of the distribution of maize availability from the normal distribution, though the plot shows the presence of extreme values. The correction is shown by the line inside the box, which is still not equidistant to the lower and upper edges of the box, but shifted from being very close to the lower edge to getting nearer to the middle of the box.

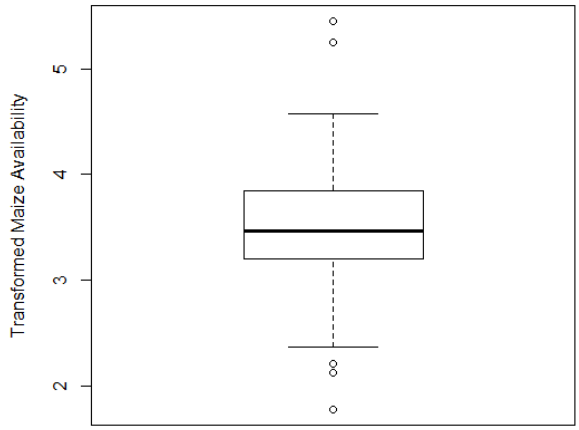

Figure 5.4: Box Plot of Transformed Household Maize Availability

The interpretation and conclusions from the results from fitting the model in which the transformed availability of maize was used as the response response variable remain the same as the the ones of the results obtained using the untransformed response variable in Table 5.2. Household size, monthly income and occupation of household head had a significant effect on the transformed availability of maize. Similarly, the transformed maize availability of households headed by casual workers and unemployed people was significantly different from that of households headed by subsistence farmers. However the strength of evidence for both categories increased, while the strength of evidence that income had an effect on maize availability decreased. These increase and decrease in the strength of evidence are shown by the p-value of casual workers that decreased from 0.0357

to 0.0053, the p-value of unemployed that decreased from 0.0281 to 0.0056 and the p-value of income that increased from 0.0018 to 0.0293. The transformation of availability of maize using the suggested value of the transformation parameter 0.20 improved the fit of the model slightly since the value of R^2 increased from 0.30 to 0.38.

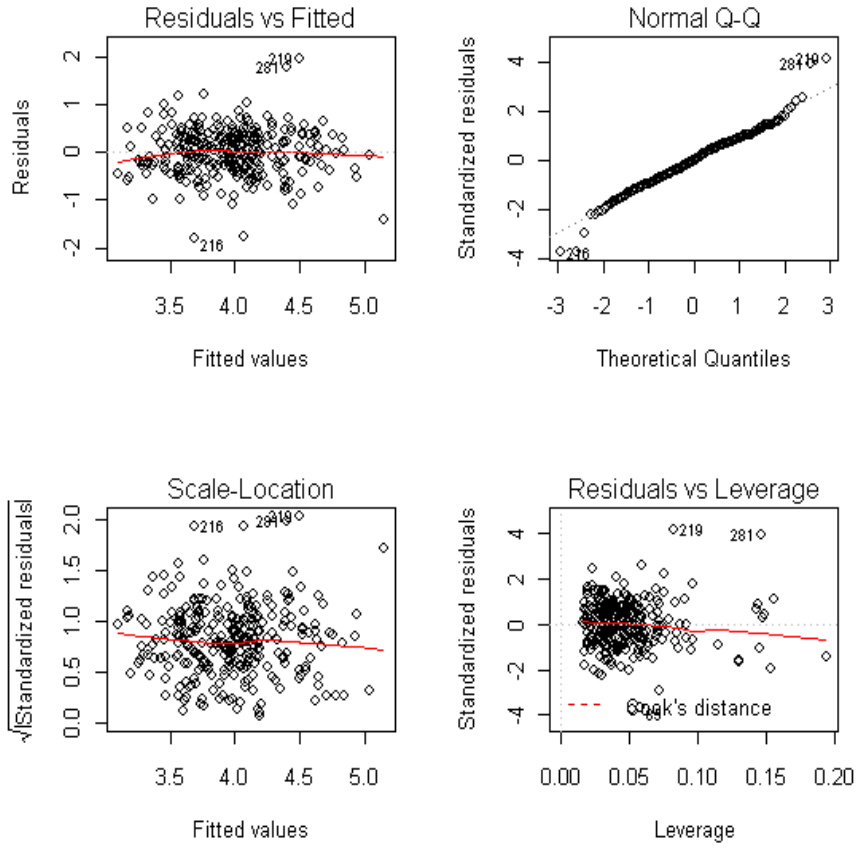

Figure 5.5: Plot of Residuals for Transformed Maize Availability

The residual plots in Figure 5.5 were used to check further if the assumptions violated when the regression model was fitted using the untransformed response variable are corrected by using the suggested transformation of the response variable. The plots on the top and bottom left show a random pattern of the points, indicating that the non constant variance observed in Figure 5.1 was corrected by transforming the data. The plot on the top right shows that the long heavy tail of the distribution is also corrected but partially. The two outlying observations, 219 and 281 are not influential anymore since they do not appear in the Cook's distance contour, as it was the case in Figure 5.1.

In the case of household sorghum data, the values of skewness and kurtosis of the transformed sorghum availability are -0.76 and -0.31 respectively, which are less than the ones of the untransformed sorghum availability [skewness = 1.78 and kurtosis = 5.96]. The values are close to zero, indicating that the transformed sorghum availability is approximately normally distributed. On the other hand, when we compare the box plots of the transformed sorghum availability in Figure 5.6 and untransformed sorghum availability in Figure 2.8, we observe that the transformation made the deviation from the assumption of normality worse, instead of correcting it. The line inside the box shifted from being nearer to the middle of the box towards the upper edge, while it was closer to being equidistant to the lower and upper edges of the box, in the case of untransformed data. The contracting observations from the two statistics (skewness and kurtosis) and the box plots suggests that several statistical tools need to be explored for detecting the distribution of data as one may not be enough to give the true picture of the distribution.

The estimated optimum value of the transformation parameter $\hat{\lambda}$ for sorghum availability is 0.30. The plot of the log likelihood against values of λ shows an approximate 95% confidence interval for λ as $[0.19, 0.42]$ (Figure 5.7). The confidence interval does not contain any of the standard power transformation values that could have been used as the nearest convenient value of $\hat{\lambda}$. Thus the estimated optimum value of 0.30 is used to transform the sorghum availability.

The interpretation of the results from fitting the model using the transformed availability of sorghum changes slightly from that of the results from fitting the model using the availability of sorghum in

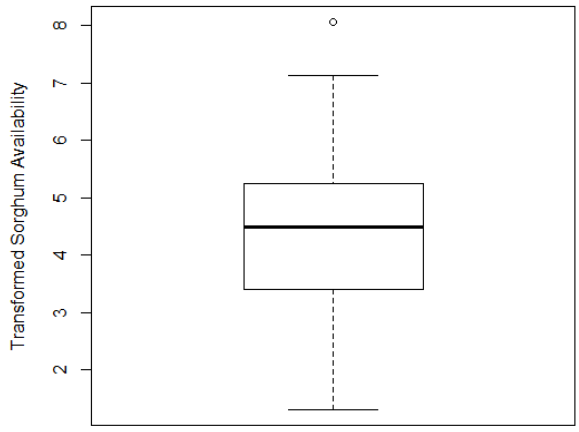

Figure 5.6: Box Plot of Transformed Household Sorghum Availability

its original form, shown in Table 5.2. Education level and occupation of a household head still had a significant effect on household sorghum availability. However, unlike in the case of the untransformed data, two categories of occupation of the head of a household are significantly different from the reference category of casual worker at 5% level of significant. The two categories are of salary earners and unemployed people, with negative estimates. This means that when the transformed data were used, sorghum availability for households headed by unemployed people, in addition to that of households headed by salary earners, was less than that of households headed by casual workers.

Furthermore, the strength of evidence for the significant difference between sorghum availability to households headed by people who attained high school education and sorghum availability to households headed by people who attained education beyond high school, decreased when the transformed data were used. However, the strength of evidence for the significant difference between sorghum availability to households headed by salary earners and sorghum availability to households headed by casual workers, increased. The transformation of the response variable using the suggested value of the transformation parameter 0.30 improved the fit of the model slightly by increasing the value of R^2 from 0.19 to 0.21.

The Residual plots in Figure 5.8 show that the heterogeneous variance of the error term is not corrected by transforming the data, while long heavy tails of the distribution of the error terms observed in Figure 5.2 are corrected. The plots on the top and bottom left of the figure show a pattern where residuals increase with the fitted values up to some point, and they start decreasing as fitted values increase. This suggests that the heterogeneity is not corrected by transforming

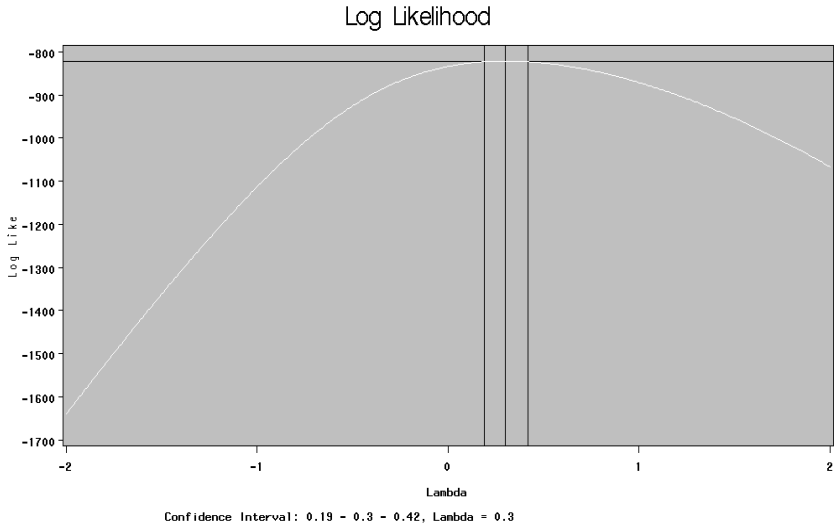

Figure 5.7: Box-Cox Plot for Sorghum Availability

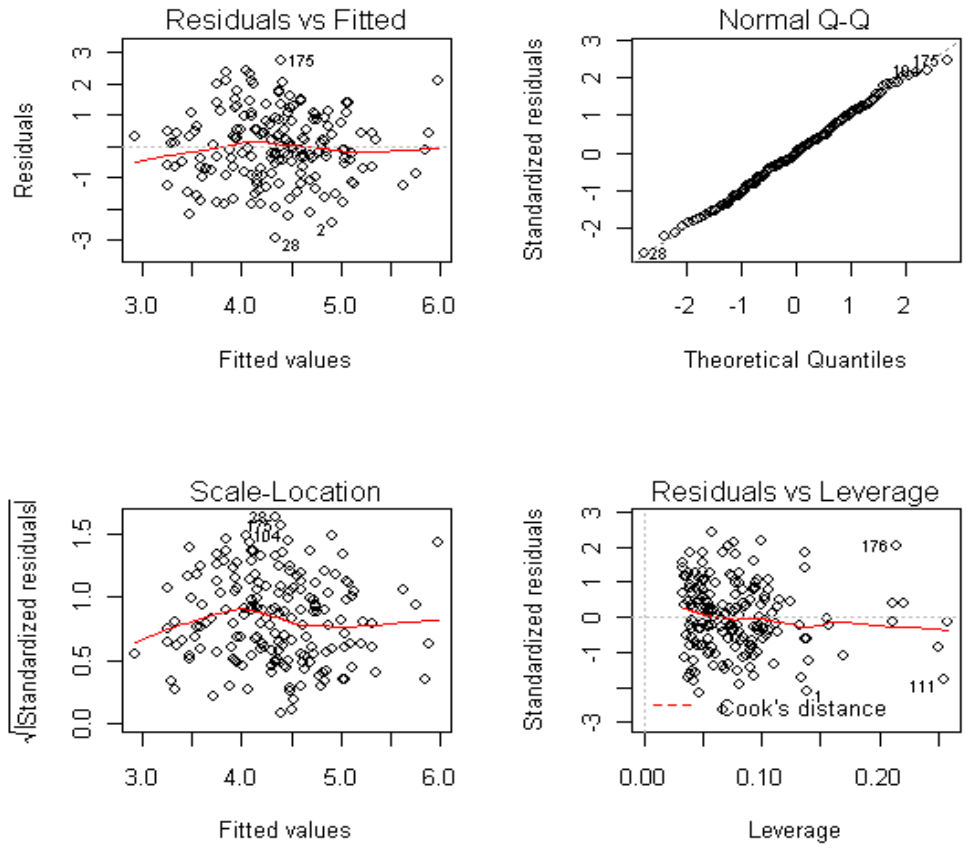

Figure 5.8: Plot of Residuals for Transformed Sorghum Availability

sorghum availability, using the estimated optimum value of 0.30. The QQ plot shows that the tail of the distribution of the error terms that was longer and heavier than the normal distribution is corrected. After transforming the data observation 176 does not appear near the Cook's distance contour in Figure 5.8, as it was the case in Figure 5.2.

5.5 Summary

The results from fitting the linear model show that both household size and monthly income had a significant positive effect on maize availability for households, while household monthly income did not have a significant effect on sorghum availability. The occupation of heads of households had a significant effect on availability of maize. Maize availability for households headed by casual workers and unemployed people was less than that of households headed by subsistence framers. This is not surprising because subsistence farming is an occupation devoted to agricultural production for subsistence including production of food cereals. On the other hand, maize availability for households headed by pensioners and salary earners was not significantly different from that of households headed by subsistence farmers. Education level of heads of households played a significant role in determining sorghum availability to households. Sorghum availability for households headed by heads with lower education qualifications was less than that of households headed by people with qualifications beyond high school such as diploma and university degrees.

Two observations that were identified as extreme with respect to maize availability influenced the fitted values, estimated regression coefficients, and decreased precision of the estimates. Statistical tools such as skewness, kurtosis, box plots and diagnostics plots were used to check if the use of Box-Cox transformation to correct the deviation of the distributions of maize availability and sorghum availability from normality worked. In some cases the tools gave agreeing results that the Box-Cox transformation worked or did not work, while in other cases they gave contradicting results. The contradicting results call for further research where these tools will be investigated to establish causes of such contradictions. The failure to correct the violation of normality and constant variance assumptions on the part of the the Box-Cox transformation called for the application of quantile regression model as a robust alternative to OLS approach, which deals with the presence of outliers, homoscedasticity and symmetric and asymmetric distributions with long and heavy tails.

Chapter 6

Robust Regression

6.1 Introduction

Robust regression refers to a general class of statistical procedures designed to reduce the sensitivity of the estimates to departures from the assumptions of the parametric model (Rawlings et al., 1998). One of the objectives of robust regression procedures is to reduce the influence of outlying influential cases in order to have a better fit for the majority of cases in the data set. The procedures outperform the ordinary least squares procedure when the data are not well behaved in the sense that there are outliers in the data and the error terms have non-normal distribution with longer and heavier tails than the normal distribution (Koenker and Bassett, 1978). However, they should perform almost as well as least squares estimation procedures when the data are well behaved. Thus the use of robust regression in an analysis will not compromise the benefit that one would get even when the data are well behaved. The advantage is that it will control for any undetected lack of adherence to ordinary least squares assumptions, if any exists. Robustness of regression estimation procedure can be from the point of view of its stability when some of the regression model assumptions are violated. An estimation procedure can be robust to an incorrectly specified model, heterogeneous variances, or the contamination of data by outliers.

Despite their appealing properties and computational simplicity, least squares estimates are notoriously known for their lack of robustness (Koenker, 2005). The application and optimality of ordinary least squares (OLS) estimation procedure in regression analysis requires the assumption that the error terms have a normal distribution, and the mean regression assumption. In practice distributions of the error terms with longer and heavier tails than that of the normal distribution are commonly encountered, and more generally the distribution of the response variable Y may not be classified as Gaussian. When the errors have a non-normal distribution, specifically a heavy-tailed distribution, least squares estimates may lose their efficiency. Least squares estimation is strongly influenced by the presence of outliers, sometimes referred to as model shifts, and heavy-tailed error

distributions. This influence is caused by the pull of the regression line towards the deviant data points. The extreme sensitivity of the least squares estimates to moderate contamination due to outlying cases, renders them to perform poorly in heavy-tailed distributions and thus outliers pose serious threat to ordinary least squares estimates (Rousseeuw and Leroy, 2003). The deficiencies in least squares estimates under the circumstances stated above call for the use of robust regression estimation procedures as alternatives.

6.2 Preliminary Concepts of Robustness

When the distribution of the error terms is normal the conditional mean of least squares is the best method of estimation, whereas the median regression is less effective (Kolmogorov, 1931). However, when there are outliers, the error distribution is not known and deviates from normality, the median regression is preferable (Yu et al., 2003). The nature of least squares that it minimizes the sum of the squared residuals, allows outliers to exert disproportionate influence on regression results. Unlike the robust regression procedures which have the ability to weight observations unequally in finding parameter estimates, the least squares weights each observation equally. On the other hand observations that gives large residuals are down-weighted under robust estimation procedures. Thus the squaring of residuals in OLS procedure gives more weight to large residuals than when their absolute values are considered. In cases where error terms have a distribution with heavier tails than the normal distribution the least squares is no longer an optimal estimation procedure and robust estimation procedures are preferred.

The classical robust estimation theory provides methods of dealing with the sensitivity of the sample mean to departures from the distributional assumptions of the observations. The theory is based on the idea of the estimating equation $\psi(y, \theta)$, which can take various forms. The objective of robust estimation is to reduce the influence of outlying observations by controlling the function $\psi(.)$, which controls the weight assigned to each residual and is sometimes called the influence function. Pawitan (2001) and Montgomery et al. (2001) showed that the least squares function is unbounded and as a result outlying observations have unlimited influence on estimates.

The theory of robust estimation of a location parameter was developed by (Huber, 1964). His estimation procedures were based on the idea of replacing the sum of squared residuals $\sum_{i=1}^{n} e_i^2$ with another function of the residuals, $\sum_{i=1}^{n} \rho(e_i)$, which weight residuals differently depending on the importance or contribution of the corresponding observation. There are a number of robust estimators which are maximum likelihood type estimators reported in literature as alternatives to ordinary least squares estimators. The least median of squares (LMS) estimator by Rousseeuw (1984) minimizes the median of the squared residuals, and yields the smallest value of the median of squared residuals computed from the entire data set. The least trimmed squares (LTS) estimator

also suggested by Rousseeuw (1984) minimizes the ordered squared residuals. The least absolute deviations (LAD) estimation procedure of the conditional median or L_1 procedure minimizes the sum of absolute values of the residuals as

$$\min_{\hat{\beta}} \sum_{i=1}^{n} |y_i - \mathbf{x}_i' \hat{\beta}| \tag{6.1}$$

The LAD procedure was introduced by Edgeworth in 1887 as the first initiative on a more robust regression estimator (Rousseeuw, 1984). This estimation procedure generalizes the median of a one-dimensional sample to the conditional median regression. The median regression model estimates the effect of predictor variables on the conditional median and thus it represents the central location even when the distribution is skewed. Koenker and Bassett (1978) noted that many graphical illustration by Gauss, Laplace and Legendre indicated that minimizing absolute deviations may be preferred to minimizing sum of squared deviations when some observations are not well behaved. The LAD estimation procedure is more robust than the OLS estimation procedure for asymmetric distribution and heavy-tailed distributions, and to influence of observations that are extreme in terms of the response variable (DasGupta and Mishra, 2004). The procedure puts less emphasis on outlying observations than the OLS procedure because it involves absolute deviations rather than squared deviations (Rousseeuw and Leroy, 2003).

The median is a special quantile that describe the central location of the distribution. Hence the conditional median regression is a special case of quantile regression where the $50th$ quantile, is modeled as a function of predictor variables. Similarly, a full range of other quantiles can be modeled as functions of predictor variables. Like other robust regression procedures, quantile regression provides more robust estimators that are as efficient as the least squares estimator for normal distributions but more efficient for continuous non-normal distributions where the least squares estimator may be seriously deficient. In the current work quantile regression was applied as an alternative to deal with the sensitivity of classical regression to long-tailed distributions and outliers.

6.3 Qauntile Regression

Quantile regression was introduced by Koenker and Bassett (1978) as an alternative robust and flexible estimation approach to classical regression approach. It estimates functional relationships between a response variable Y and predictor variables X for all parts of the response variable distribution. Unlike the classical regression model, which is confined to the estimation of the conditional mean of the distribution of Y given a set of predictor variables, quantile regression allows for the estimation of various quantile functions of the conditional distribution of Y given X and thus estimates the entire distribution of the response variable. In so doing quantile regression provides a more comprehensive explanation or view of the possible relationship between variables than when

the mean is exclusively used (Buchinsky, 1998). Koenker (2005) suggested that the conditional mean estimated by least squares approach needs to be supplemented with the estimates of conditional quantiles. The off-median conditional quantiles have the ability to distinguish between the location shift and the shape shift. This feature of conditional quantiles is critical in determining the effect of a predictor variable on the location and shape shift of the conditional distribution of the response variable.

Quantile regression is the generalization of the concept of a sample quantile to the conditional quantile of the response variable Y given one or more predictor variables X. The sample median is the $50th$ quantile denoted by $\hat{Q}_Y(0.50)$. Like the mean, which is obtained as the solution to the problem of minimizing a sum of squared deviations $\min_{\mu\in\mathbb{R}} \sum_{i=1}^{n}(y_i-\mu)^2$, the median can be obtained as the solution to the problem of minimizing the sum of absolute deviations $\min_{\xi\in\mathbb{R}} \sum_{i=1}^{n}|y_i-\xi|$. The τth sample quantile $\hat{\xi}_\tau$ can also be obtained as the solution to the problem of minimizing an asymmetric weighted absolute deviations. This problem assigns different weights to positive and negative deviations. The optimization problem is defined as

$$\min_{y_i\in\mathbb{R}}\Big(\sum_{i:y_i\geq\xi}\tau|y_i-\xi|+\sum_{i:y_i<\xi}(1-\tau)|y_i-\xi|\Big)=\min_{y_i\in\mathbb{R}}\sum_{i=1}^{n}\rho_\tau(y_i-\xi), \tag{6.2}$$

where $\rho_\tau(u)=\tau|u|I(u\geq 0)+(1-\tau)|u|I(u<0)$ is called the check function (Koenker and Bassett, 1978). The check function can also be expressed as $\rho_\tau(u)=(\tau-I(u<0))u$, where $I(.)$ is the indicator function that assign 1 to negative residuals and 0 to positive deviations. In this case finding the τth sample quantile is expressed as the solution to an optimization problem instead of the common method of ordering the sample observations.

6.3.1 Quantile Regression Model

The quantile regression model of Koenker and Bassett (1978) can be expressed as

$$\begin{aligned} y_i=\mathbf{x}_i'\boldsymbol{\beta}_\tau+u_{\tau_i} \quad \text{with} \quad Q_\tau(y_i|\mathbf{x}_i)=\mathbf{x}_i'\boldsymbol{\beta}_\tau \\ \text{and} \quad Q_\tau(u_{\tau_i}|\mathbf{x}_i)=F_u^{-1}(\tau|\mathbf{x}_i)=0, \end{aligned} \tag{6.3}$$

where y_i is the ith observation of the response variable, $\mathbf{x}_i$ is a vector of predictor variables, $\boldsymbol{\beta}_\tau$ is a vector of unknown regression parameters, and u_{τ_i} are independent identically distributed (iid) error terms with unspecified distribution. The quantities, $Q_\tau(y_i|\mathbf{x}_i)$ and $Q_\tau(u_{\tau_i}|\mathbf{x}_i)$ denote the τth conditional quantiles of y_i and u_{τ_i} given $\mathbf{x}_i$, respectively. Following that the τth sample quantile, $Q_Y(\tau)=\xi_\tau$, $(0\leq\tau\leq 1)$, of a random variable Y is the inverse of the cumulative distribution function, $F_Y(y)=\tau$, defined as

$$Q_Y(\tau)=F_Y^{-1}(\tau)=\inf\{y:F_Y(y)\geq\tau\}, \tag{6.4}$$

$Q_\tau(y_i|\mathbf{x}_i)$ can be defined as the inverse of the cumulative distribution function of the response variable conditional to the predictor variable, $F_Y^{-1}(\tau|X)$.

The model in equation (6.3) is referred to as the linear location model where predictor variables affect only the location of the conditional distribution of the response variable. When error terms are independent and identically distributed the τth regression parameter $\boldsymbol{\beta}_\tau = \boldsymbol{\beta} + (F_u^{-1}(\tau), 0, \ldots, 0)'$. This is the case where conditional quantile planes are parallel and all parameters in $\boldsymbol{\beta}$, except the intercept, are similar for every value of τ. Thus quantile regression slopes are constant for every quantile τ. However, when error terms are not independent and identically distributed the quantile regression model is the linear location-scale model of heteroscedasticity, which can be stated as

$$\begin{aligned} y_i &= \mathbf{x}_i'\boldsymbol{\beta}_\tau + (\mathbf{x}_i'\gamma)u_{\tau_i} \quad \text{with} \\ Q_\tau(y_i|\mathbf{x}_i) &= \mathbf{x}_i'\boldsymbol{\beta}_\tau + \mathbf{x}_i'\gamma F_u^{-1}(\tau), \end{aligned} \tag{6.5}$$

where γ is an unknown scale parameter. In the case of iid error terms in equation (6.3), the scale parameter $\gamma = (1, 0, \ldots, 0)$, meaning that the τth conditional quantiles of y_i given $\mathbf{x}_i$ depends on X only in location. The linear location-scale model of heteroscedasticity is an important case of general class of quantile regression models including the iid error terms case (Koenker and Basett, 1982a). In this model predictor variables affect the location as well as the scale of the response variable distribution and result into heteroscedastic error models. In such models there is no single rate of change that represents changes in the distribution since regression slopes vary across all parts of the distribution of the response variable, and thus the τth regression parameter $\boldsymbol{\beta}_\tau = \boldsymbol{\beta} + \gamma F_u^{-1}(\tau)$.

6.3.2 Estimation of Regression Quantiles

In a similar manner that the optimization problem for the sample mean can be generalized to the linear conditional mean function of Y given X, $E(y_i|\mathbf{x_i}) = \mathbf{x}_i'\boldsymbol{\beta}$, the optimization problem for sample quantiles in equation (6.2) can be generalized to the estimation of conditional quantiles. The τth quantile regression estimator $\hat{\boldsymbol{\beta}}_\tau$, also called the regression quantile, is obtained by minimizing an asymmetric sum of weighted absolute deviations for the τth regression quantile ($0 \leq \tau \leq 1$) defined as

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \Big(\sum_{i:y_i \geq \mathbf{x}_i'\boldsymbol{\beta}_\tau} \tau|y_i - \mathbf{x}_i'\boldsymbol{\beta}_\tau| + \sum_{i:y_i < \mathbf{x}_i'\boldsymbol{\beta}_\tau} (1-\tau)|y_i - \mathbf{x}_i'\boldsymbol{\beta}_\tau| \Big) = \min_{\boldsymbol{\beta}\in\mathbb{R}^p} \sum_{i=1}^{n} \rho_\tau(y_i - \mathbf{x}_i'\boldsymbol{\beta}_\tau), \tag{6.6}$$

where $\rho_\tau(u)$ is as defined under equation (6.2). Positive and negative residuals are assigned the weights of τ and $1-\tau$, respectively. The LAD estimator of $\boldsymbol{\beta}$ obtained by minimizing a symmetric sum of weighted absolute deviation is a special case of quantile regression for $\tau = 0.5$, which is the median or L_1 regression.

The minimization of the weighted sum of absolute deviations in equation (6.6) can be formulated as a linear programming (LP) problem, which can be solved using a linear programming algorithm. Literature reported a number of algorithms that can be used to solve the linear programming problems for quantile regression. The algorithms include the simplex-based algorithm for median regression of Barrodale and Roberts (1974) adapted by Koenker and D'Orey (1987) for quantile regression, the interior point algorithm by Karmarkar (1984), referred to as the Frisch-Newton algorithm by Koenker and Hallock (2000), the finite smoothing algorithm of Madsen and Nielsen (1993) extended to quantile regression by Chen (2007), and the combination of the statistical preprocessing and interior point methods proposed by Portnoy and Koenker (1997).

Interpretation of quantile regression parameter estimates is not different from that of the general linear model estimates since they are all rates of change when the effects of some variables in the model are adjusted for. The classical regression coefficient reflects the change in the mean of the distribution of the response variable Y, associated with a unit change in the predictor variable X that corresponds to the coefficient. However, the quantile regression coefficient reflects the change in a specified quantile of the response variable associated with a unit change in that predictor variable. The use of quantile regression allows for comparison of how some percentiles of the response variable may be more affected by certain predictors than other percentiles. This is reflected in the change in the size of the regression coefficients of different percentiles.

6.4 Properties of Quantile Regression Estimates

6.4.1 Properties of Equivariance

The quantile regression estimates have a number of equivariance properties, which are important for meaningful interpretation of results from regression analysis, particularly for transformed data. When the data are altered, the expectation is that regression estimates also change in such a way that the interpretation of the results remain invariant (Koenker, 2005). Thus equivariance properties play a critical role in ensuring an expressive interpretation of statistical results. Considering the τth regression quantile estimate based on observations (y_i, x_i) denoted by $\hat{\boldsymbol{\beta}}(\tau; \mathbf{y}, \mathbf{X})$, the equivariance properties by Koenker and Bassett (1978) and Bassett and Koenker (1982) are defined as

$$\hat{\boldsymbol{\beta}}(\tau; \lambda\mathbf{y}, \mathbf{X}) = \lambda\hat{\boldsymbol{\beta}}(\tau; \mathbf{y}, \mathbf{X}) \qquad \lambda \in [0, \infty) \tag{6.7}$$

$$\hat{\boldsymbol{\beta}}(1-\tau; \lambda\mathbf{y}, \mathbf{X}) = \lambda\hat{\boldsymbol{\beta}}(\tau; \mathbf{y}, \mathbf{X}) \qquad \lambda \in (-\infty, 0] \tag{6.8}$$

$$\hat{\boldsymbol{\beta}}(\tau; \mathbf{y} + \mathbf{X}\gamma, \mathbf{X}) = \hat{\boldsymbol{\beta}}(\tau; \mathbf{y}, \mathbf{X}) + \gamma \quad \gamma \in \mathbb{R}^p \tag{6.9}$$

$$\hat{\boldsymbol{\beta}}(\tau; \mathbf{y}, \mathbf{X}\mathbf{A}) = \mathbf{A}^{-1}\hat{\boldsymbol{\beta}}(\tau; \mathbf{y}, \mathbf{X}) \quad \mathbf{A}_{\mathrm{p}\times\mathrm{p}} \text{ nonsingular.} \tag{6.10}$$

The properties in equations (6.7) and (6.8) suggest that $\hat{\boldsymbol{\beta}}_\tau$ is scale equivariant. This means that if the vector of responses $\mathbf{y}$ is adjusted by a factor λ, then the quantile regression $\hat{\boldsymbol{\beta}}_\tau$ is adjusted by the same factor. The property in equation (6.9) suggests that $\hat{\boldsymbol{\beta}}_\tau$ is location, shift or regression equivariant, which means that if $\hat{\boldsymbol{\beta}}_\tau$ is a solution to $(\mathbf{y}, \mathbf{X})$, then $\hat{\boldsymbol{\beta}}_\tau + \gamma$ is the solution to $(\mathbf{y}^*, \mathbf{X})$, where $\mathbf{y}^* = \mathbf{y} + \mathbf{X}\gamma$. The property in equation (6.10) suggests that $\hat{\boldsymbol{\beta}}_\tau$ is equivariant to reparameterization of design and it means that the transformation of $\hat{\boldsymbol{\beta}}_\tau$ is given by the inverse transformation of $\mathbf{X}$. The four equivariance properties also hold for least squares estimate $\hat{\boldsymbol{\beta}}$.

6.4.2 Property of Equivariance to Monotone Transformation

Quantile regression estimates have another more powerful equivariance property (Koenker, 2005). This property is important for a complete understanding of the potential of quantile regression in studying relationships between variables. The property is called equivariance to monotone transformations and it is stated as follows: if $h(.)$ is a nondecreasing function on $\mathbb{R}$

$$Q_\tau(h(Y)|X) = h(Q_\tau(Y|X), \tag{6.11}$$

for any random variable Y. This property suggests that conditional quantiles are equivariant to monotone transformation of the response variable in the sense that the conditional quantile of the transformed random variable $h(Y)$ is the transformed conditional quantiles of the original variable Y. The equivariance to monotone transformations property follows from the fact that $P(Y \leq y) = P(h(Y) \leq h(y))$, but does not hold for the conditional mean because $E(h(y)|X) \neq h(E(y|X))$, unless the function $h(.)$ is affine. This property makes the interpretation of transformations in quantile regression easier than they are in classical regression.

6.4.3 Property of Robustness

Another important property of quantile regression estimates is that of robustness to observations that are extreme with respect to the response variable Y, referred to as outliers. This property is critical in this research as it constitute part of motivation for quantile regression estimates to serve as an alternative to ordinary least squares estimates, which are highly sensitive to outlying observations. The robustness of quantile regression with respect to outliers means that if $y_i - \mathbf{x}_i'\hat{\beta}_\tau > 0$, or $y_i - \mathbf{x}_i'\hat{\beta}_\tau < 0$, y_i can be increased or decreased toward ∞ or $-\infty$, respectively without changing the solution $\hat{\boldsymbol{\beta}}_\tau$ (Buchinsky, 1998). In other words what is important in estimating regression quantiles is whether the observation lies above or below the estimated hyperplane, not necessarily the magnitude of the observations on the response variable. This is not the case with the least squares estimates as they are overly sensitive to outliers. However, quantile regression estimates lack robustness against observations that are extreme with respect to predictor variables, called high leverage points.

The property of robustness of an estimator can be quantified using two measures of robustness, namely, the breakdown point and influence function of an estimator. The two measures are designed to assess robustness of an estimator at two different levels and thus they complement each other in quantifying robustness of an estimation method. The influence function by Hampel (1974) deals with local robustness, while the breakdown point introduced by Hampel in 1968 deals with global robustness. The influence function describes how the estimator $\hat{\theta}$ from an underlying distribution F is affected by contaminating or perturbing the distribution. It measures sensitivity to change in the distribution caused by infinitesimal contamination and indicates how quantile regression estimates, like the least squares estimates, it can be highly influenced by high leverage points. The breakdown point of an estimator measures the smallest fraction of the data contaminated with outliers that can make the estimator to take values that are far from the initial vector of regression estimates $\hat{\theta}$. The main idea in robustness is to construct high breakdown point estimators such as the least median of squares by Rousseeuw (1984) with a breakdown point of 0.50.

6.5 Goodness-of-Fit for Quantile Regression

The goodness-of-fit measure for quantile regression proposed by Koenker and Machado (1999) derives from the classical R^2 of the ordinary least squares regression estimation procedure. The measure compares the quantile regression model fitted with a given number of predictor variables including the intercept, and the model fitted with the intercept only. Consider the quantile regression model $y_i = \mathbf{x}_i'\boldsymbol{\beta}_\tau + u_{\tau_i}$, which can be partitioned and expressed as

$$Q_\tau(y_i|\mathbf{x}_i) = \mathbf{x}_{i1}'\boldsymbol{\beta}_{1\tau} + \mathbf{x}_{i2}'\boldsymbol{\beta}_{2\tau}. \tag{6.12}$$

The partitioned model presented above results from partitioning the design matrix $\mathbf{X}$ into $(\mathbf{X}_1, \mathbf{X}_2)$, and the vector of parameters $\boldsymbol{\beta}_\tau$ into $\boldsymbol{\beta}_{1\tau}$ and $\boldsymbol{\beta}_{1\tau}$. The components of the model $\mathbf{x}_{i1}$ and $\mathbf{x}_{i2}$ are the ith rows of $\mathbf{X}_1$ and $\mathbf{X}_2$, which are the $n \times (p-q)$ and $n \times q$ design matrices, respectively. The components $\boldsymbol{\beta}_{1\tau}$ and $\boldsymbol{\beta}_{2\tau}$ are $(p-q) \times 1$ and $q \times 1$ vectors of parameters, respectively.

The unrestricted τth quantile regression estimate $\hat{\boldsymbol{\beta}}_\tau$, of the full model, minimizes the weighted sum of absolute deviations given by

$$\hat{V}_\tau = \min_{\hat{\boldsymbol{\beta}}_\tau \in \mathbb{R}^p} \sum_{i=1}^{n} \rho_\tau(y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}_\tau). \tag{6.13}$$

Consider the restricted model, which can be defined as $Q_\tau(y_i|\mathbf{x}_i) = \mathbf{x}_{i1}'\boldsymbol{\beta}_{1\tau}$. Then the restricted estimator $\tilde{\boldsymbol{\beta}}_\tau = (\tilde{\boldsymbol{\beta}}_{1\tau}', \mathbf{0}')'$, which is the τth quantile regression estimate under the q dimensional linear restriction corresponding to null hypothesis $H_0 : \boldsymbol{\beta}_{2\tau} = 0$, minimizes the corresponding constrained problem with q restrictions given by

$$\tilde{V}_\tau = \min_{\hat{\boldsymbol{\beta}}_1 \in \mathbb{R}^{p-q}} \sum_{i=1}^{n} \rho_\tau(y_i - \mathbf{x}_{1i}'\hat{\boldsymbol{\beta}}_{1\tau}). \tag{6.14}$$

Thus the goodness-of-fit criterion can be defined in terms of the two objective functions, $\hat{V}_\tau$ and $\tilde{V}_\tau$ as

$$R^1_\tau = 1 - \frac{\hat{V}_\tau}{\tilde{V}_\tau}. \tag{6.15}$$

Like the classical R^2, $0 \leq R^1_\tau \leq 1$, since $\hat{V}_\tau \leq \tilde{V}_\tau$. However unlike R^2, which is a global measure of goodness-of-fit over the entire distribution of the response variable measuring the relative success of the unrestricted and restricted models, R^1_τ serves as a local measure of goodness-of-fit measuring the relative success of the regression model at a specific quantile. Under certain circumstances R^1_τ may be high at one tail of the distribution than at the other tail, which could be an indication of heteroscedasticity (Melly, 2001). If the full model in equation (6.12) is better at the τth quantile than the restricted model, $\hat{V}_\tau$ should be significantly smaller than $\tilde{V}_\tau$ and R^1_τ will be high indicating a better model fit. Better in this context means that the predictor variables $\mathbf{X}_2$ has a significant effect at the τth quantile (Koenker and Machado, 1999).

The assessment of the goodness-of-fit through the comparison of the two objective functions, $\hat{V}_\tau$ and $\tilde{V}_\tau$ forms the basis for constructing some of the tests of the linear hypothesis $H_0 : \boldsymbol{\beta}_{2\tau} = 0$, presented in the subsequent section.

6.6 Inference for Quantile Regression

In classical regression analysis, the conditional quantile functions of the response variable given predictor variables in the model are assumed to be all parallel to one another. This means that the effects of predictor variables in the model shift the location of the conditional distribution of the response variable only, but do not change its scale or shape and thus the slope coefficients of distinct quantile regressions are equal. However in many applications of quantile regression estimated slopes often differ substantially across quantiles and this makes the test of equality of slope parameters across quantiles to form a fundamental component of inference in quantile regression (Koenker, 2005). Though there are no practical statistical inference in the case of finite sample for quantile regression, like it is the case in least squares, the asymptotic theory provides practical statistical inference for quantile regression. The theory forms the basis for statistical tests such as, the Wald, rank and likelihood ratio tests and construction of some confidence intervals for regression quantiles. The asymptotic theory is based on (Koenker, 2005) work.

6.6.1 Asymptotics of Quantile Regression

The asymptotic distribution of quantile regression estimator $\hat{\boldsymbol{\beta}}_\tau$ follows from that of sample quantiles. The asymptotic distribution of the sample quantile, $\hat{\xi}_\tau$, calculated from the n independent

identically distributed (iid) observations of the response variable with the distribution function F, can be defined as

$$\sqrt{n}(\hat{\xi}_\tau - \xi_\tau) \rightarrow \mathcal{N}(0, \omega^2), \tag{6.16}$$

where $\omega^2 = \frac{\tau(1-\tau)}{f^2(F^{-1}(\tau))}$. This shows that the asymptotic precision of the quantile estimated from the sample, measured by its asymptotic variance ω^2, depends on two quantities, $\tau(1-\tau)$ and $\frac{1}{f(F^{-1}(\tau))}$. The second quantity is the reciprocal of a density function evaluated at the quantile of interest ξ_τ, which is referred to as the sparsity function by Tukey (1965) or quantile density function by (Parzen, 1979). The sparsity function, normally denoted by $s(\tau)$, reflects the density of observations near ξ_τ, such that the estimation of the quantile becomes difficult when the observations are very sparse at the close proximity of the quantile. On the contrary, the quantile is precisely estimated when the sparsity of the data near ξ_τ is low, such that there are many observations near the quantile. In other words the sparsity of the data at a specific quantile ξ_τ determines how precise is the estimated value of the quantile.

To generalize the asymptotic distribution of sample quantiles to that of regression quantiles, consider the quantile linear regression model $y_i = \mathbf{x}_i'\boldsymbol{\beta}_\tau + u_{\tau_i}$, with independent identically distributed error terms u_{τ_i}. These terms have a common distribution function F associated with the density function f, and $f(F^{-1}(\tau_i)) > 0$, for $i = 1, \ldots, m$. Then the asymptotic distribution of the quantile regression estimator $\hat{\boldsymbol{\beta}}_\tau$ can be stated as

$$\begin{aligned} \sqrt{n}(\hat{\boldsymbol{\beta}}_\tau - \boldsymbol{\beta}_\tau) &\rightarrow \mathcal{N}(0, \tau(1-\tau)\mathbf{H}_\tau^{-1}\mathbf{D}\mathbf{H}_\tau^{-1}) = \mathcal{N}(0, \boldsymbol{\Lambda}_\tau), \\ \text{where} \quad \mathbf{D} &= \lim_{n\rightarrow\infty} n^{-1} \sum_i \mathbf{x}_i\mathbf{x}_i' \\ \text{and} \quad \mathbf{H}_\tau &= \lim_{n\rightarrow\infty} n^{-1} \sum_i \mathbf{x}_i\mathbf{x}_i' f_i(\xi_{\tau_i}). \end{aligned} \tag{6.17}$$

The matrix $\mathbf{D}$ is a positive definite $p \times p$ matrix. When the error terms are assumed to be iid, the density functions $f_i(\xi_{\tau_i})$ are identical and the sandwich covariance matrix $\boldsymbol{\Lambda}_\tau$ collapses to a simplified expression given by $\boldsymbol{\Lambda}_\tau = \frac{\tau(1-\tau)}{f^2(F^{-1}(\tau))} \lim_{n\rightarrow\infty} n\Big(\sum_i \mathbf{x}_i\mathbf{x}_i'\Big)^{-1}$, such that the asymptotic distribution of $\hat{\boldsymbol{\beta}}_\tau$ becomes

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_\tau - \boldsymbol{\beta}_\tau) \rightarrow \mathcal{N}(0, \omega^2\mathbf{D}^{-1}), \tag{6.18}$$

The simplified expression of $\boldsymbol{\Lambda}_\tau$ shows that under the iid error regression model, the asymptotic precision of quantile regression estimates depends on the sparsity function and the term $\tau(1-\tau)$. Under the quantile regression model the sparsity function plays a role similar to that of the standard deviation of the error terms, σ, in the least squares estimation procedure of the iid error regression model.

However, the assumption of iid error terms is too restrictive and oftentimes it does not hold in

practical applications. When the assumption holds the conditional quantiles are simple shifts of one another since all conditional quantiles planes are parallel. Thus the application of quantile regression does not provide any additional information to that provided by the least squares estimator since estimated regression coefficients for different quantiles, $\hat{\beta}_{\tau_j}$, have a common value, $\hat{\beta}_{\tau}$. However, in real life problems it is almost impossible to justify the assumption of iid error terms. Thus the use of quantile regression in such problems should be a way to go.

The asymptotic distribution of estimated regression coefficients in equation (6.17) can be extended to several regression coefficient vectors calculated at different quantiles. Hence the joint asymptotic distribution of the $m \times p$ variate quantile regression estimators defined as $\hat{\boldsymbol{\zeta}}_n = (\hat{\boldsymbol{\beta}}'_{\tau_1}, \ldots, \hat{\boldsymbol{\beta}}'_{\tau_m})'$ can take the form

$$\sqrt{n}(\hat{\boldsymbol{\zeta}}_n - \boldsymbol{\zeta}) \rightarrow \mathcal{N}(0, \boldsymbol{\Omega} \otimes \mathbf{D}^{-1}), \tag{6.19}$$

where $\boldsymbol{\Omega}$ is an $m \times m$ matrix with the elements $(\omega_{ij}) = \frac{\min(\tau_i, \tau_j) - \tau_i \tau_j}{f(F^{-1}(\tau_i)) f(F^{-1}(\tau_j))}$, and $\mathbf{D}$ is as defined under the asymptotic distribution of $\hat{\boldsymbol{\beta}}_{\tau}$. Inference about quantile regression estimator $\hat{\boldsymbol{\beta}}_{\tau}$ requires the estimation of its asymptotic covariance matrix $\boldsymbol{\Lambda}_{\tau}$. Literature reports different estimation methods of $\boldsymbol{\Lambda}_{\tau}$ for both cases of the iid and non-iid error terms as discussed in the subsequent section.

6.6.2 Estimation of the Covariance Matrix of Regression Quantiles

The covariance matrix that measures the precision of the τth quantile can be estimated using several approaches. Some are direct and asymptotic approaches that require the estimation of the sparsity function, while other are bootstrap approaches based on resampling. Koenker and Machado (1999) obtained an estimator of the sparsity function, $s(t)$, using simple difference quotients of the empirical quantile function as

$$\hat{s}_n(t) = \big[\hat{F}_n^{-1}(t + h_n) - \hat{F}_n^{-1}(t - h_n)\big]/2h_n, \tag{6.20}$$

where $\hat{F}^{-1}$ is an estimate of F^{-1}, and h_n is a bandwidth that tends to zero as $n \rightarrow \infty$. Hall and Sheather (1988) suggested a bandwidth rule based on Edgeworth expansions for studentized sample quantiles as

$$h_n = n^{-1/3} z_{\alpha}^{2/3} \big[1.5 s(t)/s''(t)\big]^{1/3}, \tag{6.21}$$

where z_{α} satisfies $\Phi(z_{\alpha}) = 1 - \alpha/2$. After the bandwidth is selected the estimate of the quantile function, $\hat{F}^{-1}$ can be obtained as the empirical quantile function of the residuals, $\hat{u}_i = y_i - \mathbf{x}'_i \hat{\boldsymbol{\beta}}_{\tau}, \quad i = 1, \ldots, n$, from fitting the quantile regression model as suggested by (Bassett and Koenker, 1982). Substituting the estimate of the sparsity function in the simplified equation of $\boldsymbol{\Lambda}_{\tau}$, gives the estimate of the asymptotic covariance matrix of $\hat{\boldsymbol{\beta}}_{\tau}$.

The Powell (1986) estimator for censored regression quantiles can be modified and used in the quantile regression model. The estimator can be used to estimate both the sparsity function for the independent identically distributed error terms case, and $\mathbf{H}_\tau$ for the general case where error terms are not independent and identically distributed. Under the iid error terms assumption the sparsity function can be estimated by one sided estimator defined as

$$\hat{f}\big(F^{-1}(\tau)\big) = (\hat{c}_n n)^{-1} \sum_{i}^{n} I(0 \leq \hat{u}_{\tau_i} \leq \hat{c}_n), \tag{6.22}$$

where $\hat{u}_{\tau_i} = y_i - \mathbf{x}_i' \hat{\boldsymbol{\beta}}_\tau$ and c_n is the Kernel bandwidth. Cross validation methods such as, least squares and log likelihood can be used for the optimal selection of c_n. The resultant Kernel estimator of the covariance matrix for $\boldsymbol{\beta}_\tau$ can be given by

$$\hat{\boldsymbol{\Lambda}}_\tau = \frac{\tau(1-\tau)}{\hat{f}^2(F^{-1}(\tau))} \left(\frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i' \right)^{-1}. \tag{6.23}$$

The two sided Kernel estimator in which the indicator function in equation (6.22) is replaced by $I(-\hat{c}_n/2 \leq \hat{u}_{\tau_i} \leq \hat{c}_n/2)$ can be used to estimate $\boldsymbol{\Lambda}_\tau$. When the error terms are heteroscedastic, $\mathbf{H}_\tau$ can be estimated by $(\hat{c}_n n)^{-1} \sum_i^n I(0 \leq \hat{u}_{\tau_i} \leq \hat{c}_n) \mathbf{x}_i \mathbf{x}_i'$.

Instead of estimating the sparsity function, bootstrap method based on varying assumptions about error terms and the form of the asymptotic covariance matrix of $\hat{\boldsymbol{\beta}}_\tau$ can be used to estimate the covariance matrix. He and Hu (2002) suggested the Markov chain marginal bootstrap (MCMB) method that differ with other bootstrap methods in two important aspects. The aspects are that the method solves only one-dimensional equations for parameters of any dimension, and produces a Markov chain instead of an independent sequence. The purpose of the MCMB method is to reduce the problems of computations that are associated with bootstrap in high-dimensional problems.

6.6.3 Tests of Linear Hypothesis

Having discussed the estimation of the parameters for one conditional quantile as well as for several conditional quantile functions, and observed differences that may exist in their estimated regression slopes, it becomes necessary to look into statistical tests that are used to establish if the differences are significant. Koenker and Basett (1982a) proposed tests of linear hypothesis that are used to test the equality of slope parameters across quantiles slopes. Koenker and Basett (1982b) proposed tests of linear hypothesis for the linear model in the case of median (l_1) regression that include the Wald, likelihood ratio test, which were extended to other quantiles. The tests that are normally used to test equality of slopes in quantile regression are the Wald, likelihood and rank tests. The asymptotic distribution of the quantile regression estimates discussed in the preceding section serves as the basis for these tests.

1. **Wald Test**

   The Wald test is based on the regression coefficient estimated from the unrestricted model (Koenker and Basett, 1982b). It tests the general linear hypothesis for $p \times 1$ vector of parameters, $\boldsymbol{\beta}_\tau$, in the case of a single quantile regression coefficient, stated as $H_0 : \mathbf{H}\boldsymbol{\beta}_\tau = \mathbf{h}$ against $H_1 : \mathbf{H}\boldsymbol{\beta}_\tau \neq, \mathbf{h}$, where $\mathbf{H}$ is $k \times p$ matrix of coefficients defining k linear functions of the β_τ, and $\mathbf{h}$ is a $k \times 1$ vector of constants, which are frequently zeros (Melly, 2001; Koenker, 2005). The test statistic under H_0 is defined as

$$W_\tau = n(\mathbf{H}\hat{\beta}_\tau - \mathbf{h})'\big[\mathbf{H}\hat{\boldsymbol{\Lambda}}_\tau^{-1}\mathbf{H}'\big]^{-1}(\mathbf{H}\hat{\beta}_\tau - \mathbf{h}), \tag{6.24}$$

   which is asymptotically χ^2_q, where q is the rank of the matrix $\mathbf{H}$.

   Koenker and Basett (1982a) proposed tests of heteroscedasticity for quantile regression, which generalize the Wald test for the parameter $\boldsymbol{\beta}_\tau$ of one quantile to a $m \times p$ matrix of parameters $\boldsymbol{\zeta}$ in which several distinct quantiles are considered. They relaxed slightly the assumption of iid error terms and considered the asymptotic behavior of $\sqrt{n}(\hat{\boldsymbol{\beta}}_\tau - \boldsymbol{\beta}_\tau)$ under a specific form of asymptotically vanishing heteroscedasticity. In this case the general linear hypothesis can be stated as $H_0 : \mathbf{H}\boldsymbol{\zeta} = \mathbf{h}$, and the test statistic, under H_0, is given by

$$W_\tau = n(\mathbf{H}\hat{\boldsymbol{\zeta}} - \mathbf{h})'\big[\mathbf{H}(\boldsymbol{\Omega} \otimes \mathbf{D}^{-1})\mathbf{H}'\big]^{-1}(\mathbf{H}\hat{\boldsymbol{\zeta}} - \mathbf{h}). \tag{6.25}$$

   This is asymptotically non-central χ^2 with rank q degrees of freedom and non-centrality $\eta = (\mathbf{H}(Q_\tau(\mathbf{u}) \otimes \gamma_0))'\big[\mathbf{H}(\boldsymbol{\Omega} \otimes \mathbf{D}^{-1})\mathbf{H}'\big]^{-1}(\mathbf{H}(Q_\tau(\mathbf{u}) \otimes \gamma_0))$.

   In the case of homoscedastic model, the slope parameters are identical for every quantile, and the test statistic W_τ is asymptotically central χ^2 with $(m-1)(p-1)$ degrees of freedom, where p is the number of parameters in the model, and m is the number of quantiles for which the model is fitted (Koenker, 2005). This form of the Wald test accommodates a wide variety of testing situations, including joint tests that involve several predictor variables and several distinct quantiles. The tests offer a robust alternative to the conventional least-squares based tests of heteroscedasticity since they are insensitive to outlying response variable observations. In addition the tests do not require parametric assumptions about the shape of the error distribution.

2. **Likelihood Ratio Test**

   The likelihood ratio test is based on the difference between the sum of absolute residuals in the restricted and unrestricted models. The linear hypothesis to be tested using the likelihood ratio test is as stated under the Wald test. Koenker and Machado (1999) adapted the Koenker and Basett (1982b) approach of the likelihood ratio test and showed that under H_0 when the error terms are iid but drawn from the distribution function F, the test statistic is given by

$$L_{n(\tau)} = \frac{2(\tilde{V}_\tau - \hat{V}_\tau)}{\tau(1-\tau)s(\tau)}, \tag{6.26}$$

where $\tilde{V}_\tau$ and $\hat{V}_\tau$ are as defined under the section of goodness-of-fit criterion, and $s(\tau)$ is the sparsity function. Like the Wald test statistic, $L_{n(\tau)}$, is asymptotically χ^2_q.

In the case of the location-scale shift model, Koenker and Machado (1999) showed that under H_0 the test statistic becomes

$$\Lambda_{n(\tau)} = \frac{2n\sigma_\tau}{\tau(1-\tau)s(\tau)}\log(\tilde{V}_\tau/\hat{V}_\tau), \tag{6.27}$$

where the estimate of σ_τ is $\hat{\sigma}_\tau = n^{-1}\hat{V}_\tau \to \sigma_\tau$. The test statistics $\Lambda_{n(\tau)}$ is also asymptotically χ^2_q. Similarly, the likelihood ratio test can be used to test the global hypothesis that quantile regression slopes coefficients are identical across quantiles.

3. **Rank Tests of Linear Hypothesis**

Gutenbrunner et al. (1993) introduced tests of a general linear hypothesis for the linear regression model, which are based on regression rank scores of Gutenbrunner and Jurečková (1992). The tests are robust to observations that are outlying with respect to the response variable, and are asymptotically distribution free in the sense that no nuisance parameters that depends on the error term distribution need to be estimated for the computation of the test statistic.

Gutenbrunner et al. (1993) considered the general linear model with the design matrix, $\mathbf{X}$, partitioned into $(\mathbf{X}_1{:}\mathbf{X}_2)$, and the vector of parameters, $\boldsymbol{\beta}_\tau$, partitioned into $\boldsymbol{\beta}_{1\tau}$ and $\boldsymbol{\beta}_{2\tau}$. Then the linear hypothesis can be stated as $H_0 : \boldsymbol{\beta}_{2\tau} = 0; \quad \boldsymbol{\beta}_{1\tau}$ unspecified, against the local alternative $H_n : \boldsymbol{\beta}_{2n_\tau} = \boldsymbol{\beta}_{0\tau}/\sqrt{n}$; with $\boldsymbol{\beta}_{0\tau} \in \mathbb{R}^q$, fixed. The regression rank scores are a $n \times 1$ vector, $\hat{\mathbf{a}}_n(\tau) = (\hat{a}_{n1}(\tau), \ldots, \hat{a}_{nn}(\tau))$. The proposed test statistic for testing H_0 against H_n can be defined by

$$T_n = \frac{\mathbf{S}'_n\mathbf{M}_n^{-1}\mathbf{S}_n}{A^2(\varphi)}, \tag{6.28}$$

where $\mathbf{S}_n = n^{-1/2}(\mathbf{X}_{n2} - \hat{\mathbf{X}}_{n2})'\hat{\mathbf{b}}_n$, $\mathbf{M}_n = n^{-1}(\mathbf{X}_2 - \hat{\mathbf{X}}_2)'(\mathbf{X}_2 - \hat{\mathbf{X}}_2)$, $\hat{\mathbf{X}}_2 = \mathbf{X}_1(\mathbf{X}'_1\mathbf{X}_1)^{-1}\mathbf{X}'_1\mathbf{X}_2$, $\hat{\mathbf{b}}_n$ is the scores vector given by the intergral $\int_0^1 \hat{\mathbf{a}}_n(t)\,\mathrm{d}\varphi(t)$, $A^2(\varphi) = \int_0^1(\varphi(t) - \bar{\varphi})^2\,\mathrm{d}t$, $\bar{\varphi} = \int_0^1 \varphi(t)\,\mathrm{d}t$ and φ is a score function. The test is based on the asymptotic distribution of T_n under the null hypothesis H_0. Under H_0, T_n is asymptotically distributed as a central χ^2 with q degrees of freedom, while under the local alternative hypothesis H_n, T_n is a noncentral χ^2 with q degrees of freedom and non-centrality parameter η^2, defined under the Wald test.

Koenker and Machado (1999) extended the work of Gutenbrunner and Jurečková (1992) and Gutenbrunner et al. (1993) to the location-scale linear model. Their modification of the test statistic is similar to substituting an ordinary least squares fit by a weighted least squares fit at this stage. The test statistic can be defined by

$$T_n = \frac{\mathbf{S}'_n\mathbf{Q}_n^{-1}\mathbf{S}_n}{\tau(1-\tau)}. \tag{6.29}$$

Under the null hypothesis, the modified T_n has a central χ^2_q distribution, while under the local alternative hypothesis it has a noncentral χ^2_q distribution with the non-centrality parameter, $\eta(\varphi, \zeta)$. The statistic can be used to determine the global effects of the predictor variables on the response variable across quantile, or local effects by choosing the score function φ to focus exclusively on one quantile τ (Koenker, 2005).

6.6.4 Confidence Intervals of Regression Quantiles

Literature reported various approaches of constructing confidence intervals and confidence bands for regression quantiles. There are approaches based on the asymptotic distribution of the quantile regression estimator $\boldsymbol{\beta}_\tau$ that require the estimation of the sparsity function, and there are distribution free approaches that do not require the estimation of the sparsity function (Zhou and Portnoy, 1996). Some are based on the inversion of rank scores by Koenker (1994) and others are based on resampling methods (Koenker, 1994; Parzen et al., 1994; Chen and Wei, 2005).

Zhou and Portnoy (1996) proposed the studentized and direct approaches for constructing confidence intervals for regression quantiles. They noted that the two approaches are the generalization of their analogues in constructing confidence interval for sample quantiles. The studentization approach is based on the asymptotic normality of the estimated regression quantiles, and thus it requires the estimation of the sparsity function. On the contrary, the direct approach referred to as the distribution-free approach, does not require the estimation of sparsity function. The confidence interval of regression quantiles from the distribution-free approach is the generalization of the confidence interval of sample quantiles. Zhou and Portnoy (1996) showed that the studentized and distribution-free confidence intervals are asymptotically equivalent when the consistent estimator of the sparsity function is used.

The application of the tests based on rank scores involves the construction of confidence intervals for the parameters of the quantile regression model. Koenker (1994) proposed a robust approach of constructing confidence intervals for quantile regression based on the inversion of the rank score test, which does not require the estimation of the sparsity function. Unlike the confidence intervals based on the estimation of the sparsity function, the confidence intervals resulting from the inversion of rank tests are not symmetric. However, they are centered on the point estimate $\hat{\beta}_{2\tau}$ of the partitioned model consisting of one predictor variable X_2, $\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_{1\tau} + X_2\beta_{2\tau} + \mathbf{u}_\tau$, in the sense that $T_n(\hat{\beta}_{2\tau}) = 0$. This approach of constructing confidence interval possesses a good feature that it inherits the scale invariance of the test statistic T_n.

Bootstrap methods construct more reliable confidence intervals but they are computationally inefficient for moderate to large data sets (He and Hu, 2002; Kocherginsky et al., 2005). Chen and Wei (2005) noted that resampling methods are not recommended for small data sets with sample

size, $n < 5000$, and the number of predictor variables, $p < 20$, since they are stable for relatively larger data sets. Koenker (1994) proposed a resampling method that can be used to construct confidence interval for estimates of quantile regression and named it Hegf bootstrap. The method resamples directly from the full regression quantile process. Parzen et al. (1994) proposed a general and simple resampling method based on pivotal estimating functions $\mathbf{S}_x(\boldsymbol{\beta})$ for inferences about the true parameter $\boldsymbol{\beta}$. This method can be adapted and used to construct confidence intervals for quantile regression estimates. The approach achieves robustness to some heteroscedastic quantile regression models by exploiting the asymptotical pivotal role of the quantile regression (Koenker, 1994).

Kocherginsky et al. (2005) adapted the Markov chain marginal bootstrap (MCMB) method proposed by He and Hu (2002) that aims to provide faster computations, to construct confidence intervals for quantile regression and called it MCMB-A method. The MCMB-A method is an affine transformation of the parameter space that helps to eliminate the problems of autocorrelation in the MCMB sequence generated by repeating the sampling process B times, denoted by $\boldsymbol{\beta}_\tau^{(B)}$.

Chapter 7

Fitting the Quantile Regression Model to Household Data

7.1 Introduction

Quantile regression model was applied as a robust alternative to the linear model fitted using least squares procedure, to deal with asymmetrical distributions of maize and sorghum household availability, including the distributions with longer and heavier tails than the normal distribution. Such distributions are the same as the ones observed from the values of skewness and kurtosis, box plots in Figures 2.7 and 2.8, and diagnostics plots in Figures 5.1 and 5.2. The quantile regression package in R, quantreg, by Koenker (2005) was used to fit the quantile regression model as well as to plot the quantile plots. The quantreg uses the command rq() to fit the quantile regression model.

The effects of predictor variables used in Chapter 4 to fit the classical linear regression model on maize availability and sorghum availability for households were investigated at different positions of the distribution of the response variable. As it was discussed in the preceding chapter, the use of quantile regression in modeling availability of the two cereals provides a more comprehensive explanation of the relationships that exist between availability and predictor variables in the model than in the classical regression model where the mean response was exclusively used. This was done by fitting the quantile regression model at the selected three quantiles, the $25th$, $50th$ and $75th$ quantiles. The simplex based algorithm adapted for quantile regression by Koenker and D'Orey (1987) was used to compute the quantile regression parameter estimates given in Tables 7.1 and 7.2. Since the sample size is not very large the simplex algorithm performed well. Following Koenker (2005) that the introduction of the quardratic component alleviates some nonmonotonicity in the qauntile regression estimates that may occur at the lower and upper tails of the distribution of the response variable, the quadratic term of household size was added as one of the predictor variables to cater for possible nonlinearities in the data.

7.2 Regression Quantiles

The goodness-of-fit of the quantile regression model fitted to the household maize data at each of the three quantiles was assessed using the goodness-of-fit measure, R^1_τ by Koenker and Machado (1999). The values of R^1_τ at the $25th$, $50th$ and $75th$ quantiles together with the value of the measure of goodness-of-fit for the classical regression models, R^2, are shown in the last row of Tables 7.1 and 7.2. The value of R^1_τ for maize data, increased by almost the same amount from one quantile to another. This is an indication that in some cases the effects of the predictor variables in the quantile regression model are different at different quantiles and they become more significant as values of quantiles increase. The asterisks next to some estimates indicate the statistical significance of a variable corresponding to the estimates, and the level of significance as indicated below the table.

Table 7.1: Quantile Regression Parameter Estimates for Maize Availability

Predictor Variable	Category	Quantile Regressions			OLS Estimate
		0.25	0.50	0.75	
Intercept		205.46 ***	270.37 ***	571.23 ***	265.20 ***
HHSize		54.52 **	72.45 ***	73.49 ***	87.22 ***
Household Size^2		0.69	2.75 **	3.46	-0.03
Income		0.03	0.01	0.08 **	0.14 ***
Sex	Ref=Male				
	Female	-4.72	24.03	18.74	8.43
Location	Ref=Mafeteng				
	Berea	11.46	-59.77	-53.29	-13.99
	Maseru Foothills	10.28	-165.69 ***	-61.08	-61.26
	Maseru Lowland	-44.08	-130.89 ***	-67.04	-88.54
Education	Ref=No educ				
	High School	28.79	-4.37	-16.39	29.27
	Post High School	-45.64	295.11 ***	131.86	131.98
	Primary	-9.31	-55.49 *	-162.04 ***	-53.35
Occupation	Ref=Farmer				
	Casual Worker	-146.36 **	-82.10 *	-248.82 ***	-195.97 **
	Pensioner	-86.75	-78.64	-218.84 **	92.74
	Salary Earner	-124.20 **	-13.55	-228.00 ***	-105.42
	Unemployed	-153.44 ***	-86.06 **	-279.37 ***	-159.55 **
R^2 and R^1_τ		0.18	0.23	0.27	0.30

Asterisks denote the significance level as ***: 1%, **: 5%, *: 10%.

Household size is significant throughout the three quantiles of maize availability as well as at the conditional mean of maize availability from the ordinary least squares procedure. Significance of household size is observed with an increasing effect from 54.52 of the $25th$ quantile to 73.49 of the $75th$ quantile, and an increasing strength of evidence shown by the level of significance that decreases from 5% of the $25th$ quantile to 1% of the median, $75th$ quantile, and under the mean response (Table 7.1). Thus the effect of household size is higher at the $50th$ and $75th$ quantiles

as well as the mean response than at the $25th$ quantile. The results suggest that an increase in the size of households by one member, when other variables are held constant, increased the $25th$ percentile, median regression, $75th$ percentile and mean response of maize availability by 54.52, 72.45, 73.49 and 87.22 kilograms respectively. The effect of household size at the median and $75th$ quantile of maize availability does not differ that much. The quadratic term of household size is significant at the 5% level of significance, for the median regression only. Household monthly income is significant at the $75th$ quantile and the mean response only, with very small effects of less than half a kilogram. Like under ordinary least squares, sex of household head is not significant across the entire distribution of maize availability. This suggests that household maize availability for households headed by females was not significantly different from that of households headed by males.

The location of a household is statistically significant at 1%, under the median regression only. Maize availability for households that resided in Maseru foothill and Maseru lowland was significantly different from that of households that resided in Mafeteng at the $50th$ quantile, when other variables are held constant. Maize availability of households from both locations was less than that of households in Mafeteng, with households in Maseru foothill having a bigger difference of 165.69 kilograms. On the other hand, maize availability of households that resided in Berea was not significantly different from that of households in Mafeteng, throughout the three quantiles as well as at the mean response.

In the case of education level of the household head, maize availability for households headed by people who attained post high school education was significantly higher than that of households headed by people without any form of formal education, at the $50th$ quantile, by 295.11 kilograms. The reason behind this could be that household heads with post high school qualifications such as diploma and university degrees had skills that got them better jobs, and hence earned income that enabled them to acquire more maize for their households, either through purchases or production from own land. Maize availability of households headed by people with primary education was significantly less than that of households headed by heads with no formal education by 55.49 and 162.04 kilograms, at the $50th$ and $75th$ quantiles respectively. Though the differential effect of education level at the median regression, measured by the size of the regression quantile, is less than that at the $75th$ percentile. The strength of evidence for the significance of the differential effect, shown by the number of asterisks, is also lower at the median than at the $75th$ percentile.

Occupation of the head of a household played an important role in determining availability of maize for households because almost all occupations in Table 7.1, except pensioners, are significantly different from subsistence farmers at the $25th$ and $75th$ quantiles. According to the regression quantiles, availability of maize for households headed by casual workers, salary earners, and unemployed people was significantly less than that of households headed by subsistence farmers at the

two quantiles. In addition, maize availability of households headed by casual workers and unemployed heads was also significantly less than that of households headed by subsistence farmers at the median regression, though the differential effects are low when they are compared with their counterparts at the other two quantiles and the mean response.

The size of the regression quantiles show that the differential effect of occupation on maize availability is the highest at the $75th$ quantile. Households headed by casual workers and unemployed people maintained the highest differences throughout the quantiles as well as at the mean response, with the biggest differences of -248.82 kilograms for casual workers and -279.37 kilograms for unemployed people observed at the $75th$ quantile. The differences suggest that maize availability for households headed by casual workers and unemployed heads was less than that of households headed by subsistence farmers by 248.82 and 279.37 kilograms respectively. Normally, unemployed heads do not earn any income and casual workers do not have stable income that enable them to purchase food or finance food production for their households and thus the likelihood of maize availability for their households being far less than that of households headed by subsistence farmers is high. An observation that households headed by casual workers, salary earners, and unemployed people were worse off than those headed by subsistence farmers in terms of maize availability is not surprising because subsistence farming is an occupation devoted to production of food for households.

Table 7.2: Quantile Regression Parameter Estimates for Sorghum Availability

Predictor Variable	Category	Quantile Regressions			OLS
		0.25	0.50	0.75	
Intercept		300.39***	436.72***	399.18	429.13***
Household Size		-0.67	-22.74	5.16	1.35
Household Size^2		0.81	2.96***	0.94	1.03
Income		-0.01	0.01	0.03	0.01
Sex	Ref=Male				
	Female	-0.18	15.75	7.86	3.67
Location	Ref=Berea				
	Mafeteng	-71.84**	-25.02	-13.17	-41.78
	Maseru Foothills	-71.82*	-72.78*	-6.03	-78.41
	Maseru Lowland	-28.42	36.32	52.36	-21.78
Education	Ref=Post COSC				
	High School	-140.65**	-231.00***	-123.80	-215.90***
	No Formal Education	-132.12**	-226.08***	-277.27***	-255.20***
	Primary	-132.68**	-225.03***	-203.13**	-247.50***
Occupation	Ref=Casual Worker				
	Subsistence farmer	-62.75*	-50.53	-64.78	-24.29
	Pensioner	-89.08**	-127.43***	-38.50	-29.78
	Salary Earner	-80.28**	-124.88***	-95.92	-84.42*
	Unemployed	-94.67**	-85.41***	-87.77	-63.50
R^2 and R^1_τ		0.11	0.12	0.13	0.19

Asterisks denote the significance level. ***: 1%, **: 5%, *: 10%.

The value of R^1_τ for sorghum availability increased by 0.01 from one quantile to another (Table 7.2). The small increment in R^1_τ values shows that the effect of predictor variables does differ but slightly, at the three quantiles. The results from the goodness-of-fit measure for the two models fitted using maize and sorghum availability will be confirmed by the test of equality of slopes in the subsequent analyses.

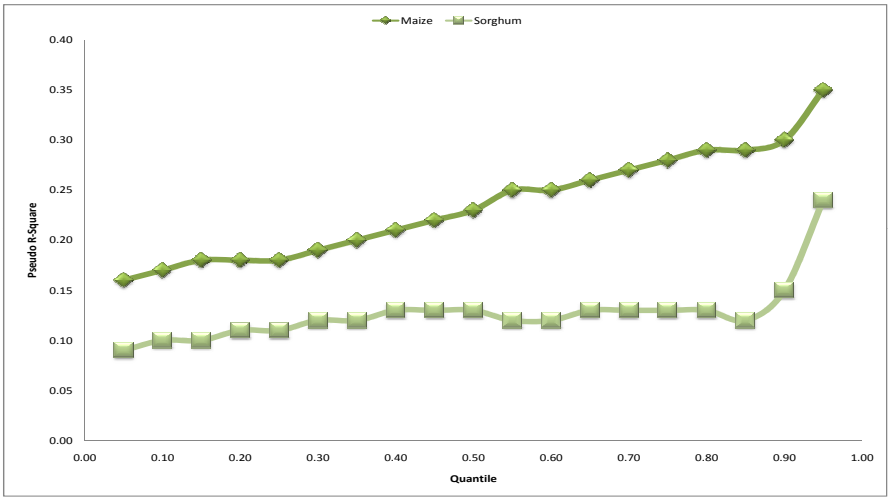

Figure 7.1: Goodness-of-fit of Quantile Regression Model for Maize and Sorghum Availability

The two plots of R^1_τ (pseudo R square) values against quantiles at which the quantile regression model was fitted were used to illustrate the goodness-of-fit of maize and sorghum availability models (Figure 7.1). The plots, at a glance, give an idea of how regression slopes differ across quantiles. Both plots have small values of R^1_τ at the lower tails of the conditional maize and sorghum availability distribution than at the upper tails. This indicates the poor fit of the model at the lower tails than at the upper tails of the distributions. The plot for maize availability shows a better fit than that of sorghum availability since it has higher values of pseudo R square across quantiles. The plot also shows that the value of pseudo R square increased with the quantile at which the model was fitted. This suggests that the effects of the predictor variables varies across quantiles, with more significant effects in the upper tail of the distribution. On the contrary, the flat slope of the plot for sorghum availability, up to the $85th$ quantile, is indicative of the regression slopes that remain the same across quantiles up to the quantile. This suggests that the effects of the predictor variables remains the same up to the $85th$ quantile, and begin to vary at quantiles beyond this quantile.

The results from the quantile regression model for sorghum availability show that household size,

household monthly income and sex are not significant throughout the three quantiles (Table 7.2). However, the quadratic term of household size is significant at 1% for the $50th$ quantile. Two categories of location of households, Mafeteng and Maseru foothills, are significantly different from Berea. Mafeteng is significant at 5% level of significance at the $25th$ quantile, while Maseru foothills is significant at 10% level at both the $25th$ quantile and median regression. The differential effects of location of a household shown by the regression quantiles are almost the same in the two locations. The quantiles suggest that sorghum availability for households from Mafeteng was less than that of households from Berea by 71.84 kilograms at the $25th$ quantile. Sorghum availability of households from Maseru foothills was less than that of households from Berea by 71.82 and 72.78 kilograms at the $25th$ quantile and median regression respectively.

Education level of household head played an important role in determining availability of sorghum to households. Sorghum availability for households headed by people with no formal education, primary, and high school education was significantly different from that of households headed by people with post high school qualifications such as diploma and university degrees, at the $25th$ quantile, median regression, as well as at the mean response of ordinary least squares (Table 7.2). Though the strength of evidence at the $25th$ percentile is weaker than at the median regression. At the $75th$ quantile, sorghum availability for households headed by people with no formal education and those headed by people with primary education was significantly different from that of households headed by diploma and degree holders. Negative regression quantiles show that sorghum availability for all these households was less than that of households headed by people who attained diploma and university degrees. The differential effect of education level is bigger at $75th$ quantile and under the mean response for households headed by people with no formal education, where sorghum availability was less by 277.27 and 255.20 kilograms, respectively.

Occupation of household head is significant at both the $25th$ quantile and median regression, except for one category of subsistence farmer. The regression quantiles show that differential effect of occupation on sorghum availability for households headed by pensioners and salary earners are increasing with the quantile from -89.08 to -127.43 and from -80.28 to -124.88, respectively. These quantiles indicate that at the median regression, sorghum availability of households headed by pensioners and salary earners was less than that of households headed by casual workers by 127.43 and 124.88 kilograms respectively. The decrease in the significance level from 5% to 1% for the categories of pensioners, salary earners, and unemployed shows an increase in the strength of evidence as the quantile increases from 0.25 to 0.50.

7.3 Standard Errors for Regression Quantiles for Household Maize

Standard errors of regression quantiles reported in Tables 7.3, 7.4, 7.5 and Tables in Appendix J were computed at the $25th$, $50th$ and $75th$ of maize and sorghum availability. They were computed using two asymptotic methods and one bootstrap method of computing the covariance matrix of the quantile regression estimates. The asymptotic methods used are the sparsity function estimation under the assumption of independent identically distributed error terms using the Hall and Sheather (1988) bandwidth rule and the Kernel estimator. The boostrap estimator used the He and Hu (2002) MCMB resampling method.

Table 7.3: Standard Errors of the $25th$ Regression Quantile for Maize Availability

Predictor Variable	Category	Estimate	Standard Errors		
			Sparsity	Kernel	Bootstrap
Intercept		205.46	83.02	131.14	120.15
Household Size		54.52	22.68	53.05	50.20
Household Size^2		0.69	1.80	5.32	5.00
Income		0.03	0.03	0.04	0.04
Sex	Ref=Male				
	Female	-4.72	36.06	50.50	41.59
Location	Ref=Mafeteng				
	Berea	11.46	62.99	85.46	71.71
	Maseru Foothills	10.28	63.87	98.28	90.21
	Maseru Lowland	-44.08	42.99	70.76	62.08
Education	Ref=No educ				
	High School	28.79	60.92	89.49	78.58
	Post High School	-45.64	100.59	173.80	189.32
	Primary	-9.31	44.56	65.44	56.06
Occupation	Ref=Farmer				
	Casual Worker	-146.36	59.59	86.81	80.73
	Pensioner	-86.75	72.02	113.15	83.89
	Salary Earner	-124.20	61.24	94.28	74.59
	Unemployed	-153.44	46.31	71.83	58.43

The sparsity function estimation using Hall and Sheather (1988) bandwidth rule reports a higher precision of all estimates of regression quantiles across the three quantiles for maize and sorghum data. The high precision is shown by standard errors that are relatively low when they are compared with the ones estimated by the Kernel estimator and MCMB resampling method. However, standard errors estimated by the Kernel estimator show varying patterns of the precision of estimates depending on the quantile and the data used.

Standard errors estimated from the Kernel estimator at the $25th$ and $50th$ quantile of maize availability show a lower precision of estimates in the sense that they are greater than standard errors estimated by MCMB resampling method, except in the case of one level of education of household

Table 7.4: Standard Errors of the $50th$ Regression Quantile for Maize Availability

Predictor Variable	Category	Estimate	Standard Errors		
			Sparsity	Kernel	Bootstrap
Intercept		270.37	63.26	124.02	98.93
Household Size		72.46	17.28	40.46	31.97
Household Size^2		2.75	1.37	3.70	3.12
Income		0.01	0.02	0.05	0.05
Sex	Ref=Male				
	Female	24.03	27.48	52.21	39.60
Location	Ref=Mafeteng				
	Berea	-59.77	48.00	86.12	69.55
	Maseru Foothills	-165.69	48.67	93.45	72.98
	Maseru Lowland	-130.89	32.76	61.93	52.26
Education	Ref=No educ				
	High School	-4.37	46.42	98.47	92.72
	Post High School	295.11	76.65	186.40	254.79
	Primary	-55.49	33.95	65.75	53.78
Occupation	Ref=Farmer				
	Casual Worker	-82.10	45.40	82.10	65.69
	Pensioner	-78.64	54.88	108.58	94.20
	Salary Earner	-13.58	46.66	91.43	75.06
	Unemployed	-86.07	35.29	71.38	65.12

head, post high school education, where the Kernel estimator shows a higher precision than the MCMB resampling method (Tables 7.3 and 7.4). On the contrary, at the $50th$ quantile of sorghum data, the Kernel estimator show a higher precision than the MCMB resampling method for all estimates under the education level of household head and one estimate under occupation of household head, salary earner (Table J.2).

On the basis of the empirical evidence from the application of quantile regression model on household data, standard errors of quantiles estimated by sparsity function using the Hall and Sheather (1988) bandwidth rule are recommended since they are the smallest when they are compared with standard errors estimated from the Kernel estimator and MCMB resampling method. This recommendation follows from the fact that small standard errors of regression estimates imply high precision of the estimates in estimating regression parameters. However, we should note that larger standard errors are not always a disadvantage. This might indicate the capacity to a statistical method to capture inherent additional variability in the estimation of a parameter, which protects against committing type I error too frequently, compared to a method that does not have such a strength.

Table 7.5: Standard Errors of the $75th$ Regression Quantile for Maize Availability

Predictor Variable	Category	Estimate	Standard Errors		
			Sparsity	Kernel	Bootstrap
Intercept		571.24	102.88	144.48	109.20
Household Size		73.49	28.11	40.82	32.75
Household Size^2		3.47	2.23	3.74	3.13
Income		0.08	0.04	0.07	0.08
Sex	Ref=Male				
	Female	18.74	44.69	60.80	63.77
Location	Ref=Mafeteng				
	Berea	-53.29	78.05	98.58	73.11
	Maseru Foothills	-61.08	79.14	107.26	122.65
	Maseru Lowland	-67.04	53.27	63.10	67.30
Education	Ref=No educ				
	High School	-16.39	75.49	115.62	126.80
	Post High School	131.86	124.65	185.42	679.19
	Primary	-162.04	55.21	80.67	85.33
Occupation	Ref=Farmer				
	Casual Worker	-248.82	73.83	94.48	72.58
	Pensioner	-218.84	89.24	130.04	193.31
	Salary Earner	-228.00	75.88	108.60	105.02
	Unemployed	-279.37	57.39	82.53	68.39

7.4 Test of Equality of Slopes for Household Data

The Wald test was used to test the hypothesis of pure location shift that all the slope coefficients of the quantile regression model fitted to the household data are the same across the three quantiles. The joint test of equality of slope coefficients of maize data for the quantiles 0.25, 0.50 and 0.75 is significant at 1% level of significant. This means that not all slope coefficients across the three quantiles are the same and thus the effects of the predictor variables on maize availability are not the same across the three quantiles.

Since the test of joint slopes shows that the effect of predictor variables in the model are not the same across quantiles, then the next step is to establish predictor variables that are associated with the effects differences. The test of equality of distinct slopes was used to find out predictor variables whose effects do not remain the same across quantiles. The results in Table 7.6 show that the differences in the effects on maize availability across quantile are associated with each of the categories marked with asterisks, in relation to their reference categories. The categories are of households that resided in Maseru foothill, households headed by someone who attained primary education, someone who was a casual worker, salary earner or an unemployed head. These categories are of the variables, location of a household, education level of head of household, and occupation of head of household, respectively. However, the category of households headed by unemployed heads has

Table 7.6: Test of Equality of Distinct Slopes for Maize Availability

Quantiles = 0.25, 0.50, 0.75

Variable	Category	F value	Pr>F	
Household Size		0.2946	0.744876	
Household Size^2		0.3324	0.717322	
Income		1.7398	0.176163	
Sex	Ref=Male			
	Female	0.3572	0.699731	
Location	Ref=Mafeteng			
	Berea	0.7023	0.495711	
	Maseru Foothills	4.0566	0.017631	**
	Maseru Lowland	2.0175	0.133602	
Education	Ref=Post COSC			
	High School	0.1956	0.822346	
	Post High School	1.4700	0.230478	
	Primary	3.5055	0.030450	**
Occupation	Ref=Farmer			
	Casual Worker	3.7056	0.024966	**
	Pensioner	0.1988	0.819791	
	Salary Earner	4.0411	0.017903	**
	Unemployed	5.6469	0.003657	***

Asterisks denote the significance level. ***: 1%, **: 5%, *: 10%.

the strongest evidence that its estimates vary across quantiles, shown by three asterisks that denote the significance level of 1%.

In the case of sorghum availability the joint test of equality of slope coefficients for the three quantiles is significant at 10%. This suggests the effect of the predictor variables on sorghum availability that are not the same across the three quantiles, though the evidence for significance is weaker than that of the test for maize availability. The difference in the effects of predictor variables on sorghum availability are associated with each of the two categories marked with asterisks, in relation to their reference categories. The two categories are of the location of households, namely, Mafeteng and Maseru Lowlands (table 7.7).

Quantile plots in Figure 7.2, and Figure K.1 of Appendix K present a summary of quantile regression results that show quantile regression estimates for the entire distribution and mean response of maize availability and sorghum availability, respectively. It shows their confidence bands of quantile regression estimates and confidence intervals of the ordinary least squares estimates of the mean effect. Each plot illustrates one regression coefficient for distinct regression estimates for the values of τ in the range [0.02, 0.98]. The dotted line represents point estimates for different values of τ, $\hat{\beta}_j(\tau)$, for $j = 1, \ldots, 15$, the solid line superimposed on the quantile plot shows the ordinary least squares estimate of the mean effect. The two dashed lines represent 90% confidence intervals for the least squares estimates and the shaded grey area illustrates the 90% confidence band for the

Table 7.7: Test of Equality of Distinct Slopes for Sorghum Availability

Quantiles = 0.25, 0.50, 0.75

Variable	Category	F value	Pr>F	
Household Size		1.4309	0.24001	
Household Size^2		1.2903	0.27606	
Income		0.6942	0.49992	
Sex	Ref=Male			
	Female	0.3720	0.68955	
Location	Ref=Berea			
	Mafeteng	2.6464	0.07185	*
	Maseru Foothills	0.4376	0.64583	
	Maseru Lowland	4.0716	0.01759	**
Education	Ref=Post COSC			
	High School	0.3558	0.70077	
	No Formal Education	0.4055	0.66684	
	Primary	0.3722	0.68941	
Occupation	Ref=Casual Worker			
	Subsistence Farmer	0.1228	0.88448	
	Pensioner	1.0292	0.35802	
	Salary Earner	0.4135	0.66155	
	Unemployed	0.0463	0.95476	

Asterisks denote the significance level. ***: 1%, **: 5%, *: 10%.

quantile regression estimates. The plots shows how estimates for different regression coefficients change across quantiles. The computation of the confidence intervals for the regression quantiles was based on the asymptotic distribution of the estimates, which requires the estimation of the sparsity function.

The first plot in Figure 7.2, which shows the intercept of the model, may be interpreted as the estimated conditional quantile function of the distribution of maize availability of a household which had five members, and monthly income of R587.40, which resided in Mafeteng and headed by a male who did not have formal education and whose occupation was of a subsistence farmer. The number of members of households and monthly income are chosen to reflect the sample means of these two variables given in Table 2.10. The rest of the plots in Figure 7.2 show the effects of different variables and categories of categorical variables, in relation to their reference categories, across the distribution of maize availability.

The plots confirm what was shown by the results from the test of equality of distinct slopes. The plot of the category of households that resided in Maseru Foot Hills shows that the slope coefficients at quantiles 0.25, 0.50 and 0.75 are not constant (Figure 7.2). The plots of the categories of households headed by casual workers, salary earners and unemployed heads also show that slope coefficients are changing across quantiles, and thus the effect of these categories, in relation to their respective reference categories, is different across quantiles. The plot of a category of households

that resided in Maseru lowlands shows changing slope coefficients across quantiles (Figure K.1). Slope coefficients that are constant across quantiles, such as that of high school level of education, are around the ordinary least squares estimate shown by the solid line (Figure 7.2).

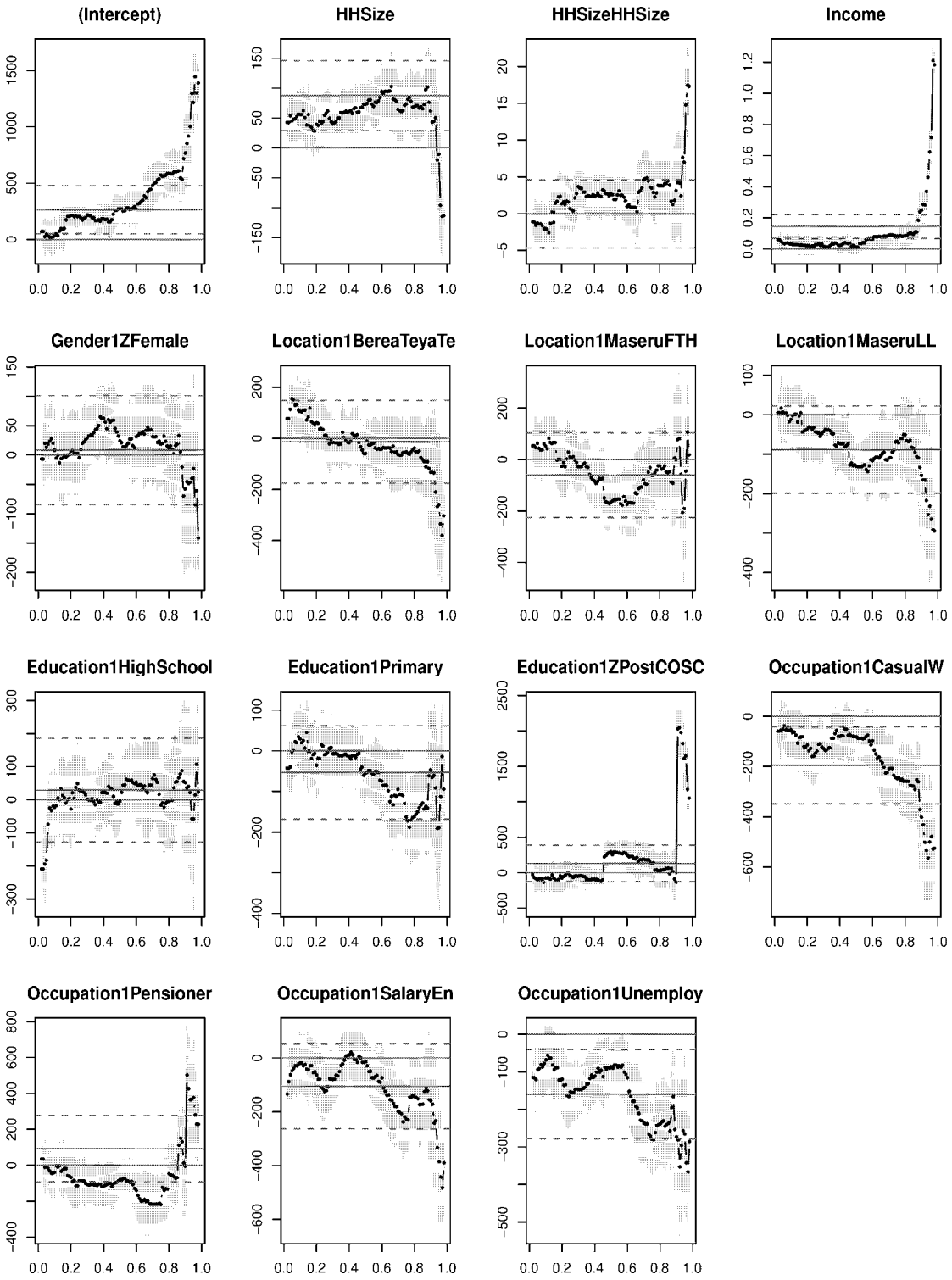

Figure 7.2: Quantile Plot of Maize Availability

7.5 Summary

The plots of the local measure of goodness of fit, pseudo R square, show a poor fit of the quantile regression model at lower tails of the conditional distributions of both maize and sorghum availability, with maize availability having a better fit across quantiles. At a glance, the plots show that the effects of predictor variables on maize availability varies across quantiles, with more significant effects at the upper tail of the distribution, while the effects on sorghum availability remained constant up to $85th$ quantile, and started varying beyond this quantile.

Fitting the regression model at different quantiles shows the effects of predictor variables that were not shown under the mean response of OLS procedure. This is in line with the observation by Baur et al. (2004) that OLS procedure conceals a lot of information about the dependence of the conditional distribution of the response variable on predictor variables in the model, which can be revealed by applying quantile regression. Occupation of heads of households had a significant effect on availability of maize for a household, particularly at the $75th$ percentile. It also had significant effect on sorghum availability at the median regression, while it did not have effect under the mean response. Household size had a positive effect on maize availability under the three quantiles. The significant role played by education status of head of household in determining sorghum availability shown under the mean response regression is also shown by the quantile regression model of the three quantiles.

Standard errors of quantiles estimated by sparsity function using the Hall Sheather bandwidth rule are recommended since they are the smallest when they are compared with those estimated from the Kernel estimator and MCMB resampling, and hence they give high precision of the regression quantiles in estimating regression parameters. The joint test of equality of slope coefficients confirmed what was illustrated by the pseudo R square that the effects of predictor variables on maize availability are not the same across the three quantiles. The results from the test of equality of distinct slopes show that the differences in the effects are associated with one category of each of the two variables, location of households and education status of heads of households, and three categories of the occupation of households heads. Quantile plots confirmed variables that are associated with the differences in the effects. The modelling availability of cereals for households in this chapter and preceding chapters used availability of a given cereal as the continuous variable. In trying to understand availability of cereals further, households were categorized according to specific cereals that were available to them, and the GLM was applied to model availability of cereals in the subsequent chapter.

Chapter 8

Logistic Regression Models and Diagnostics

8.1 Introduction

Sometimes applied research requires generalized linear models (GLMs) such as the logistic regression for analysis of non-Guassian data, not linear regression models. This is due to the nature of the response variables, which is categorical but not continuous. The generalized linear model is a natural generalization of classical linear models (McCullagh and Nelder, 1989), of which the simple logistic regression model is one of the special cases. The choice of the model in studying the relationships that exist between the response variable and a set of predictor variables is largely determined by the scale of measurement of the response variable (Greenland, 1985). Logistic regression is a statistical tool with three variants used to model data where the response variable is a categorical variable with two categories, normally referred to as a binary outcome variable, or its extension to categorical variables with more than two categories that can be nominal or ordinal. Thus logistic regression and its extensions are an alternative to the classical regression used when the response variable is not measured on a ratio scale but takes on limited number of discrete values within a specified range. It is used when the distributional assumptions required for the classical linear regression model do not hold. Simple logistic regression with a dichotomous response, multinomial or polytomous logistic regression model with a nominal or ordinal response are discussed in the subsequent sections.

8.2 Binary Response Logistic Regression Model

Binary logistic regression normally referred to as logistic regression is commonly applied to model data with a binary response (Hosmer and Lemeshow, 2000). The binary response variable takes on

values one for the outcome of interest normally called a success, and 0 for the other outcome normally called a failure. Predictor variables in the model can be continuous or categorical. Generally the probability that the value of the response variable is a success, given values of the predictor variables, is given by $P(Y = 1|X = x) = \pi(x)$ hence the probability that it is a failure is $P(Y = 0|X = x) = 1 - \pi(x)$. Unlike in the general linear regression model where the interest is to study the relationship between the response variable Y and the predictor variables $(X_1, X_2, \ldots, X_p)$, the interest in logistic regression centers on the relationship between the probability of the response variable being a success or equivalently being a failure.

The logistic regression model is used to model the probability of occurrence of an outcome of interest, $\pi(x)$, which is the conditional mean of Y given x for the binomial distribution. Like in the classical regression model, we can consider two cases of the logistic regression model, namely a model with one predictor variable and a model with multiple predictor variables. The logistic model with one predictor variable can be defined in terms of the odds of the outcome of interest as

$$\frac{\pi}{1-\pi} = \exp(\beta_0 + \beta_1 x_1). \tag{8.1}$$

If the predictor variable X is a dichotomous predictor variable with values 0 and 1, the model defined in equation (8.1) can be presented as

$$\frac{\pi(1)/[1-\pi(1)]}{\pi(0)/[1-\pi(0)]} = \exp(\beta_1), \tag{8.2}$$

which indicates that the odds ratio (OR) depends on the regression parameter β_1. Alternatively the logistic model can refer directly to the probability of the outcome of interest and be presented as

$$\pi = \frac{e^{\beta_0+\beta_1 x_1}}{1 + \exp(\beta_0 + \beta_1 x_1)}, \tag{8.3}$$

where π is the expected response, $E(Y|X)$ and β_1 is the regression coefficient.

The link function, $g(\pi) = \log\{\pi/(1-\pi)\}$, called the logit or logistic function is used to transform the model in (8.1). The transformation changes the range of π from [0 to 1] to $[-\infty$ to $+\infty]$ and yields a linear logistic model given by

$$\log\Big(\frac{\pi}{1-\pi}\Big) = \beta_0 + \beta_1 x_1, \tag{8.4}$$

for the log odds of the outcome of interest. This model states that the log odds of the outcome of interest is linearly related with the predictor variable X_1. The parameters β_0 and β_1 are the intercept and slope coefficient respectively. The regression coefficient measures the effect of a unit change in X_1 on the log odds of the probability of the outcome of interest. The sign of β_1 indicates the direction of the change in π. When $\beta_1 > 0$, π increases as X_1 increases and when $\beta_1 < 0$, π decreases as X_1 increases.

The logistic regression model can be generalized to a case with p predictor variables, $X_1, X_2, \ldots, X_p$, so that the odds of the outcome of interest are defined as

$$\frac{\pi}{1-\pi} = \exp(\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p). \tag{8.5}$$

Interaction terms can be included in the model and when an interaction of X_1 and X_2 is included, the model becomes

$$\frac{\pi}{1-\pi} = \exp(\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \beta_{p+1} x_1 x_2). \tag{8.6}$$

Alternatively the logistic model refers directly to the probability of the outcome of interest as in the single predictor case presented in equation (8.3) and can be presented as

$$\pi = \frac{\exp(\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p)}{1 + \exp(\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p)}, \tag{8.7}$$

where β_j is the regression parameter that shows the effect of the jth predictor variable on the log odds that $Y = 1$ when other predictor variables in the model are held constant. In equation (8.7), e^{β_j} is the multiplicative effect of a one unit increase in X_j, on the odds , when other predictor variables are fixed.

Using the link function, $g(\pi)$, to transform the model in equation (8.7) gives the linear logistic model or logit model expressed as

$$\log\Big(\frac{\pi}{1-\pi}\Big) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p. \tag{8.8}$$

The multiple logistic regression model is based on the same assumptions underlying the logistic regression model with one predictor variable. The model states that the log odds of the outcome of interest is linearly related with predictor variables in the model, where β_0 is an intercept and $\beta_1, \beta_2, \ldots, \beta_p$ are slope coefficients. The logistic model in equation (8.8) can also be expressed in a matrix form as

$$\log\Big(\frac{\boldsymbol{\pi}}{1-\boldsymbol{\pi}}\Big) = \mathbf{X}\boldsymbol{\beta}, \tag{8.9}$$

where $\mathbf{X}$ is a matrix of predictor variables or the design matrix that includes a column of ones as the initial column to signify the constant term β_0, and thus $\boldsymbol{\beta}$ is a vector of model parameters including the constant.

The specification of the logistic regression model assumes that the logit of the outcome of interest is a linear combination of the predictor variables in the model. This involves two aspects of the model as shown in equations (8.4) and (8.8). The first aspect is about the link function of the outcome where the assumption is that the logit function is the correct function to use. The second aspect is about the assumptions that all relevant variables are included in the model, no irrelevant

variables are included, and the logit function is a linear combination of the predictor variables in the model. It can happen that the logit function as the link function is not the correct choice or the relationship between the logit and the predictor variables is not linear. The appropriateness of the link function can be tested by computing the square of the linear predictor after fitting the generalized linear model, and refitting the model with a quadratic term (Hinkley, 1985). If the linear predictor is statistically significant and the square of the linear predictor is not, there is evidence that the link function is appropriate.

Estimation of the unknown parameters for logistic regression model is done using the method of maximum likelihood discussed under section 2.4 (Hosmer and Lemeshow, 2000). This method yields values of the parameters that maximize the likelihood or log likelihood of the parameters for the observed data. The log likelihood of a set of independent observations $y_1, y_2, \ldots, y_n$ may be expressed as

$$\ell(\boldsymbol{\beta}; \mathbf{y}) = \sum_i \sum_i y_i x_{ij} \beta_j - \sum_i n_i \log\Big[1 + \exp\Big(\sum_i \beta_j x_{ij}\Big)\Big]. \tag{8.10}$$

The likelihood depends on $\mathbf{y}$ only, through the p linear combinations $\mathbf{X}'\mathbf{y}$, which are the sufficient statistics for the model parameters, $\boldsymbol{\beta}$. The likelihood equations that result from differentiating the log likelihood with respect to the vector $\boldsymbol{\beta}$ can be given by

$$\mathbf{X}'\mathbf{y} = \mathbf{X}'\hat{\boldsymbol{\mu}}, \tag{8.11}$$

where $\hat{\mu}_i = n_i \hat{\pi}_i$. The likelihood equations equate the sufficient statistics to their expected value. Maximum likelihood estimates satisfy the equation

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{z}, \tag{8.12}$$

where $\mathbf{W} = \mathbf{diag}\big[n_i \hat{\pi}_i (1 - \hat{\pi}_i)\big]$ is the $n \times n$ diagonal matrix of weights. The maximum likelihood estimates are unbiased to the first order of approximation, with the asymptotic covariance matrix that is equivalent to the inverse of the Fisher information matrix. The estimated covariance matrix of the estimates can be presented as

$$\widehat{\mathrm{cov}}(\hat{\boldsymbol{\beta}}) = \left\{\mathbf{X}'\mathbf{W}\mathbf{X}\right\}^{-1}. \tag{8.13}$$

The square roots of the main diagonal elements of the matrix are the estimated standard errors of $\hat{\boldsymbol{\beta}}$.

The maximum likelihood estimates of the parameters can be obtained by the Newton-Raphson, Fisher's scoring or iterative weighted least squares methods. The Newton-Raphson method is an iterative method for solving nonlinear equations whose solution determines the point at which a function reaches its maximum (Agresti, 2002). In the iterative weighted least squares, both the adjusted response variable Z and the weight W depend on the fitted values, for which only current estimates are available. The adjusted response variable Z_0 is developed using the current estimate

of the linear predictor $\hat{\boldsymbol{\eta}}_0$ and the corresponding fitted value $\hat{\boldsymbol{\mu}}_0$, derived from the link function $\eta = g(\mu)$.

Once the maximum likelihood estimates are obtained, they can be used to make statistical inferences concerning the relationship between the response variable and predictor variables. These inferences involve assessment of the significance of predictor variables in the logistic regression model. The assessment is done by formulating and testing a statistical hypothesis that the predictor variables in the model are significantly related to the response variable. Three tests, namely, the Wald, likelihood ratio and score tests can be used to test the hypothesis in logistic regression (Agresti, 2002; Kleinbaum and Klein, 2010).

The Wald test statistic that is used to test the significance of each coefficient β in the model or the null hypothesis, $H_0 : \beta = 0$, is the square of the approximate z test statistic given by

$$z = \frac{\hat{\beta}}{\widehat{\mathbf{SE}}(\hat{\beta})}, \tag{8.14}$$

where $\hat{\beta}$ is the maximum likelihood estimate of β and $\hat{SE}(\hat{\beta})$ is the estimated standard error of the estimate. The z statistic follows a standard normal distribution and its square, z^2, which is the Wald statistic is asymptotically chi-square with one degree of freedom (χ^2_1). In the case of multiple logistic regression, the Wald test has the test statistic

$$\begin{aligned} W &= \hat{\boldsymbol{\beta}}' \big[\text{cov}(\hat{\boldsymbol{\beta}})\big]^{-1} \hat{\boldsymbol{\beta}} && (8.15) \\ &= \hat{\boldsymbol{\beta}}' \big(\mathbf{X}'\mathbf{V}\mathbf{X}\big)^{-1} \hat{\boldsymbol{\beta}}. && (8.16) \end{aligned}$$

This has a chi-square distribution with k degrees of freedom, where k is the rank of the covariance matrix, $\text{cov}(\hat{\boldsymbol{\beta}})$.

The likelihood ratio test compares two models, where one model is a special case of another. The larger model is normally referred to as the full model and the smaller model is the reduced model, obtained by setting certain parameters in the full model to zero. The likelihood ratio test tests the hypothesis that extra parameters in the full model are equal to zero. The likelihood ratio test statistic is given by

$$-2\ln\big[\ell(\hat{\boldsymbol{\mu}}_0; \mathbf{y}) - \ell(\hat{\boldsymbol{\mu}}_A; \mathbf{y})\big] = -2\ln\left(\frac{\ell(\hat{\boldsymbol{\mu}}_0; \mathbf{y})}{\ell(\hat{\boldsymbol{\mu}}_A; \mathbf{y})}\right), \tag{8.17}$$

where $\ell((\hat{\boldsymbol{\mu}}_A; \mathbf{y})$ and $\ell(\hat{\boldsymbol{\mu}}_0; \mathbf{y})$ are maximum likelihoods for the full and reduced models respectively. Under the null hypothesis, H_0, the likelihood ratio statistic has approximately a chi-square distribution with the degrees of freedom equal to the difference between the number of parameters in the full and reduced models. When the sample size is large the Wald and likelihood ratio tests give similar results. In the case of small to moderate sample sizes, the likelihood ratio statistic tend to be larger and gives more powerful test than the Wald test (Agresti, 2002).

The score test is used to test the significance of a variable in the model. It uses the size of the score function given by $\mathrm{U}(\beta) = \frac{\partial \ell}{\partial \beta}$. The score statistic is the ratio of the score function to its standard error and it can be expressed as

$$\frac{\mathrm{U}(\beta)}{\sqrt{\mathfrak{I}}} = \frac{\partial(\beta)/\partial\beta_0}{\sqrt{(-E\left[\partial^2 L(\beta)/\partial\beta_0^2\right]}}, \tag{8.18}$$

where $\mathfrak{I}$ is the information matrix. The statistic has an approximate standard normal distribution and its square is a χ_1^2. In the case of multiple logistic regression, the score statistic can be presented as

$$\mathrm{U}'(\boldsymbol{\beta})\mathfrak{I}^{-1}\mathrm{U}(\boldsymbol{\beta}) \tag{8.19}$$

Under H_0, the score statistic is approximately distributed as a chi-square with degrees of freedom equal to the number of parameters tested.

There are other criteria that can be used to select a good model in terms of estimating quantities of interest (Agresti, 2002). The commonly known is the Akaike Information Criterion (AIC) introduced by Akaike in 1971 (Akaike, 1974). The criteria assesses the model fit by comparing its values from fitting the model with an intercept only and the model with predictor variables. The Akaike Information Criterion can be given by

$$AIC = -2\ln(L) + 2p, \tag{8.20}$$

where L is the likelihood function and p is the number of parameters in the model. According to Akaike (1974), a model with a smaller value of AIC is considered to be a better model.

8.3 Multinomial Logistic Regression Model

Logistic regression with binary response variable can be generalized to handle cases where the response variable has more than two categories (Kleinbaum and Klein, 2010). This case is sometimes referred to as logistic regression with polychotomous response variable or multinomial logistic regression model. Multinomial logistic regression model is applicable when the response variable Y has more than two categories that are nominal, meaning that they do not have inherent ordering. One of the categories is designated as a reference or base category and each of the rest of the categories is compared with the reference category (Agresti, 2002). The comparison gives pairs of each category with the reference category, which provide several logistic regression models. When the response variable has m categories, the multinomial logistic regression model consists of $m-1$ logit functions, also called pairwise or generalized logits. The model is based on the assumption that the categories are independent.

The probability that the response variable takes on the value j given a set of predictor variables is given by $P(Y = j \mid \mathbf{x}) = \pi_j(\mathbf{x})$ and $\sum_{j=1}^{m} \pi_j(\mathbf{x}) = 1$. The counts in the m categories of Y are considered as multinomial with probabilities $\pi_1(\mathbf{x}), \pi_2(\mathbf{x}), \ldots, \pi_m(\mathbf{x})$. The multinomial logistic model can be presented in terms of these probabilities as

$$\pi_j(\mathbf{x}) = \frac{\exp(\alpha_j + \boldsymbol{\beta}_j'\mathbf{x})}{1 + \sum_{j=1}^{m-1} \exp(\alpha_j + \boldsymbol{\beta}_j'\mathbf{x})}. \tag{8.21}$$

The probability for the reference category, is given by

$$\pi_j(\mathbf{x}) = \frac{1}{1 + \sum_{j=1}^{m-1} \exp(\alpha_j + \boldsymbol{\beta}_j'\mathbf{x})}, \tag{8.22}$$

where α_j is the intercept common to all the $m-1$ logit functions, and $\boldsymbol{\beta}_j$ are regression parameters that vary from one logit function to another. By comparing each response variable with the reference category, the multinomial logistic regression provides $m-1$ logit functions. The use of the logit functions results with the multinomial logistic regression model consisting of $m-1$ logistic regression equations that are estimated simultaneously. Thus multinomial logistic regression is a statistical tool that fits multiple logistic regression models. When the reference category is the last category, which is the mth category, the jth logistic regression equation can be presented as

$$\log\Big(\frac{\pi_j(\mathbf{x})}{\pi_m(\mathbf{x})}\Big) = \alpha + \boldsymbol{\beta}_j'\mathbf{x}, \tag{8.23}$$

where $\pi_j(\mathbf{x})$ is as defined earlier in this section and $\pi_m(\mathbf{x})$ is the probability of the reference category. The probability distribution of the response variable is multinomial distribution instead of the binomial distribution of the case in binary logistic regression. The multinomial logistic regression describes the effects of predictor variables in the model on the $m-1$ logits, where the effects vary depending on the category paired with the reference category.

Multinomial logistic regression model uses maximum likelihood procedure to estimate regression coefficients, as it is the case with the binary logistic regression. The likelihood function in equation (8.10) can be generalized to accommodate j categories of the response variable.

8.4 Ordinal Logistic Regression Model

In some application of multinomial logistic regression, the categories of the response variable are ordered in nature or the data can be grouped by the researcher in such a way that the categories are ordinal not nominal. The regression model applicable in such situations is multinomial logistic model for ordinal responses normally called ordinal logistic regression model. Like in binary and multinomial logistic regression models, predictor variables may be categorical and/or continuous.

The ordinal logistic regression model takes the ordering of the categories into account and hence it makes full use of the information about the ordering (Kleinbaum and Klein, 2010).

Literature reports a number of logistic regression models that can be used for ordinal responses, such as the cumulative logit model by Walker and Duncan (1967), continuation-ratio model by Feinberg (1980), stereotype logistic model by Anderson (1984) and partial proportional odds model by Peterson and Harrel (1990). The most commonly used model is the cumulative logistic model also called the proportional odds model (McCullagh, 1980). When the response variable have ordered categories, the category with $Y = y_i$ is considered to be in the lower rank than the category with $Y = y_j$, if $i < j$.

Consider m categories of an ordinal response variable Y, the proportional odds model is based on the cumulative probability that the value of the response variable falls in category j or below, $P(Y \leq j)$. According to Agresti (2002), the cumulative probabilities given a vector of predictor variables, $\mathbf{x}$, can be defined by

$$P(Y \leq j \mid \mathbf{x}) = \frac{\exp(\alpha_j - \boldsymbol{\beta}'\mathbf{x})}{1 + \exp(\alpha_j - \boldsymbol{\beta}'\mathbf{x})}, \quad j = 1, 2, \ldots, m \tag{8.24}$$

The proportional odds model aims to model several logits of cumulative probabilities called cumulative logits. These logits can be presented as

$$\begin{aligned} \text{logit}(\gamma_j) &= \log \frac{P(Y \leq j \mid \mathbf{x})}{1 - P(Y \leq j \mid \mathbf{x})} && (8.25) \\ &= \log \frac{\pi_1(\mathbf{x}) + \ldots + \pi_j(\mathbf{x})}{\pi_{j+1}(\mathbf{x}) + \ldots + \pi_m(\mathbf{x})}, \quad j = 1, 2, \ldots, m-1. && (8.26) \end{aligned}$$

Each of the cumulative logits uses all the m categories of the response variable and $m - 1$ logits are modeled, one for each of the cut points of the response variable, 1 versus $2, 3, \ldots, m - 1$; $1, 2$ versus $3, \ldots, m - 1$ and others. The proportional odds model uses all the cumulative logits and can be presented by

$$\text{logit}(\gamma_j) = \alpha_j + \boldsymbol{\beta}'\mathbf{x}, \quad j = 1, 2, \ldots, m-1, \tag{8.27}$$

where each cumulative logit has its own intercept α_j, satisfying the inequality that $\alpha_1 \leq \alpha_2 \leq \ldots \leq \alpha_m$. The intercept depends on j and it increases with j, since the probability $P(Y \leq j \mid \mathbf{x})$ increases with j for given values of predictor variables. The proportional odds model assumes that the cumulative logits can be represented as parallel linear functions of predictor variables. In other words for each cumulative logit the slopes $\boldsymbol{\beta}$ remain the same. This assumption is referred to as the assumption of proportional or parallel odds. The proportional odds model is based on the assumption that the observations of the response variable are independent and have multinomial distribution. Maximum likelihood estimation procedure can be used to estimate parameters in the proportional odds model.

The score test can be used to test the assumption of proportional odds, so that the proportional odds is valid. It tests the null hypothesis $H_0 : \beta_j = \boldsymbol{\beta}$, that is the effect of the predictor variables on the odds of the category j of a response variable or below is the same for all j. This test compares p parameters for p predictor variables across $(m-1)$ logits, where m is the number of categories of a response variable. Let the parameter vector $\boldsymbol{\theta}$ of the proportional odds model consists of the parameters $[\alpha_1, \alpha_2, \ldots, \alpha_{m-1}, \boldsymbol{\beta}'_1, \boldsymbol{\beta}'_2, \ldots, \boldsymbol{\beta}'_{m-1}]$, where $\boldsymbol{\beta}'_j = \beta_{1j}, \beta_{2j}$ when there are two predictor variables in the model. Then Abeyskera and Sooriyarachchi (2008) presented the score statistic that tests the hypothesis, $H_0 : \beta_1 = \beta_2 = \ldots = \beta_{m-1}$, as

$$S = U'(\hat{\boldsymbol{\beta}}_0)W^{-1}(\hat{\boldsymbol{\beta}}_0)U(\hat{\boldsymbol{\beta}}_0), \tag{8.28}$$

where U is the multivariate analogue of the quasi-score function given by

$$U = \sum_{i=1}^{n} \Big(\frac{\partial \mu_i}{\partial \boldsymbol{\beta}}\Big) V_i^{-1}(Y_i - \mu_i). \tag{8.29}$$

This function is asymptotically multivariate normally distributed with μ_i and V_i, the mean and the working covariance matrix of Y_i, respectively. The statistic S is asymptotically χ^2 with $p(m-2)$ degrees of freedom, where p is the number of parameters corresponding to the predictor variables in the model. If the score test is not significant, there is evidence that the effect of the predictor variables on the odds of the category j of a response variable or below is the same for all j. Thus the odds ratios can be considered constant across all possible cut points of the response variable.

When the assumption of proportional odds does not hold, alternative ordinal regression models based on alternative assumptions about the ordinal nature of the response variable may be used (Kleinbaum and Klein, 2010). Examples of such models are the continuation-ratio model, stereotype logistic model and partial proportional odds model.

8.5 Goodness-of-Fit in Logistic Regression

The assessment of goodness-of-fit for any generalized linear model should start with examining the deviance, Pearson chi-square statistic and, if possible, a goodness of fit statistic based on deciles (Hosmer et al., 1991). The first two measures are functions of residuals, which can also be used to assess goodness-of-fit in multinomial and ordinal logistic regression models. The Pearson statistic is defined as

$$\chi^2 = \sum_{i=1}^{\mathrm{n}} \chi_i^2 = \sum_{i=1}^{n} \frac{(y_i - n_i\hat{\pi}_i)^2}{n_i\hat{\pi}_i(1-\hat{\pi}_i)}, \tag{8.30}$$

where χ_i is the Pearson residuals, which can be defined as

$$\chi_i = \frac{y_i - n_i\hat{\pi}_i}{\sqrt{n_i\hat{\pi}_i(1-\hat{\pi}_i)}}. \tag{8.31}$$

The larger the value of the Pearson statistic, the worse the model fits the data. The Pearson statistic is asymptotically χ^2 with $(n-p)$ degrees of freedom, where n is the number of subjects, experimental or observational units and p is the number of parameters in the logistic regression model.

The diagonal elements of the hat matrix, h_{ii} defined under diagnostics for linear regression model, can be used to standardize the Pearson residual. The standardized Pearson residuals is given by

$$r_i^P = \frac{\chi_i}{\sqrt{1-h_{ii}}} = \frac{y_i - n_i\hat{\pi}_i}{\sqrt{n_i\hat{\pi}_i(1-\hat{\pi}_i)(1-h_{ii})}}, \tag{8.32}$$

An absolute value of r_i^P larger than 2 and 3 shows that there is lack of fit (Agresti, 2002).

The deviance is the log-likelihood ratio test statistic of a saturated model with n parameters against the fitted model with p parameters. The deviance for a logistic model is an analogy of the residual sum of squares in ordinary least squares regression for the linear model. The larger the deviance, the worse the logistic model fits the data, and the smaller the deviance the better the fit of the logistic model. Like the Pearson statistic, the deviance is asymptotically χ^2 with $(n-p)$ degrees of freedom. It is worth noting that when the asymptotic approximation is doubted, such as in binary data, deviance cannot be used to provide an absolute goodness-of-fit test because the Pearson chi-square and deviance statistics are identical and become the scaled residuals sum of squares with χ^2_{n-p} distribution for normal models (Lee et al., 2006). In performing tests of hypotheses regarding the fit of the model, the deviance is compared with the percentiles of a χ^2 distribution. The degrees of freedom is determined by the number of observations less the number of parameters estimated.

In the case of the binomial GLM, where the logistic regression model is a special case with $n_i = 1$, the deviance statistic is

$$D = \sum_{i=1}^{\mathrm{n}} t_i^2 = 2\sum_{i=1}^{n} y_i \log\frac{y_i}{n_i\hat{\pi}_i} + (n_i - y_i)\log\frac{n_i - y_i}{n_i - n_i\hat{\pi}_i} \tag{8.33}$$

where $t_i = 2\left(y_i \log\frac{y_i}{n_i\hat{\pi}_i} + (n_i\ -\ y_i)\log\frac{n_i - y_i}{n_i - n_i\hat{\pi}_i}\right)^{\frac{1}{2}}$.
The deviance statistic is build from the deviance residual, which can be given by

$$d_i = \sqrt{(t_i)} \times (y_i - n_i\hat{\pi}_i). \tag{8.34}$$

Similarly, the deviance residuals can be standardized and be presented as

$$r_i^D = \frac{d_i}{\sqrt{1-h_{ii}}} \tag{8.35}$$

Plots of residuals against predictor variables or linear predictor values may detect lack of fit of the model. The residuals lose relevance in detecting lack of fit when fitted values are small, so do the

Pearson statistic X^2 and deviance G^2 (Agresti, 2002).

Hosmer and Lemeshow (1980) used a contingency table approach type and developed seven goodness-of-fit test statistics that involve grouping based on estimated probabilities obtained from the fitted logistic regression model, and grouping based on the fixed pre-determined cutoff points. Four of these statistics have a $\chi^2(g-2)$ distribution and the other three have a $\chi^2(2g-5)$ distribution, where g is the number of groups. The majority of the applications of logistic regression use $g = 10$, however, less or more than 10 groups can be used (Lemeshow and Hosmer (1982)). In developing the test statistic, Hosmer and Lemeshow (1980) required that $g > (p+1)$, where p is number of predictor variables in the model.

Two of the seven test statistics are commonly used. These are $\hat{C}_g^*$, based on grouping the data according to the probabilities of the n observations, $\hat{\pi}(x_1), \hat{\pi}(x_2), \ldots, \hat{\pi}(x_n)$, estimated from the fitted model, and $\hat{H}_g^*$ based on grouping the data according to the fixed cutoff points. Under the null hypothesis H_0, if the number of predictor variables plus one is less than the number of groups, $(p+1 < g)$, and the logistic regression model is the correct model, $\hat{C}_g^*$ has a distribution closely approximated by χ^2 with $(g-2)$ degrees of freedom. This statistics can be defined as

$$C_g^* = \sum_{k=0}^{1} \sum_{r=1}^{g} \frac{(o_{kr} - e_{kr})^2}{e_{kr}}, \tag{8.36}$$

where n_r is the number of observations in the rth group, k is 0 for observations without the attribute of interest and 1 for observations with the attribute of interest Hosmer and Lemeshow (1980). The quantities,

$$o_{1r} = \sum_{i=1}^{n_r} y_i \quad \text{and} \quad o_{0r} = \sum_{i=1}^{n_r} (1 - y_i) \tag{8.37}$$

denote the observed number of observations with the attribute and without the attribute of interest, respectively. The quantities,

$$e_{1r} = \sum_{i=1}^{n_r} \hat{\pi}(x_i) \quad \text{and} \quad e_{0r} = \sum_{i=1}^{n_r} (1 - \hat{\pi}(x_i)) \tag{8.38}$$

denote the expected number of observations with and without the attribute of interest, respectively. The first group of the g groups contains approximately the smallest $n_1' = \frac{n}{g}$ values of $\hat{\pi}(x_i)$, the second group contains approximately the second smallest $n_2' = \frac{n}{g}$ values of $\hat{\pi}(x_i)$, up to the last group that contains approximately the highest $n_g' = \frac{n}{g}$ values of $\hat{\pi}(x_i)$.

The test statistic $\hat{C}_g^*$ is computed using empirical deciles. The approach that uses deciles has an advantage that it ensures that each group has a fair number, $(\frac{n}{g})$, of observations. However, it has a disadvantage that the actual values of the estimated probabilities of the outcome of interest are not included. Hosmer and Lemeshow (1980) developed $\hat{H}_g^*$ as an alternative goodness-of-fit

test, where grouping is done according to the fixed cutoff points. The cutoff points are the predetermined values of estimated probabilities from the fitted model given by $0.0 \leq \hat{\pi}(x_i) < 0.1,\ 0.1 \leq \hat{\pi}(x_i) < 0.2,\ \ldots,\ 0.9 \leq \hat{\pi}(x_i) < 1.0$. The test statistic $\hat{H}_g^*$ can be defined as

$$H_g^* = \sum_{k=0}^{1} \sum_{r=1}^{g} \frac{(o'_{kj} - e'_{kr})^2}{e'_{kr}}, \tag{8.39}$$

where o'_{kr} and e'_{kr} are as defined in equations (8.37) to (8.38), n_r in this case is the number of observations whose estimated probabilities fall in the rth cutoff point, and $r = 1, 2, \ldots, g$. Like C_g^*, the statistic H_g^* has a distribution closely approximated by χ^2 with $(g-2)$ degrees of freedom. Hosmer and Lemeshow (1980) noted that H_g^* was found to be more powerful than C_g^*, however, H_g^* has a disadvantage that when the sample size n is small n_r could be small for some cutoff points.

Fagerland et al. (2008) derived three goodness-of-fit test statistics for multinomial regression model. The statistics are the Pearson chi-square statistic, standardized statistic derived from the work of Osius and Rojek (1992), and the modification of Hosmer and Lemeshow (1980) test statistic to suit multinomial logistic model. The Pearson chi-square statistic compares binary indicator variables $\mathbf{y}_i$ with the predicted probabilities, $\hat{\pi}_i$, estimated from the fitted model. The statistic can be defined as

$$\chi^2 = \sum_{i=1}^{\text{n}} \sum_{j=0}^{\text{m-1}} \frac{(y_{ij} - \hat{\pi}_{ij})^2}{\hat{\pi}_{ij}}, \tag{8.40}$$

where y_{ij} is the observed frequency in cell (i, j), and $\hat{\pi}_{ij}$ is the expected frequency. The Pearson chi-square statistic is asymptotically χ^2 with $n \times (m-1)$ degrees of freedom. Fagerland et al. (2008) noted that if the model is fitted with more than few binary predictor variables, the expected frequencies will be too small for the asymptotic χ^2 distribution to hold, though they did not indicate how much is considered to be more than few and to be small in this case.

A standardized statistic with p-values computed using the standard normal distribution suggested by Osius and Rojek (1992), was used to derive the goodness-of-fit statistic. The standardized statistic is defined by

$$z = \frac{(X^2 - \hat{\mu})}{\hat{\sigma}}, \tag{8.41}$$

where the asymptotic mean $\hat{\mu} = n \times (m-1)$, and the estimator of the asymptotic-variance was obtained using the estimator $\hat{\boldsymbol{\beta}}$. The statistic z has the asymptotic standard normal distribution, under the hypothesis that the correct model is fitted.

Fagerland et al. (2008) adapted the Hosmer and Lemeshow (1980) statistic based on the deciles formed from the estimated probabilities and proposed the statistic based on the deciles formed from the sum of estimated probabilities, $\sum_{j=1}^{m-1} \hat{\pi}_{ij} = 1 - \hat{\pi}_{i0}$, for testing goodness of fit for multinomial

logistic regression. The sizes of g groups $n_1, \ldots, n_g$ are determined using a similar procedure as that used for C_g^*. When $\frac{n}{g}$ is not an integer, all groups will not have the same number of observations. However, the value of the statistic is not affected by the small imbalances in the size of groups, unless there are a number of tied values. The test statistic for multinomial logistic regression is given by

$$C_g^* = \sum_{r=1}^{g} \sum_{j=1}^{m-1} \frac{(O_{rj} - E_{rj})^2}{E_{rj}}, \tag{8.42}$$

where O_{rj} and E_{rj} are the sums of the observed frequencies and estimated probabilities for each category of the response variable in each group. The two sums are defined as

$$O_{rj} = \sum_{l \in \Omega_r} y_{li} \quad \text{and} \quad E_{rj} = \sum_{l \in \Omega_r} \hat{\pi}_{lj}, \tag{8.43}$$

for r=1, 2, ..., g, j=0, 1, ..., m-1, where Ω_r denote the indices of the $\frac{n}{g}$ observations in group r. The statistic C_g has an approximate χ^2 distribution with $(g-2) \times (c-1)$ degrees of freedom.

Lipsitz et al. (1996) noted that the deviance and Pearson chi square statistics can be used to assess goodness-of-fit for ordinal logistic regression models with categorical predictor variables. However, the two statistics require that most of the expected counts formed by cross classifying the categories of the response variable are greater than 5. If more than 20% of the expected counts are less than 5, they may not be appropriate for testing the goodness-of-fit for ordinal logistic regression model. The authors then proposed an extension of Hosmer and Lemeshow (1980) goodness-of-fit test that suits ordinal logistic regression model under the circumstance.

The proposed goodness-of-fit statistic is based on scores and it starts by assigning a score s_j to category j of the response variable. The score can be the actual numerical value of the response or the midpoint of the interval when the response variable is a grouped continuous variable. In cases where the response variable has no underlying numerical scale, integer scores, such as $1 \equiv \text{poor}, 2 \equiv \text{moderate}$ and $3 \equiv \text{good}$, can be used. Then the fitted score or predicted mean score is calculated as

$$\hat{\mu}_i = \sum_{j=1}^{m} s_j \hat{\pi}_{ij}; \quad i = 1, 2, \ldots, n, \tag{8.44}$$

where $\hat{\pi}_{x_{i1}}, \hat{\pi}_{x_{i2}}, \ldots, \hat{\pi}_{x_{im}}$ are the predicted probabilities for the ith observation in category j of the response variable. The observed score for the ith observation is given by

$$Z_i = \sum_{j=1}^{m} s_j Y_{ij} = \mathbf{s}'\mathbf{Y}_i, \tag{8.45}$$

where $\mathbf{s} = (s_1, s_2, \ldots, s_m)'$. The goodness of fit statistic is formed by partitioning subjects into regions based on the percentiles of the mean scores $\hat{\mu}_i$. Lipsitz et al. (1996) followed Hosmer and

Lemeshow (1980) approach for a binary response variable and suggested 10 groups of approximately equal size. The grouping is done according to the mean scores, where size is determined as in C_g^*. Abeyskera and Sooriyarachchi (2008) suggested that as a general rule, the value of g should satisfy the inequality $6 \leq G < n/5m$.

Given the partitioning of the data, the goodness-of-fit statistic is developed by defining the $g-1$ group indicators and an alternative model is used to assess the goodness-of-fit of the proportional odds model. If the proportional odds model is correctly specified, the likelihood ratio, Wald and score statistics have an approximate χ^2 distribution with $(g-1)$ degrees of freedom when the sample size n is large.

8.6 Logistic Regression Model Diagnostics

The step that naturally follows the fitting of the regression model is that of assessing the fit critically (Pregibon, 1981). The fit of the model is normally assessed using diagnostics, which are used to identify observations that are not well explained by the fitted model, as well as observations that unduly influence some important aspects of the fit of the model. Diagnostics used in logistic regression are analogues of some of the linear regression model diagnostics. However, an observation can have a big influence in ordinary regression than a binary observation can have in logistic regression (Agresti, 2002).

Other regression diagnostics tools that include plots of ordered residuals against normal percentiles, and case deletion diagnostics can be used to assess the fit of the model. However, when there are more than two predictor variables in the logistic regression model, plots of residuals are not adequate in identifying influential outliers Sarkar et al. (2011), and case deletion diagnostics can be used. Case deletion diagnostics measure an influence of an observation on parameter estimates and different quantities of the fitted model.

In logistic regression as it is the case in the classical regression, the basic components that build diagnostics measures used to identify outlying and influential points, are residuals and diagonal elements of the projection matrix. Residuals for the logistic regression are components of the Pearson chi-square and deviance as discussed under the section of goodness-of-fit in logistic regression. The projection matrix for the logistic regression model can be denoted by $\mathbf{M}$, which can be defined as

$$\mathbf{M} = \mathbf{I} - \mathbf{H} = \mathbf{I} - \mathbf{W}^{1/2}\mathbf{X}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}^{1/2}. \tag{8.46}$$

The examination of the Pearson residuals χ_i, deviance residuals d_i and diagonal elements of the projection matrix m_{ii}, can point to outlying and influential points in the data set (Pregibon, 1981). Pregibon recommended the use of index plots of the three quantities $\chi_i, d_i,$ and m_{ii}, and plots of

each of the quantities against fitted values, for specific cases. The index plots are based on plotting each of the quantities against the observations index numbers. Another plot proposed by Pregibon (1981) is based on the weighted hat matrix

$$\mathbf{H}^* = \mathbf{W}^{1/2}\mathbf{X}^*(\mathbf{X}^{*\prime}\mathbf{W}\mathbf{X}^*)^{-1}\mathbf{X}^{*\prime}\mathbf{W}^{1/2}, \tag{8.47}$$

where $\mathbf{X}^* = (\mathbf{X}; \mathbf{z})$ with diagonal elements $h^*_{ii} = h_{ii} + \chi^2_i/\chi^2$, which are called leverages as h_{ii} in the linear regression model, χ^2 and χ^2_i are defined in equations 8.30 and 8.31. Values of h^*_{ii} near one correspond to observations that are either poorly fit by the model and extreme with respect to the predictor variables or both. These will be shown by large values of the Pearson residuals and leverage, respectively. The plot of χ^2_i/χ^2 against h_{ii} can be used to detect poorly fit observations and high leverage points.

The discussed quantities identify observations that are not well explained by the model or have unduly effect on some quantities of the fitted model, but cannot measure the influence of such observations on quantities of the fitted model. Pregibon (1981) addressed this limitation by generalizing diagnostics for linear regression models by Welsch and Kuh (1977) and developing diagnostics for logistic regression models. The diagnostics measure the influence of the ith observation on regression coefficients, goodness-of-fit, and the remaining $n-1$ observations.

The suggested diagnostic that measures the influence of the ith observation on the model coefficients is given by

$$\Delta_i\hat{\boldsymbol{\beta}}^1 = \frac{(\mathbf{X}'\mathbf{V}\mathbf{X})^{-1}\mathbf{x}_i s_i}{(1-h_{ii})}, \tag{8.48}$$

where $\Delta_i\hat{\boldsymbol{\beta}}^1 = \hat{\boldsymbol{\beta}}'(\bar{w})$ is the derivative of the one-step estimate, $\hat{\boldsymbol{\beta}}(\bar{w})$, evaluated at $\bar{w}$ and $s_i = y_i - \hat{y}_i = y_i - n_i\hat{p}_i$. Plots of $\frac{\Delta_i\hat{\boldsymbol{\beta}}^1_j}{\mathrm{SE}(\hat{\boldsymbol{\beta}}^1_j)}$ against the observations index numbers can be used to identify observations that influence some coefficients.

In realizing that it becomes cumbersome to look at the index plots of each coefficient when there are many predictor variables in the model, Pregibon (1981) adapted the aggregate measure of influence by Cook (1977) to develop the confidence interval displacement diagnostic. The developed diagnostics is given by

$$c_i = \frac{\chi^2_i h_{ii}}{(1-h_{ii})^2}. \tag{8.49}$$

This diagnostic is based on an approximate ellipsoidal confidence region. The plots of the confidence interval displacement against the index numbers or predicted values $\hat{\theta} = \mathbf{x}\hat{\boldsymbol{\beta}}$ can be used to identify influential observations. Another confidence interval displacement diagnostic similar to c_i was suggested by Pregibon (1981) as

$$\bar{c}_i = \frac{\chi^2_i h_{ii}}{(1-h_{ii})}. \tag{8.50}$$

This measure of influence has a smaller value, when it is compared with c_i in equation (8.49). It measures the overall influence of the ith observation in the fitted values for all cases in the data set except the deleted case, while c_i include all the cases in the data set.

Pregibon (1981) noted that observations with a big influence on the model are likely to have a big effect on the quality of the fit of the model, which is determined by the Pearson and deviance statistics. Hence the proposal of measures of influence that identify observations that have substantial influence on the two statistics, D and χ^2. The diagnostic proposed for the deviance statistic D is given by

$$\Delta_i D = d_1^2 + \frac{\chi_i^2 h_{ii}}{(1 - h_{ii})}. \tag{8.51}$$

It measures a change in deviance associated with deleting the ith observation. Index plot of $\Delta_i D$ or plot of square of the components of the deviance t_i^2 against $\bar{c}_i$ can be used for identifying observations that influence the deviance statistic. Normal probability plot based on the ordered values of normal studentized residuals, $\frac{t_i}{\sqrt{m_{ii}}}$, was suggested as an alternative plot for identifying influential observations.

The diagnostic proposed for identifying observations that influence the Pearson Statistic χ^2 is given by

$$\Delta_i \chi^2 = \frac{\chi_i^2}{(1 - h_{ii})}. \tag{8.52}$$

This is the logistic regression analogue to $\Delta_i RSS$, the change in residual sum of squares, of a linear regression model. Similarly, index plots of $\Delta_i \chi^2$ can be used to identify observations that affect the Pearson statistic.

Pregibon (1981) also developed a diagnostic that measures the extent to which the ith observation influence the fit of the remaining $n - 1$ observations. This analysis is normally based on a subset of $\lambda \leq 0.05n$ observations that clearly or marginally stood out in the diagnostics plots that were discussed earlier. This diagnostic can only show the magnitude of the change in the fit not the direction as whether the fit is moving toward or away from y_i. The diagnostic is given by

$$\Delta_i d_i^2 = \frac{2\chi_i^2 h_{li}\chi_l}{(1 - h_{ii})} + \frac{\chi_l^2 h_{li}^2}{(1 - h_{ll})^2}, \tag{8.53}$$

which is a function of χ_i, h_{li}, and χ_l. The index plot of $\Delta_i d_i^2$ against i for a subset of observations with indices $L = (l_1, l_2, \ldots, l_\lambda)$ can be used to identify influential observations.

Diagnostics have not yet been extended to multinomial and ordinal logistic regression models, and thus diagnostics for logistic regression model can be used where separate binary responses are used (Hosmer and Lemeshow, 2000).

Chapter 9

Application of Logistic Regression Models to Households Data

9.1 Introduction

We applied the three variants of logistic regression to model availability of the three major food cereals in Lesotho, namely, maize, sorghum and wheat. In the application of linear regression models to households data, availability of a given cereal, for example sorghum, used as the response variable is a continuous variable. It was defined as the amount of sorghum available to a household, measured in kilograms. In this chapter, we further investigated availability of cereals to households by creating three categorical response variables. These are a binary response variable, a polychotomous or multiple response variable with nominal and ordinal categories, respectively.

The binary response variable is created for sorghum and wheat only, where the two responses are "availability of sorghum or wheat" and "non-availability of sorghum or wheat". The binary response variable is not created for maize because the cereal was available to all households included in the studied sample, thus there were no households without maize. In creating a multiple response variable with nominal categories, all the three food cereals were taken into consideration and four categories of households were created. The first category is of households with availability of maize only, the second is of households with availability of maize and sorghum, the third is of households with availability of maize and wheat, and the last category is of households with availability of all the three food cereals. The creation of multiple responses compensated for the loss of information that resulted from dichotomizing households as households with availability of a given food cereal and households without availability, without indicating if households had availability of one, two or all the three cereals.

The third categorical variable has ordinal categories in the sense that households were ranked

according to whether one cereal, two or all the three cereals were available to them. Thus the created ordinal categorical variable has three categories as; households with availability of maize only, households with availability of maize and either sorghum or wheat, and households with availability of maize, sorghum and wheat. This ranking was done with the realization that a household with availability of all the three food cereals might be better off in terms of access to a variety of cereals than a household with availability of maize only, or availability of maize and one of sorghum or wheat. On the other hand a household with availability of maize and one of sorghum or wheat was better off than a household with availability of maize only.

Ordering of categories gives additional information about which households had better entitlement or access to food cereals than when categories are nominal without considering order. The difference between the three ordinal categories and the four nominal categories is that in ranking the responses, the second and third nominal categories were collapsed to have a category that consists of households with maize and one additional food cereal. In the case of nominal categories, the point was to show that there were households with access to maize and sorghum and those with access to maize and wheat. These two nominal categories do not say much about which category of households was better off in terms of variety of food cereals available to households.

9.2 Logistic Regression Modeling of Wheat Availability

The LOGISTIC procedure was used to fit the simple logistic regression model as well as to perform logistic regression diagnostics. The procedure has the INFLUENCE and IPLOTS options that produce diagnostics plots. The fitted model compares availability and non-availability of wheat for households. The predictor variables in the model are the characteristics of households that were used to fit the linear and quantile regression models. The response variable is binary with categories, '0 = availability of wheat to a household' and '1 = non availability of wheat to a household'.

Out of the 296 households in the sample 152 (51%) had wheat available to them and 144 (49%) did not, during the observational period. The iterative method used to estimate the logistic regression coefficients is Newton-Raphson method. The Akaike Information Criterion (AIC) and the Log-Likelihood (-2 Log L) tests were used to assess the model fit by comparing their values from fitting the model with an intercept only and the model with predictor variables. The values for the model with an intercept only are 412.127 and 410.127 for AIC and -2 Log L, respectively, while the values from the model fitted with predictor variables are 377.396 and 347.396 respectively. The smaller values of the test statistics associated with the fitted model than those from the model with the intercept only suggest that the model fits the data well.

The appropriateness of the logit link function used in fitting the logistic regression model was

checked by computing the square of the linear predictor and refitting the model with a quadratic term. The results show a small p-value of the linear predictor and a large p-value of the quadratic term (Table 9.1). These values suggest that the linear predictor is significant while the quadratic term is not significant. Thus the logit is the correct link function for this model.

Table 9.1: Results of the Link Function Test for Wheat Data

Variable	DF	Chi-square	P-value
Constant	1	-0.25	0.804
Linear Predictor	1	6.33	<.0001
Squared Linear Predictor	1	0.42	0.676

The likelihood ratio, score and Wald tests, that test the global null hypothesis that all the regression coefficients associated with the predictor variables in the model are equal to zero against the alternative hypothesis that at least one of the coefficients is not equal to zero, are all significant at 5% level. This indicates that there is enough evidence that at least one of the predictor variables in the model predicts the probability of availability of wheat to households.

Table 9.2: Type III Analysis of Effects on Wheat Availability

Effect	DF	Wald Chi-square	P-value
Household Size	1	0.0000	0.9962
Income	1	0.8752	0.3495
Sex	1	1.6472	0.1993
Location	3	8.5468	0.0360
Education	3	7.2824	0.0634
Occupation	4	12.7860	0.0124

The Wald test with p-values that are smaller than 0.05 show that only two variables are significant at 5% level of significance (Table 9.2). The variables that correspond to the small p-values are the location of a household and education level of head of household. The significance of the variables shows that there is evidence that the two variables predicts the probability of availability of wheat to a household. In particular one category of location, Berea with a p-value of 0.008, and one category of occupation of unemployed household heads with a p-value of 0.006, are significant at 5% level of significance (Table 9.3). This suggests that the odds of availability of wheat for households which resided in Berea was significantly different from that of households which resided in Mafeteng. The results further suggest that the odds of availability of wheat for households headed by unemployed people was significantly different from that of households headed by subsistence farmers.

Table 9.3 further presents 95% confidence intervals of odds ratios, obtained by exponentiating con-

Table 9.3: Logistic Regression Estimates for Wheat Availability

Variable	Estimate	Standard Error	Wald Chi-square	P-value	Odds Ratio	95% Conf. Interval of Odds Ratio	
Household Size	-0.00	0.06	-0.00	0.996	1.00	0.897	1.114
Income	0.00	0.00	0.94	0.350	1.00	1.000	1.001
Sex (ref=Male)							
Female	-0.38	0.30	-1.28	0.199	0.68	0.383	1.222
Location (ref=Mafeteng)							
Berea	1.56	0.59	2.64	0.008	4.78	1.499	15.234
Maseru Foothill	-0.33	0.54	-0.61	0.545	0.72	0.249	2.083
Maseru Lowland	0.27	0.35	0.78	0.438	1.31	0.664	2.576
Education (ref=No educ)							
Primary	-0.00	0.36	-0.00	0.998	1.00	0.497	2.010
High School	0.42	0.44	0.91	0.340	2.97	1.061	8.324
Post High School	1.57	1.18	1.34	0.181	4.82	0.481	48.380
Occupation (ref=Farmer)							
Casual Worker	-0.25	0.48	-0.52	0.603	0.78	0.303	2.000
Pensioner	-0.28	0.58	-0.49	0.625	0.76	0.244	2.333
Salary Earner	0.12	0.51	0.23	0.820	1.12	0.415	3.039
Unemployed	-1.03	0.37	-2.77	0.006	0.36	0.171	0.740

fidence intervals of parameter estimates shown in the first column. The confidence intervals for categories of households that resided in Berea [1.499, 15.234], households headed by unemployed people [0.171, 0.740] and households headed by people with high school education [1.06, 8.32] do not contain the value one. This indicates that the results of the confidence intervals are in agreement with the results from the Wald test that availability of wheat for households in these categories was significantly different from that of their counterparts in the respective reference categories.

Considering the effect of the location of a household on availability of wheat, households that resided in Berea with an odds ratio of 4.78 were 4.78 more likely to have availability of wheat than households that resided in Mafeteng, when other variables in the model are adjusted for. In the case of the effect of education status of household head, households headed by people who attained high school education with an odds ratio of 2.97 were 2.97 times more likely to portray availability of wheat than households headed by people with no formal education. Concerning the effect of occupation of household head, households headed by unemployed people have an odds ratio of 0.36, which is less than one, meaning such households were less likely to portray availability of wheat than their counterparts headed by subsistence farmers. The effect in each category was observed when the rest of the variables in the model were held constant.

The goodness-of-fit of the logistic regression model was assessed using the Pearson chi-square, deviance and Hosmer and Lemeshow test. The tests were used to test the null hypothesis that the logistic regression model fits the data against the alternative that the model does not fit the data.

Table 9.4: Grouping for the Hosmer and Lemeshow Goodness-of-Fit Test for Wheat Availability

Group	Total	Success = Wheat		Failure = No Wheat	
		Observed	Expected	Observed	Expected
1	30	5	5.74	25	24.26
2	30	12	7.96	18	22.04
3	31	8	10.27	23	20.73
4	30	11	12.19	19	17.81
5	30	17	14.23	13	15.77
6	31	12	16.87	19	14.13
7	30	18	18.32	12	11.68
8	30	24	20.64	6	9.36
9	30	24	23.96	6	6.04
10	24	21	21.83	3	2.17

The value of Pearson chi-square statistic is 278.58 with a p-value of 0.2857. The large p-value shows that the null hypothesis cannot be rejected and hence the model fits the data well. The value of deviance statistic is 348.65 with a p-value of 0.0041. Contrary to the results of the Pearson statistics, the p-value of the deviance statistic is small leading to the rejection of the null hypothesis that the model fits the data. However, the results of the Hosmer and Lemeshow test having a value 10.07 with 8 degrees of freedom and the large p-value of 0.2599, are in agreement with the the results of the Pearson test that the model fits the data well, The disagreement between the results from the Pearson chi-square and deviance statistics is explained by Lee et al. (2006) noting that in the case where the asymptotic approximation is doubted, like in the binary data that we are dealing with in this section, deviance cannot be used to provide an absolute goodness-of-fit test. This is so because the two statistics are identical and become the scaled residuals sum of squares with χ^2_{n-p} in the case of normal models. Hence we conclude on the basis of the results from the Pearson chi-square statistic and Hosmer and Lemeshow test that the model fits the data well. The grouping of observations for the Hosmer and Lemeshow test according to the probabilities of the observations is presented in Table 9.4.

Influential observations were identified using index plots of diagnostics measures presented in Figures 9.1, 9.2, 9.3, and Figures L.1, L.2, L.3 of Appendix L. The index plots of Pearson and deviance residuals in Figure 9.1 suggest that observations 166, 168, 177 and 179 are poorly fitted by the model. These are outlying observations with large values of the residuals. The bottom left plot of diagonal elements of the hat matrix shows observation 169 as a high leverage point that is extreme with respect to predictor variables.

The top left and right plots of confidence interval displacement C and CBar in Figure 9.2 show that observation 122 [displacement C = 1.26 and displacement C = 1.07] and observation 129 [displacement C = 1.25 and displacement CBar = 0.87], have undue influence on individual parameter

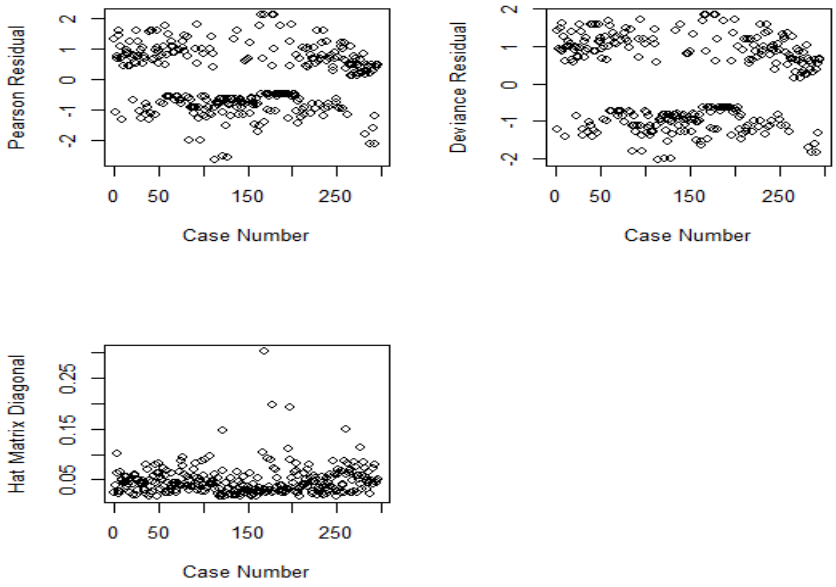

Figure 9.1: Plots of Residuals and Hat Matrix Diagonal for Wheat Availability

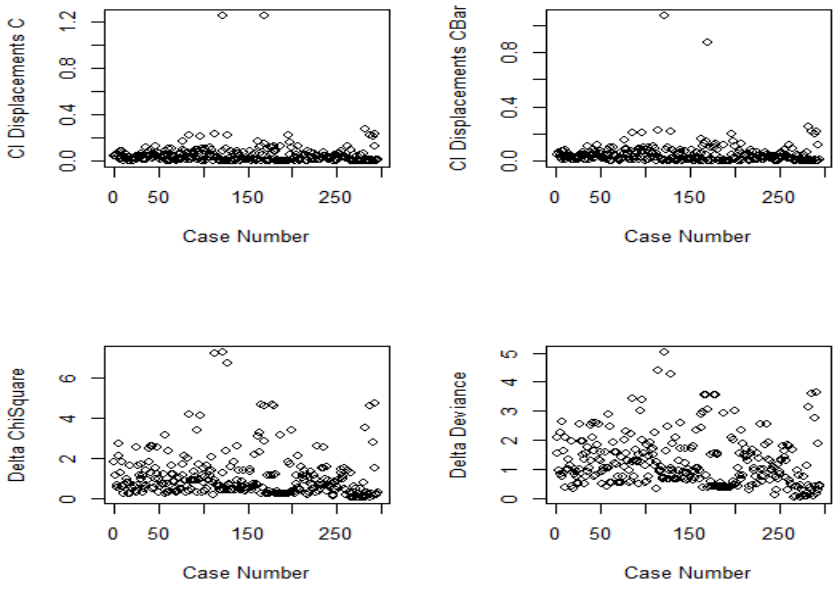

Figure 9.2: Plots of CI Displacements C and CBar, and Change in ChiSquare and Deviance for Wheat Availability

estimates and the fit of the model. The plots can only identify observations with undue influence on estimates but cannot show how each estimate is being affected. Observations 122 also have undue influence on the Pearson chi-square and deviance goodness of fit measures since it appears outstanding in the bottom left and right plots. Similarly, observation 114 [Pearson 7.25 and Deviance = 4.39] have a substantial influence on the two goodness-of-fit measures and hence a big effect on the quality of the fit of the model.

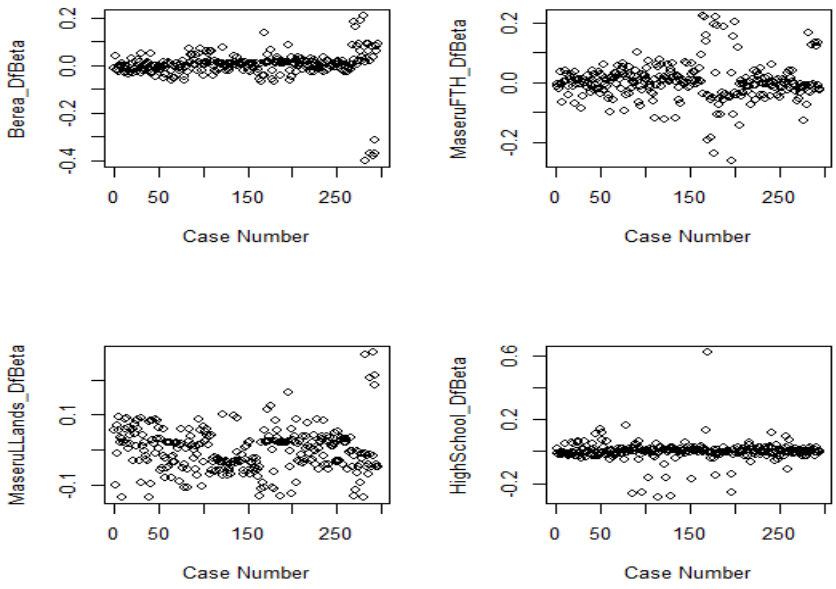

Figure 9.3: Plots of DBetas for Berea, Maseru Foot Hills, Maseru Low Lands and High School for Wheat Availability

Observation 169 has an effect on estimated coefficients of three levels of education, namely, primary, high school and post high school education (Figures 9.3 and L.2). In the case of occupation, observations 73 and 92 have an impact on the parameter estimate of casual workers, while observations 125 and 169 have an impact on the parameter estimate of salary earners (Figures L.2 and L.3).

9.3 Logistic Regression Modelling of Sorghum Availability

The second logistic regression model fitted compares availability and non-availability of sorghum for households. More than half of the households 162 [55%] had availability of sorghum and 134 [45%] did not have. The iterative method used to estimate the logistic regression coefficients is Newton-Raphson method. The values of AIC and -2 Log L for the model with an intercept only are 409.691 and 407.691 respectively, while their respective values for the fitted model are 380.755

and 352.755. The smaller values from the fitted model than the values from the model with the intercept only show that the model fits the data well.

Table 9.5: Results of the Link Function Test for Sorghum Data

Variable	DF	Chi-square	P-value
Constant	1	-0.25	0.804
Linear Predictor	1	6.33	<.0001
Squared Linear Predictor	1	0.42	0.676

The results from testing the appropriateness of the logit link function suggest that the logit is the correct link function for this model since the linear predictor is significant while the quadratic term is not (Table 9.5). The likelihood ratio, score and Wald tests are all significant at 5% level. This suggests that there is evidence that at least one of the predictor variables in the model predicts the probability of availability of sorghum to household.

Table 9.6: Type III Analysis of Effects on Sorghum Availability

Effect	DF	Wald Chi-square	P-value
Household Size	1	0.2936	0.5879
Income	1	0.3561	0.5507
Sex	1	1.1298	0.2878
Location	3	6.2572	0.0997
Education	3	6.9933	0.0721
Occupation	4	22.7183	0.0001

The results of the wald test from the Type III analysis of effects on sorghum availability suggest that there is evidence that only occupation of head of a household predicts the probability of availability of sorghum to a household (Table 9.6). Specifically, two categories, of subsistence farmers [p-value is <.0001] and salary earners [p-value = 0.0079], are significantly different from the category of casual workers, in terms of availability of sorghum (Table 9.7). In addition, a category of location, Mafeteng [p-value = 0.0249], and a category of education level, no formal education [p-value = 0.0269] are significantly different from their respective reference categories. This is contrary to the results from the type III analysis that location and education level are not significant at 5% level.

The 95% confidence interval for a category of households headed by subsistence farmers [3.015, 26.436] suggests that sorghum availability for households headed by subsistence farmers was significantly different from that of households headed by casual workers as indicated by the Wald test. The confidence intervals of categories of households that resided in Mafeteng and households headed by people with no formal education contain the value one, suggesting that the location

Table 9.7: Logistic Regression Estimates for Sorghum Availability

Variable	Estimate	Standard Error	Wald Chi-square	P-value	Odds Ratio	95% Conf. Interval of Odds Ratio	
Household Size	0.03	0.05	0.29	0.5879	1.03	0.925	1.147
Income	0.00	0.00	0.36	0.5507	1.00	1.000	1.001
Sex (ref=Male)							
Female	-0.16	0.15	1.13	0.2878	0.73	0.401	1.311
Location (ref=Berea)							
Mafeteng	0.65	0.29	5.03	0.0249	2.20	0.719	6.719
Maseru Foothill	-0.26	0.35	0.56	0.4559	0.89	0.256	3.056
Maseru Lowland	-0.25	0.22	1.31	0.2529	0.89	0.337	2.346
Education (ref=Post COSC)							
No formal educ	-0.70	0.32	4.90	0.0269	0.43	0.092	2.037
Primary	-0.07	0.27	0.08	0.7782	0.81	0.184	3.583
High School	0.64	0.37	3.08	0.0790	1.67	0.326	8.499
Occupation (ref=Casual Worker)							
Pensioner	0.46	0.43	1.16	0.2811	3.19	0.95	10.684
Salary Earner	-0.89	0.33	7.06	0.0079	0.83	0.320	2.128
Subsistence farmer	1.50	0.38	15.89	<.0001	8.93	3.015	26.436
Unemployed	-0.38	0.23	2.74	0.0981	1.37	0.622	3.021

and education level are not good predictors of the probability of availability of sorghum. Considering the effect of occupation of household head on sorghum availability, the odds ratio of 8.93 for households headed by subsistence farmers suggests that the households were 8.93 times more likely to portray availability of sorghum than their counterparts headed by casual workers, when other variables in the model are held constant.

The value of the Pearson chi-square statistic is 284.69 with a p-value of 0.4439. The large p-value suggests that the model fits the data well. The value of the deviance statistic is 352.76 with a p-value of 0.0026. This statistic contradicts what was shown by the Pearson statistics since the small p-value suggests that the model does not fit the data well. However, the Hosmer and Lemeshow test with a value 11.36, 8 degrees of freedom and a large p-value of 0.1818, is in agreement with the Pearson statistic that the model fits the data well. The disagreement between the results from the Pearson chi-square and deviance statistics is explained by Lee et al. (2006). Hence we conclude on the basis of the results from the Pearson chi-square statistic and Hosmer and Lemeshow test that the model fits the data well.

Index plots in Figures L.4, L.5, L.6, L.7, L.8, L.9 of Appendix L are used to identify observations that are influential. The index plots of Pearson and deviance residuals suggest that observation 251 is an outlying observation with large respective values of the residuals. Observation 122 is identified from the bottom left plot of Figure L.4 as a high leverage point that is extreme in the design space.

The index plots of confidence interval displacement C and CBar show that observation 122 [displacement C = 0.37 and displacement C = 0.29] and observation 275 [displacement C = 0.36 and displacement CBar = 0.31], have undue influence on individual parameter estimates and the fit of the model. Observation 286 is outstanding in the bottom left and right plots of Figure L.5, with values 7.05 for the change in Pearson chi-square statistic and 4.39 for the change in deviance statistic. Thus the observation has substantial influence on the two goodness-of-fit measures as well as a big effect on the quality of the fit of the model.

Observations 277, 219 and 286 have undue influence on estimated coefficients of household size, income and a category of households headed by females, respectively (Figure L.6). In the case of location, observation 286 has an effect on parameter estimates of two categories, Mafeteng and Maseru lowlands, while observations 188, 197 and 198 have an effect on the estimate of Maseru foothills (Figure L.7). Observation 260 has undue influence on estimated coefficients of two categories of education level, namely, primary and high school, whereas observation 169 has influence on an estimate of the category of no formal education (Figures L.7 and L.8). When it comes to occupation, observation 76 has undue influence on the parameter estimates of subsistence farmers and unemployed, whereas observation 251 has undue influence on an estimate of salary earners (Figures L.8 and L.9). The parameter estimate of pensioners is unduly influenced by observations 38, 252 and 265 (Figure L.8).

9.4 Fitting Multinomial Logistic Regression Model to Household Data

The multinomial logistic regression model fitted three logits that are based on four nominal categories of households according to cereals that were available to them, using the CATMOD procedure of SAS. The three logits were formed from designating the category of households with availability of all the three cereals, maize, wheat and sorghum, as the base category and comparing each of the other three categories with it. The first logit is of the category of households with availability of maize only versus the category of households with availability of all the three cereals. The second logit is of the category of households with availability of maize and sorghum versus the category of households with availability of all the three cereals. The third logit is of a the category of households with availability of maize and wheat versus the base category.

The profile of availability of cereals with nominal categories is given in Table 9.8. The category of households with availability of all the three cereals had the highest number of households, 103 that constitutes 35%. The category with availability of maize only was the second highest with 85 households, making 39%. The category of households with availability of maize and wheat was the

lowest with 49 households, making 16%.

Table 9.8: Profile of the Nominal Availability of Cereals

Ordered Availability	Households	Percentage
Maize Only	85	29
Maize Wheat	49	16
Maize Sorghum	59	20
Maize Wheat Sorghum	103	35

The iterative method used to estimate the regression coefficients is Newton Raphson method. The values of AIC and -2 Log L for the model with an intercept only are 802.14 and 796.14 respectively, while the respective values from the model with predictor variables are 756.36 and 672.36. The smaller values from the model with predictor variables than those from the model with the intercept only show that the model fits the data well. The three tests, likelihood ratio, score and Wald tests are all significant at 5% level of significance, suggesting that at least one of the predictor variables in the three logits, predicts the probability of availability of cereals with nominal categories of households.

Table 9.9: Multinomial Logistic Regression Estimates for Availability of Cereals (Maize Only)

Comparison: Maize Only and the Base Category (maize, wheat and sorghum)

Variable	Estimate	Standard Error	Wald Chi-square	P-value	Odds Ratio	95% Conf. Interval of Odds Ratio	
Intercept	-2.35	0.96	-2.44	0.015			
Household Size	0.02	0.07	-0.23	0.819	0.98	0.858	1.128
Income	0.00	0.00	-1.37	0.171	1.00	0.999	1.000
Sex (ref=Male)							
Female	0.50	0.39	1.30	0.193	1.66	0.775	3.540
Location (ref=Mafeteng)							
Berea	-0.80	0.91	-0.87	0.383	0.45	0.075	2.696
Maseru Foothill	0.93	0.68	1.36	0.172	2.53	0.667	9.570
Maseru Lowland	0.59	0.48	1.23	0.218	1.81	0.705	4.632
Education (ref=No educ)							
Primary	-0.47	0.47	-1.00	0.316	0.62	0.246	1.571
High School	-1.90	0.71	-2.67	0.008	0.15	0.037	0.604
Post High School	-1.77	1.24	-1.43	0.154	0.17	0.015	1.944
Occupation (ref=Farmer)							
Casual Worker	2.65	.88	3.02	0.003	14.15	2.536	78.943
Pensioner	1.72	0.98	1.75	0.080	5.58	0.816	38.243
Salary Earner	2.55	0.89	2.87	0.004	12.86	2.248	73.602
Unemployed	2.96	0.79	3.77	0.000	19.29	4.133	90.049

The results from fitting the three logits are presented in three different tables, where Table 9.9

presents the results from the logit that compares the category of households with maize only and the base category. Tables 9.10 and 9.11 present the results from fitting the logits that compare the base category and each of the categories of households with maize and wheat, and households with maize and sorghum. The Type III analysis of effects show that three variables, location of a household, education level and occupation of household heads predicts the probabilities of the three categories of households when they are compared with the base category.

Considering availability of maize only versus availability of all the three cereals in Table 9.9, households headed by people who obtained high school education [odds ratio = 0.15] were less likely to have maize only than to have all the three cereals, when they are compared with their counterparts headed by people with no formal education. Households headed by casual workers were 14.15 times more likely to have maize only than to have all the three cereals, when they are compared with their counterparts headed by subsistence farmers. Further more, households headed by salary earners and unemployed people with odds ratios of 12.86 and 19.29, respectively, were more likely to have maize only than to have all the three cereals, when they are compared with their counterparts headed by subsistence farmers. The effect in each category is observed when other variables are adjusted for. All these suggest that households headed by subsistence farmers and those headed by people with some education were relatively better off in terms of availability of food cereals.

Table 9.10: Multinomial Logistic Regression Estimates for Availability of Cereals (Maize and Wheat)

Comparison: Maize and Wheat, and the Base Category (maize, wheat and sorghum)

Variable	Estimate	Standard Error	Wald Chi-square	P-value	Odds Ratio	95% Conf. Interval of Odds Ratio	
Intercept	-1.38	0.86	-1.60	0.110			
Household Size	-0.06	0.08	-0.79	0.431	0.94	0.800	1.100
Income	0.00	0.00	0.50	0.615	1.00	0.999	1.001
Sex (ref=Male)							
Female	0.34	0.45	0.76	0.450	1.40	0.584	3.365
Location (ref=Mafeteng)							
Berea	0.80	0.73	1.10	0.273	2.23	0.531	9.409
Maseru Foothill	-0.06	0.87	-0.07	0.940	0.94	0.171	5.133
Maseru Lowland	0.65	0.58	1.13	0.259	1.92	0.618	5.987
Education (ref=No educ)							
Primary	-1.03	0.53	-1.95	0.052	0.36	0.128	1.007
High School	-1.48	0.68	-2.20	0.028	0.23	0.060	0.853
Post High School	-1.11	0.97	-1.15	0.252	0.33	0.050	2.197
Occupation (ref=Farmer)							
Casual Worker	1.73	0.69	2.52	0.012	5.65	1.469	21.713
Pensioner	0.34	0.96	0.36	0.719	1.40	0.217	9.173
Salary Earner	1.76	0.71	2.47	0.014	5.81	1.438	23.503
Unemployed	1.10	0.62	1.78	0.075	3.00	0.896	10.048

Concerning availability of maize and wheat versus availability of all the three cereals, households headed by people who obtained high school education [odds ratio = 0.028] were less likely to have maize and wheat than to have all the three cereals, when they are compared with their counterparts headed by people with no formal education (Table 9.10). Further more, households headed by casual workers and salary earners with respective odds ratios 5.65 and 5.81 were more likely to have maize and wheat than to have all the three cereals, when they are compared with their counterparts headed by subsistence farmers.

Table 9.11: Multinomial Logistic Regression Estimates for Availability of Cereals (Maize and Sorghum)

Comparison: Maize and Sorghum, and the Base Category (maize, wheat and sorghum)

Variable	Estimate	Standard Error	Wald Chi-square	P-value	Odds Ratio	95% Conf. Interval of Odds Ratio	
Intercept	0.06	0.67	0.09	0.928			
Household Size	-0.01	0.08	-0.17	0.862	0.99	0.852	1.143
Income	0.00	0.00	0.73	0.463	1.00	0.999	1.001
Sex (ref=Male)							
Female	0.451	0.385	1.17	0.241	1.57	0.738	3.336
Location (ref=Mafeteng)							
Berea	-1.68	0.73	-2.32	0.020	0.19	0.045	0.771
Maseru Foothill	-0.59	0.76	-0.77	0.442	0.56	0.125	2.479
Maseru Lowland	-0.66	0.43	-1.55	0.121	0.52	0.224	1.192
Education (ref=No educ)							
Primary	-0.26	0.50	-0.52	0.606	0.77	0.289	2.060
High School	-1.23	0.72	-1.71	0.088	0.29	0.071	1.200
Post High School	-13.25	551.85	-0.02	0.981	0.00	0.00	-
Occupation (ref=Farmer)							
Casual Worker	-0.37	0.70	-0.53	0.593	0.69	0.176	2.697
Pensioner	0.00	0.69	0.00	1.000	1.00	0.259	3.857
Salary Earner	-1.69	0.92	-1.85	0.065	0.18	0.031	1.110
Unemployed	0.55	0.44	1.26	0.206	1.74	0.738	4.083

In the case of availability of maize and sorghum as opposed to availability of all the three cereals, households that resided in Berea with odds ratio of 0.19 were less likely to have maize only than to have all the three cereals, when they are compared with their counterparts that resided in Mafeteng (Table 9.11).

The goodness-of-fit of the multinomial logistic regression model was assessed using the Pearson chi-square and deviance tests. The value of Pearson chi-square statistic is 804.47 with a p-value of 0.4293, and the value of deviance statistic is 639.77 with a p-value of 1.0000. The large p-values of both tests suggest that the model fits the data well.

9.5 Fitting Ordinal Logistic Regression Model to Household Data

The ordinal logistic regression model that we applied is the proportional odds model. The LOGISTIC procedure of SAS was used to fit the proportional odds model for ordinal responses. This model fitted two cumulative logits formed on the basis of the two cut points of the data. The first cumulative logit based on the first cut point is of the high ranking category of households with availability of all the three cereals as opposed to the middle category of households with availability of maize and either wheat or sorghum, and the low ranking category of households with availability of maize only. The second cumulative logit based on the first cut point is of high and middle ranking categories as opposed to the low raking category.

Table 9.12: Profile of the Ordered Availability of Cereals

Ordered Availability	Households	Percentage
Maize Only	85	29
Maize with Wheat or Sorghum	108	36
Maize with Wheat and Sorghum	103	35

The profile of the ordered availability of the three cereals is given in Table 9.12. The ordered categories of households are presented in their order of importance in terms of availability of cereals. The category that ranked highest with respect to availability of cereals, which is of 103 households with availability of all the three cereals constitutes 35% of the studied households. The middle category of households with availability of maize and either wheat or sorghum had 108 households, making 29%. The category of households that ranked lowest with availability of maize only had 85 households, making 29%.

The score test that tests the assumption of proportional or parallel odds that $H_0 : \beta_j = \boldsymbol{\beta}$, has a value 14.46 with 13 degrees of freedom and a large p-value of 0.3423. The large p-value indicates that the null hypothesis is not rejected and the test supports the assumption of proportional odds as valid. This suggests that the proportional odds model adequately fits the data and the effect of predictor variables on the proportional odds of the jth category of ordered availability of cereals or below is the same for all j. Hence there is only one parameter estimated for the two cumulative logits that corresponds to each predictor variable .

The iterative method used to estimate the regression coefficients is Fisher's scoring method. The values of AIC and -2 Log L for the model with an intercept only are 651.35 and 647.35 respectively, while the values from the model with predictor variables are 617.55 and 587.55 respectively. The smaller values from the model with predictor variables than the values from the model with an intercept only show that the model fits the data well. The three tests of significance of the predic-

tor variables in the model, likelihood ratio, score and Wald tests are all significant at 5% level of significance. This suggests that there is enough evidence that at least one of the predictor variables in the model predicts the cumulative probabilities of ordered availability of cereals to households.

The results of the Wald test from Type III analysis of effects suggest that only two variables, education level and occupation of head of household predicts the cumulative probabilities of availability of the three cereals to a household. Specifically, one category of education level, high school, with a p-value of 0.0309, and one category of occupation, unemployed heads of households, with a p-value of 0.0031, are significant at 5% level of significance (Table 9.13).

Table 9.13: Ordinal Logistic Regression Estimates for Availability of Cereals

Variable	Estimate	Standard Error	Wald Chi-square	P-value	Odds Ratio	95% Conf. Interval of Odds Ratio	
Intercept							
1st Cumulative Logit	-0.50	0.33	2.20	0.1382			
2nd Cumulative Logit	1.31	0.34	14.65	0.0001			
Household Size	0.02	0.05	0.13	0.7136	1.02	0.927	1.118
Income	0.00	0.00	0.61	0.4363	1.00	1.000	1.001
Sex (ref=Male)							
Female	-0.20	0.13	2.45	0.1176	0.67	0.402	1.107
Location (ref=Mafeteng)							
Berea	0.53	0.32	2.64	0.1042	1.53	0.610	3.828
Maseru Foothill	-0.49	0.31	2.50	0.1136	0.55	0.226	1.347
Maseru Lowland	-0.13	0.19	0.49	0.4856	0.79	0.434	1.437
Education (ref=No educ)							
Primary	-0.28	0.24	1.40	0.2362	1.45	0.786	2.669
High School	0.70	0.33	4.66	0.0309	3.89	1.607	9.402
Post High School	0.24	0.52	0.20	0.6533	2.43	0.585	10.128
Occupation (ref=Farmer)							
Casual Worker	-0.32	0.28	1.36	0.2444	0.31	0.134	0.727
Pensioner	0.32	0.37	0.75	0.3850	0.60	0.215	1.643
Salary Earner	-0.24	0.29	0.69	0.4052	0.34	0.142	0.812
Unemployed	-0.59	0.20	8.72	0.0031	0.24	0.123	0.467

The two intercepts are the estimated ordered logits for the two cut points of the response variable when the predictor variables are evaluated at zero. The first intercept with log odds of -0.50 corresponds to the first cumulative logit while the second intercept with log odds of 1.31 corresponds to the second cumulative logit. The 95% confidence intervals of the proportional odds ratios for categories of households headed by people with high school education [1.607, 9.402] and households headed by unemployed people [0.123, 0.467] do not contain the value 1. This suggests that the confidence intervals of proportional odds ratio are in agreement with the Wald test.

Considering the first cumulative logit, households headed by people with high school education

were 3.89 times more likely to have availability of all the three cereals as opposed to availability of maize and either wheat or sorghum, or availability of maize only, when they are compared with their counterparts headed by people who had no formal education (Table 9.13). Similarly, for the second cumulative logit, households headed by people with high school education were 3.89 times more likely to have availability of all the three cereals and availability of maize and either wheat or sorghum as opposed to having availability of maize only, when they are compared with households headed by people with no formal education. Further more, households headed by unemployed people with a proportional odds ratio of 0.24 were less likely to portray availability of all the three cereals as opposed to availability of maize and either wheat or sorghum, or availability of maize only, when they are compared with their counterparts headed by subsistence farmers. The same odds ratio and the respective interpretation apply in the case of the second cumulative logit of availability of all the three cereals and availability of maize and either wheat or sorghum as opposed to availability of maize only.

The goodness-of-fit of the ordinal logistic regression model was assessed using the Pearson chi-square and deviance. The value of the Pearson chi-square statistic is 536.48 with a p-value of 0.5946, and the value of the deviance statistic is 554.95 with a p-value of 0.3745. The large p-values suggest that the null hypothesis that the ordinal logistic regression model fits the data, cannot be rejected and thus the model fits the data well.

9.6 Summary

The results from fitting the simple logistic regression model for both wheat and sorghum availability show that the logit was found to be an appropriate link function. The type III analysis of effects show only two variables, location and occupation, with significant effects on wheat availability, while the Wald test of individual regression coefficients and confidence intervals show three variables, location, occupation and education level, with significant effects. Households that resided in Berea were likely to portray wheat availability than households from Mafeteng. Households headed by someone with high school education were likely to portray wheat availability than households headed by someone with no formal education, while households headed by unemployed people were less likely to portray wheat availability than those headed by subsistence farmers.

Pearson chi-square and Hosmer and Lemeshow tests are in agreement and show that the model for availability of wheat and availability of sorghum fits the data well, while the deviance statistic shows contradicting results that the model does not fit the data. Hence we go by the results of the agreeing two tests and conclude that the model fits the data well. Index plots of different quantities of the fitted logistic regression model pointed to outlying observations, high leverage points, observations with undue influence on individual parameters, the Pearson chi-square and deviance,

and hence a substantial effect on the quality of the fit of the model.

The important role played by the occupation of heads of households in determining availability of cereals, shown by the results from the quantile regression model is also shown by the results from the multinomial regression model with nominal categories. In general, the results from the three variants of logistic regression model are in agreement, and they are also in agreement with the results from the OLS procedure and quantile regression model that households headed by subsistence farmers were better off in terms of availability of cereals than households headed by people with other occupations. Further more, the logistic regression results are in agreement with the results from the quantile regression model that households headed by people with higher education qualifications were better off in terms of availability of cereals than households headed by people with lower education qualifications. Probably, higher education qualifications equipped heads of households with skills that enabled them to earn better salaries. Better salaries helped them acquire cereals for their households, either through production from own land or purchases, and the know-how necessary for cereals production.

Chapter 10

Discussions and Conclusions

Given the importance attached to availability of food cereals as one of the main components of food security, its study and understanding through statical modelling has been given little or no attention. Modelling of food cereals has always focussed more on crop production and yield, where regression models were applied without much consideration of problems and challenges that may emerge with the application of the models to real life data, and thus compromising the validity of the results. Hence the focuss in this thesis was to apply a series of regression models with increasing complexity to model availability of the three main cereals in Lesotho, at both the national and household levels. In addition, the work involved the application of diagnostics to identify problems that accompany this application, and to use appropriate corrective measures that address the problems.

Modelling of availability of the three main cereals in Lesotho at both the national and household data was preceded by exploration of data. At the national level, the exploratory analysis detected preliminary relationships between variables and it further helped to identify one of the supplies of cereals, which is production of a given cereal, that was used as the response variable. As in most application problems some data limitations were encountered. The national data on maize, sorghum and wheat do not contain some of the variables that are considered to be important in studying the national food cereals demand and supplies. The variables are; cereals domestic stocks, domestic requirements, surplus or deficit, and consumption per capita. This is due to the scantiness and incompleteness of the data on these variables, which resulted with gaps of data in some years, and hence insufficient observations. The scantiness and incompleteness of the data also necessitated that the data on each cereal be sub-divided into three subsets, based on time intervals with informative data. These sub-divisions were based on availability of data under each variable in the data, since there were data gaps in some years within the observational period. Thus three subsets of data were created for each cereal, which varied in terms of the number of observations and one additional variable. These subsets were identified by the years in which the data were available.

The relationships that were identified through exploration of the national data formed the basis for the in-depth analysis where two forms of the linear regression model were applied. The first form is of the classical linear model with continuous predictor variables, while the second is of the linear regression model with first-order autoregressive process AR(1). The two forms of the linear model are similar except that parameters of the model with the first order autoregressive process AR(1) includes an autoregressive coefficient associated with the autoregressive term. Thus this model did not only measure the effects of predictor variables on production of a given cereal, but it also reflected the dependence of production of a given cereal in the current year on production of the cereal in the past immediate year, which is commonly referred to as serial correlation.

In the case of the household data, the general linear regression model with categorical predictor variables was applied. The application of this model was justified by the composition of predictor variables in these data, which consist of a mixture of continuous, discrete and categorical variables. The use of categorical predictor variables increased flexibility of the linear regression model in estimating and predicting availability of cereals by measuring how the conditional availability of a given cereal varied as categorical variables changed from one category to another. For example, availability of sorghum for households headed by someone with no formal education, primary education or high school education was significantly different from availability of sorghum for households headed by someone with diplomas and university degrees, by varying amounts.

Collinearity diagnostics identified collinear relationships among predictor variables in the national data, and the variables that were involved in such relationships. These helped to understand inconsistencies that emerged in some of the results, such as some variables that were identified to have correlation with production of a given cereal, by preliminary analysis, but did not have significant effects on production when the linear regression model was fitted. A typical example is of a case where the correlation coefficient showed a negative linear relationship between population size and production of wheat, yet the results from fitting the model showed that population size had no significant effects on wheat production. This is one of the collinearity problems, where a severe collinearity identified between population size and time in the subset of 1976/1977 to 2006/2007 wheat data concealed the possible effect of population size on production of wheat. Empirically, in some subsets of sorghum and wheat data, condition index and variance-decomposition proportion showed a peculiar pattern that instead of the highest condition index indicating the presence of a collinear relationship, the second highest did by being associated with high variance-decomposition proportions of coefficients of two or more variables. Though the ridge regression allows for a small bias in the estimates, it remedied collinearity problems and controlled the instability in parameter estimates by reducing inflation factors to values that are less than 10, and reducing standard errors of the estimates.

Case deletion diagnostics identified observations that were influential on different quantities of the fitted model, such as the goodness-of-fit measure (R-squared), parameter estimates, and their standard errors. Diagnostics plots highlighted outliers and high leverage points that were influential, and showed violation of normality and constant error assumptions. The use of the Box-Cox transformation corrected violation of the assumptions, increased the strength of evidence for significance of some variables in predicting the response variable, and improved the goodness of fit of the model in some cases. In cases where the Box-Cox transformation did not correct violation of the assumptions completely, quantile regression as an alternative that is more robust to outlying observations and distributions that deviate from the normal distribution, applies. However, the limited number of observations in the national data could not allow the application of quantile regression, and hence it was applied to the household data only.

The application of quantile regression to the household data helped to have a comprehensive investigations of effects of predictor variables by estimating various regression quantiles at different parts of the distribution of the response variable. It showed that the effects of predictor variables varied across quantiles because regression slopes are not constant across quantiles. The most important strength of quantile regression is its capacity to assess the significant effects at different positions of the distribution and by so doing it deals with skewed data and reveals information about the dependence of the conditional distribution of the response variable on predictor variables, that otherwise could be concealed when OLS procedure is applied.

Further investigations of availability of cereals to households was done by categorizing households according to availability of specific cereals, and applying the logistic regression model and its two extensions of polytomous logistic regression models with nominal and ordinal responses, as special cases of GLMs. The application of GLMs gave a different perspective of modelling data on food cereals, which gave a broader understanding of how availability of specific cereals can relate with other characteristics of households. The results from the three logistic regression models are in agreement, and they are also in agreement with the results from the mean response regression model estimated by OLS procedure and the results from the quantile regression model. It was quite interesting to assess the effects of different factors on availability of cereals based on continuous and discrete scales and yet lead to the same conclusions. It was assuring to get such results because this increased our confidence in the way the problem was approached.

In general, linear and generalized linear regression models were reviewed and applied, within the scope and limitations of the data, to model availability of the three main cereals in Lesotho. The models were used in consideration of problems that may come with their applications to real life data, and may consequently compromise the validity of the results. In the light of the application of regression models with varying complexities and approaches of handling problems that accompany their application to real life data, we recommend that statistical modelling of data of this nature

be preceded by detailed exploratory data analysis. This will help to understand the structure of variables in a set of data, and establish preliminarily, the distribution of data and any accompanying problems in the data that may interfere with modelling of such data and compromise validity of the results. We further recommend that for statistical analyses to produce useful information, modelling of real life data should be done in consideration of problems that exist in the data.

There are problems that were not addressed in this thesis and call for further research. The sensitivity of the ridge estimator to the presence of extreme observations in the data was not addressed, hence there is a need for application of diagnostics that can detect extreme cases and their influence in ridge regression. The peculiar pattern observed from condition index and variance proportions in identifying collinear relationships needs to be investigated further to establish the statistical theory behind it. There is a need for the development of distributional theory for ridge regression, which is required for making inferences about ridge estimates. There is also a need for diagnostics that detect influence of leverage points on estimates of quantile regression model since quantile regression estimates are robust to outliers but they lack robustness to high leverage points. Diagnostics for logistic regression are developed for a case of binary response, however, they have not yet been extended to polytomous logistic regression models for both nominal and ordinal responses. Thus logistic regression diagnostics need to be developed further and generalized to polytomous logistic regression models.

Bibliography

Abeyskera, W. W. M. and Sooriyarachchi, R. (2008). A novel method for testing goodness of fit of a proportional odds model: an application to AIDS study. *Journal of National Science Foundation Sri Lanka*, 36(2):125–135.

Agresti, A. (2002). *Categorical Data Analysis*. John Wiley and Sons, Inc., Hoboken, New Jersey.

Akaike, H. (1974). A new look at the statistical model identification. IEEE TRans. Automat, Contr. AC-19: 716-23. Institute of Statistical Mathematics, Minato-ku, Tokyo, Japan.

Anderson, J. A. (1984). Regression and ordered categorical variables. *Journal of the Royal Statistical Society, Series B*, 46(1):1–30.

Balaghi, R., Tychon, B., Eerens, H., and Jlibene, M. (2008). Empirical regression models using NDVI, rainfall and temperature data for the early prediction of wheat grain yields in Morocco. *International Journal of Applied Earth Observation and Geoinformation*, 10:438–452.

Barrodale, I. and Roberts, F. (1974). Solution of an overdetermined system of equations in the ℓ_1 norm. *Communications of the ACM*, 17:319–320.

Bassett, G. and Koenker, R. (1982). An empirical quantile function for linear models with iid errors. *Journal of the American Statistical Association*, 77(378):407–415.

Baur, D., Saisana, M., and SChulze, N. (2004). Modelling the effects of meteorological variables on ozone concentration a quantile regression approach. *Atmospheric Environment*, 38:4689–4699.

Bella, C. M. D., Margin, G. O., Boullon, D. R., Grondoma, M. O., and Rebella, C. M. (1996). Zea mays prediction using satellite information and a simulation model. Anais VIII Simpósio de Sensoriamentob Remoto, Salvador, Brasil.

Belsley, D. A. (1991). *Conditioning Diagnostics: Collinearity and Weak Data in Regression*. John Wiley and Sons.

Belsley, D. A., Kuh, E., and Welsch, R. E. (1980). *Regression Diagnostics: Identifying Influential Data nad Sources of Collinearity*. John Wiley and Sons.

Box, G. E. P. and Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society, Series B*, 26(2):211–243.

Buchinsky (1998). Recent advances in quantile regression models: A practical guideline for empirical research. *The Journal of Human Resources*, 33(1):88–126.

Cade, B. S. (2003). *Quantile Regression Models of Animal Habitat Relationships*. Unpublised phd dissertitaion, Colorado State University, Fort Collins, Colorado.

Chatterjee, S. and Hadi, A. S. (2006). *Regression Analysis by Example*. John Wiley and Sons, fourth edition.

Chatterjee, S. and Hadi, S. (1996). Influential observations, high leverage points, and outliers in linear regression. *Statistical Science*, 1(3):379–416.

Chen, C. (2007). A finite smoothing algorithm for quantile regression. *Journal of Computational and Graphical Statistics*, 16(1):136–164.

Chen, C. and Wei, Y. (2005). Computational issues for quantile regression. *The Indian Journal of Statistics*, 67:399–417.

Cohen, M. J. (2005). Food supply, factors afecting production, trade and access. Technical report, International Food Policy Institute (IFPRI), Washington, DC.

Cook, R. D. (1977). Detection of influential observation in linear regression. *Technometrics*, 19(1):15–18.

Cook, R. D. and Weisberg, S. (1980). Characterization of an empirical influence function for detecting influential cases in regression. *Technometrics*, 22(4):495–508.

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

Cutts, M. and Hassan, R. (2003). An econometric model of the SADC maize sector. Contributed paper presented at the 41st annual conference of the Agricultural Economic Association of South Africa (AEASA), Pretoria, South Africa.

DasGupta, M. and Mishra, S. K. (2004). Least absolute estimation of linear econometric models: A literature review. MPRA Paper.

Draper, N. R. and Smith, H. (1998). *Applied Regression*. John Wiley and Sons, Inc, third edition.

Evenson, R. E. and Mwabu, G. (1998). The effects of agricultural extension on farm yields in kenya. Center Discussion Paper No. 798.

Everitt, B. S. and Dunn, G. (1983). *Advanced Methods of Data Exploration and Modeling*. Heinemann Educational Books Ltd, London.

Fagerland, M. W., Hosmer, D. W., and Bofin, A. M. (2008). Multinomial goodness-of-fit tests for logistic regression models. *Statistics in Medicine*, 27:4238–4253.

FAO (1997). *Agriculture food and nutrition for Africa - A resource book for teachers of agriculture*. Food and Agriculture Organization Information Division, Rome.

FAO and WFP (2005). FAO/WFP crop and food supply assessment mission of lesotho. Technical report, Food and Agriculture Ogarnization and World Food Programm, Lesotho.

FAO and WFP (2007). FAO/WFP crop and food supply assessment mission to lesotho. Special report, Food and Agriculture Ogarnization and World Food Programm, Lesotho, Rome.

Feinberg, B. (1980). *Analysis of Cross-Classified Categorical Data*. Massachusetts Institute of Technology Press, Cambrige.

Freund, R. J. and Wilson, W. J. (1998). *Regresssion Analysis: Statistical Modeling of a Response Variable*. Academic Press.

Gervais, S. (2004). Local capacity building in tittle II food security projects: A framework. *Food and Nutrition Technical Assistance Project*. Academy for Educational Development (AED), Washington, D. C.

Greenland, S. (1985). An application of logistic models to the analysis of ordinal responses. *Biomedical Journal*, 27:189–197.

Gunst, R. F. and Mason, R. L. (1980). *Regression Analysis and Its Application: A Data-Oriented Approach*. Marcel Dekker, Inc.

Gutenbrunner, C. and Jurečková, J. (1992). Regression rank scores and regression quantiles. *The Annals of Statistics*, 20(1):305–330.

Gutenbrunner, C., Jurečková, J., Koenker, R., and Portnoy, S. (1993). Tests of linear hypothesis based on regression rank scores. *Nonparametric Statistics*, 2:307–331.

Hadi, A. S. and Ling, R. F. (1998). Some cautionary notes on the use of principal components regression. *The American Statistician*, 52(1):15–19.

Hall, P. and Sheather, S. (1988). On the distribution of a studentized quantile. *Journal of the Royal Statistical Society, Series B*, 50:381–391.

Hampel, F. R. (1974). The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, 69(346):383–393.

Hansen, J. W. and Indeje, M. (2004). Linking dynamics seasonal climate forecasts with crop simulation for maize yield prediction in semi-arid Kenya. *Agriculture and Forest Meteorology*, 125:143–157.

He, X. and Hu, F. (2002). Markov chain marginal bootstrap. *Journal of the American Statistical Association*, 97(459):783–795.

Hinkley, D. (1985). Transformation diagnostics for linear models. *Biometrika*, 72(3):487–496.

Hocking, R. R. (1996). *Methods and Applications of Linear Models*. John Wiley and Sons.

Hoerl, A. E. and Kennard, R. W. (1970a). Ridge regresssion: Applicatios to nonorthogonal problems. *Technometrics*, 12(1):69–82.

Hoerl, A. E. and Kennard, R. W. (1970b). Ridge regresssion: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67.

Hoerl, A. E. and Kennard, R. W. (1976). Ridge regression: iterative estimation of the biasing factor. *Communication in Statistics - Theory and Methods*, 5(1):77–88.

Hoerl, A. E., Kennard, R. W., and Baldwin, K. F. (1975). Ridge regression: some simulations. *Communication in Statistics - Theory and Methods*, 4(2):105–123.

Hosmer, A. H., Taber, S., and Lemeshow, S. (1991). The importance of assessing the fit of logistic regression models: A case study. *American Journal of Public Health*, 81(12):1630–1635.

Hosmer, D. W. and Lemeshow, S. (1980). A goodness-of-fit tests for the multiple logistic regression model. *Communications in Statistics*, A10:1043–1069.

Hosmer, D. W. and Lemeshow, S. (2000). *Applied Logistic Regression*. John Wiley and Sons, Inc., New York, second edition.

Huber, P. J. (1964). Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35:73–101.

Jones, D. R. (1982). A statistical inquiry into crop-weather dependence. *Agricultural Meteorology*, 26:91–104.

Karmarkar, N. (1984). A new polynomial-time algorithm for linear programming. *Combinatorica*, 65:507–534.

Kleinbaum, D. G. and Klein, M. (2010). *Logistic Regression A Self-Learning Text*. Springer, New York, London, third edition.

Kocherginsky, M., He, X., and Hu, F. (2005). Practical confidence intervals for regression quantiles. *Journal of Computational and Graphical Statistics*, 14(1):41–55.

Koenker, R. (1994). Confidence intervals for regression quantiles. In Mandl, P. and Hušková, M., editors, *Asymptotic Statistics: Proceedings of the 5th Prague Symposium*, pages 349–359, Heidleberg. Physica-Verlag.

Koenker, R. (2005). *Quantile Regression*. Cambridge University Press.

Koenker, R. and Basett, G. (1982a). Robust tests for heteroscedasticity based on regression quantiles. *Econometrica*, 50(1):43–61.

Koenker, R. and Basett, G. (1982b). Tests of linear hypothesis and l_1 estimation. *Econometrica*, 50(1):1577–1584.

Koenker, R. and Bassett, G. (1978). Regression quantiles. *Econometrica*, 46(1):33–50.

Koenker, R. and D'Orey, V. (1987). Algorithm as 229: Computing regression quantiles. *Journal of the Royal Statistical, Society Series C (Applied Statistics)*, 36(3):383–393.

Koenker, R. and Hallock, K. A. (2000). Quantile regression: An introduction. University of Illinois, Urbana-Champaign.

Koenker, R. and Machado, A. F. (1999). Goodnoess of fit and related inference processes for quantile regression. *Journal of the American Statisticsl Association*, 94(448):1296–1310.

Kolmogorov, A. N. (1931). The method of the median in the theory of errors. In Shiryayev, A. N., editor, *Selected Works of A. N. Kolmogorov*, volume II. Kluwer.

Kutner, M. H., Nachtsheim, C. J., Neter, J., and Li, W. (2005). *Applied Linear Statistical Models*. McGrwa-Hill, fifth edition edition.

Lee, Y., Nelder, J. A., and Pawitan, Y. (2006). *Generalized Linear Models with Random Effects Unified Analysis via H-likelihood*. Chapman and Hall/CRC.

Lemeshow, S. and Hosmer, D. W. (1982). A review of goodness of fit statistics for use in the development of logistic regression models. *American Journal of Epidemiology*, 115:92–106.

Lewis, J. E., Rowland, J., and Nadeau, A. (1998). Estimating maize production in Kenya using NDVI: Some statistical considerations. *Journal of Remote Sensing*, 19(13):2609–2617.

Lipsitz, S. R., Fitzmaurice, G. M., and Molenberghs, G. (1996). Goodness-of-fit tests for ordinal response regression models. *Applied Statistics*, 45(2):175–190.

Madsen, K. and Nielsen, H. B. (1993). A finite smoothing algorithm for linear l_1 estimation. *SIAM Journal of Optimization*, 3:223–235.

Marquardt, D. and Snee, R. D. (1975). Ridge regression in practice. *The American Statistician, Vol. 29, No.1*, 29(1):3–20.

Maxwell, S. and Frankenberger, T. R. (1992). Household food security: Concepts, indicators and measurements. Technical report, Uinted Nations Children's Fund (UNICEF) and International Fund for Agricultural Development (IFAD).

McCullagh, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society, Series B*, 42(2):109–142.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. Chapman and Hall, second edition.

Melly, B. (2001). *The Theory and Practice of Quantile Regression*. Unpublished Ph.D thesis, University of St. Gallen.

Montgomery, D. C., Peck, E. A., and Vining, G. (2012). *Introduction to Linear Regression Analysis*. John Wiley and Sons, Inc., Hoboken, New Jersey, fifth edition.

Montgomery, D. C., Peck, E. A., and Vining, G. G. (2001). *Introduction to Linear Regression Analysis*. John Wiley and Sons Inc.

Mosteller, F. and Tukey, J. W. (1977). *Data Analysis and Regression*. Addison Welsley.

Osius, G. and Rojek, D. (1992). Normal goodness-of-fit tests for multinomial models with large degrees of freedom. *Journal of the American Statistical Association*, 87(420):1145–1152.

Parzen, E. (1979). Nonparametric statistical data modeling. *Journal of the American Statistical Association*, 74:105–121.

Parzen, M. I., Wei, L. J., and Ying, Z. (1994). A resampling method based on pivotal estimating equations. *Biometrika*, 81(2):341–350.

Pawitan, Y. (2001). *In All Likelihood: Statistical Modeling and Inference Using Likelihood*. Oxford, first edition.

Peterson, B. and Harrel, F. E. (1990). Partial proportional odds for ordinal response variables. *Applied Statistics*, 39(2):205–217.

Portnoy, S. and Koenker, R. (1997). The Gaussian Hare and the Laplacian Tortoise: Computability of squared-error versus absolute-error estimators. *Statistical Science*, 12(4):279–300.

Powell, J. L. (1986). Censored regression quantile. *Journal of Econometrics*, 32:143–155.

Prasad, A. K., Chai, L., Singh, R. P., and Kafatos, M. (2005). Crop yield estimation model for Iowa using remote sensing and surface parameters. *International Journal of Applied Earth Observation and Geoinformation*, 8:26–33.

Pregibon, D. (1981). Logistic regression diagnostics. *The Annals of Statistics*, 9(4):705–724.

Rawlings, J. O., Pantula, S. G., and Dickey, D. A. (1998). *Applied Regression Analysis: A Research Tool*. Springer, second edition.

Ren, J., Chen, Z., Zhou, Q., and Tang, H. (2008). Regional yield estimation for winter wheat with MODIS-NDVI data in Shandong, China. *International Journal of Applied Earth Observation and Geoinformation*, 10:403–413.

Riely, F., Mock, N., Cogill, B., Bailey, L., and Kenefick, E. (1999). Food security indicators and framework for use in the minitoring and evaluation of food aid programs. *Food and Nutrition Technical Assistance Project*. Academy for Educational Development (AED), Washington, D. C.

Rousseeuw, P. J. (1984). Least median of squares regression. *Journal of the American Statistical Association*, 79(388):871–880.

Rousseeuw, P. J. and Leroy, A. M. (2003). *Robust Regression and Outlier Detection*. John Wiley and Sons, Inc.

Sarkar, S. K., Midi, H., and Rana, S. (2011). Detection of outliers and influential observations in binary logistic regression: An empirical study. *Journal of Applied Sciences*, 11(1):26–35.

Schabenberger, O. and Pierce, F. J. (2002). *Contemporary Statistical Models for the Plant and Soil Sciences*. CRC Press LLC, New York.

Schillinger, W. F., Schofstoll, S. E., and Alldredge, J. R. (2008). Available water and grain yield relations in a Mediterranean climate. *Field Crops Research*, 109:45–49.

Sen, A. (1981). *Poverty and Famines: An Essay on Entittlement and Deprivation*. Oxford Clarendon Press.

Sen, A. and Srvastava, M. (1990). *Regression Analysis: Theory, Methods and Applications*. Springer.

Silvey, S. D. (1969). Multicollinearity and imprecise estimation. *Journal of the Royal Statistical Society, Series B*, 31:539–552.

Singh, N. T., Vig, A. C., and Singh, R. (1985). Nitrogen response of maize under temporary flooding. *Nutrient Cycling in Agroecosystems*, 6(2):111–120.

Tukey, J. W. (1965). What part of the sample contains the information. *Proceedings of the National Academy of Sciences*, 53:127–134.

Walker, S. H. and Duncan, D. E. (1967). Estimation of the probability of an event as a function of several independent variables. *Biometrika*, 54:167–179.

Weisberg, S. (2005). *Applied Linear Regression*. John Wiley and Sons, Inc., third edition.

Welsch, R. E. and Kuh, E. (1977). Linear regression diagnostics. Nat. Bur. Econ. Res. Inc. Working Paper 173.

Yu, K., Lu, Z., and Stander, J. (2003). Quantile regression: Applications and current research areas. *Journal of the Royal Statistical Society*, 52(3):331–350.

Zhou, K. Q. and Portnoy, S. L. (1996). Direct use of regression quantiles to construction confidence sets in linear models. *The Annals of Statistics*, 24(1):287–306.

Appendix A

Scatter Plot Matrices for National Data

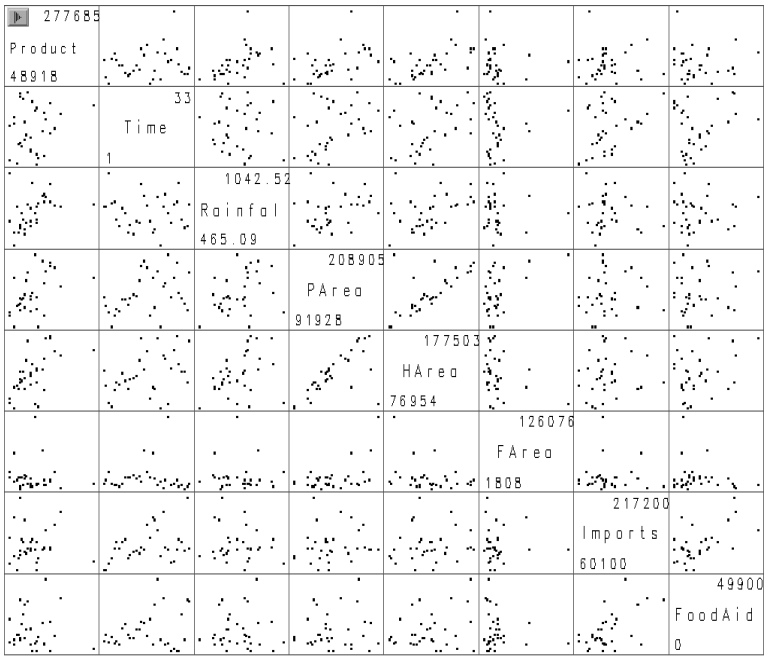

Figure A.1: Scatter Plot Matrix for 1973 - 2007 Maize Data

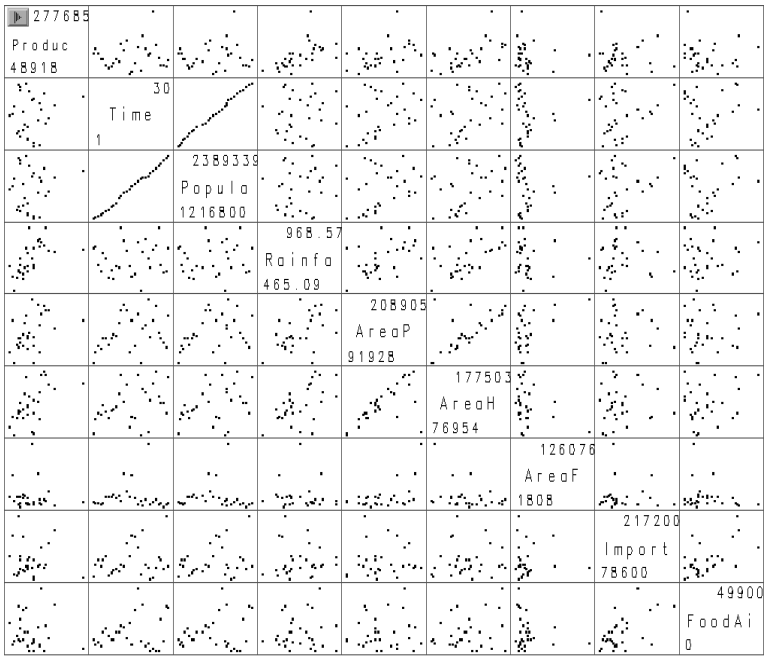

Figure A.2: Scatter Plot Matrix for 1976 - 2007 Maize Data

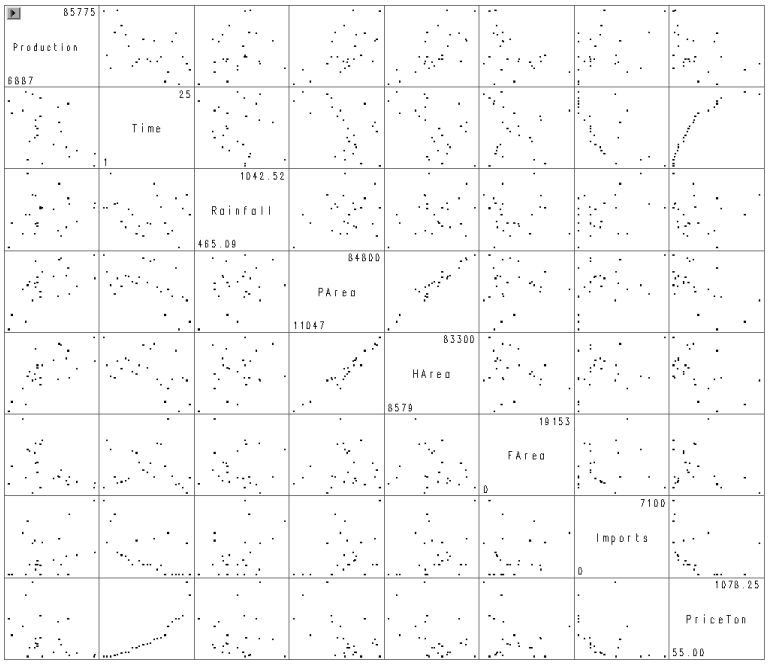

Figure A.3: Scatter Plot Matrix for 1973 - 1998 Sorghum Data

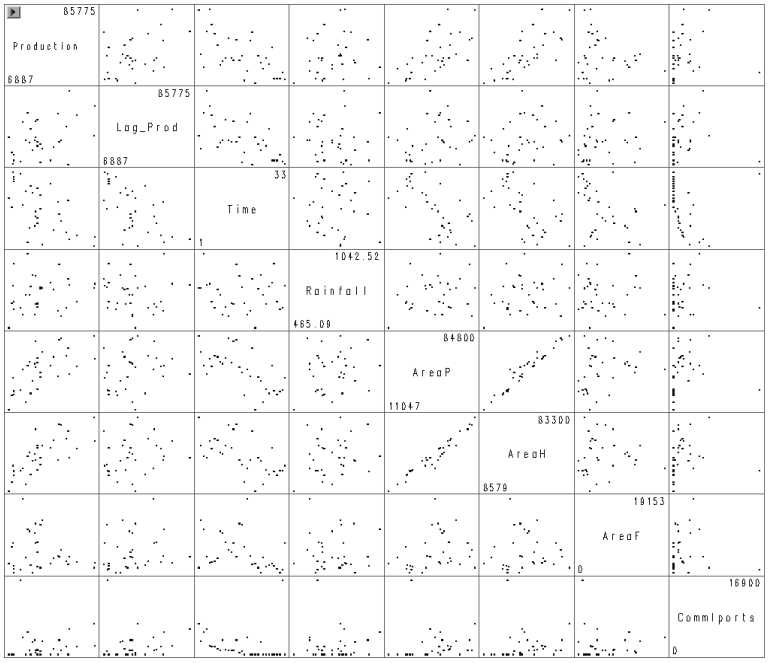

Figure A.4: Scatter Plot Matrix for 1973 - 2007 Sorghum Data

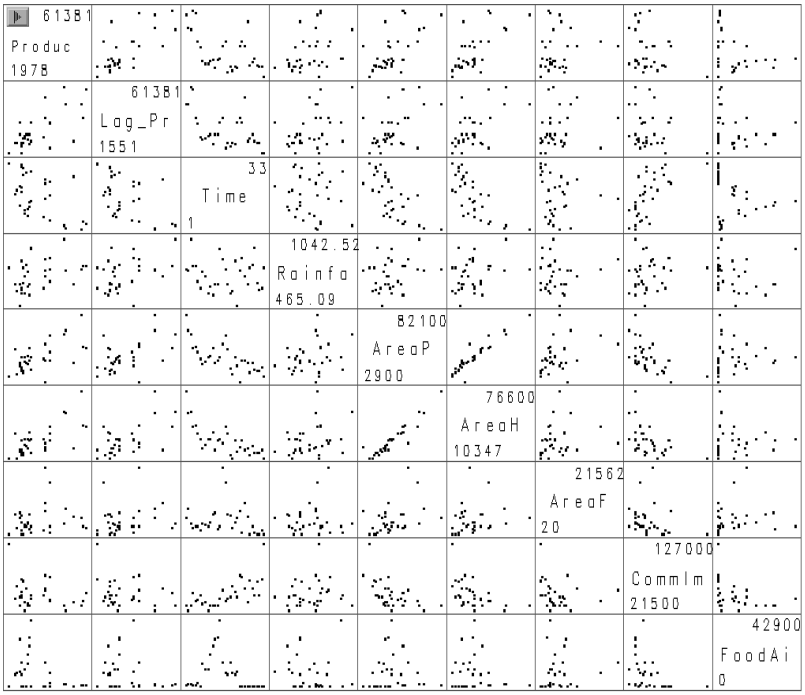

Figure A.5: Scatter Plot Matrix for 1973 - 2007 Wheat Data

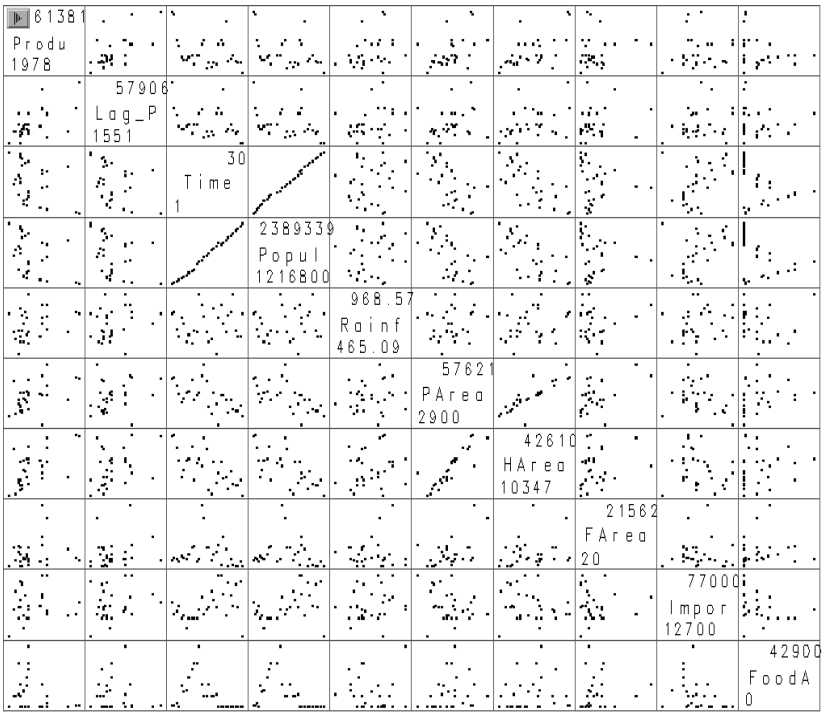

Figure A.6: Scatter Plot Matrix for 1976 - 2007 Wheat Data

Appendix B

Correlation Matrices

Table B.1: Correlation Matrix for 1973 - 2007 Maize Data

	Production	Time	Rainfall	PArea	HArea	FArea	Imports	Food aid
Production	1.0000	0.1896	0.3906	0.4862	0.6143	-0.0926	0.0128	-0.0420
Time	0.1896	1.0000	-0.0845	0.3836	0.4122	0.0127	0.3970	0.2338
Rainfall	0.3906	-0.0845	1.0000	0.3518	0.3974	-0.0112	-0.1393	-0.0004
PArea	0.4862	0.3836	0.3518	1.0000	0.6806	0.4991	0.1139	0.0984
HArea	0.6143	0.4122	0.3974	0.6806	1.0000	-0.2952	0.1208	0.2479
FArea	-0.0926	0.0127	-0.0112	0.4991	-0.2952	1.0000	0.0056	-0.1648
Imports	0.0128	0.3970	-0.1393	0.1139	0.1208	0.0056	1.0000	0.5688
Food aid	-0.0420	0.2338	-0.0004	0.0984	0.2479	-0.1648	0.5688	1.0000

Table B.2: Correlation Matrix for 1976 - 2007 Maize Data

	Production	Time	Population	Rainfall	PArea	HArea	FArea	Imports	Food aid
Production	1.0000	0.0997	0.1003	0.5962	0.4536	0.5802	-0.0824	-0.0908	-0.0793
Time	0.0997	1.0000	0.9975	0.0941	0.3429	0.3860	-0.0021	0.2245	0.1856
Population	0.1003	0.9975	1.0000	0.1011	0.3353	0.3840	-0.0094	0.2525	0.2156
Rainfall	0.5962	0.0941	0.1011	1.0000	0.5042	0.5965	-0.0361	-0.0138	0.0675
PArea	0.4536	0.3429	0.3353	0.5042	1.0000	0.6644	0.5208	0.0387	0.0717
HArea	0.5802	0.3860	0.3840	0.5965	0.6644	1.0000	-0.2921	0.0544	0.2293
FArea	-0.0824	-0.0021	-0.0094	-0.0361	0.5208	-0.2921	1.0000	-0.0127	-0.1701
Imports	-0.0908	0.2245	0.2525	-0.0138	0.0387	0.0544	-0.0127	1.0000	0.5709
Food aid	-0.0793	0.1856	0.2156	0.0675	0.0717	0.2293	-0.1701	0.5709	1.0000

Table B.3: Correlation Matrix for 1973 - 1998 Sorghum Data

	Production	Time	Rainfall	PArea	HArea	FArea	Imports	Price/ton
Production	1.0000	-0.5181	0.2251	0.5995	0.7032	-0.4475	0.3558	-0.4127
Time	-0.5181	1.0000	-0.1895	-0.5578	-0.5166	-0.1303	-0.6296	0.9023
Rainfall	0.2251	-0.1895	1.0000	0.2432	0.2532	-0.0541	0.3596	-0.0409
PArea	0.5995	-0.5578	0.2432	1.0000	0.9679	0.0677	0.3711	-0.5667
HArea	0.7032	-0.5166	0.2532	0.9679	1.0000	-0.1853	0.3359	-0.4897
FArea	-0.4475	-0.1303	-0.0541	0.0677	-0.1853	1.0000	0.1176	-0.2715
Imports	0.3558	-0.6296	0.3596	0.3711	0.3359	0.1176	1.0000	-0.4042
Price/Ton	-0.41267	0.9023	-0.0409	-0.5667	-0.4897	-0.2715	-0.4042	1.0000

Table B.4: Correlation Matrix for 1973 - 2007 Sorghum Data

	Production	Production-t	Time	Rainfall	PArea	HArea	FArea	Imports
Production	1.0000	0.4479	-0.6296	0.1599	0.6796	0.7513	-0.1828	0.1281
Production-t	0.4479	1.0000	-0.6122	-0.0699	0.2246	0.1922	0.1633	0.1062
Time	-0.6296	-0.6122	1.0000	-0.0845	-0.6682	-0.6033	-0.3582	-0.1574
Rainfall	0.1599	-0.0697	-0.0845	1.0000	0.1924	0.2052	-0.0221	0.01974
PArea	0.6796	0.2246	-0.6682	0.1924	1.0000	0.9702	0.26970	0.1219
HArea	0.7513	0.1920	-0.6033	0.2052	0.9702	1.0000	0.0274	0.1266
FArea	-0.1828	0.1633	-0.3582	-0.0221	0.2690	0.0274	1.0000	-0.0000
Imports	0.1281	0.1062	-0.1574	0.0197	0.1219	0.1266	-0.0000	1.0000

Table B.5: Correlation Matrix for 1973 - 2007 Wheat data

	Production	Production-t	Time	Rainfall	PArea	HArea	FArea	Imports	Food aid
Production	1.0000	0.7170	-0.5550	0.3569	0.6540	0.7176	0.2094	-0.3129	-0.3020
Production-t	0.7170	1.0000	-0.5370	0.4124	0.5363	0.6102	0.1855	-0.2444	-0.1407
Time	-0.5550	-0.5370	1.0000	-0.0845	-0.7138	-0.7440	-0.1727	0.6756	-0.2560
Rainfall	0.3569	0.4124	-0.0845	1.0000	0.2655	0.3112	0.0532	0.1479	-0.1178
PArea	0.6540	0.5363	-0.7138	0.2655	1.0000	0.9334	0.4251	-0.5503	0.0400
HArea	0.7176	0.6102	-0.7440	0.3112	0.9334	1.0000	0.2127	-0.4910	-0.0551
FArea	0.2094	0.1855	-0.1727	0.0532	0.4251	0.2127	1.0000	-0.2193	0.1819
Imports	-0.3129	-0.2444	0.6756	0.1479	-0.5503	-0.4910	-0.2193	1.0000	-0.3790
Food aid	-0.3020	-0.1407	-0.2560	-0.1178	0.0400	-0.0551	0.1819	-0.3790	1.0000

Table B.6: Correlation Matrix for 1976 - 2007 Wheat data

	Production	Production-t	Time	Population	Rainfall	PArea	HArea	FArea	Imports	Food aid
Production	1.0000	0.6177	-0.4459	-0.4490	0.2321	0.4952	0.5993	0.2746	0.2268	-0.0979
Production-t	0.6177	1.0000	-0.3637	-0.3532	0.2451	0.2679	0.3702	0.1907	0.2340	-0.0345
Time	-0.4459	-0.3637	1.0000	0.9975	0.0941	-0.6020	-0.6818	-0.1500	0.4142	-0.4172
Population	-0.4490	-0.3532	0.9975	1.0000	0.1011	-0.5999	-0.6738	-0.1426	0.4213	-0.4232
Rainfall	0.2321	0.2451	0.0941	0.1011	1.0000	0.1107	0.1610	0.0611	0.2063	-0.0619
PArea	0.4952	0.2679	-0.6020	-0.5999	0.1107	1.0000	0.8604	0.5309	-0.3036	0.2631
HArea	0.5994	0.3702	-0.6818	-0.6738	0.1610	0.8604	1.0000	0.2532	-0.1970	0.1673
FArea	0.2746	0.1907	-0.1500	-0.1426	0.0611	0.5309	0.2532	1.0000	-0.0221	0.2050
Imports	0.2268	0.2340	0.4142	0.4213	0.2063	-0.3037	-0.1970	-0.0221	1.0000	-0.4199
Food aid	-0.0979	-0.0345	-0.4172	-0.4232	-0.0619	0.2631	0.1673	0.2050	-0.4199	1.0000

Appendix C

Parameter Estimates and Collinear Diagnostics

Table C.1: Parameter Estimates for 1973 - 2007 Full Maize Data

Response Variable: maize production

Variable	DF	Estimate	Std. error	t value	Pr > |t|	Tolerance	VIF
Intercept	1	-44129.00	40380.00	-1.09	0.2832	.	0
Rainfall	1	61.76	54.60	1.13	0.2670	0.84	1.19
Harvested Area	1	0.90	0.25	3.55	0.0013	0.84	1.19

Table C.2: Collinear Diagnostics for 1973 - 2007 Maize Data

Number	Eigenvalue	Condition Index	Proportions of Variance	
			Rainfall	Harvested Area
1	2.96	1.00	0.00281	0.00438
2	0.03	10.63	0.15224	0.99546
3	0.02	14.11	0.84495	0.00016

Table C.3: Parameter Estimates for 1976 - 2007 Maize Data

Response Variable: maize production

Variable	DF	Estimate	Std. error	t value	Pr > |t|	Tolerance	VIF
Intercept	1	-60968.00	40606.00	-1.50	0.1448	.	0
Rainfall	1	146.53	68.10	2.15	0.0405	0.64	1.55
Harvested Area	1	0.57	0.30	1.93	0.0638	0.64	1.55

Table C.4: Collinear Diagnostics for 1976 - 2007 Maize Data

Number	Eigenvalue	Condition Index	Proportions of Variance	
			Rainfall	Harvested Area
1	2.97	1.00	0.00197	0.00331
2	0.02	11.28	0.00416	0.69769
3	0.01	15.99	0.99387	0.29900

Table C.5: Parameter Estimates for 1973 - 2007 Sorghum Data Using All Variables

Response Variable: sorghum production

Variable	DF	Estimate	Std. error	t value	Pr > |t|	Tolerance	VIF
Intercept	1	-5788.36	15685.00	-0.37	0.7148	.	0
Time	1	-112.22	378.29	-0.30	0.7688	0.38021	2.63
Production-t	1	0.31	0.15	2.03	0.0521	0.57577	1.74
Harvested Area	1	0.76	0.16	4.71	<.0001	0.58572	1.71

Table C.6: Collinear Diagnostics for 1973 - 2007 Sorghum Data Using All Observations

Number	Eigenvalue	Condition Index	Proportions of Variance		
			Time	Production-t	Harvested Area
1	3.42	1.00	0.00570	0.00964	0.00622
2	0.42	2.87	0.13459	0.10045	0.01476
3	0.15	4.83	0.02311	0.38733	0.30837
4	0.01	15.56	0.83659	0.50258	0.67064

Table C.7: Parameter Estimates for 1973 - 1998 Sorghum Data Using All Variables

Response Variable: sorghum production

Variable	DF	Estimate	Std. error	t value	Pr > |t|	Tolerance	VIF
Intercept	1	40805.00	14451.00	2.82	0.0105	.	0
Time	1	-991.28	871.46	-1.14	0.2688	0.16973	5.89
Harvested Area	1	0.50	0.17	2.87	0.0095	0.64359	1.55
Failed Area	1	-1.77	0.64	-2.76	0.0122	0.75400	1.33
Price/Ton	1	1.39	25.05	0.06	0.9563	0.15576	6.42

Table C.8: Collinear Diagnostics for 1973 - 1998 Sorghum Data Using All Observations

Number	Eigenvalue	Condition Index	Proportions of Variance			
			Time	Harvested Area	Failed Area	Price/Ton
1	4.06	1.00	0.00215	0.00387	0.01065	0.00284
2	0.61	2.58	0.01004	0.01702	0.12222	0.04212
3	0.294	3.76	0.00217	0.11921	0.45198	0.00222
4	0.03	12.24	0.57438	0.24153	0.31529	0.86987
5	0.02	13.67	0.41126	0.61838	0.09986	0.08295

Table C.9: Parameter Estimates for 1973 - 2007 Wheat Data Using All Variables

Response Variable: wheat production

Variable	DF	Estimate	Std. error	t value	Pr > |t|	Tolerance	VIF
Intercept	1	-4407.79	12450.00	-0.35	0.7260	.	0
Time	1	36.33	298.94	0.12	0.9041	0.39749	2.52
Rainfall	1	4.64	16.22	0.29	0.7769	0.75258	1.33
Production-t	1	0.45	0.16	2.82	0.0088	0.53882	1.86
Harvested Area	1	0.51	0.21	2.42	0.0222	0.36008	2.78

Table C.10: Collinear Diagnostics for 1973 - 2007 Wheat Data Using All Observations

Number	Eigenvalue	Condition Index	Proportions of Variance			
			Time	Rainfall	Production-t	Harvested Area
1	4.33	1.00	0.00357	0.00116	0.00721	0.00335
2	0.52	2.89	0.09950	0.00026	0.08499	0.02334
3	0.12	6.21	0.04403	0.00207	0.81208	0.23121
4	0.03	12.55	0.81554	0.25793	0.04251	0.70964
5	0.01	18.13	0.03735	0.73857	0.05320	0.03246

Table C.11: Parameter Estimates for 1976 - 2007 Wheat Data Using All Variables

Response Variable: wheat production

Variable	DF	Estimate	Std. error	t value	Pr > |t|	Tolerance	VIF
Intercept	1	-7060.86	86757	-0.08	0.9358	.	0
Time	1	347.88	3182.94	0.11	0.9138	0.00466	214.41
Production-t	1	0.55	0.16	3.41	0.0022	0.82609	1.21
Population	1	-0.004	0.08	-0.05	0.9579	0.00478	209.28
Harvested Area	1	0.80	0.29	2.72	0.0117	0.51274	1.95

Table C.12: Collinear Diagnostics for 1976 - 2007 Wheat

Number	Eigenvalue	Condition Index	Proportions of Variance			
			Time	Production-t	Population	Harvested Area
1	4.39	1.00	0.00005	0.00971	0.000009	0.00270
2	0.44	3.15	0.00107	0.20172	0.000031	0.01758
3	0.15	5.49	0.00067	0.75864	0.000001	0.16087
4	0.02	14.90	0.01114	0.01511	0.000545	0.81266
5	0.00	192.58	0.98707	0.01483	0.999410	0.00619

Appendix D

Case Deletion Diagnostics for National Data

Table D.1: Results of Regression Diagnostic for 1973 - 2007 Maize Data

Diagnostic Measure	3rd Case Diagnostics	Cutoff Point
RStudent t_i	-1.76	2
leverage h_{ii}	0.42	$2p/n = 0.12$
DFFITS	-1.49	$2\sqrt{p/n} = 0.49$
Cook's Distance D_i	0.69	$F_{(0.5,p,n-p)} = 1.41$
COVRATIO	1.41	$1 \pm 3p/n = 1 \pm 0.18$
DFBETAS		$2/\sqrt{n} = 0.35$
Rainfall	-1.32	0.35
Harvested Area	1.05	0.35

Table D.2: Effect of the $3rd$ Case on the Fitted Model for 1973 - 2007 Maize Data

Predictor Variable	With the 3rd case $R^2 = 0.40$				Without the $3rd$ Case $R^2 = 0.42$			
	Estimate	Std. error	t value	Pr $> \|t\|$	Estimate	Std. error	t value	Pr $> \|t\|$
Intercept	-44129	40380	-1.09	0.2832	-60363	40116	-1.50	0.1432
Rainfall	61.76	54.60	1.13	0.2670	131.30	65.95	1.99	0.0560
Harvested Area	0.90	0.25	3.55	0.0013	0.64	0.29	2.24	0.0327

Table D.3: Effect of the 27th Case on the Fitted Model for 1973 - 2007 Maize Data

Predictor Variable	With the 27th case $R^2 = 0.40$				Without the 27th Case $R^2 = 0.44$			
	Estimate	Std. error	t value	Pr $>\|t\|$	Estimate	Std. error	t value	Pr $>\|t\|$
Intercept	-44129	40380	-1.09	0.2832	-24803	30548	-0.81	0.4234
Rainfall	61.76	54.60	1.13	0.2670	62.46	40.96	1.52	0.1381
Harvested Area	0.90	0.25	3.55	0.0013	0.71	0.20	3.63	0.0011

Table D.4: Case Deletion Diagnostic for 1973 - 2007 Sorghum Data

Diagnostic Measure	14th Case Diagnostics	Cutoff Point
RStudent t_i	-3.08	2
leverage h_{ii}	0.18	$2p/n = 0.18$
DFFITS	-1.43	$2\sqrt{p/n} = 0.60$
Cook's Distance D_i	0.40	$F_{(0.5,p,n-p)} = 1.24$
COVRATIO	0.43	$1 \pm 3p/n = 1 \pm 0.27$
DFBETAS		$2/\sqrt{n} = 0.35$
Time	-0.88	0.35
Production-t	-0.89	0.35
Harvested Area	-1.14	0.35

Table D.5: Effects of the 14th Case on the Fitted Model for 1973 - 2007 Sorghum Data

Predictor Variable	With the 14th case $R^2 = 0.66$				Without the 14th Case $R^2 = 0.75$			
	Estimate	Std. error	t value	Pr $>\|t\|$	Estimate	Std. error	t value	Pr $>\|t\|$
Intercept	-5788.36	15685	-0.37	0.7148	-20940	14650	-1.43	0.1640
Time	-112.22	378.29	-0.30	0.7688	182.24652	346.26	0.53	0.6028
Production-t	0.31	0.15	2.03	0.0521	0.43	0.14	3.07	0.0047
Harvested Area	0.76	0.16	4.71	<.0001	0.92	0.15	6.09	<.0001

Table D.6: Case Deletion Diagnostic for 1973 - 1998 Sorghum Data

Diagnostic Measure	24th Case Diagnostics	Cutoff Point
RStudent t_i	-1.24	2
leverage h_{ii}	0.65	$2p/n = 0.32$
DFFITS	-1.70	$2\sqrt{p/n} = 0.80$
Cook's Distance D_i	0.56	$F_{(0.5,p,n-p)} = 1.15$
COVRATIO	2.52	$1 \pm 3p/n = 1 \pm 0.48$
DFBETAS		$2/\sqrt{n} = 0.40$
Time	1.00	0.40
Harvested Area	-0.26	0.40
Failed Area	-0.12	0.40
Price/Ton	-1.43	0.40

Table D.7: Effects of the $24th$ Case on the Fitted Model for 1973 - 1998 Sorghum Data

Predictor Variable	With the 24th case $R^2 = 0.68$				Without the 24th Case $R^2 = 0.70$			
	Estimate	Std. error	t value	Pr $> \|t\|$	Estimate	Std. error	t value	Pr $> \|t\|$
Intercept	40805	14451	2.82	0.0105	39167	14323	2.73	0.0132
Time	-991.28	871.46	-1.14	0.2688	1853.27	1106.12	-1.68	0.1102
Harvested Area	0.50	0.17	2.87	0.0095	0.55	0.18	3.10	0.0059
Failed Area	-1.77	0.64	-2.76	0.0122	-1.69	0.64	-2.65	0.0156
Price/Ton	1.39	25.05	0.06	0.9563	36.65	37.69	0.97	0.3431

Table D.8: Case Deletion Diagnostics for 1973 - 2007 Wheat Data

Diagnostic Measure	26th Case Diagnostics	Cutoff Point
RStudent t_i	4.61	2
leverage h_{ii}	0.07	$2p/n = 0.24$
DFFITS	1.28	$2\sqrt{p/n} = 0.70$
Cook's Distance D_i	0.25	$F_{(0.5,p,n-p)} = 1.16$
COVRATIO	0.19	$1 \pm 3p/n = 1 \pm 0.36$
DFBETAS		$2/\sqrt{n} = 0.35$
Time	0.74	0.35
Rainfall	-0.51	0.35
Production-t	-0.03	0.35
Harvested Area	0.38	0.35

Table D.9: Effects of the $26th$ Case on the Fitted Regression for 1973 - 2007 Wheat Data

Predictor Variable	With the 26rd case $R^2 = 0.64$				Without the 26th Case $R^2 = 0.78$			
	Estimate	Std. error	t value	Pr $> \|t\|$	Estimate	Std. error	t value	Pr $> \|t\|$
Intercept	-4407.79	12450	-0.35	0.7260	-5599.63	9487.80	-0.59	0.5600
Time	36.33	298.94	0.12	0.9041	-133.84	230.70	-0.58	0.5666
Rainfall	4.63921	16.22	0.29	0.7769	10.90	12.43	0.88	0.3881
Production-t	0.45	0.16	2.82	0.0088	0.46	0.12	3.72	0.0009
Harvested Area	0.51	0.21	2.42	0.0222	0.45	0.16	2.79	0.0095

Table D.10: Case Deletion Diagnostics for 1976 - 2007 Wheat Data

Diagnostic Measure	23rd Case Diagnostics	Cutoff Point
RStudent t_i	4.92	2
leverage h_{ii}	0.09	$2p/n = 0.27$
DFFITS	1.57	$2\sqrt{p/n} = 0.73$
Cook's Distance D_i	0.25	$F_{(0.5,p,n-p)} = 1.16$
COVRATIO	0.04	$1 \pm 3p/n = 1 \pm 0.40$
DFBETAS		$2/\sqrt{n} = 0.36$
Time	0.96	0.36
Production-t	-0.14	0.36
Population	-0.91	0.36
Harvested Area	0.27	0.36

Table D.11: Effects of the $23rd$ Case on the Fitted Model for 1976 - 2007 Wheat Data

Predictor Variable	With the 23rd case $R^2 = 0.57$				Without the 23rd Case $R^2 = 0.75$			
	Estimate	Std. error	t value	Pr $> \|t\|$	Estimate	Std. error	t value	Pr $> \|t\|$
Intercept	-7060.86	86757	-0.08	0.9358	-62193	63475	-0.98	0.3370
Time	347.88	3182.94	0.11	0.91	-1866.19	2335.95	-0.80	0.4322
Production-t	0.55	0.16	3.41	0.0022	0.57	0.12	4.87	<.0001
Population	-0.004	0.08	-0.05	0.9579	0.05	0.06	0.82	0.4194
Harvested Area	0.80	0.29	2.72	0.0117	0.74	0.21	3.50	0.0018

Appendix E

Model Fit Diagnostics Plots for Untransformed National Data

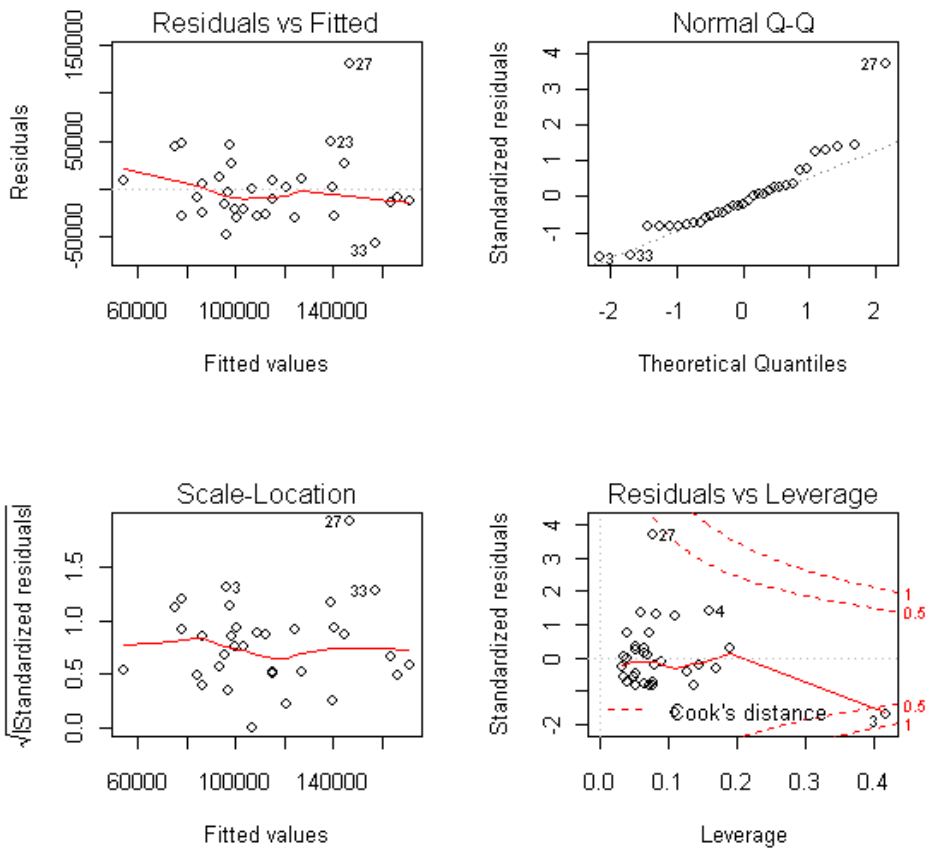

Figure E.1: Residual Plots for 1973 - 2007 Maize Data

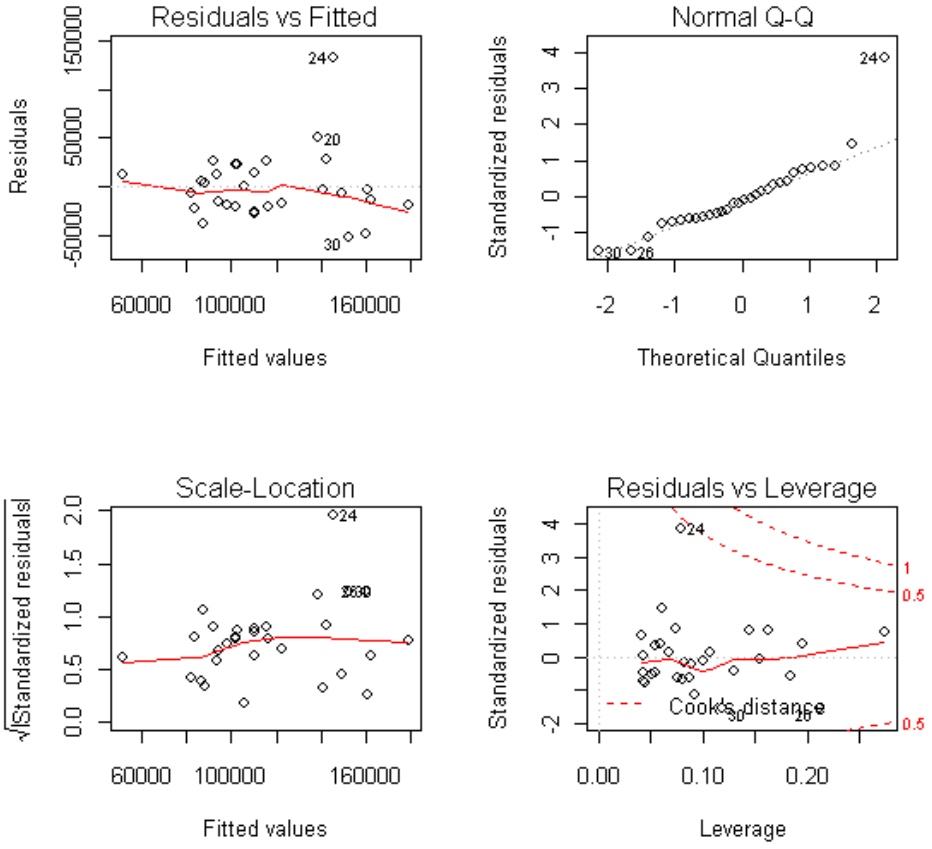

Figure E.2: Residual Plots for for 1976 - 2007 Maize Data

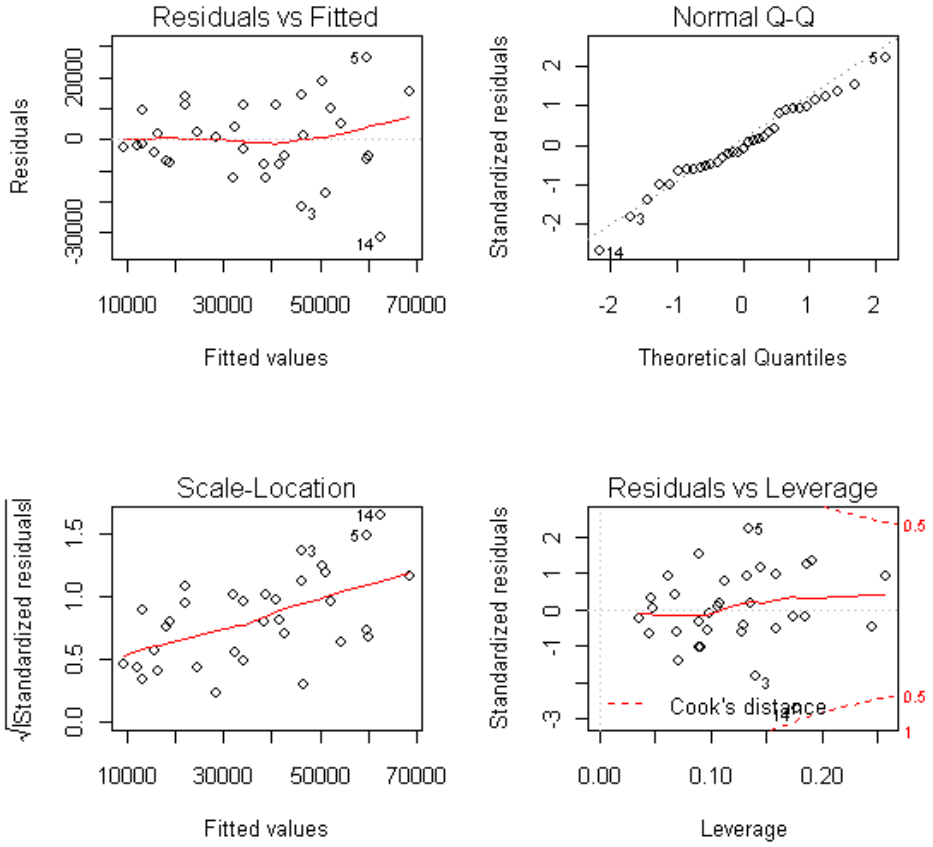

Figure E.3: Residual Plots for 1973 - 2007 Sorghum Data

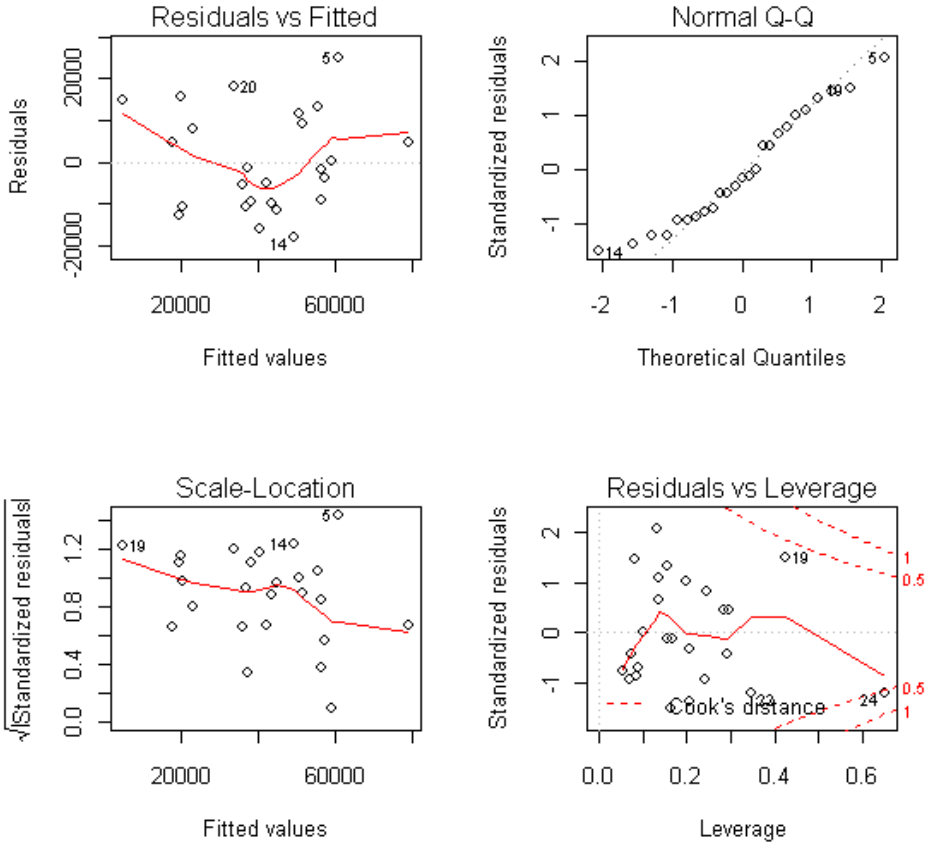

Figure E.4: Residual Plots for 1973 - 1999 Sorghum Data

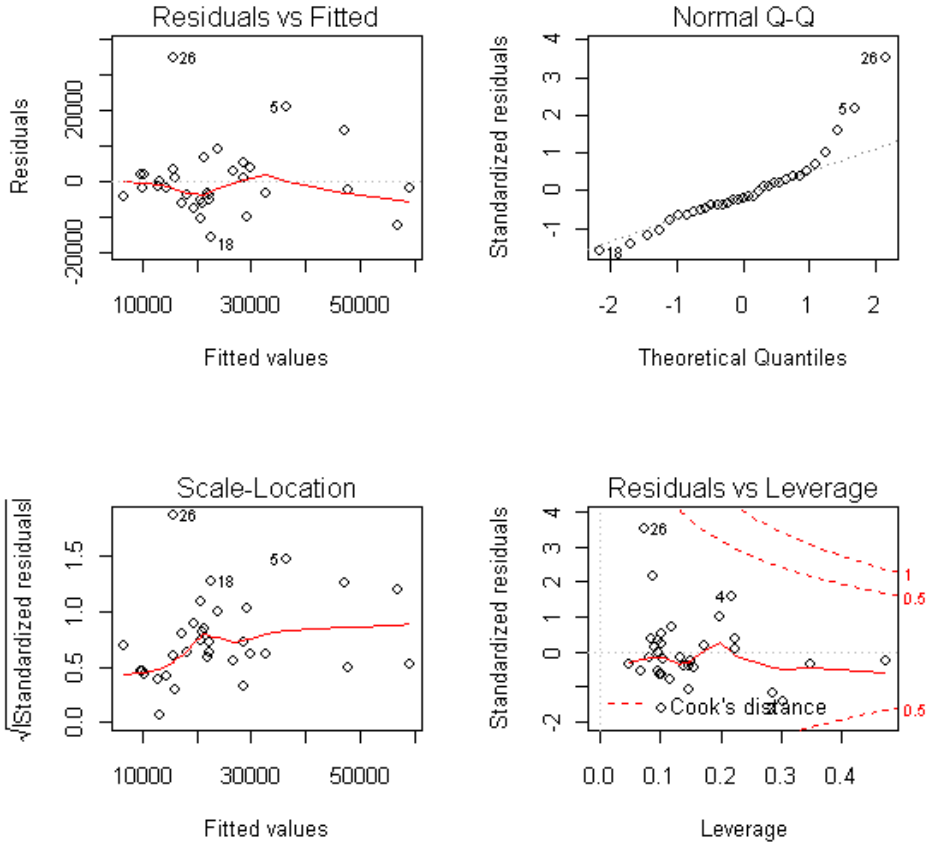

Figure E.5: Residual Plots for 1973 - 2007 Wheat Data

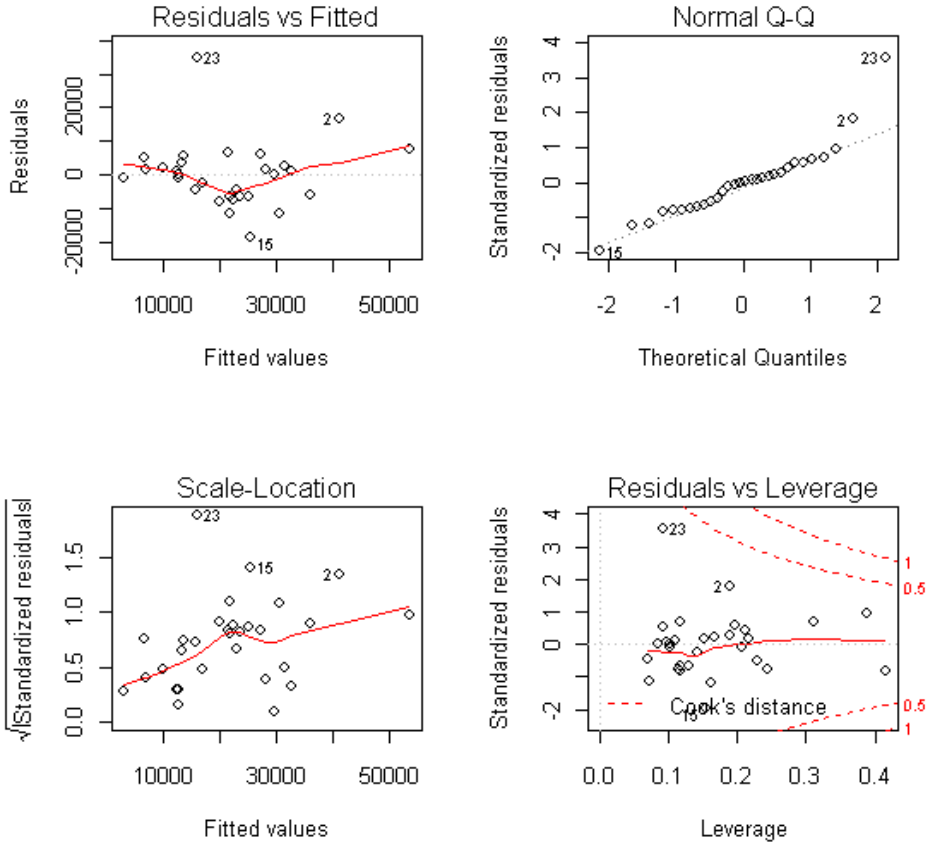

Figure E.6: Residual Plots for 1976 - 2007 Wheat Data

Appendix F

Box-Cox Transformation Log Likelihood Plots for National Data

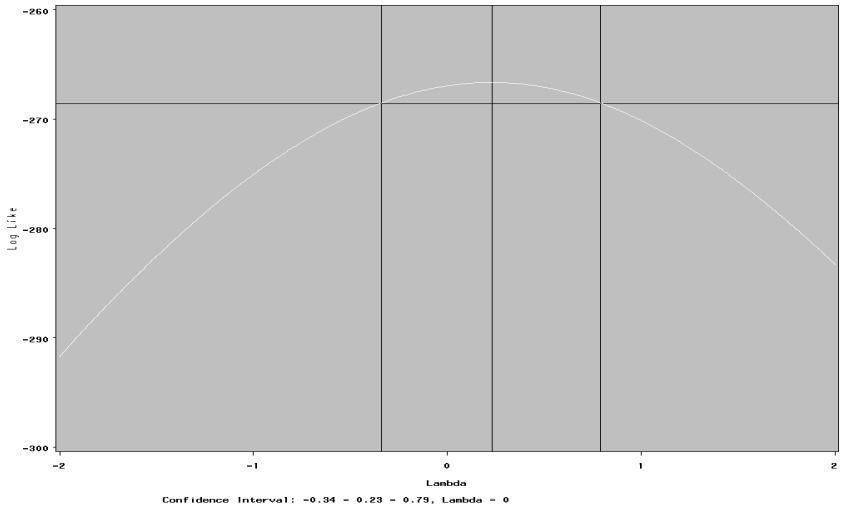

Figure F.1: Log Likelihood Plot for 1973 - 2002 Wheat Data

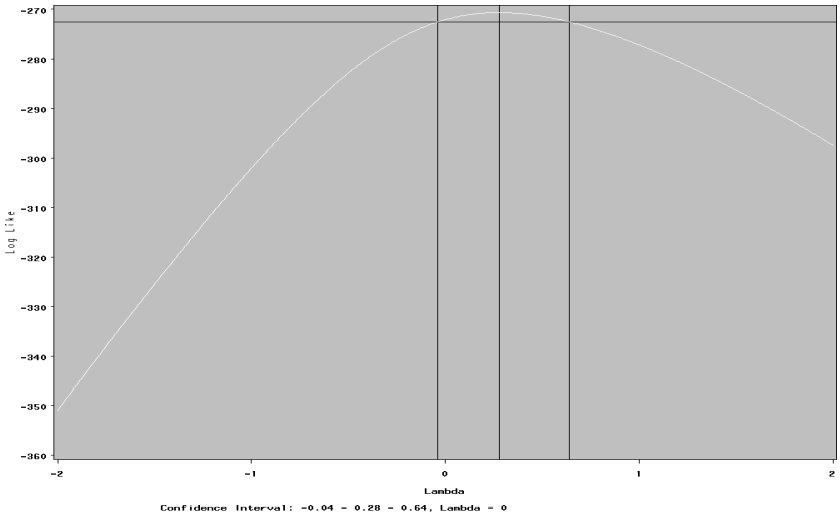

Figure F.2: Log Likelihood Plot for 1976 - 2007 Wheat Data

Appendix G

Model Fit Diagnostics Plots for Transformed Data

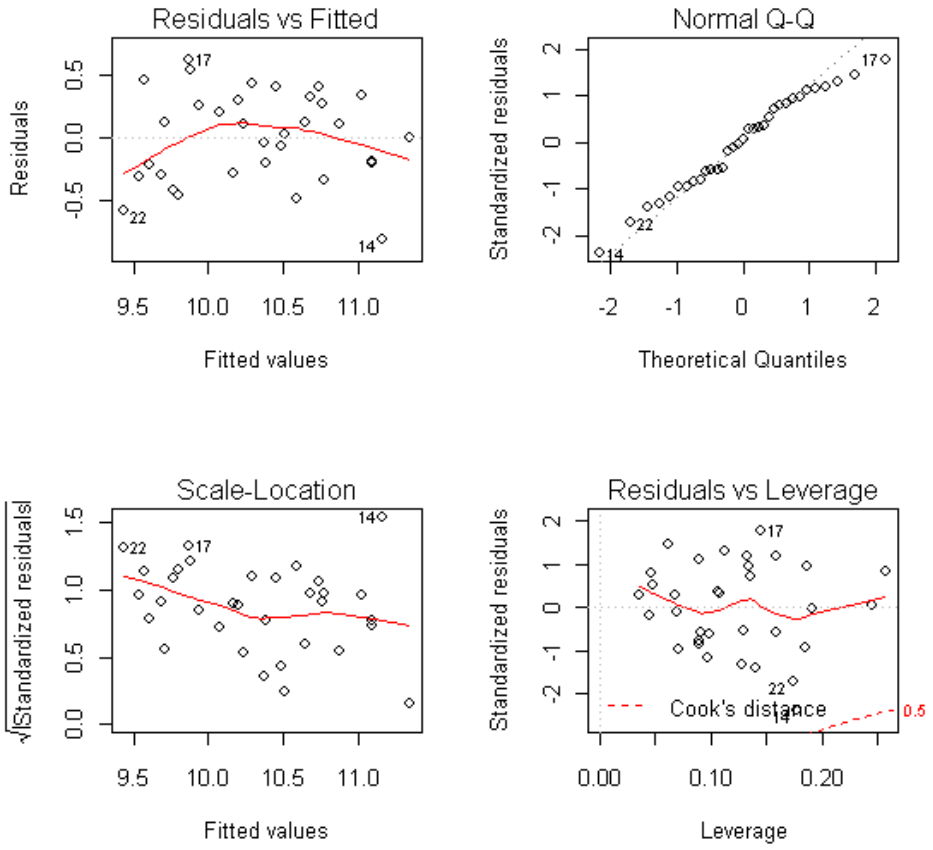

Figure G.1: Plot of Residuals for Transformed 1973 - 2007 Maize Data

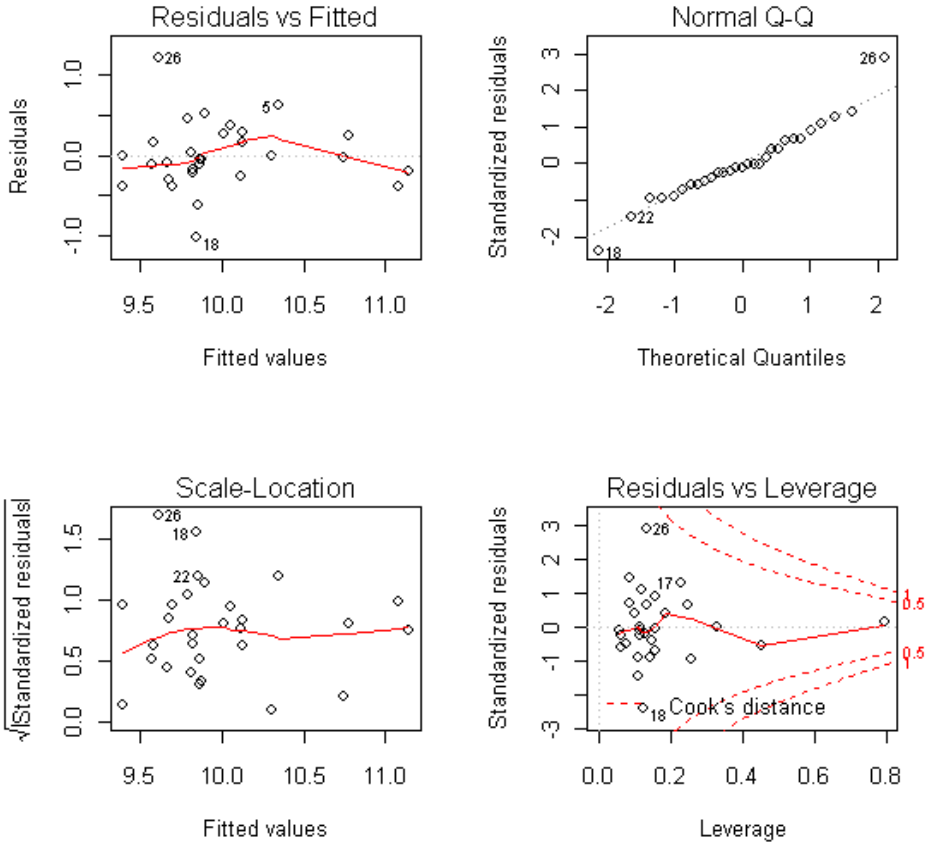

Figure G.2: Plot of Residuals for Transformed 1973 - 2002 Wheat Data

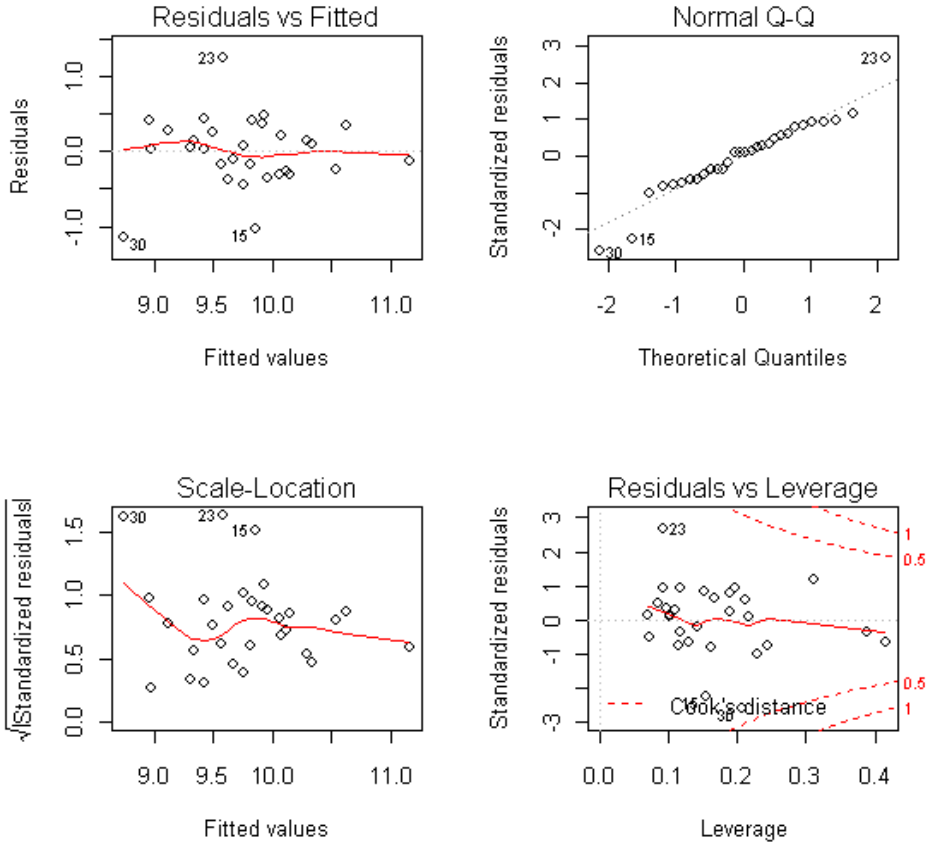

Figure G.3: Plot of Residuals for Transformed 1976 - 2007 Wheat Data

Appendix H

Ridge Traces for National Data

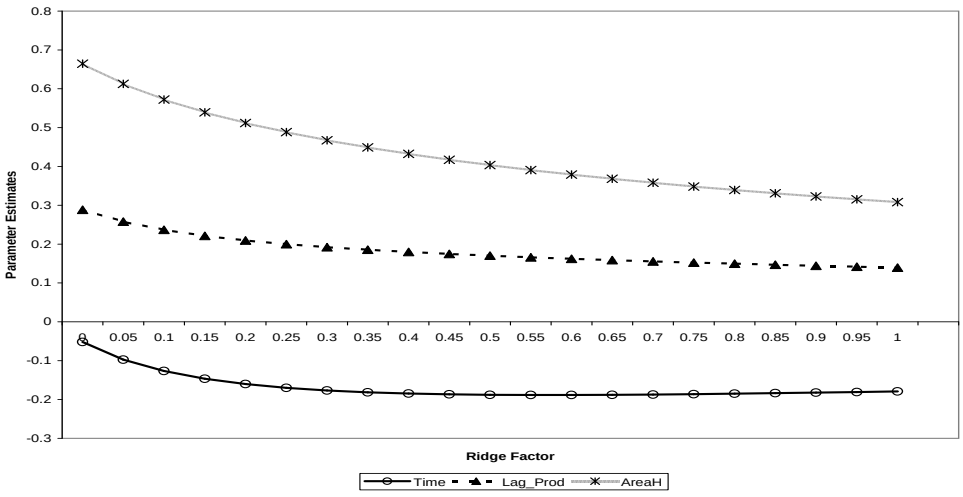

Figure H.1: Ridge Trace for 1973 - 2007 Sorghum Data

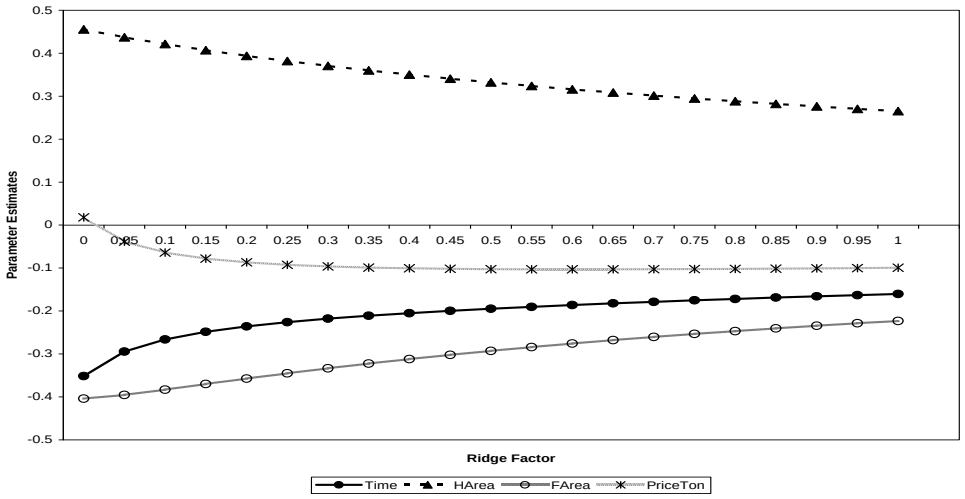

Figure H.2: Ridge Trace for 1973 - 1998 Sorghum Data

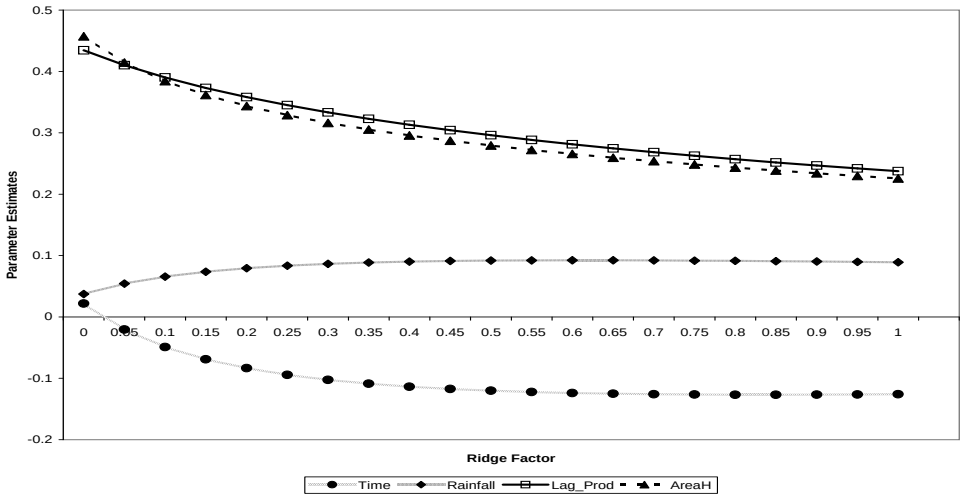

Figure H.3: Ridge Trace for 1973 - 2007 Wheat Data

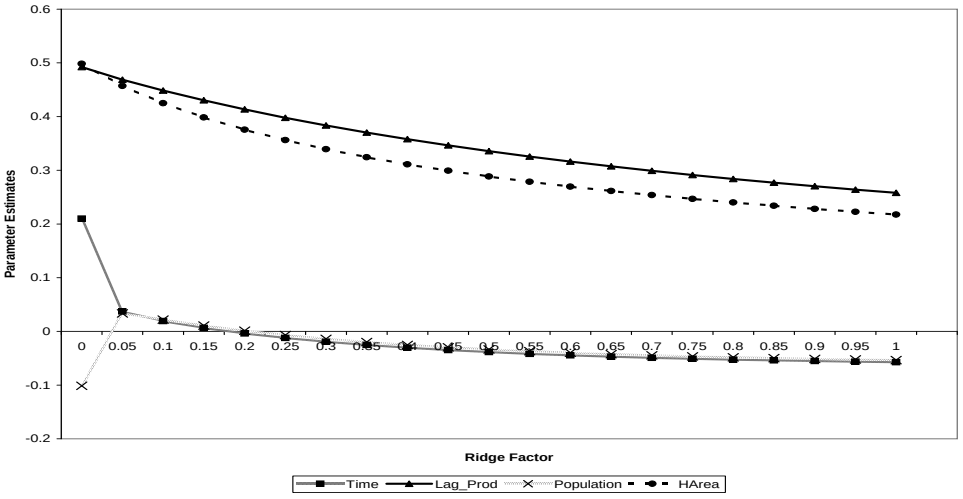

Figure H.4: Ridge Trace for 1976 - 2007 Wheat Data

Appendix I

Ridge Regression Parameter Estimates for National Data

Table I.1: Parameter Estimates for 1973 - 2007 Sorghum Data at $\delta = 0$ and $\delta = 0.05$

Variable	$\delta = 0$			$\delta = 0.05$		
	Estimate	Std. Error	VIF_i	Estimate	Std. Error	VIF_i
Time	-112.22	378.29	2.63	-210.01	317.24	1.84
Production-t	0.31	0.15	1.74	0.27	0.13	1.34
Harvested area	0.76	0.16	0.71	0.70	0.14	1.32

Table I.2: Parameter Estimates for 1973 - 1998 Sorghum Data at $\delta = 0$ and $\delta = 0.20$

Variable	$\delta = 0$			$\delta = 0.20$		
	Estimate	Std. Error	VIF_i	Estimate	Std. Error	VIF_i
Time	-991.28	871.46	5.89	-664.63	324.70	0.78
Harvested area	0.50	0.17	1.55	0.43	0.12	0.82
Failed area	-1.77	0.64	1.33	-1.56	0.48	0.73
Price/Ton	1.39	25.05	6.42	-6.73	8.87	0.78

Table I.3: Parameter Estimates for 1973 - 2007 Wheat Data at $\delta = 0$ and $\delta = 0.15$

Variable	$\delta = 0$			$\delta = 0.15$		
	Estimate	Std. Error	VIF_i	Estimate	Std. Error	VIF_i
Time	36.33	298.94	2.52	-114.61	196.62	1.07
Rainfall	4.64	16.22	1.33	9.15	13.06	0.84
Production-t	0.45	0.16	1.86	0.39	0.12	1.03
Harvested area	0.51	0.21	2.78	0.41	0.14	1.13

Table I.4: Parameter Estimates for 1976 - 2007 Wheat Data at $\delta = 0$ and $\delta = 0.25$

Variable	$\delta = 0$			$\delta = 0.25$		
	Estimate	Std. Error	VIF_i	Estimate	Std. Error	VIF_i
Time	347.88	3182.94	214.41	-204.1	114.42	0.26
Production-t	0.55	0.16	1.21	0.45	0.12	0.68
Population	-0.004	0.08	209.28	-0.0003	0.003	0.28
Harvested area	0.80	0.29	1.95	0.57	0.19	0.78

Appendix J

Standard Errors of Regression Quantiles for Sorghum Availability

Table J.1: Standard Errors of the $25th$ Regression Quantile for Sorghum Availability

Predictor Variable	Category	Estimate	Standard Errors		
			Sparsity	Kernel	Bootstrap
Intercept		300.39	75.60	131.53	117.08
Household Size		-0.67	15.33	28.40	22.79
Household Size^2		0.81	1.28	2.56	2.15
Income		-0.01	0.02	0.02	0.02
Sex	(Ref=Male)				
	Female	-0.18	20.87	32.89	21.67
Location	(Ref=Berea)				
	Mafeteng	-71.84	29.77	39.80	25.21
	Maseru Foothills	-71.82	43.37	59.37	41.53
	Maseru Lowland	-28.42	28.01	38.13	29.58
Education	(Ref=Post COSC)				
	High School	-140.65	60.15	109.19	87.95
	No Formal Education	-132.12	58.44	106.28	89.34
	Primary	-132.68	57.23	107.08	88.05
Occupation	(Ref=Casual Worker)				
	Subsistence Farmer	-62.75	33.62	46.50	30.63
	Pensioner	-89.08	43.65	66.54	41.00
	Salary Earner	-80.28	37.57	63.33	33.61
	Unemployed	-94.67	33.00	45.06	27.10

Table J.2: Standard Errors of the $50th$ Regression Quantile for Sorghum Availability

Predictor Variable	Category	Estimate	Standard Errors		
			Sparsity	Kernel	Bootstrap
Intercept		436.72	70.84	140.01	163.00
Household Size		-22.74	14.36	31.26	26.82
Household Size^2		2.96	1.19	2.80	2.48
Income		0.01	0.02	0.03	0.03
Sex	(Ref=Male)				
	Female	15.75	19.56	33.97	24.53
Location	(Ref=Berea)				
	Mafeteng	-25.02	27.90	42.39	30.21
	Maseru Foothills	-72.78	40.64	62.95	46.14
	Maseru Lowland	36.32	26.25	42.75	31.20
Education	(Ref=Post COSC)				
	High School	-231.00	56.36	109.16	133.83
	No Formal Education	-226.08	54.76	106.90	139.10
	Primary	-225.04	53.63	103.99	131.74
Occupation	(Ref=Casual Worker)				
	Subsistence Farmer	-50.53	31.50	55.35	46.47
	Pensioner	-127.43	40.90	77.09	72.85
	Salary Earner	-124.88	35.21	69.75	74.14
	Unemployed	-85.41	30.93	53.79	47.14

Table J.3: Standard Errors of the $75th$ Regression Quantile for Sorghum Availability

Predictor Variable	Category	Estimate	Standard Errors		
			Sparsity	Kernel	Bootstrap
Intercept		399.18	127.14	168.59	295.61
Household Size		5.16	25.78	31.80	34.56
Household Size^2		0.94	2.14	2.91	2.93
Income		0.03	0.03	0.06	0.06
Sex	(Ref=Male)				
	Female	7.86	35.10	37.99	34.69
Location	(Ref=Berea)				
	Mafeteng	-13.17	50.07	54.14	60.14
	Maseru Foothills	-6.03	72.93	86.48	94.15
	Maseru Lowland	52.36	47.11	53.25	58.29
Education	(Ref=Post COSC)				
	High School	-123.80	101.16	151.65	277.36
	No Formal Education	-277.27	98.29	148.37	270.82
	Primary	-203.13	96.25	133.94	262.82
Occupation	(Ref=Casual Worker)				
	Subsistence Farmer	-64.78	56.54	72.91	81.82
	Pensioner	-38.50	73.41	93.10	101.28
	Salary Earner	-95.92	63.19	80.39	89.27
	Unemployed	-87.77	55.51	74.92	83.87

Appendix K

Quantile Plots of Sorghum Household Data

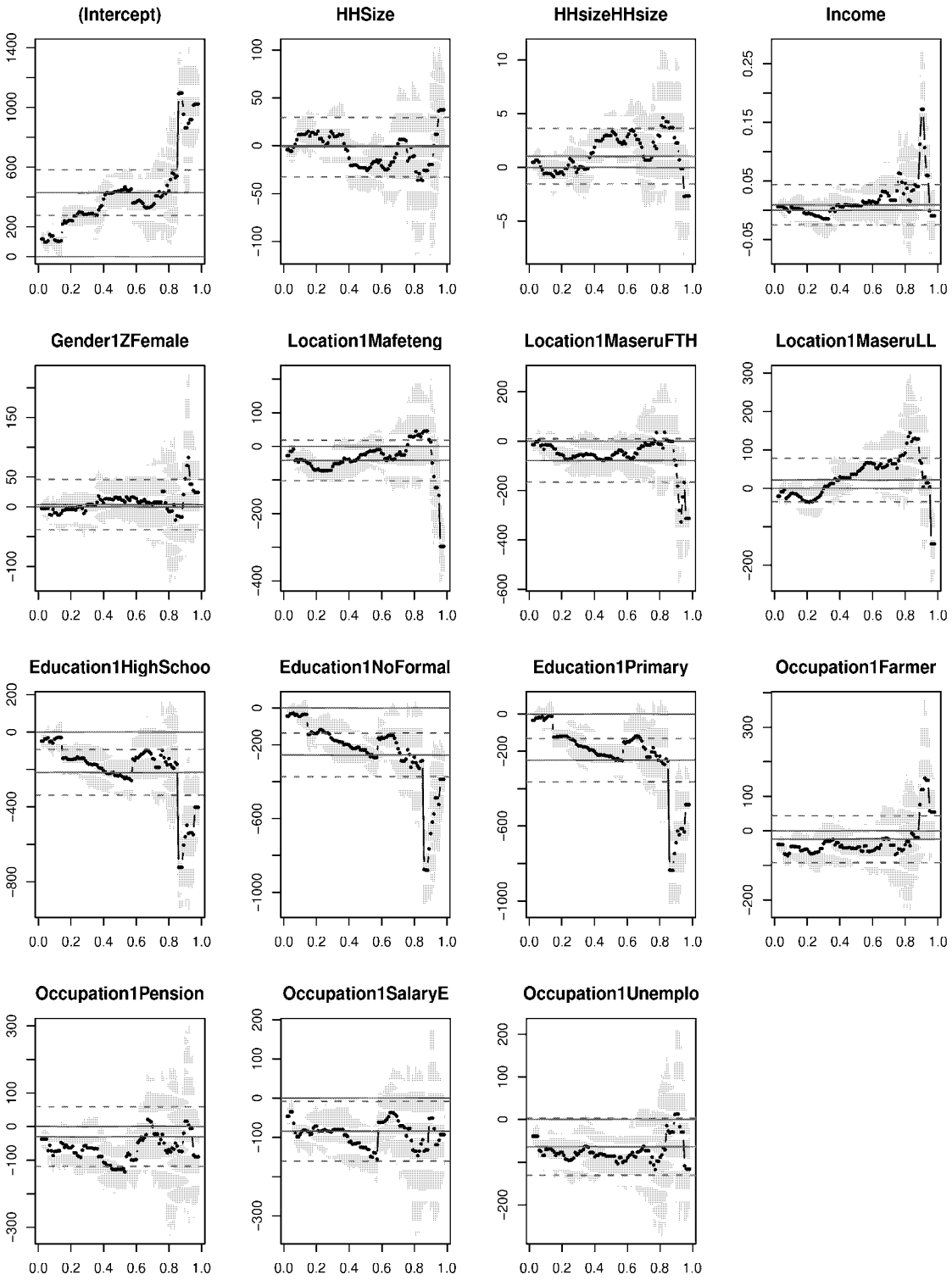

Figure K.1: Quantile Plot of Sorghum Availability

Appendix L

Diagnostics Plots For Logistic Regression

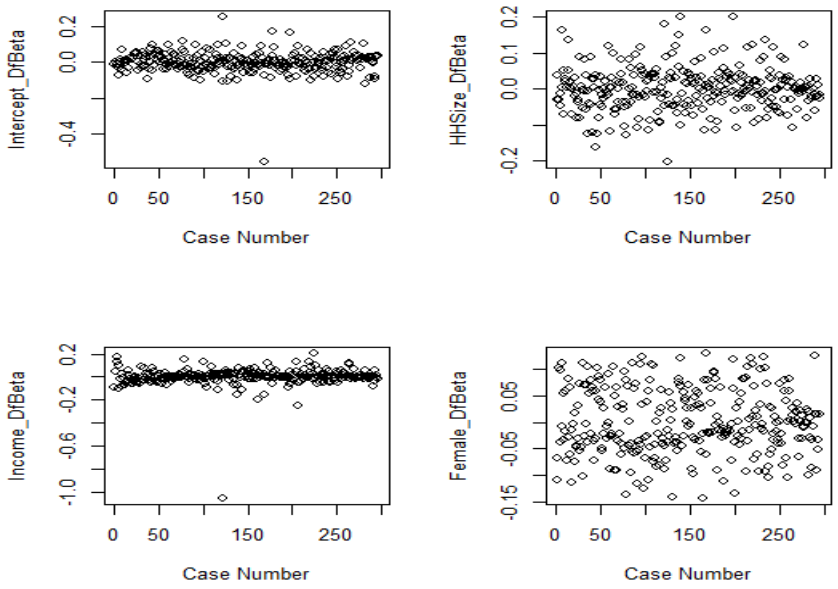

Figure L.1: Plots of DBetas for Intercept, Household Size, Income and Females for Wheat Availability

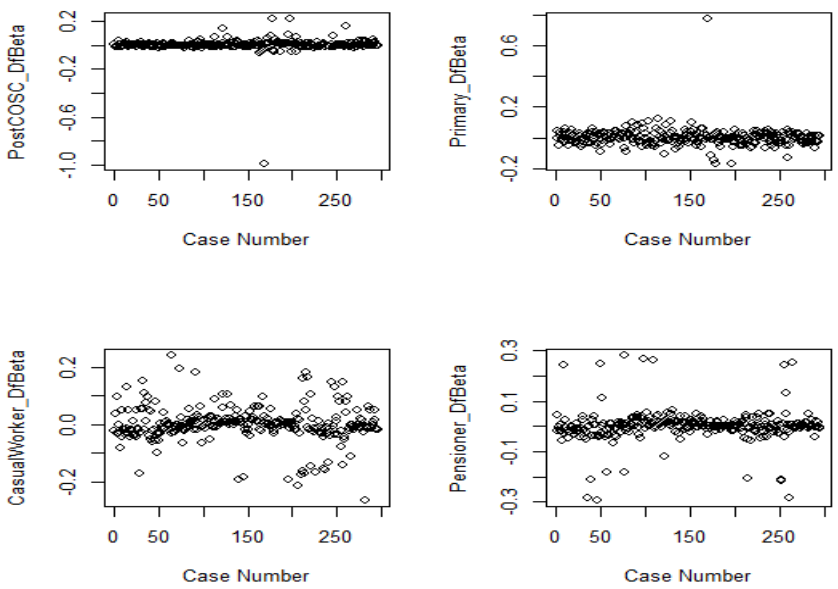

Figure L.2: Plots of DBetas for PostCOSC, Primary, Casual Worker and Pensioner for Wheat Availability

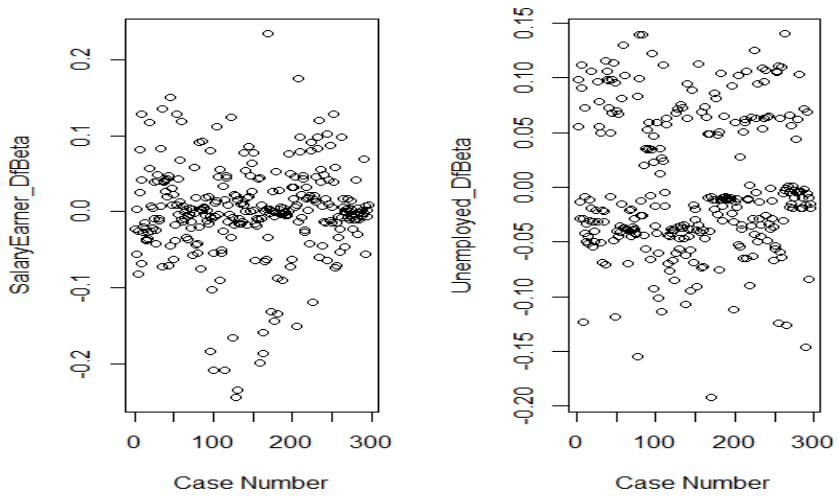

Figure L.3: Plots of DBetas for Salary Earner and Unemployed for Wheat Availability

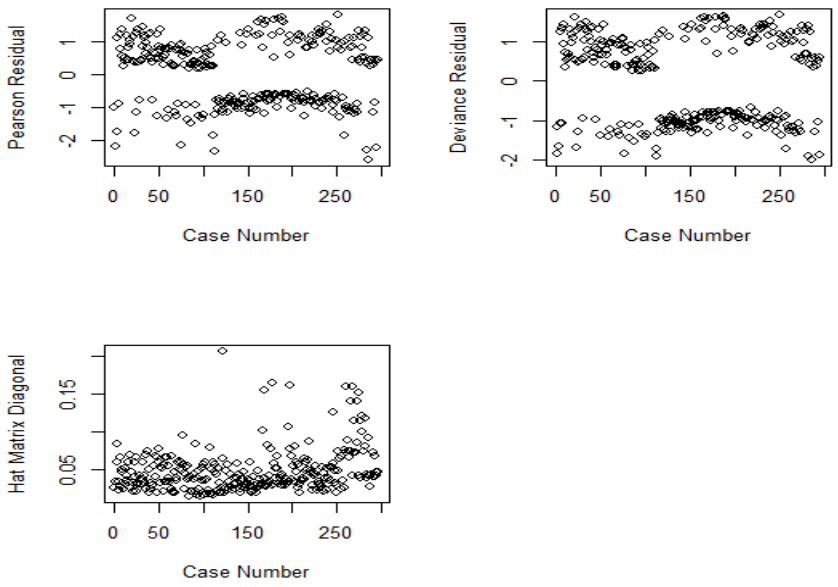

Figure L.4: Plots of Residuals and Hat Matrix Diagonal for Sorghum Availability

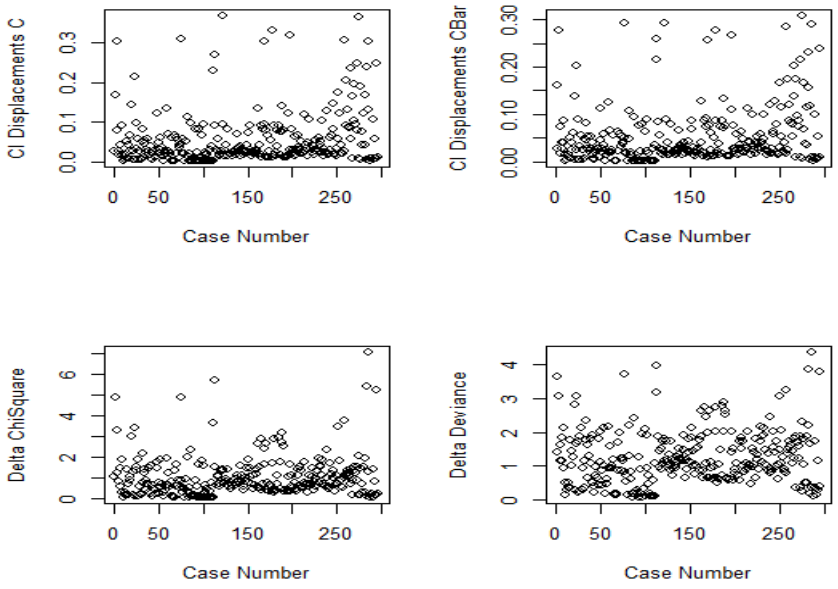

Figure L.5: Plots of CI Displacements C and CBar, and Change in ChiSquare and Deviance for Sorghum Availability

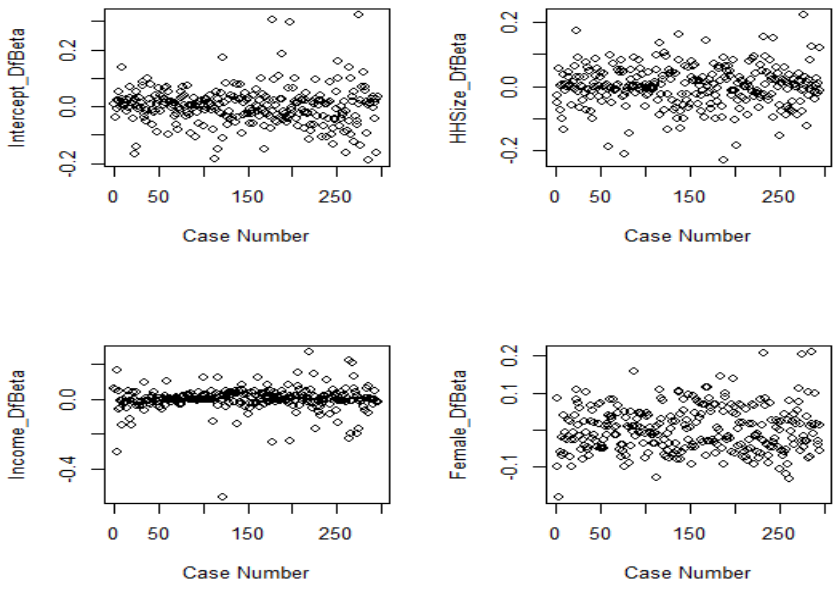

Figure L.6: Plots of DBetas for Intercept, Household Size, Income and Females for Sorghum Availability

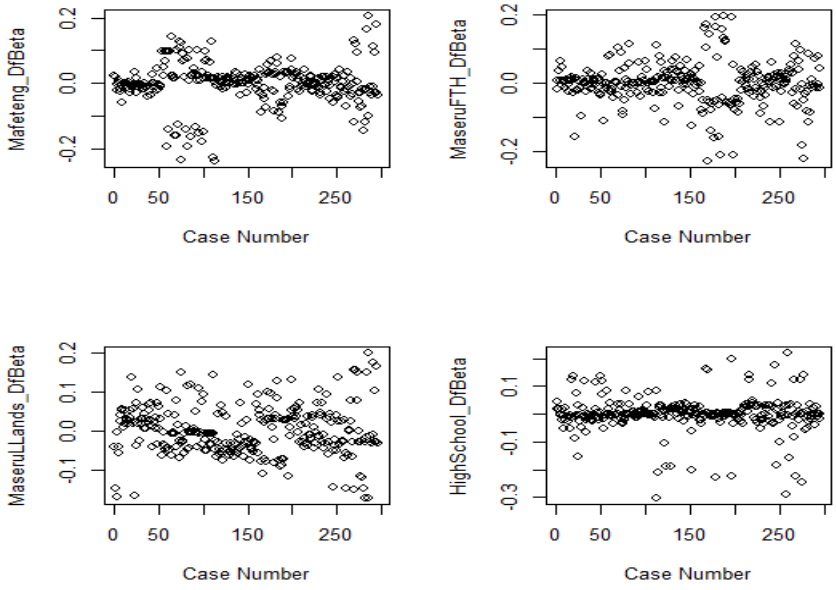

Figure L.7: Plots of DBetas for Mafeteng, Maseru Foot Hills, Maseru Low Lands and High School for Sorghum Availability

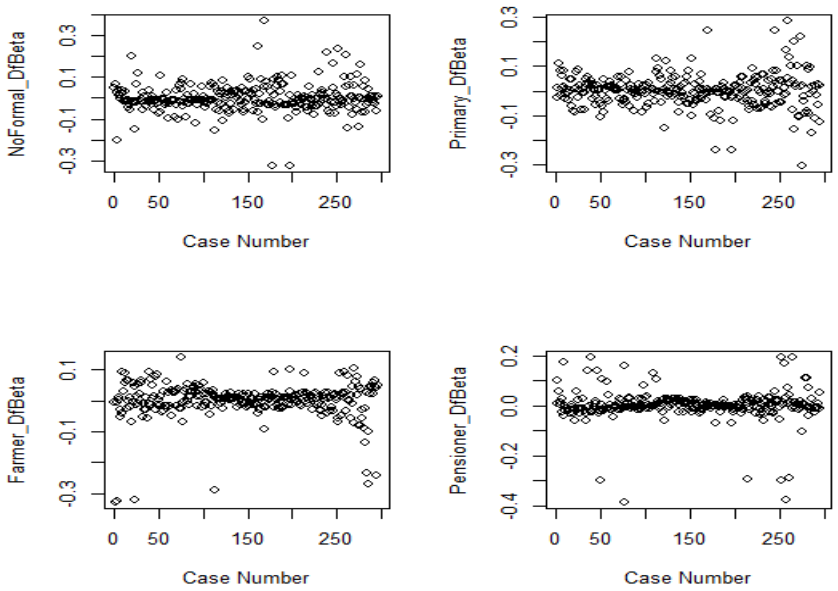

Figure L.8: Plots of DBetas for No Formal, Primary, Casual Worker and Pensioner for Sorghum Availability

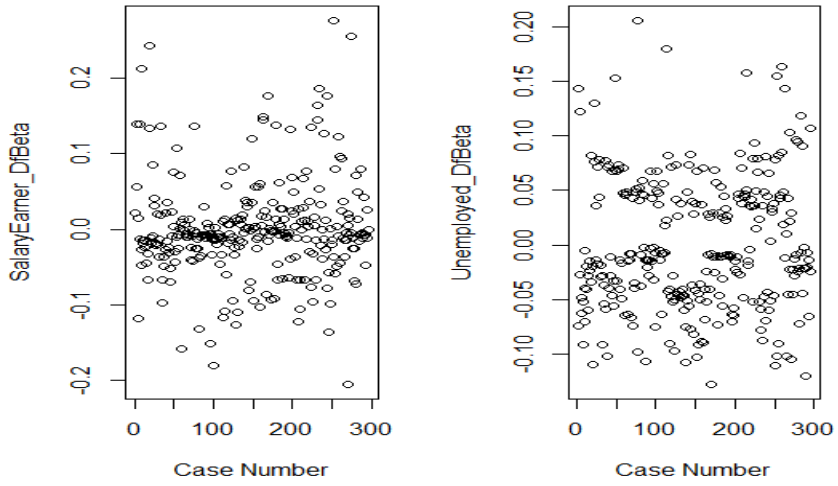

Figure L.9: Plots of DBetas for Salary Earner and Unemployed for Sorghum Availability

Appendix M

Box Plots for National Data

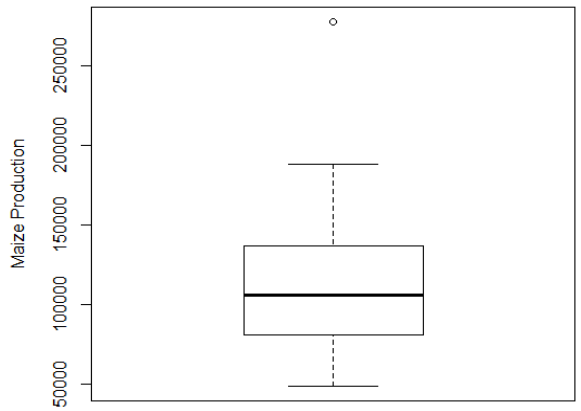

Figure M.1: Boxplot for 1973 - 2007 Maize Data

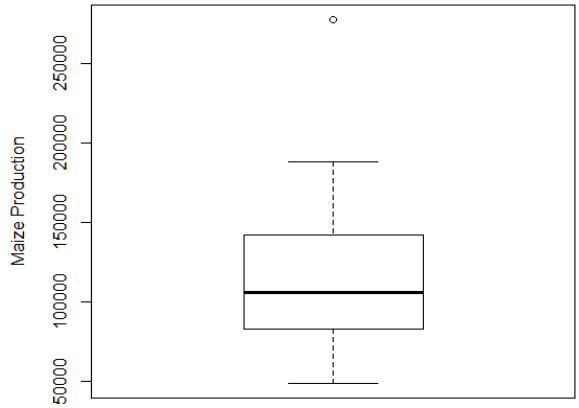

Figure M.2: Boxplot for 1976 - 2007 Maize Data

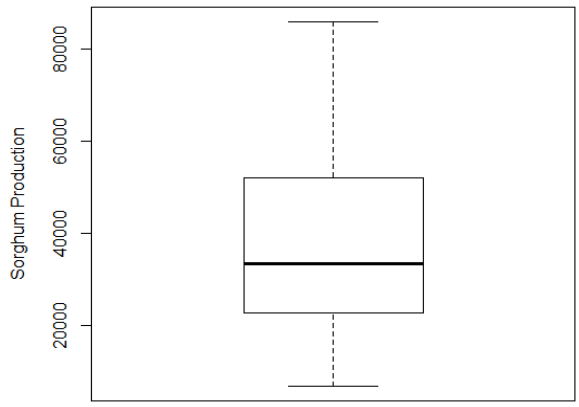

Figure M.3: Boxplot for 1973 - 2007 Sorghum Data

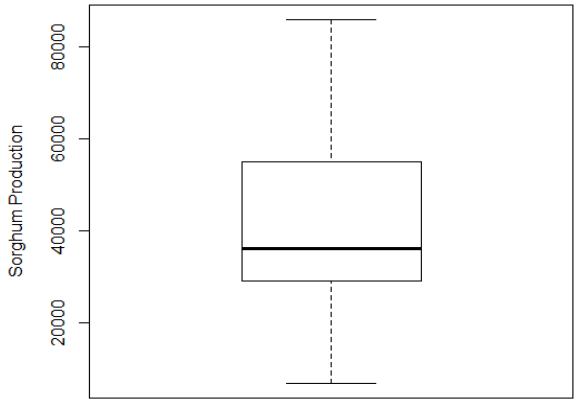

Figure M.4: Boxplot for 1973 - 1998 Sorghum Data

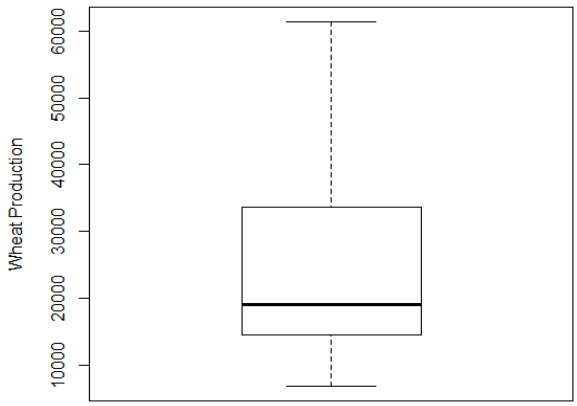

Figure M.5: Boxplot for 1973 - 2007 Wheat Data

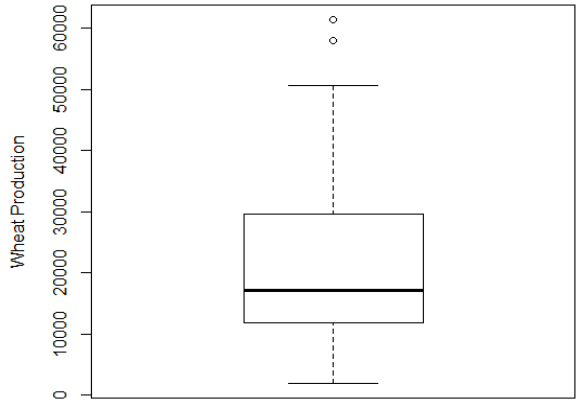

Figure M.6: Boxplot for 1976 - 2007 Wheat Data

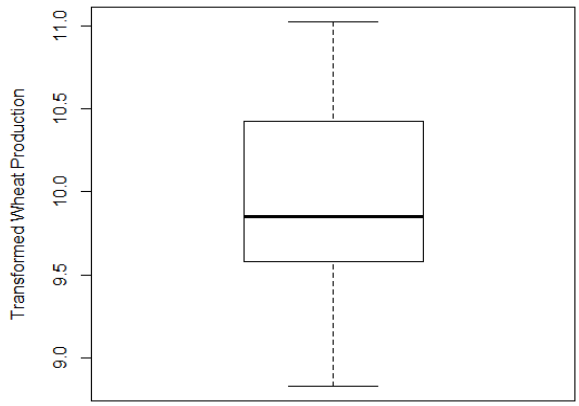

Figure M.7: Boxplot for the Transformed 1973 - 2002 Wheat Data

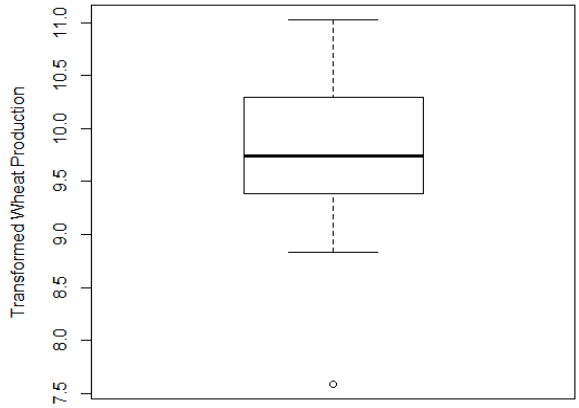

Figure M.8: Boxplot for the Transformed 1976 - 2007 Wheat Data