32

ADDIS ABABA UNIVERSITY

GRADUATE STUDIES PROGRAM

FACULTY OF SCIENCE

DEPARTMENT OF STATISTICS

A PRODUCTION FUNCTION ANALYSIS FOR PRIVATE PEASANT HOLDINGS CROP FARMS IN ETHIOPIA: AN APPLICATION OF ROBUST REGRESSION

By

Taddesse Kassahun

A thesis submitted to the Graduate Studies Program of Addis Ababa University in partial fulfillment of the requirements for the Degree of Master of Science in Applied Statistics

JUNE, 2009

ADDIS ABABA UNIVERSITY

GRADUATE STUDIES PROGRAM

FACULTY OF SCIENCE

DEPARTMENT OF STATISTICS

A PRODUCTION FUNCTION ANALYSIS FOR PRIVATE PEASANT HOLDINGS CROP FARMS IN ETHIOPIA: AN APPLICATION OF ROBUST REGRESSION

By

Taddesse Kassahun

Approved by the Board of Examiners:

Sileshi Fanta (Mr.) ____________________ ____________________

Department Head Signature

Eshetu Wencheko (Prof.) ____________________ ____________________

Examiner Signature

M.K. Sharma (Ph.D) ____________________ ____________________

Examiner Signature

Tel. 251-(0)11-1239463 P.O.Box: 1176 Addis Ababa, Ethiopia

ACKNOWLEDGEMENTS

First and foremost, my profound and special gratefulness go to my instructor and thesis advisor, Dr. Fentaw Abegaz for his undeniable and efficient way of guiding throughout the entire thesis work time. His wholehearted encouragement to inspire me spotlight on this area as well as his gracious and due response for each of my problems as of the day I have known him are countless in that I can hardly have the aptitude to articulate in words. All of his contributions have significantly positive consequence on the completion of this research.

I am in high spirits to express my sincere enjoyment to my friend, Girma Minalu for his valuable assistance in the course of the study by providing me various learning materials including the most recent journals and statistical packages. I would also like to articulate my gratitude to Belay Anagaw for all responsibilities that he accomplished during my study leave from my workplace.

I am very grateful to the staff members of Central Statistical Agency of Ethiopia, for consenting me to access the data used in this research without fee, especially to W/ro Eleni for her genuine and polite way of reception when I went to take data at her office.

I further extend my gratitude to all of my instructors for their regularity and welcoming approach they have shown me during the teaching – learning seasons.

Praise the Lord for opening all possible gates to my further education as well as giving me a lot that I pray!

Finally, my heartfelt thanks go to every member of my beloved family for their day to day and cherished enthusiasm not only in my academic triumph but also in my personality at large. Dad, your persistent curious about every of my steps forward inspired me to complete my tasks competently. I also extend my earnest appreciation to my elder sister, Ejeyensu Kassahun, who played a pivotal role to build the base for pursuing my current career.

TABLE OF CONTENTS

Contents **Page**

LIST OF TABLES

Table **Page**

LIST OF FIGURES

Figure **Page**

Acronyms and Abbreviations

ADLI	Agriculture Development Led Industrialization
ANOVA	Analysis of Variance
AR (p)	Autoregressive of order p.
CES	Constant Elasticity of Substitution
CN	Condition Number
CSA	Central Statistical Agency
CV	Coefficient of Variation
EC	Ethiopian Calendar
FAO	Food and Agriculture Organization of the United Nations
FDREPCC	Federal Democratic Republic of Ethiopia Population Census Commission
GDP	Gross Domestic Product
LAD	Least Absolute Deviation
LTS	Least Trimmed of Squares
MA (q)	Moving Average of order q.
MGQ	Modified Goldfeld-Quandt
MLE	Maximum Likelihood Estimation
MOFED	Ministry of Finance and Economic Development
MSDR	Median of the Squared Deletion Residuals
MSE	Mean Square Error
NPK	Nitrogen-Phosphorus-Potassium
OLS	Ordinary Least Squares
Q-Q	Qunatile Quantile
RLS	Reweighted Least Squares
SAS	Statistical Analysis Software
SNNP	South Nations, Nationalities and Peoples
translog	Transcendental Logarithmic
UNESC	United Nations Economic and Social Council
VIF	Variance Inflation Factor
WFP	World Food Programme
WLS	Weighted Least Squares

ABSTRACT

This study applied production function analysis for private peasant holdings crop farms in Ethiopia. Four major crop producing regions viz., Tigray, Amhara, Oromia and SNNP were included in the study. Three regression models for production function namely, linear, exponential and Cobb Douglas were considered and thoroughly assessed for statistical model diagnostics. The statistical model diagnostics and checking suggested that crop production function for each of the regions was found to be appropriately represented by the Cobb-Douglas production function based on data from the 2007/08 (2000 EC) agricultural sample survey. The Cobb-Douglas production function was first fitted for each region using ordinary least squares (OLS) regression. As expected, the parameter estimates using OLS were misleading due to the occurrence of several outlying cases and hence robust regression was taken as a viable alternative. Based on the results of robust regression, many of the parameter estimates took on the expected signs, the R^2 values were substantially increased and the standard errors of parameter estimates were decreased at large.

In general, robust regression results indicated that farm size, fertilizer, seed, oxen power and human labor were playing a pivotal role for the maximization of crop yield in each of the studied regions. From among these variables, the great contribution was found to be due to farm size in each of the regions with SNNP an exception in which the great share was due to human labor. However, the contribution of education variable for crop yield was found to be statistically insignificant and received negative sign in Tigray and Amhara regions. Counter to expectations, the coefficient estimate for irrigation variable in Tigray, Amhara and SNNP regions had come to obtain negative sign though it was not found to be statistically significant. The production elasticities for each of the inputs at each region except farm size in Tigray, Amhara and Oromia suggested that the relation between inputs and output was inelastic, i.e., holding other factors constant, the marginal return to each factor will decrease as more of the factors are used. Additionally, crop production functions revealed that returns to scale were estimated to be greater than unity in each of the regions indicating increasing returns to scale.

CHAPTER ONE

1. INTRODUCTION

1.1 Background

Agriculture in Ethiopia is the leading activity of the nation and hence the country's economy is predominantly agrarian. It has on average accounted for about 50% of the overall GDP, generates 90% of export earnings and supplies about 70% of the country's raw material to the secondary activities (MOFED, 2007). The major food crops are produced in almost all regions of the country in spite of the variation in volume of production across the regions. To this end, main season ('Meher' season, i.e., September to December) production of major crops by private peasant holdings accounted on average for over 90% of total output of major crops and 93% of cultivated area in any one year (MOFED, 2007).

Ethiopia grows large varieties of crops, which include cereals ('teff', corn, wheat, barley, sorghum, millet, oats, etc.); pulses (horse beans, chick-peas, haricot beans, field peas, lentils, soybean, and vetch); oilseeds (linseed, fenugreek, 'noug', rapeseed, sunflower, castor bean, groundnuts, etc.); stimulants (coffee, tea, 'chat', tobacco, etc.); fibers (cotton, sisal, flax, etc.); fruits (banana, orange, grape, papaya, lemon, 'menderin', apple, pineapple, mango, avocado, etc.); vegetables (onion, tomato, carrot, cabbage, etc.); root and tuber (potato, 'enset', sweet-potatoes, beets, yams, etc.) and sugarcane. According to the 2007/08 (2000 EC) annual crop production forecast survey, which was conducted by Central Statistical Agency (CSA) of Ethiopia, a total area of about 11 million hectares were covered by grain crops, i.e., cereals, pulses and oil seeds, from which a total of about 164.51 million quintals of grains were expected to be produced, from private peasant holdings.

Ethiopia has great agricultural potential because of its vast areas of fertile land, diverse climate, generally adequate rainfall, and large labor pool. Despite this potential, however, agriculture in Ethiopia has remained underdeveloped because it is plagued by periodic drought; soil degradation (which is caused by overgrazing, deforestation and high population density)[1] and a poor economic base (low productivity, low land management, weak infrastructure and low level of technology). The existing backward and traditional farm tools and very limited use of modern farm inputs together with dominating rain fed agriculture, (where the performance of the sector is highly dependent on the timely onset, duration, amount and distribution of rainfall) have contributed a lot for the existing subsistence / hand to mouth / farming system.

In a country with dominating agrarian economy like Ethiopia, due attentions, (including extensive research works) have to be given for increasing the performance of agriculture, in terms of total volume of production in order to secure domestic food availability for its population. To this verity, the government of Ethiopia has been devising and implementing various economy wide and sectoral policies and strategies. During the last one and half decade, the government has identified agriculture as a priority sector for development, and hence, devised the Agriculture Development Led Industrialization (ADLI) strategy. In order to realize the millennium development goals, agricultural extension services were expanded and supplies and applications of modern inputs increased leading to some improvements of aggregate production particularly of cereals, pulses, and oil seeds (UNESC, 2007).

1.2: Crop Characteristics of Ethiopia

Ethiopia grows a large variety of crops in which grains are the most important field crops and being the chief element in the diet of most Ethiopians. The principal grain crops are teff, wheat, barley, which are primarily cool-weather crops; and corn, sorghum, and millet which are warm weather grain crops. Teff is the most preferred crop grown in the cooler highlands, while sorghum is the principal lowland crop because it thrives well in semi-arid environments due to its hardy and drought resistant properties.

[1] Wikipedia: the free encyclopedia, last modified on 5 March 2009 at 03: 09pm, Wikimedia Foundation Inc. encyclopedia on-line, available from http://en.wikipedia.org/wiki/Economy_of_Ethiopia. Internet retrieved 06 March 2009.

Cool Weather Cereal Crops: Teff, Wheat and Barley

Teff, wheat, and barley are cool weather crops grown predominantly in the Ethiopian highlands at optimum altitude range of 1800 to 2200 meters. Teff, indigenous to Ethiopia, forms the staple diet of many Ethiopians and it furnishes the flour to make 'injera', unleavened bread that is consumed in the highlands and in urban centers throughout the country. Teff is, however, a very delicate and fragile crop that requires a lot of work and care, and it has one of the lowest yields of the cereal crops. Wheat is grown primarily as a rainfed crop by smallholders in the highlands of Ethiopia. In most parts of the country, wheat crop is grown during the longer rainy season (meher), which usually starts in June than the short rains season (belg), starting in March. Wheat is cultivated on more than 980,000 ha and its total production in 2008 was 23.1million quintals, which amounted about 14.36% of the grain production (CSA, 2008). Barley, another major subsistence crop, grown mostly between 2,000 and 3,500 meters is one of the most important staple food crops in the highlands of Ethiopia. Currently, barley grain is used for the preparation of different foodstuffs, such as 'injera', porridge, 'kolo', and local drinks, such as 'tela', 'borde', and beer. Barley is cropped twice a year. The major barley producing regions are Oromiya, Amhara, Tigray, and SNNP.

Warm Weather Cereal Crops: Corn, Sorghum and Millet

Common warm weather cereal crops in Ethiopia are corn, sorghum, and millet, where they are cultivated mostly at lower altitudes along the country's western, southwestern, and eastern peripheries. These three grains are the staple foods for a large part of the population and are major items in the diet for pastoralists. Sorghum and millet are drought resistant and grow well at low elevations where rainfall is less reliable. Sorghum is particularly important in northern Ethiopia, including in the highland areas of western Tigray. Corn is grown chiefly between elevations of 1500 and 2200 meters and requires large amounts of rainfall to ensure good harvests. Corn is particularly important in southwest Ethiopia, with the Oromiya Region producing the largest amount of corn.

Pulses

Pulses are the second most important element in the national diet, providing principal protein source and important dietary supplement to cereal consumption. Though pluses are widely grown in the highlands, they are more common in northern Ethiopia. Pulses recently have regained significance as export commodities.

Cash Crops

Coffee and chat are Ethiopia's major cash crops, with coffee cultivation in direct competition with chat, the second major agricultural export. Domestically, chat is a major source of revenue in the southeastern areas of Ethiopia. Coffee grows best at altitudes between 1000 and 2000 meters and it grows wild in many parts of Ethiopia, although most Ethiopian coffee is produced in the southern and western regions of SNNP, Oromia and Harrari.

1.3: Statement of the Problem

The sound performance of agriculture warrants the availability of food crops. The principal role that agriculture plays in Ethiopia's political, economic and social stability makes measures of agricultural productions extremely sensitive. Agriculture in Ethiopia is characterized by its low productivity. The reason for this is the use of limited modern agricultural techniques and traditional practices as well as the declining soil fertility due to continuous cropping. Among traditional practices used to increase the crop productivity, the most widely used practice has been and still maintaining is soil fertility through long fallow periods and the use of dung and crop residue. These gradually become impossible due to the prevailing rapid and uncontrolled population growth, which led to the reduction of the fallow lands and fuel wood deficit in the country. The other practice to increase crop production was based on expanding cultivable cropland. However, this scheme has been in practice for a long time and as a result of high population growth in the country most of the highlands suitable for cropping have already been exhausted.

Therefore, the only realistic option to raise the living standards of the rural population, to ensure food security and poverty alleviation is to focus on methods of increasing productivity of land and other resources while conserving those which are over-utilized.

Farmers in Ethiopia are faced with key decisions on how best to produce crops and how much to produce, given their limited resources. The problems of low crop production include unavailability of enough crop land, the use of traditional agricultural technology (such as inappropriate application of chemicals and fertilizer amount) and poor distribution of other agricultural inputs. In connection with this, a study conducted by Addis et.al. (2001) revealed that the main crop production problems in some selected woredas of the central highlands of Ethiopia were the lack of land, shortage of family household labor, high price of inputs, lack of loans from formal and informal sources, poor access to markets, shortage of appropriate storage facilities, and lack of extension services.

In view of the above issues, it is extremely imperative to deal with optimal production of crops. In general, several studies in the literature have employed production function analyses for the sake of handling problems of optimal production. A production function relates a single output y to a series of factors of production $x_1, x_2, \ldots, x_n$. In particular, crop production function relates the amount of crop yield per household to factors of production such as area of crop land, labour force participation, amount of fertilizer employed, amount of seed applied and amount of water applied.

In agricultural production function analysis, the marginal product forthcoming from a decision to increase or decrease a factor level depends on the available quantities of the other factors; e.g., the additional product yielded by an additional unit of fertilizer applied depends greatly upon the quantities of land, labor etc., combined with it. That is, one expects production inputs to be technically interdependent; specifically, for normal inputs it is expected that an increase in an input level increases the marginal and average productivities of other inputs in the production process. In such cases, regression parameter estimation should address the problem of collinearity so that better trustworthy estimates of the parameters will be obtained.

Exceptionally low or high crop yields are ordinary in many crop production schemes. Additionally, there may happen to observe extreme values from among factors of crop production such as crop area, amount of fertilizer applied and number of plowing oxen in any farming system. The impact of exceptional values of such observations (outliers) is that the classical methods of obtaining estimates of parameters like least squares fitting criteria can produce misleading results. The problem of outliers has been treated by using robust regression techniques (e.g. Finger and Hediger, 2007). Thus, robust regression method is applied as a means of addressing the problem of outliers in this thesis work so that the parameter estimates, which are obtained from a given production function, will no longer be misleading. Moreover, the adequate representation of production or crop yield functions is crucial for modeling purposes in agricultural and environmental economic analyses. To this end, the discussion and estimation of different functional forms of production function has gained much attention in agronomic and agricultural economics literatures.

Though there are few studies, which have been conducted on agricultural production and efficiency of farmers in some parts of Ethiopia (e.g., Pender and Gebremedhin (2007); Addis et.al.(2001) and Yohannes and Coffin (1993)), much attention has not been given to the estimation of crop production function in Ethiopia with the applications of statistical techniques such as robust regression. Therefore, this study attempts to show the application of robust regression on production function analyses for private peasant holdings crop farming system in Ethiopia, focusing on the most important factors of production affecting crop yield, such as crop area in hectares, labour, fertilizer, and seed.

1.4: Objectives of the Study

The objectives of conducting this research are outlined below.

- To apply production function analysis for private peasant holdings crop farms in Ethiopia using robust regression.
- To fit different crop production functions and choose the appropriate one for private peasant holdings crop farms based on data from the 2007/08 (2000 EC) agricultural sample survey in Ethiopia.
- To analyze factors of crop production at different regions of private peasant holdings in Ethiopia.

1.5: The Study Area

One of the most crucial aspects of any functional analysis is the selection of study areas. In general, the production function relationship for the entire region under consideration must refer to factors, which are homogenous so that reliable estimates will be obtained (Frank and Beattie, 1979 citing Heady, 1978). Therefore, especially in cross-sectional analysis, great care should be given to the selection of an area which is relatively homogenous, but at the same time has sufficient variation in the explanatory variables to yield accurate statistical results.

In this study, a relative homogeneity will be achieved by considering regions, which comprise relatively uniform crop mixes, and somewhat similar topographies and climatic features. A combination of these characteristics should yield a study area, which employs essentially the same technology for the production of a relatively homogeneous mix. With the help of the preceding criteria, four major crop producing regions in Ethiopia, namely Tigray, Amhara, Oromia and SNNP are identified for further investigations.

1.6: Applications of the Results

The results derived from the analyses will assist agricultural economic advisors and crop production extension agents to advise farmers on which factors of production should they spend maximum effort and money so that production can be maximized. The results may also provide private peasant crop producers and policy-makers with a better insight into the optimal allocation of scarce farming resources.

The results of this study can also be a starting point for further studies in the area of crop production technology with the application of advanced statistical techniques.

1.7: Limitations of the Study

A number of limitations in this research need to be acknowledged, these are:

1. Accurate data with the appropriate unit of measurement and without measurement error is needed in order to get a reliable result from any empirical analysis. However, in view of the fact that the data used here are secondary any measurement errors associated with the data are beyond the scope of this study. In connection with this, the study has limitations

on some of the explanatory variables, e.g., human labour and oxen labour are usually measured in terms of man hours/ days and oxen hours/ days spend in agricultural production, but in this study we took the size of the household as human labour and the number of plowing oxen as oxen labour because in the survey man hours/ days and oxen hours/ days were not incorporated.

2. Owing to the lack of detailed data on amount of rainfall for the year 2007/08 (2000 EC) at each of the study site, it was not possible to include this variable in our regression analysis although it might have had a powerful effect on crop productivity.

3. The study was limited to deal with resource utilization from a profit maximization perspective for the reason that data on unit price of input factors were not available.

CHAPTER TWO

2. LITERATURE REVIEW

Several studies have been conducted to empirically estimate the parameters of different functional forms of production functions, such as the transcendental logarithmic (translog) production function, stochastic frontier production function and Cobb-Douglas production function using econometric analysis across a range of production industries. Some of the studies revealed that the estimations of these production functions present unexpected signs as well as non-significance of parameters due to a serious problem of correlations among the explanatory variables. In order to circumvent the multicollinearity problem, which would have resulted from the ordinary least squares (OLS) estimation of the functions, few of them (e.g. Olarinde et al., 2007; Ghebremariam, et al, 2006; Lawrence and Marsh, 1984; Madariaga and McConnell, 1984; Frank and Beattie, 1979) have used ridge regression methods.

The other potential problem, which exists in the estimation and analysis of crop production functions, but not given much consideration in the economic literature, is the occurrence of exceptionally low or high values of observations both in the response and factor variables. Such phenomenon has led researchers such as Finger and Hediger (2007) to employ robust estimation techniques.

Pender and Gebremedhin (2007) preferred to make use of a logarithmic Cobb–Douglas specification than a more general functional form, such as a translog production function due to multicollinearity problems on their recent study. Their finding using OLS showed that the amount of seed and oxen power used had relatively large and statistically significant positive impacts in the production function. In contrast, the impact of human labour was quantitatively small and statistically insignificant. In addition, population pressure, farm sizes and household sizes had small and statistically insignificant impacts on crop production per hectare in their reduced form yield regressions. Several land investments and land management practices had large and statistically significant influences on the value of crop production. According to Pender and Gebremedhin (2007), most income strategies had an insignificant impact on crop production. One exception was households dependent on food aid or other assistance, whose yields were surprisingly significantly higher than those of other households. Other factors being equal, the effect of irrigation on the value of crop production was not statistically significant. However, irrigation increased crop production indirectly by increasing the use of inputs, including labor, oxen, fertilizer, and improved seeds. Use of credit (formal or informal) was not associated with significant increases in crop production. Contact with the agricultural extension program also had insignificant impact on crop production.

Bakhsh et al. (2006) adopted stochastic frontier production function method in their study to estimate technical efficiency and determinants of potato production in Pakistan. The researchers preferred to apply Cobb-Douglas stochastic frontier production function than the translog functional form for the reason that the latter contains serious issues of multicollinearity. The estimates of the parameters of the maximum likelihood (MLE) methods show that the coefficients of land preparation, seed, plant protection measures and labour were statistically different from zero. On the other hand, only two variables, i.e., irrigation and labour had statistically significant effect when parameters were estimated using ordinary least squares (OLS) method. Though irrigation variable was not significantly different from zero in the MLE method, Bakhsh et al. (2006) claimed that the MLE model was well representative of data set for the potato growers as compared to OLS method. Their study revealed that production elasticity for labour variable was very high compared to other variables used in potato production. However, the unexpected elasticity estimate (negative sign) was found for the variable plant protection measure.

Agricultural production was assumed to be the function of water availability, agricultural labour force, cropped area, and agricultural credit in a study carried out by Iqbal et.al (2003). A time series data of 30 years was considered in order to estimate a Cobb-Douglas production function using the ordinary least squares (OLS) method. Estimation of the production function using original variables showed moderate to strong multicollinearity among the independent variables. Thus Iqbal et al (2003) transformed the dependent and all the explanatory variables to per cultivated hectare as a means of avoiding the problem of multicollinearity. However the estimated transformed equation showed the presence of autocorrelation, an AR (1) and MA (1) processes. Therefore, the regression equation was re-estimated by adjusting for AR (1) and MA (1). The final estimates of the equation showed that the coefficient for agricultural credit, water availability and labour had positive and statistically significant effect on agricultural production at the 5 percent level. Similarly, the coefficient for the cropping intensity variable had positive and is statistically significant effect on agricultural production at 10 percent level.

A Cobb-Douglas production function was used to assess the extent in which productivity was affected by various farm inputs in the study by Addis et.al. (2001). Their study revealed that family labor, farm size, number of livestock and amount of inorganic fertilizer used had a significant and positive impact on the gross value of farm output whereas, farmer's age had a significant and negative impact in the male headed households (MHHs). In the female headed households (FHHs), family labor, farm size, number of livestock, the use of hired labor for agricultural production and inorganic fertilizer had a positive and significant impact on the gross value of farm output; however, extension contact had a negative and significant impact. Above all, herbicide, insecticide and education of head had no statistically significant impact for both MHHs and FHHs. The study also showed that MHHs could increase productivity by using more labor and fertilizer, while FHHs could do so by using more land and fertilizer.

Qunaibet et.al (1995) estimated a Cobb-Douglas type production function with the assumption that the functional form for milk production is homogenous using a cross sectional data from 25 dairy farms of Saudi Arabia. In the process of estimating the production function, the inputs (total milk production, number of dairy cows, metabolic energy, operating cost and labor cost) were positively correlated with highest correlation being 0.93. As a result, the researchers used a modified farm milk production function by considering the proportions of the farm inputs required to each production unit of X_p, i.e., X_1/X_p, X_2/X_p, $\ldots X_{p-1}/X_p$, 1. Their finding showed that farm productivity, or milk production per cow, is monotonically increasing with dairy farm size.

Yohannes and Coffin (1993) conducted a study using a Cobb-Douglas function on crop and milk production structure of smallholders in Ethiopia. Since there were differences in the impacts of inputs on outputs and that the structure of crops and milk production across regions, the researchers proceeded the analysis of production decisions by region and farmer group. The study showed that physical factors such as land, labour, oxen and seed rate exert positive and significant impact on the amount of crop output obtained. Management related variables such as schooling and crop production knowledge exerted positive and significant effect on production of most crops while the effect of extension education was not significant. Black soil exerted a positive and significant effect only on teff and peas cultivated by Ada farmers. Moreover, the intra-region analysis of crop production revealed that plot size and oxen labour had statistically significant and positive effect on the level of production of most crops. Fertilizer exerted a positive and significant effect on crops cultivated by Ada farmers as do pesticides on crops cultivated by Selale farmers. Production knowledge and schooling exerted a positive and significant effect on all crops.

Being aware that corn yields are driven by numerous factors, Finger and Hediger (2007) focused on two crucial production factors: nitrogen fertilizer and irrigation water for the production function comparison of Swiss corn. Three types of crop production functions were analyzed in their study: two polynomial specifications (the quadratic and the square root function) and the Mitscherlich-Baule function. The quadratic form consists of an additive composition of the input factors, their squared values, and an additional interaction term. The square root function consists of an additive composition of the input factors, their square root values, and an additional square root of interaction term. They found that input levels were uncorrelated with other variables that also influence corn yields using simulated corn yield data that was generated with the CropSyst model. Because Finger and Hediger (2007) observed exceptional crop yield levels in the data set, they applied reweighted least squares (RLS) regression in the estimation of parameters for the quadratic and square root functions. Accordingly, each estimated coefficient had the expected sign and was statistically significant except the coefficient for Applied Nitrogen * Irrigation Water. Mitscherlich-Baule production function reached higher goodness of fit than the respective estimates of the quadratic and square root forms. However, upon applying optimal input levels and costs of misspecification, the authors found that the square root function was the most appropriate form to represent the data generated with corn yield simulations for Switzerland.

Chunkwuji, et al (2006) employed panel data of three batch broiler production cycles in order to describe the production technology. Linear, semi log and double log functional forms of the production function were estimated. Though all models were significant ($P < 0.05$) and the coefficients had the expected signs except for variable and operating expense, the linear functional form was adopted as the best fit to the data set used for the study on the basis of R^2. The estimates indicated that while stock size and expenses of feed were positively related to output and revenue of the farms, variable and operating costs were negatively related to them. The results showed that the farmers were efficient in the allocation of their resources except in the case of fixed capital items. However, the farmers appear to be underutilizing their resources.

Yilmaz and Ozkan (2004) studied the effect of land tenure systems on cotton production in Turkey by using production functions. The models used in the study to estimate production functions were: Linear, Cobb-Douglas, Semi Log and Exponential. All of these models were estimated using ordinary least square (OLS). The OLS estimation revealed that the number of statistically significant parameters varied among the functional forms used in the study. However, the F-test results showed that overall regression models were statistically significant. Adjusted R^2 was higher in the linear and Cobb-Douglas production functions than others. Moreover, their finding confirmed that all estimated equations had multicollinearity problems. Though the estimated regressions were not suitable to analyze the structure of production because of multicollinearity problem, the researchers claimed that linear model seems to yield better results in term of R^2, number of statistically significant parameters and expected signs of parameters. From among the explanatory variables i.e. land tenure type, parcel size, labour, tractor, seed, nitrogen and chemicals considered in the study, the variables parcel size, labour and nitrogen variables were statistically significant in the linear model. Only two variables, parcel size and labour, were found statistically significant and have expected signs in the Cobb-Douglas functional form. However, semi log and exponential forms produced poor results in terms of significance level of the parameters.

Mbanasor and Obioha (2003) studied resource productivity under Fadamas cropping system in Umuahia north local government area of Abia state, Nigeria by estimating three functional forms of a production function: linear, Cobb-Douglas and semi-log. The Cobb Douglas form provided the best fit among the rest due to high value of R^2, the number of regression coefficients that were significant and the significant level of the F-ratio. Accordingly, the variables labour, farm size and value of other inputs were significant while value of fixed assets and value of fertilizer used were not. From among the significant variables, labour and farm size had a positive relationship with the value of outputs. It was also indicated that the value of other inputs used had negative influence on the value of output. The result further confirmed that farm size and labour use were more productive than other resources. The relationship between the resources (labour and value of other inputs) and output was inelastic except that of the farm size which was at the maximum technical efficiency i.e elasticity greater than unity. Also the coefficient of returns to scale implied that the farmers were operating at the region of maximum technical efficiency, an irrational region of production.

A recent study was conducted by Olarinde et al. (2007) to quantitatively determine the individual risk attitudes of maize farmers in the dry savanna zone of Nigeria. The researchers considered two functional forms: the exponential and power functions for the analysis of their data. After transforming these functions into their linear forms using logarithms, Olarinde et al. (2007) found that the double logarithm (log-linear) function had the best fit with $R^2 = 0.8117$. Despite the high value of R^2, some of the parameters were poorly estimated due to multicollinearity, which existed between two variables (fertilizer-NPK and fertilizer-Urea). Consequently, a ridge regression analysis with ridge parameter = 0.1 was performed to re-estimate the parameters of the double logarithmic production function. The results of ridge regression showed that the variables: quantity of seed planted in kg/ha, fertilizer (NPK) in kg/ha, labour utilization in labour-day/ha, herbicide in litre/ha and tractor in hour/ha had statistically significant effects in maize yield. But the variables fertilizer (Urea), insecticide in litre/ha and animal (traction) in hour/ha were found statistically insignificant at the standard level of significance ($P < 0.05$). After all, the variable quantity of seed planted in kg/ha was found to be the most important input affecting maize yield.

Ghebremariam, et al. (2006)conducted a production function analysis of commercial dairy farms in the Highlands of Eritrea by using a sample of 120 respondents at a dairy farm level data. The correlation matrix of explanatory variables depicted high correlations among some factors of production, i.e., annual purchased concentrates, annual purchased forage, annual labour cost and number of milking cows. Moreover, some of the estimated parameters had negative production coefficients where a priori all such coefficients were assumed to be non-negative. Consequently, ordinary least squares (OLS) estimates were found to be very unstable. However, upon applying the technique of ridge regression, the parameter estimates were found to be more precise.

Ghebremariam, et al. (2006) basically, made an attempt to pool the data of the three study areas (central zone, Mendefera and Dekemhare), using dummy variables to test if in the three study areas the regressions have a common intercept and a common slope. From these analyses, the intercept and slope dummy coefficients for the pooled data were statistically different from zero.

As a result, it was not statistically appropriate to pool the data from the three regions to estimate a common function that represents the sample of dairy farmers of the highlands of Eritrea as a whole. For this reason, separate analyses were conducted for each study area. The analyses results using ridge regression portray that the variables annual purchased concentrates, annual purchased forage, annual labour cost and number of milking cows were statistically significant for all of the three study areas. In addition, the variable annual operating and mechanical cost was significant in Mendefera and Dekemhare. However, the variable annual veterinary and medicine cost was statistically significant only in Dekemhare. Their study also showed that most of the resources are under-utilized, with the exception of concentrates, which are over-utilized by dairy farmer respondents in Mendefera and Dekemhare.

The study by Lopez (1997) on the structure of production of the Spanish telecommunications sector has provided empirical measures of the returns to scale coefficient, substitution elasticities between capital and labor, their marginal productivities, and the profitability of inputs in the Spanish telecommunications industry. Both trans-log and Cobb-Douglas production functions were estimated utilizing time series data. Estimation of the translog production function presented a serious problem of multicollinearity among the variables. As a result, ridge regression method was employed to estimate the parameters of translog production function. Both specifications, the translog and the Cobb Douglas model, yielded similar results for the Spanish telecommunications sector. In addition, the estimates of parameters (capital, labour and proxy for technological change) were significant in both cases. The values of the marginal productivities of the Cobb Douglas specification for capital and labor were slightly lower than those of the translog function, but in both cases they were steadily increasing throughout the period under study. However, F-test statistic did not provide a strong support for the translog specification when it was tested against the Cobb Douglas function.

As an estimation of the economic value of irrigation water for five regions of the Middle Atlantic States Madariaga and McConnell (1984) employed a Cobb-Douglas production function model. The ordinary least squares (OLS), ridge regression and covariance analysis were employed separately to estimate the production function. The OLS results making use of an aggregated expenditure input variables (labour; fertilizer; seed, bulbs, plants and trees; machinery; other chemicals; and petroleum) data provided the correct sign of estimated coefficients expected a priori for all variables except for the signs of the year dummy terms. On the other hand, OLS regression on the disaggregated input variables, i.e., incorporating fertilizer expenditure, labour expenditure and machinery expenditures separately yielded the signs of estimated coefficients for the variables labour and machinery, which is different from expected a priori. Additionally, the grouped expenditure estimated coefficient was unreasonably large. As a result, Madariaga and McConnell (1984) employed ridge regression in such a way that the regression coefficients for all variables included in their study had the expected signs. According to their finding, though ridge regression greatly affected most of the estimated coefficients, the ridge estimates of the irrigation water coefficients were similar to the OLS estimates.

Frank and Beattie (1979) applied ridge regression as a viable alternative to OLS under multicollinearity in order to obtain better estimates of the underlying structural parameters and hence better description of water demand characteristics. Using 1969 Census of Agricultural data, Frank and Beattie (1979) found that the parameter estimates under OLS were contrary to a priori reasoning due to high correlations among the explanatory variables: values of livestock inventory, value of machinery inventory, feed expenditures, fertilizer and lime expenditures and irrigation water applied. One third of the estimated coefficients for the production function using OLS took on unexpected signs and the standard errors were generally high. In addition, all the coefficients in the Cobb-Douglas production function were found to be insignificant though the value of R^2 was high. On the other hand, upon applying the technique of ridge regression, all parameter estimates except for one out of 99, took on the expected sign and the standard errors were decreased in every sense. Overall, the ill-effects of multicollinearity appear to have been substantially lessened.

The idea of combing robust regression and ridge regression to deal simultaneously with outliers and multicollinearity has been discussed by Lawrence and Marsh (1984). The initial ordinary least squares estimation based on a generalized Cobb-Douglas production function showed that the overall model was highly adequate. However, only one of the estimated regression coefficients had a t statistic that was significant at any reasonable level of significance. In addition, four of the six variables had high corresponding variance inflation factors, which are viewed as an indicator of harmful multicollinearity. The other problem they faced in the study was the presence of outliers in the data. For providing credible estimates of the coal mining fatality model, Lawrence and Marsh (1984) found the Lawless and Wang (L-W) method, which consistently produced the expected signs for the regression coefficient estimates under the robust weighting schemes the best. They finally used the Huber weighting method with L-W estimator so that the estimated coefficients were found to be as expected a priori in terms of both sign and significance.

The system of seemingly unrelated regression was employed by Yao (1996) in the study of the determinants of cereal productivity of the peasant farm sector in Ethiopia, from 1981 - 1987 in order to capture the relationship in between different inputs of a particular crop: labour, fertilizer and rainfall. He also made experiments to estimate the CES and translog functions, but since these did not generate any better results, he presented only the classical Cobb-Douglas production function to estimate the elasticities of inputs with respect to crop production. Yao (1996) estimated the same production function for each of the major crops: teff, wheat, maize, barley and sorghum. Yao (1996) concluded that crop production is mainly determined by two major traditional factors, land and labour. His results also revealed that the contributions of fertilizer and rainfall had a significant and sizeable effect on food production. He also reported that a dramatic decline in food production is mainly due to reduction of yield rather than area. Lastly, for all the crops, the elasticities of returns to scale were almost equal to 1. This suggests that output can be expected to increase in proportion with the increase of all inputs in a normal year, i.e. if the effect of rainfall is ignored.

Despite such immense works in the estimation and analysis of production functions, little has been done in the Ethiopian context, i.e., only few of them e.g., Pender and Gebremedhin (2007); Addis et.al. (2001); Yohannes and Coffin (1993) and Yao (1969) among others have employed Cobb-Douglas production function. Unlike other research works considered in this chapter, the researches in Ethiopia did not give much attention to discuss and tackle the occurrence of extreme values of observations using statistically recommended techniques such as robust regression. Consequently, this study attempts to fill this gap by focusing on statistical methods of ameliorating problems related to the occurrence of extreme values of observations, which are common to happen in the agricultural productions.

CHAPTER THREE

3. DATA AND METHODOLOGY

3.1: The Data

The analysis and estimation of crop production function in this study employed data from the 2007/08 (2000 EC) agricultural sample survey ('Meher' season) conducted by Central Statistical Agency of Ethiopia. The agricultural sample survey covered the entire rural parts of the country except the non-sedentary populations of three zones of Afar and six zones of Somali regions. However, this research considered only the four major crop producing regions namely: Tigray, Amhara, Oromia and SNNP.

3.2: The Variables in the Study

For the sake of obtaining meaningful results, a careful selection of relevant variables is the central concern in any study. The omission of any relevant variable will result in a model which is biased in an economic sense. Some variables, although relevant are impossible (e.g. rainfall in this study) or too costly to be included. However, the economic bias can be lessened by specifying an appropriate model. Accordingly, the lists of variables that were considered in the crop production function analyses are the following.

The Dependent Variable:

Total crop production, denoted by Y is defined as the total amount of crop yield in quintals per private peasant holding.

Independent Variables:

The following independent variables were hypothesized to influence the crop yield in each of the study regions either positively (+), negatively (-), or positively and/or negatively (+/-).

- X_1: Amount of chemical fertilizer employed in kilogram (+).
- X_2: Weight of improved and/or non-improved seed in kilogram (+).
- X_3: Area of agricultural land in hectares (+).

- X_4: Human labour, which is calculated based on the household size of the farmer (+).
- X_5: Oxen labour: it is defined as the total number of plowing oxen a household own (+).
- X_6: Education (highest grade) attained by the head of the household (+).
- X_7: A dummy variable scoring 1 for respondents having extension contact and 0 otherwise (+).
- X_8: A dummy variable scoring 1 for damaged crops and 0 otherwise (-).
- X_9: A dummy variable scoring 1 for irrigated crop field and 0 otherwise (+).
- X_{10}: A dummy variable scoring 1 if the crop land ownership type is private and 0 if the crop land ownership type is Rent/Leased (+/ -).

3.3: The Methodology

3.3.1: The Production Functions

A production function is a function that specifies the output of a firm, an industry, or an entire economy for all combinations of inputs. Given the set of all technically feasible combination of output and inputs, only the combinations encompassing a maximum output for a specified set of inputs would constitute the production function. Alternatively, a production function can be defined as the specification of the minimum input requirements needed to produce designated quantities of output, given available technology. It is usually presumed that unique production functions can be constructed for every production technology.

The relationship of output to inputs is non-monetary, that is, a production function relates physical inputs to physical outputs, and prices and costs are not considered. But, the production function is not a full model of the production process: it deliberately abstracts away from essential and inherent aspects of physical production processes, including error and waste. The primary purpose of production function is to address allocative efficiency in the use of factor inputs in production and the resulting distribution of income to those factors. Under certain assumptions, the production function can be used to derive a marginal product for each factor, which implies an ideal division of the income generated from output into an income due to each input factor of production. [2]

An appropriate production function (model) should exhibit several technical characteristics generally believed true of production processes. On an a priori basis, agricultural production function can be hypothesized to exhibit three essential characteristics. First, like most production processes, inputs in an agricultural process will likely follow the law of diminishing marginal productivity, i.e., as successive units of a variable input are applied to a given quantity of other resources, the resultant increments to output (marginal product) will decline. Secondly, the marginal product forthcoming from a decision to increase or decrease a factor level depends on the available quantities of the other factors. Finally, if one of the requisites for production is absent, i.e., a zero input level, the process would yield no output.

The technical characteristics of a production function, which are stated above, may not be strictly valid for any functional form considered in a given study; because model specifications having desirable technical properties are sometimes accompanied by statistical problems or simply not supported by the data. As a result, other functional forms, though they do not satisfy the stated technical characteristics should be considered for further analysis in view of statistical and mathematical desirability. Relaxing one or more of the above characteristics and based on a review of traditional and popular literatures, Griffin, et al. (1987) identified twenty functional forms including linear, quadratic, square root, translog and Cobb-Douglas. The models considered in the estimation of production function in this research work, are: linear, Cobb-Douglas and exponential.

A production function can be expressed in the implicit form as:

$$Y_i = f(X_{1i}, X_{2i}, X_{3i}, \ldots, X_{Ki}) + \varepsilon_i \quad i = 1, 2, 3, \ldots, n \qquad \textbf{(3.1)}$$

where,

- Y_i is the i^{th} output;
- $X_{1i}, X_{2i}, \ldots, X_{Ki}$ are K explanatory variables (inputs);

[2] Wikipedia, the free encyclopedia, Wikimedia Foundation Inc. encyclopedia on-line, available from: http://en.wikipedia.org/wiki/Production_function, Internet accessed on 13 November 2008.

- ε_i is the i^{th} error term
- n is the number of observations (cases) included in the study.
- f(.) is a known function of explanatory variables.

The different inputs for crop production function can be labor, land, seed, fertilizer, chemicals, and tractor (plowing oxen).

3.3.1.1: The Linear Production Function

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_K X_{Ki} + \varepsilon_i \quad \text{(3.2)}$$

where,

- Y is the output or response (dependent) variable.
- $X_1, \ldots, X_K$ are the K factors of production (inputs).
- n = the number of cases considered in the estimation of production function.
- i = 1, 2, 3, …, n
- β_j's are the parameters to be estimated.
- ε_i = The random disturbance term and is distributed independent-normal with mean zero and constant variance σ^2.

Properties:

1. The input value(s) for one or all explanatory variables, X's is zero, does not imply that the output value, Y is zero.
2. The first order partial derivative of Y with respect to each X_K is unrestricted in sign but has a constant value.

   i.e., $\frac{\partial Y}{\partial X_K}$ can be zero, positive or negative.

3. The linear functional form has no asymptotic convergence.

4. If $\beta_0 = 0$, then the function is linearly homogeneous.

5. The function is linearly separable even if any X_i is zero.

3.3.1.2: The Cobb-Douglas Production Function

The Cobb-Douglas functional form given hereunder is frequently used in the literature and proved to accurately capture the underlying relationship (e.g., Ghebremariam, et. al (2006), Pender and Gebremedhin (2007), Addis et.al (2001) and Qunaibet, et.al (1995)).

Model: $$Y_i = \beta_0 X_{1i}^{\beta_1} X_{2i}^{\beta_2} \ldots X_{Ki}^{\beta_K} \exp(\varepsilon_i) \quad \text{.......... (3.3)}$$

where,

- $X_1, \ldots, X_K$, Y, i and ε are as defined above.

The model can be changed into linear form as follows so that the parameters will be estimated using the method of estimation for linear regression.

$$\ln Y_i = \ln\beta_0 + \beta_1 \ln X_{1i} + \beta_2 \ln X_{2i} + \ldots + \beta_K \ln X_{Ki} + \varepsilon_i \quad \text{.......... (3.4)}$$

Properties:

1. When $X_i = 0$, $Y = 0$, which satisfies the condition of zero product intercept.

2. The function is not asymptotically convergent.

3. The first order partial derivative of Y with respect to each X_j, i.e., the equation for marginal productivity is unrestricted but non-switching in sign.

i.e., $\dfrac{\partial Y}{\partial X_j} = \beta_j \beta_0 X_j^{(\beta_j - 1)} \prod_{\substack{i=1 \\ i \neq j}}^{K} X_i^{\beta_i}$

- Decreasing marginal productivity is exhibited for the case of $\beta_0 > 0,\ 0 < \beta_j < 1$ and $0 < \beta_i < 1$.

- $\dfrac{\partial Y}{\partial X_j}$ increases without bound as X_j approaches zero while $\dfrac{\partial Y}{\partial X_j}$ approaches to zero when X_j tends to infinity for $\beta_0 > 0,\ 0 < \beta_j < 1$ and $0 < \beta_i < 1$.

4. The sign of the derivative of the marginal productivity of X_j with respect to X_j is negative,

   i.e., $\dfrac{\partial^2 Y}{\partial X_j^2} = \beta_j(\beta_j - 1)\beta_0 X_j^{(\beta_j - 2)} \prod_{\substack{i=1 \\ i \neq j}}^{K} X_i^{\beta_i} < 0$ for $\beta_0 > 0,\ 0 < \beta_j < 1$ and $0 < \beta_i < 1$ implying diminishing marginal returns.

5. The second order mixed derivative yields a positive value for $\beta_0 > 0,\ 0 < \beta_j < 1$ and $0 < \beta_i < 1$.

   i.e., $\dfrac{\partial^2 Y}{\partial X_m \partial X_j} = \beta_m \beta_j \beta_0 X_j^{(\beta_j - 1)} X_m^{(\beta_m - 1)} \prod_{\substack{i=1 \\ i \neq m,\, j}}^{K} X_i^{\beta_i} > 0$ which implies an increase in the marginal productivity of X_j as X_m is increased.

3.3.1.3: The Exponential Production Function

$$Y_i = \beta_0 \exp(\beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_k X_{Ki} + \varepsilon_i) \quad \text{.......... (3.5)}$$

where,

- $X_1, \ldots, X_K$, Y, i and ε are as defined above.

This model can also be transformed into linear form so that the parameters will be estimated using the method of estimation for linear regression.

$$\ln Y_i = \ln\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{Ki} + \varepsilon_i \quad \text{……………………………………… (3.6)}$$

Properties:

1. Since $X_i = 0$ does not imply $Y = 0$, the condition of zero product intercept is not satisfied.
2. The first order partial derivative of Y with respect to each X_k has either positive or negative sign (+ or -).

   i.e., $\frac{\partial Y}{\partial X_K}$ can either be positive or negative.

3.3.2: The Multiple Linear Regression Model and Statistical Assumptions

The empirical investigation of problems that result from a given data set should begin only after the model has been satisfactorily specified and further checked for statistical assumptions. Consider the following multiple linear regression model:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad \text{…………………………………………………………… (3.7)}$$

where,

- **Y** is an n x 1 vector of dependent variable;
- n is the number of cases/ subjects considered in the regression;
- **X** is an n x (K+1) matrix of independent variables;
- K is the number of explanatory variables;
- $\boldsymbol{\beta}$ is a (K+1) x 1 vector of regression coefficients and

- ε is an n x 1 vector of random disturbances.

Upon considering the above regression model, the following assumptions should be satisfied by the data set at hand, detail discussion can be found in Gujarati (1995) among others.

1. Linearity - the relationship between the predictors and the outcome variable should be linear. This can be checked by the partial regression plot of an outcome variable versus each of the predictor variable. If the plot suggests that linearity assumption is violated, the possible remedies are transformation of variables into such as log or polynomials and inclusion of the cross product terms of predictor variables. This is equivalent to approximating any continuous function by a polynomial or logarithms and use linear regression methods. A preferable method of detection is examination of residual plots, i.e., plots of the unstandardized residuals as a function of predicted values. If such a plot has no visible curvature (scattered randomly), then linearity assumption is satisfied. However, in practice this assumption can virtually never be confirmed; fortunately, multiple regression procedures are not greatly affected by minor deviations from this assumption.

2. Normality - the errors should be normally distributed. Technically normality is necessary only for the t-tests to be valid, but estimation of the coefficients only requires that the errors be identically and independently distributed. The assumption of normality can be assessed by either histograms for bell shaped distribution of residuals or normal quantile plot (Q-Q plot) of residuals for approximately straight line. The Q-Q plot is a plot of the quantiles of residuals against the theoretical quantiles if the residuals arose from a normal distribution. If the residuals come from a normal distribution, then the points on this plot should fall close to the diagonal line. A bow-shaped pattern of deviations from the diagonal indicates that the residuals have excessive skewness (i.e., they are not symmetrically distributed, with too many large errors in the same direction). An S-shaped pattern of deviations indicates that the residuals have excessive kurtosis i.e., there are either too many or too few large errors in both directions. One remedial measure for non-normality is transformation of the response variable. The other remedial approach is to allow the response variable to have non-normal distribution (e.g., Poisson, Binomial,

Gamma, etc.) depending on the data observed (e.g., count, binary, skewed continuous, etc.).

3. Homogeneity of variance (homoscedasticity) - the error variance should be constant. Often the existence of a few extreme or unusual observations (outliers) in a homoscedastic model makes the model heteroscedastic (Rana, et.al, 2008). The plot of residuals versus predicted values are important to see whether the assumption of homoscedasticity is violated or not. If the residuals have some patterns as a function of predicted values, then there will be an evidence of non constant variance (heteroscedasticity). When the plot of residuals appears to deviate substantially from normal, more formal tests for heteroscedasticity should be performed (Jason and Waters (2002)). Possible tests for this are the Goldfeld-Quandt test, the Breusch-Pagan test and the White test. However, Rana, et.al (2008) confirmed that the above tests suffer a huge setback in detecting heteroscedasticity when outliers are present. Accordingly, they proposed a modified Goldfeld-Quandt (MGQ) test with the computational steps given hereunder.
   i. Likewise the classical Goldfeld-Quandt test, order or rank the observations according to the value of X that supposed to cause heteroscedasticity, beginning with the lowest X value.
   ii. Omit central c observations, where c is specified a priori and then divide the remaining (n-c) observations into two groups each of (n-c)/2 observations.
   iii. Check for the outliers by any robust regression technique, preferably use the robust Least Trimmed of Squares, LTS (see the detail on LTS at the end of this chapter) to fit the regression line. Then compute the deleted residuals for the entire data set based on a fit without the points identified as outliers by the LTS fit.
   iv. For both the groups compute the Median of the Squared Deletion Residuals (MSDR) and compute the ratio:

      MGQ = MSDR2/ MSDR1

      Where,

      - MSDR2 and MSDR1 are the medians of the squared deletion residuals for the smaller and the larger groups respectively. Under normality, the MGQ statistic follows an F distribution with numerator and denominator degrees of freedom each of (n-c-2(K+1))/2.

- K is the number of explanatory variables in the model.

The problem of heteroscedasticity can be fixed by transforming variables, applying weighted least squares (WLS), i.e. reweighting the cases with large residuals or use of White's heteroscedastcity consistent covariance matrix. Heteroscedasticity can also be a byproduct of a significant violation of the linearity and/or independence assumptions, in which case it may also be fixed as a byproduct of fixing those problems.

4. Independence - the error associated with one observation is not correlated with the errors of any other observations (no autocorrelation). Autocorrelation is more common in time series data where the error terms of one or more consecutive periods are correlated. With cross sectional data, however, random sampling guarantees that different error terms are mutually independent, and autocorrelation is not an issue (Verbeek, 2008, pp. 105). But, when the sample is constructed in a particular (nonrandom) way correlation between different observations may arise. In some cases, for instance, if one studies the consumption patterns of households, the error terms for households in the same neighborhood can be correlated. This is because the error term picks up the effect of omitted variables and these variables tend to be correlated for households in the same neighborhood (Maddala, 1992, pp. 230). Similarly, in this research error terms of closer households within the same region may correlate and hence there is a need to test for autocorrelation. Durbin-Watson Statistic is the most popular measure used to detect such correlations. The values of Durbin-Watson Statistic, ranges from 0 to 4 with an ideal value of 2 indicating that errors are not correlated (although values from 1.75 to 2.25 may be considered acceptable). A value significantly below 2 indicates a positive correlation and a value significantly greater than 2 suggests negative correlation. One remedial measure to this problem is trying to find out if autocorrelation is pure or resulted from misspecification of the model. If the autocorrelation is pure, then use appropriate transformation and some type of linear and nonlinear mixed effects models.

Consider the following multiple linear regression model:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_K X_{Ki} + \varepsilon_i \qquad (3.8)$$

Define the Durbin-Watson, d statistic as:

$$d = \frac{\sum_{i=2}^{n} (\hat{\varepsilon}_i - \hat{\varepsilon}_{i-1})^2}{\sum_{t=1}^{n} \hat{\varepsilon}_i^2} \text{; Where } \hat{\varepsilon}_i = Y_i - \mathbf{X}_i^T \hat{\boldsymbol{\beta}}$$

$$d = \frac{\sum_{i=1}^{n} \hat{\varepsilon}_i^2 + \sum_{i=2}^{n} \hat{\varepsilon}_{i=1}^2 - 2\sum_{i=2}^{n} \hat{\varepsilon}_i \hat{\varepsilon}_{i-1}}{\sum_{i=1}^{n} \hat{\varepsilon}_i^2}$$

$$d \approx 2\left(1 - \frac{\sum_{i=2}^{n} \hat{\varepsilon}_i \hat{\varepsilon}_{i-1}}{\sum_{i=1}^{n} \hat{\varepsilon}_i^2}\right)$$

Define also the correlation between errors, $\hat{\rho}$ as: $\hat{\rho} = \dfrac{\sum_{i=2}^{n} \hat{\varepsilon}_i \hat{\varepsilon}_{i-1}}{\sum_{i=1}^{n} \hat{\varepsilon}_i^2}$

Thus, $d \approx 2(1 - \hat{\rho})$ .... **(3. 9)**

Since $-1 \le \hat{\rho} \le 1$, the d statistic is in the interval $0 \le d \le 4$

Positive autocorrelation	Test is inconclusive	No autocorrelation	Test is inconclusive	Negative autocorrelation
$0 < d < dL$	$dL < d < dU$	$dU < d < 4 - dU$	$4 - dU < d < 4 - dL$	$4 - dL < d < 4$

Where, dL and dU respectively are the lower and upper critical values obtained from the Durbin-Watson table with K (the number of independent variables in the model), n (total number of subjects) at a given level of significance α .

5. The x's are linearly independent (no multicollinearity) and hence rank ($\mathbf{X}^T\mathbf{X}$) = rank ($\mathbf{X}$) = K, which implies that $(\mathbf{X}^T\mathbf{X})^{-1}$ exists. Multicollinearity is neither a specification error that may be uncovered by exploring regression residuals nor a modeling error but is a condition of deficient data (Chatterjee and Price, 1977, pp. 143-144). The various techniques that are useful for detecting multicollinearity are the following.

- **Examination of Correlation Matrix:** A very simple measure of multicollinearity is inspection of the off-diagonal elements of the correlation coefficient, r_{ij} in $\mathbf{X}^T\mathbf{X}$. If regressors x_i and x_j are nearly linearly dependent, then $|r_{ij}|$ will be near unity.

- **Variance Inflation Factor (VIF):** VIF = $\frac{1}{1 - R_i^2}$ where R_i^2 is the squared multiple correlation coefficient between x_i and other explanatory variables. As R_i^2 tends toward 1 indicating the presence of a linear relationship in the x's, the VIF for the estimated coefficient of x_i tends to infinity. It is suggested that a VIF in excess of 10 is an indication that multicollinearity may be causing problems in estimation (Chatterjee and Price, 1977, pp. 182).

- **Eigenvalues and Condition Number (CN):** If there are one or more near-linear dependences in the data, then one or more eigenvalues of $\mathbf{X}^T\mathbf{X}$, say $\lambda_1, \lambda_2, \ldots, \lambda_k$ will be small. The CN is supposed to measure the sensitivity of the regression estimates to small changes in the data and is defined as the ratio of the largest to smallest eigenvalue of the matrix $\mathbf{X}^T\mathbf{X}$ of the explanatory variables, i.e.,

  CN = $\frac{\lambda_{max}}{\lambda_{min}}$. Generally, if the condition number is less than 100, there is no serious problem with multicollinearity. CN between 100 and 1000 imply moderate to strong multicollinearity, and if it exceeds 1000, severe multicollinearity is indicated.

There are several approaches to combat multicollinearity, these are: use a priori information (impose restrictions), combine cross sectional and time series data, creating a new independent variable by dividing one of the inter-correlated independent variables by another, drop a variable until multicollinearity problem is solved (this may however lead to specification error into the model, Crown, 1998, pp. 75), get bigger sample size and use of biased regression method like ridge regression.

In addition to the above assumptions, there are issues that can arise during the analysis that, while strictly speaking, are not assumptions of regression, are none the less, of great concern to regression analysts. These are unusual observations, which may be outliers or leverage points that exert undue influence on the coefficients. The issue of satisfying the above assumptions and detecting outlying cases are intertwined. For example, if a case has a value on the dependent variable that is an outlier, it will affect the skew and hence the normality of the distribution. Detail discussion on outliers is given in the next section.

Whenever the above assumptions are entirely satisfied, the best (minimum variance) and unbiased linear estimator of $\boldsymbol{\beta}$ in equation (3.7) is obtained by the method of ordinary least squares (OLS). The method of OLS is based on minimizing the error sum of squares (Q):

$Q = \boldsymbol{\varepsilon}^T \boldsymbol{\varepsilon} = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$, where T stands for transpose.

$$\text{i.e, } Q = \mathbf{Y}^T\mathbf{Y} - 2\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{Y} + \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} \quad \text{.......... (3. 10)}$$

Minimization of Q can be achieved by taking the partial derivative of Q with respect to each β_i and equating to zero,

$$\text{i.e, } \frac{\partial Q}{\partial \boldsymbol{\beta}} = \mathbf{0} \text{ gives } \hat{\boldsymbol{\beta}}_{OLS} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \quad \text{.......... (3. 11)}$$

$$\hat{\boldsymbol{\beta}}_{OLS} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}), \text{ since } \mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

$$\hat{\boldsymbol{\beta}}_{OLS} = \boldsymbol{\beta} + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\varepsilon}$$

The above implies that $\hat{\boldsymbol{\beta}}_{OLS} - \boldsymbol{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\varepsilon}$

The expectation of $\hat{\boldsymbol{\beta}}_{OLS} = E(\hat{\boldsymbol{\beta}}_{OLS}) = \boldsymbol{\beta}$, i.e., $\hat{\boldsymbol{\beta}}_{OLS}$ is unbiased since $E(\boldsymbol{\varepsilon}) = \mathbf{0}$

The variance-covariance matrix of $\hat{\boldsymbol{\beta}}_{OLS}$ is given by:

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}_{OLS}) = E[(\hat{\boldsymbol{\beta}}_{OLS} - E(\hat{\boldsymbol{\beta}}_{OLS}))(\hat{\boldsymbol{\beta}}_{OLS} - E(\hat{\boldsymbol{\beta}}_{OLS}))^T]$$

$$= E[(\hat{\boldsymbol{\beta}}_{OLS} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}}_{OLS} - \boldsymbol{\beta})^T]$$

$$= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T)\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}$$

$$= (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2, \text{ since } E(\boldsymbol{\varepsilon}\,\boldsymbol{\varepsilon}^T) = \sigma^2\mathbf{I}. \quad \ldots\ldots (3.\ 12)$$

where σ^2 is estimated by the mean square error, $\hat{\sigma}^2$ which is given by:

$$\hat{\sigma}^2 = \frac{\hat{\boldsymbol{\varepsilon}}^T\hat{\boldsymbol{\varepsilon}}}{n - (K + 1)} \quad \ldots\ldots (3.\ 13)$$

where, $\hat{\boldsymbol{\varepsilon}} = \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{OLS}$ and n, K, T and $\hat{\boldsymbol{\beta}}_{OLS}$ are as defined above.

However, we are expecting that the crop production data in this research is vulnerable to unusual observations in such a way that the parameter estimates of OLS regression are misleading. As a result, a special attention is given to deal with regression analysis in the presence of unusual observations.

3.3.3: Unusual Observations and Robust Regression

3.3.3.1: Outliers

Outliers are observations that are numerically distant from the rest of the data. An observation with an extreme value on a predictor variable is a point with high leverage. Leverage is a measure of how far an independent variable deviates from its mean. If such leverage point deviates from the linear relationship described by the majority of observations it is called 'bad leverage point'. In contrast, a leverage point is called 'good leverage point' if it does not deviate from the typical relationship. Good leverage points are not outliers and even improve the regression inference as these points reduce standard errors of coefficient estimates. Outliers can occur by chance in any distribution, but they are often indicative either of measurement error or that the population has a heavy tailed distribution.

A frequent cause of outliers is a mixture of two distributions, which may be two distinct sub-populations, or may indicate "correct trial" versus "measurement error" this is modeled by a mixture model. Outliers, being the most extreme observations, will include the sample maximum or sample minimum or both, depending on whether they are extremely high or low. However, the sample maximum and minimum need not be outliers, unless they are unusually far from other observations. Moreover, exceptionally low or high crop yields as well as extreme input values are ordinary in many crop production schemes. We particularly expect that production inputs are influential. For instance, the level of education can influence the appropriate use of modern agricultural practices and thus indirectly increases yield level. The response to plowing oxen and human labor can also change under greater and smaller number of cases. The impact of exceptional observation values (outliers) is that the least squares estimation is inefficient and can be biased.

Automatic rejection of outliers is not always a very wise procedure. Sometimes the outlier is providing information which other data points cannot due to the fact that it arises from an unusual combination of circumstances which may be of vital interest and requires further investigation rather than rejection. As a general rule, outliers should be rejected out of hand only if they can be traced to causes such as errors in recording the observations or in setting up apparatus (Draper and Smith, 1966, pp. 76). The problem of outliers has been treated by using robust regression techniques (e.g. Finger and Hediger, 2007). Thus, robust regression method is applied as a means of addressing the problem of outliers/ leverage points in this thesis work so that the parameter estimates will no longer be vulnerable as least squares estimates to unusual data.

3.3.3.2: Robust Regression

A statistical procedure is regarded as robust if it performs reasonably well even when the assumptions of the statistical model are not true. Robust regression procedure generally refers to one that not only performs well if the population of errors is normally distributed but also is insensitive to small departures from the normality assumption. In the estimation of statistical regression models and testing the assumptions, one frequently finds that the assumptions are substantially violated. Sometimes the variables can be transformed as a means to conform the assumptions. Often, however, a transformation will not eliminate or attenuate the leverage of influential outliers that bias the prediction and distort the significance of parameter estimates. In such cases, robust regression that is resistant to the influence of outliers may be the only reasonable recourse.

3.3.3.3: Outlier Detection:

Outlier detection, one purpose of robust regression involves the determination whether the residual is an extreme negative or positive value. In the case of simple linear regression analysis, outliers can be detected using the scatter plot. However, this becomes impossible if the dimension of the problem exceeds the simple linear regression case and the number of observations is very large. Making use of residual plots as outlier diagnostic is a bad practice since residual plots might suffer from outliers (Finger and Hediger, 2007 citing Rousseeuw and Leroy, 1987), especially in the case of bad leverage points, i.e., outliers can tilt the regression line and have small regression residuals. As a result, other diagnostic tools are required to identify outlying or influential observations. In practical considerations, one often tries to detect outliers using diagnostics from a least squares procedure. Such procedures, however, are susceptible to the so called masking effect since they can be affected by extreme observations so strongly that the fitted model will fail to detect observations, which deviate from others. Outlier diagnostic procedures such as Studentized and jackknifed residuals, Cooks distances and Hat matrix elements also suffer from masking effect. Above all, in cases when two or more outliers are present, these outlier diagnostics may be able to detect only one since one outlier can be masked by other(s). To avoid this effect, robust methods of outlier detection have been employed in the literature.

3.3.3.4: Types of Robust Regression

There are many types of robust regression methods. Although they work in different ways, they all give less weight to observations that would otherwise influence the regression line. The main purpose of robust regression is to detect outliers and provide resistant (stable) results in the presence of outliers. Three classes of problems have been addressed with robust regression techniques:

- Problems with outliers in the y-direction (response direction)
- Problems with multivariate outliers in the x-space (i.e., outliers in the covariate space, which are also referred to as leverage points)
- Problems with outliers in both the y-direction and the x-space.

Methods of robust regression in response to the above problems include: least absolute values regression or least absolute deviation (LAD) regression, M-Estimation (Huber estimators and Bisquare estimators) and Bounded Influence Regression (least median of squares and least-trimmed squares).

High Breakdown Point

The robustness of an estimate against heavier data contamination is measured by its breakdown point, which is the largest proportion of outliers that can occur in a sample without entailing the possibility of arbitrarily large bias. Since one unusual observation (outlier) is enough to influence the coefficient estimates of OLS regression, the breakdown point for OLS regression is 0%. The maximum possible breakdown point is 50%. This is achieved by the least-trimmed-squares (LTS) estimate and least median of squares estimate, which is the estimate that minimizes the median of the squared residuals.

Least Absolute Deviation (LAD) Regression

LAD regression unlike least squares regression minimizes the sum of the absolute values of residuals in order to estimate the regression coefficients. The strength of LAD estimation is that its robustness to the distribution of the response variable (although not with respect to the explanatory variable). For this reason, LAD estimates are sometimes recommended as starting values for iterative estimation algorithms. LAD method is especially suitable when the distribution produces a larger proportion of outliers than normal or when the sample is very large. Though LAD regression is robust to outliers, it is typically worse than OLS for cases with high leverage. If a leverage point is very far away, the LAD regression line will pass through it i.e., its breakdown point is also 0%.

M-Estimates

The main advantage of LAD estimates over OLS is that they are not so sensitive to outliers. When there are no outliers, however, OLS estimates may be more accurate. M-estimates combine the advantages of both methods. In the case of high leverage, the performance of M-estimates falls down, meaning that the breakdown point is also 0%. M-estimation is a commonly used method for outlier detection and robust regression when contamination is mainly in the response direction.

Bounded Influence (BI) Methods

Despite M estimates are more efficient than LAD estimates; a single leverage point can completely dominate the ensuing estimate. This limitation has led to the development of estimates that bound the influence of any single element or row of X so that they guard against leverage points as well as regression outliers. Furthermore, these methods, which comprise least-trimmed-squares (LTS) estimates and least median of squares estimates have a much higher breakdown point as high as 50%. However, the efficiency of BI methods is less when the sample size under consideration is small.

Least-Trimmed Squares (LTS)

LTS regression is based on the subset of h observations (out of a total of n observations) whose least squares fit possesses the smallest sum of squared residuals. The method of LTS regression is performed as follows.

- Order the squared robust residuals from smallest to largest, i.e., $(r^2)_{(1)}, (r^2)_{(2)}, \dots, (r^2)_{(h)}$, where $r_i = Y_i - \mathbf{X}_i^T\hat{\boldsymbol{\beta}}$
- Calculate $\hat{\beta}_{LTS}$, the LTS estimates of the regression coefficients by minimizing the sum of the squared robust residuals: $\underset{\hat{\beta}}{\mathrm{Min}} \sum_{i=1}^{h} (r^2)_i$ where h is defined in the range: $\frac{n}{2} + 1 \le h \le \frac{3n + K + 1}{4}$ and by default SAS takes h as $\frac{3n + K + 1}{4}$ (SAS Institute, 2008) and K is the number of independent variables.

Above all, iteratively reweighted least square regression is applied for the estimation of production function (models) considered in this study, using the robustreg procedure of SAS9.2.

The data contamination in the crop production function is mainly in the response direction and hence the following M-estimation method can be employed. In M-estimation, the goal is to choose the regression coefficients that minimize some function of the residuals. The method of ordinary least squares has as its solution the coefficients that minimize the sum of squared residuals. This solution is undesirable when the data contain outliers, since an observation with a large error term will have a much larger effect (relative to the other observations) on the estimated coefficients.

Consider the linear regression model: $y_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_K x_{Ki} + \varepsilon_i$ .... **(3.14)**

The most commonly used M-estimates, the Huber M-estimates, $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_K$ are the values of $\beta_0, \beta_1, \ldots, \beta_K$ that minimize the function:

$\Sigma \rho(y_i - (\beta_0 + \beta_1 x_{i1} + \ldots + \beta_K x_{iK}))$ .... **(3. 15)**

$$\text{where } \rho(\hat{\varepsilon}) = \begin{cases} \hat{\varepsilon}^2 & \text{if } -c \le \hat{\varepsilon} \le c \\ 2c|\hat{\varepsilon}| - c^2 & \text{if } \hat{\varepsilon} < -c \text{ or } c < \hat{\varepsilon} \end{cases} \qquad \text{(3. 16)}$$

It is suggested to take $c = 1.5\hat{\sigma}$, where $\hat{\sigma}$ is an estimate of the standard deviation σ of the population of random errors.

In order to find the minimum of (3.15), for a fixed value of $\hat{\sigma}$, take the derivative of (3.15) with respect to $\beta_0, \beta_1, \ldots, \beta_K$ and set each of them equal to zero. This yields $K+1$ equations in $K+1$ unknowns:

$$\Sigma x_{ij} \rho(y_i - (\beta_0 + \beta_1 x_{i1} + \ldots + \beta_K x_{iK})) = 0 \qquad \text{(3. 17)}$$

for j = 0, 1, …., K, where one lets $x_{i0} = 1$ for all i. These are non-linear equations in the unknowns $\beta_0, \beta_1, \ldots, \beta_K$ but they can be approximated by linear equations as follows.

Consider an iterative procedure in which $\hat{\beta}_0^0, \hat{\beta}_1^0, \ldots, \hat{\beta}_K^0$ are current estimates and $\hat{\beta}_0^1, \hat{\beta}_1^1, \ldots, \hat{\beta}_K^1$ represent improved estimates.

Let $\hat{\varepsilon}_i^0 = y_i - (\hat{\beta}_0^0 + \hat{\beta}_1^0 x_{i1} + \ldots + \hat{\beta}_K^0 x_{iK})$ and $\hat{\varepsilon}_i^1 = y_i - (\hat{\beta}_0^1 + \hat{\beta}_1^1 x_{i1} + \ldots + \hat{\beta}_K^1 x_{iK})$.

In order to solve for the improved estimates, write $\rho'(\hat{\varepsilon}_i^1) = [\frac{\rho'(\hat{\varepsilon}_i^1)}{\hat{\varepsilon}_i^1}]\hat{\varepsilon}_i^1 \approx [\frac{\rho'(\hat{\varepsilon}_i^0)}{\hat{\varepsilon}_i^0}]\hat{\varepsilon}_i^1$.

Let further that $w_i = \frac{\rho'(\hat{\varepsilon}_i^0)}{\hat{\varepsilon}_i^0}$, that is,

$$w_i = \begin{cases} 2 & \text{if } \left|\hat{\varepsilon}_i^0\right| \le 1.5\hat{\sigma} \\ 3\hat{\sigma}/\left|\hat{\varepsilon}_i^0\right| & \text{if } \left|\hat{\varepsilon}_i^0\right| > 1.5\hat{\sigma} \end{cases}$$

Then $\rho'(\hat{\varepsilon}_i^1) \approx w_i \hat{\varepsilon}_i^1$ and one can estimate equation (3.17) by the linear equations:

$$\Sigma x_{ij} w_i [y_i - (\hat{\beta}_0^1 + \hat{\beta}_1^1 x_{i1} + \ldots + \hat{\beta}_K^1 x_{iK})] = 0 \quad \textbf{(3. 18)}$$

Let $\mathbf{W}$ be the diagonal matrix with diagonal entries w_i. Then equation (3.17) can further be expressed in terms of matrix as: $\mathbf{X}^T\mathbf{W}(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = \mathbf{0}$.

Where,

- $\hat{\boldsymbol{\beta}}$ is a K+1 by 1 vector of estimated regression coefficients;
- $\mathbf{X}$ is an n by K + 1 design matrix;
- $\mathbf{Y}$ is an n by 1 vector of response variable.

Solving the above equation for $\hat{\boldsymbol{\beta}}$, one can obtain the following weighted least squares estimator.

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{Y} \quad \textbf{(3. 19)}$$

The iterative reweighted least squares can be started by setting $\hat{\boldsymbol{\beta}}^0$ the vector of least squares estimates. At each iterative step, the vector $\hat{\boldsymbol{\beta}}^0$ of current estimates is used to calculate the vector $\hat{\boldsymbol{\varepsilon}}^0 = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}^0$ of residuals. Then use residuals to obtain $\hat{\sigma}$ and weights w_i. The vector $\hat{\boldsymbol{\beta}}$ of improved estimates can now be computed as in equation (3.19). The iterative procedure continues until convergence.

In cases when data are contaminated in the x-space (this is what we normally anticipate in this research), M estimation does not do well (SAS Institute, 2008). As a result, the coefficient estimates of the iterative reweighted least squares regressions are obtained from the method of least trimmed squares (LTS). Since the efficiency of LTS estimation is low, the estimates obtained from this method can no longer be reliable and hence LTS estimation is only used as a means of outlier detection. Consequently, the final estimates of parameters are obtained from the weighted least squares fit in which weights can be determined as follows.

Upon employing the method of iteratively reweighted least squares using LTS estimation, the diagonal elements of the weighting matrix ($\mathbf{W} = \mathrm{diag}\{w_1,...w_h\}$) are generated by an indicator function, $I_{Outlier}$:

$$w_i = I_{outlier}\left[\left|\frac{r_i}{\hat{\sigma}_{LTS}}\right| \leq 3.0\right] \quad \text{(3. 20)}$$

where,

- $\hat{\sigma}_{LTS}$ is the scale estimate which can be obtained as follows.

  $\hat{\sigma}_{LTS} = d_{h,n}\sqrt{\frac{1}{h}\sum_{i=1}^{h} r_{(i)}^2}$ and $d_{h,n}$ is chosen to make $\hat{\sigma}_{LTS}$ consistent assuming a Gaussian model. Specifically,

  $$d_{h,n} = \frac{1}{\sqrt{1 - \frac{2n}{h\,c_{h,n}}\varphi(1/c_{h,n})}} \quad \text{and} \quad c_{h,n} = \frac{1}{\Phi^{-1}\left(\frac{h+n}{2n}\right)}$$

With Φ and φ being the distribution function and the density function of standard normal distribution, respectively.

- r_i is the residual obtained from LTS estimation.
- The above indicator function generates weights of zero for observations that are identified as outliers and weights of one otherwise.

- The cutoff value 3 is chosen from the fact that if the residuals are normally distributed, then roughly 99% of the standardized residuals will lie in the interval [-3.0, 3.0]

Finally, the coefficient estimates for weighted least squares are obtained with the help of equation (3.19) and these estimates are considered further for the interpretation of the final model.

CHAPTER FOUR

4. RESULTS AND DISSCUSSION

This chapter presents the statistical data analysis with detail discussion including descriptive analysis of the characteristics of inputs and output for crop production function. The descriptive analysis here is to summarize the overall records of each of the variables incorporated in the research before a full examination of the production functions proposed in the previous chapter using robust regression methods. The statistical package used for data analysis is SAS 9.2.

4.1: Description of the Study Variables:

The inputs and output characteristics of crop production for each region considered in this study are summarized in Table 4.1 and Table 4.2 below. The average amounts of crop production by private peasant holdings of Tigray, Amhara, Oromia and SNNP regions in the surveyed year were 17.36; 22.02; 32.89 and 30.74 quintals respectively with respective mean agricultural land areas of 1.24; 1.49; 2.22 and 1.25 hectares used for production. Despite the mean crop land owned by private peasant holdings in SNNP region was less than Amhara region and nearly the same as Tigray region, the above results show that the average crop production (yield) per peasant was the highest in SNNP region from among the three. This further indicates that average productivity was higher in SNNP region compared to Amhara region with a given mean area of crop land. The above results also imply that farmers in each region owned very small areas of agricultural land in which the mean area was less than 1.50 hectares at each of the regions except Oromia and this probably brought about the small mean crop yield per peasant in general. Table 4.1 also shows that the coefficients of variation (CVs%) for crop yield in Tigray, Amhara, Oromai and SNNP regions were 64.86; 65.85; 57.09 and 77.62 respectively so that the yield data were highly variable in SNNP region while they were the most stable in Oromia region. The maximum amount of crop yield in quintals was recorded at SNNP region whereas the minimum was at region Tigray (see Table 4.1). Although the minimum area of agricultural land in hectare was almost the same at each region, the maximum being at Oromia region was exceedingly the highest as it is revealed in Table 4.1.

Table 4.1: Summary statistics for continuous predictors and the dependent variable of crop production function in each region

Region	Variable	Number of cases	Minimum	Maximum	Mean	Median	Std.[3] Deviation	CV%[4]
Tigray	Education	1676	0	12	2.29	1.00	2.28	99.56
	Family size	1676	1	11	3.4	3.00	1.87	55
	Fertilizer	1676	0.00	335	45.04	30.44	41.31	91.76
	Weight of seed	1676	2.50	425	88.07	69.79	71.67	81.38
	Area of land	1676	0.03	6	1.24	1.02	0.86	69.35
	Production	1676	0.17	34.75	17.36	9.79	11.26	64.86
	Number of oxen	1676	0	6	1.74	2.00	0.75	43.1
Amhara	Education	6157	0	12	1.81	1.00	1.93	106.63
	Family size	6157	1	12	5.21	5.00	2.05	39.35
	Fertilizer	6157	0.00	710	56.19	40.00	99.52	177.11
	Weight of seed	6157	3.00	625	110.3	72.84	121.98	110.59
	Area of land	6157	0.26	8.25	1.49	1.26	1.08	72.48
	Production	6157	1.007	45.5	22.02	14.06	14.39	65.35
	Number of oxen	6157	0	9	1.5	1.00	0.87	58
Oromia	Education	6371	0	12	2.87	1.00	3.17	110.45
	Family size	6371	1	13	6.04	6.00	2.4	39.74
	Fertilizer	6371	0.00	550	125.8	90.00	115.72	91.99
	Weight of seed	6371	1.00	600	164.6	133.00	127.18	72.27
	Area of land	6371	0.05	11.67	2.22	1.89	1.52	68.47
	Production	6371	1.14	60.5	32.89	17.71	18.78	57.09
	Number of oxen	6371	1	12	2.03	2.00	1.09	53.69
SNNP	Education	8084	0	12	2.76	1.00	3.07	111.23
	Family size	8084	1	14	4.79	4.00	2.33	48.64
	Fertilizer	8084	0.00	658	27.04	15.65	45.98	170.04
	Weight of seed	8084	1.00	681.6	45.14	25.00	60.11	133.16
	Area of land	8084	0.05	10.35	1.25	0.93	1.56	124.8
	Production	8084	0.86	66.43	30.74	16.34	23.86	77.62
	Number of oxen	8084	0	14	0.85	2.00	1.42	167.06

Peasants at each region completed an average of less than 3 years of schooling; the highest being 12 complete in all cases (Table 4.1). This indicates that farmers were relatively of low

[3] Std.: represents standard

[4] CV%: represents percentage of coefficient of variation

educational status with mean level attaining below second cycle primary education (grades 5 - 8). Table 4.1 also reveals that the mean value of weight of seed employed in Kg was the highest at Oromia region (164.64 Kg) but the lowest was observed at SNNP region (45.14 Kg). Yet the data on weight of seed employed in Kg were highly stable at Oromia region from among the entire regions since the CV% for weight of seed employed in Kg at Oromia region was the minimum (72.27). This implies that farmers in Oromia region applied relatively uniform amounts of seed as compared to farmers in the remaining regions. On the other hand, the data on weight of seed employed in Kg were highly variable at SNNP region, which further indicates that some peasants in the region applied high amounts of seed while others applied relatively small (Table 4.1).

In terms of the mean amounts of fertilizer employed in Kg that are given in Table 4.1, Oromia region took the highest (125.79 Kg) compared to the remaining regions. On the other hand, the lowest mean amount of fertilizer employed in Kg was corresponding to SNNP region (27.04 Kg). This by and large indicates that farmers applied small amounts of fertilizer in their crop production process. Furthermore, the values of coefficients of variation given at Table 4.1 reveal that the amounts of fertilizer used in Kg were greatly variable within peasants at Amhara region (CV% = 177.11) in which some of the peasants used high amount while others used small or not at all. Whereas, the data on amounts of fertilizer applied in Kg were relatively stable at Tigray region (CV% = 91.76), indicating that farmers in Tigray region used relatively similar amounts of fertilizer than farmers in the other regions.

Although the median number of individuals per household was greatest at Oromia region (6 persons), the maximum number of individuals in a household was observed at SNNP region (14 persons). Additionally, the values of coefficients of variation for number of individuals in a household show that, the data were relatively stable at Amhara region (CV% = 39.35), which indicates comparatively uniform family size per household. The maximum number of plowing oxen owned by a peasant was observed at SNNP region (14 plowing oxen) followed sequentially by Oromia (12), Amhara (9) and Tigray (6). The median number of plowing oxen owned by a peasant as indicated in Table 4.1 was small at each region in which it did not exceed two in any of the regions. This shows that number of plowing oxen, which is generally believed as being one of the most important inputs of crop farming in Ethiopia, was not fairly adequate. Moreover, the data on number of plowing oxen were greatly variable with coefficient of variation 167.06% at

SNNP region, which reveals that some of the peasants owned greater number of plowing oxen while others possessed few or zero (see Table 4.1 for detail).

Table 4.2: Summary of dummy variables included in each region production function

		Region							
		Tigray		Amhara		Oromia		SNNP	
Variable		**N**	**Percent**	**N**	**Percent**	**N**	**Percent**	**N**	**Percent**
Extension	Yes	862	51.4	2792	45.3	4068	63.9	5050	62.5
Contact	No	814	48.6	3365	54.7	2303	36.1	3034	37.5
	Total	1676	100	6157	100	6371	100	8084	100
Irrigation	Yes	717	42.8	2498	40.6	2392	37.5	3415	42.2
Applied	No	959	57.2	3659	59.4	3979	62.5	4669	57.8
	Total	1676	100	6157	100	6371	100	8084	100
Crop damage	Yes	1107	66.1	3721	60.4	3950	62	5152	63.7
	No	569	33.9	2436	39.6	2421	38	2932	36.3
	Total	1676	100	6157	100	6371	100	8084	100
Land	Private	1167	69.6	4541	73.8	4552	71.4	5556	68.7
Ownership type	Rent/ Leased	509	30.4	1616	26.2	1819	28.6	2528	31.3
	Total	1676	100	6157	100	6371	100	8084	100

Generally speaking, more than half of the peasants in each of the regions except Amhara (only 45.3%) had extension contacts (see Table 4.2). Inspection of Table 4.2 also shows that the proportion of peasants, who applied irrigation, was generally low in each of the regions from which the highest proportion being observed at Tigray region (42.8%). Although Ethiopia has a good potential for developing irrigation, the above results indicate that irrigation was not highly practiced by peasants in each region. Perhaps, this is partly because irrigation requires a long-term effort and substantial investment, which is unlikely to implement it at the private peasant level owing to shortage of technical and financial resources. The descriptive results given above also reveal that the proportions of private agricultural land ownership by farmers in each region were high as compared to rented and/or leased agricultural land. Though the mean size of crop land was generally small, the results in Table 4.2 reveal that most of the peasants owned agricultural land. In addition, the results in Table 4.2 show that crop damage was generally high in each of the regions. This probably led to minimum crop yield per peasant at each of the regions.

4.2: Associations of the Dependent and Independent Variables

The correlations between each of the predictor variables and the response variable were performed using the Pearson's correlation coefficient. This was done in order to check whether there is significant association between the dependent variable and each of the predictors prior to considering the complete production functions. The correlation matrices, which are displayed in Annex A reveal that each of the predictor variables except education had a highly significant ($p < 0.0001$) positive correlation with the response variable. In fact, the education variable at the SNNP region had a highly significant ($p < 0.0001$) positive correlation with the response variable. The education variable for Tigray region had no statistically significant correlation with the response at the 5% level of significance. However, the same variable for Amhara and Oromia regions had a statistically significant ($p < 0.01$) positive correlation with the response variable.

Though correlation coefficients greater than 0.80 were observed for Tigray region among few predictor variables, i.e. in between family size, crop area and weight of seed, the correlation matrices reveal that the associations between each pair of predictor variables were not generally high in each of the regions.

4.3: Production Function Estimation Results

Since it may be difficult to accomplish other regression techniques such as robust regression without at least implicitly involving OLS regression, the OLS estimation results are presented. Owing to the distinctness of parameter estimates (both in signs and values), the estimates of production functions for each of the regions was found to be different from region to region. As a result, the statistical model fitted for each region is presented separately.

4.3.1: Ordinary Least Squares (OLS) Estimates

The OLS estimates and related statistics for each of the regions except Oromia region production function are reported in Annex B. For the sake of providing general insights on estimation and further consequences, the OLS estimates for Oromia region crop production function are presented in this section. In all cases, the F test results confirm that the models were statistically highly significant ($p < 0.0001$). On the other hand, a cursory examination of OLS results reveals that some of the parameter estimates were inconsistent with theoretical expectations. In addition, the confidence intervals for each of the parameter estimates were generally wider and the standard errors were also large. These collectively promoted the coefficient estimates for some of the important variables to be statistically insignificant.

Table 4.3: ANOVA and model summary for Oromia region crop production function

Source	DF	Sum of Squares	Mean Squares	F Value	Pr > F
Model	10	1547.87370	154.78737	562.66	<.0001
Error	6360	1749.68430	0.27510		
Corrected Total	6370	3297.55800			
Root MSE	0.52449	**R-Square** = 0.4694	**Durbin-Watson D** = 1.842		
Dependent Mean	3.28340	**Adj R-Sq** = 0.4686			
Coeff Var	15.97399				

Table 4.4: OLS parameter estimates for Oromia region crop production function

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits		VIF
Intercept	Intercept	1	2.37315	0.05662	41.91	<.0001	2.2622	2.48415	0.0000
LEDUC	Log of education	1	0.01294	0.00963	1.34	0.1790	-0.0059	0.03182	1.0182
LHS	Log of family size	1	0.04968	0.02128	2.34	0.0196	0.0079	0.09139	1.1230
LFERT	Log of fertilizer	1	0.05862	0.00880	6.66	<.0001	0.0414	0.07588	1.6867

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits		VIF
LWT	Log of seed in Kg	1	-0.11250	0.01065	-10.57	<.0001	-0.1334	-0.09163	2.0248
LAREA	Log of crop area	1	1.04057	0.02374	43.84	<.0001	0.9940	1.08710	1.8779
LOX	Log of oxen	1	0.00380	0.02772	0.14	0.8909	-0.0505	0.05815	1.3444
EXT2	Extension= 1	1	-0.08659	0.01558	-5.56	<.0001	-0.1171	-0.05604	1.0586
IRRIG2	Irrigation= 1	1	-0.00824	0.01515	-0.54	0.5865	-0.0379	0.02146	1.0161
DAM2	Damage= 1	1	0.00898	0.01508	0.60	0.5515	-0.0208	0.03853	1.0112
OWN2	ownership= 1	1	0.00824	0.01615	0.51	0.6098	-0.0234	0.03991	1.0050

According to OLS estimation, the proportions of variation explained by the included explanatory variables in the crop production functions for Oromia and SNNP regions were less than 50% ($R^2 < 0.50$) of the total variations. Though not statistically significant, the coefficient estimate for the education variable in Tigray region production function received a sign different from expected a priori. The OLS estimate of extension contact variable in Oromia region was unexpectedly negative. Also the coefficient estimates for the variable number of plowing oxen were found to be statistically insignificant in Amhara and Oromia regions. Although not statistically significant, the OLS estimates of irrigation variable in Tigray, Amhara and Oromia regions had signs different from theoretical expectations. These justifications suggested that OLS estimation could not be the preferred method to express the actual input-output relationships. As a result, statistical model diagnostics and checking was performed in order to identify which statistical assumptions were violated and accordingly go through the possible remedial measures.

4.4: Statistical Model Diagnostics and Checking

In any regression analysis, model diagnostics is an integral part that should be done prior to coming up to the final model for interpretation. As it was proposed in the methodology section of this thesis work, three distinct production functions (models), namely the linear, exponential and Cobb-Douglas were thoroughly assessed using crop production data of each study region. For simplicity as well as step by step method of model diagnostics and checking, the model being considered first was the linear production function (see equation 3. 2). Accordingly, the linear model was fitted and diagnosed for the statistical assumptions behind it. Upon employing the linear model, almost all the assumptions of linear regression, which were stated in the methodology section, were violated in many of the regions production functions except the assumptions of no autocorrelation and no multicollinearity. This is of course in line with the general literature on production function in such a way that the linear production function is not mostly recommended as viable alternative.

Consequently, various mathematical transformations (the log, square root and quadratic) were made to at least make the assumptions of linearity, normality and homogeneous variance true. From among these transformations the log form was preferred to others for its outperformance in response to the violated assumptions, ease of interpretation and conformity with the proposed models. Whenever the response variable in the linear production function of each region is transformed to log scale and keeping the predictors as they are, the model becomes an exponential production function (see equation 3.5). In this model, the plot of residuals versus predicted value revealed that heteroscedasticity has decreased to a certain extent and the model approached to retain linearity, but yet the problems of extreme observations were highly identified. Therefore, further transformations were mandatory in order to possibly alleviate the problems of heteroscedasticity, non-linearity and outliers. As a result, the transformations being made were taking the log of each predictor variables, considering cross products of predictors, including polynomials of predictors, incorporating square root and cube roots of predictors. Comparable to the reason described above, the log of the response variable regressed on logs of each predictor variable was chosen. That is, the transformations of the dependent and explanatory variables led to the log form of Cobb-Douglas production function (see equation 3.4). Though these transformations did not lessen the problem of outliers, they helped to further minimize the visible heterogeneity of error variances and hence led to near linearity. Additionally, the values of coefficient of determination (R^2) had come to be higher in the log transformed cases of both predictor and response variables. In view of the above explanations, the Cobb-Douglas model was chosen for further investigation of crop production functions at each of the regions.

Therefore, all the assumptions discussed in the previous chapter, were checked for the log form of Cobb-Douglas model (see equation 3.4). The statistical assumption of no multicollinearity was verified using the values of VIFs. The values of VIFs, which are displayed at the final columns of

Table 4.3, Table B2, Table B4 and Table B6, were each less than 10, which confirm that multicollinearity was not severe in each of the regions crop production functions. Checking for the existence of autocorrelation between random disturbance terms in each region's production function was made by using the Durbin-Watson statistic. The Durbin-Watson statistic estimates for the OLS regression of regions Tigray, Amhara, Oromia and SNNP were 1.81, 1.78, 1.842 and 1.768 respectively. Since all these values are hovering around 2, the error components were not autocorrelated. The normal quantile-quantile plots of each region's production function, which are displayed in Annex C confirm that the assumption of normality was not satisfied for the log form of Cobb-Douglas model. This is perhaps due to the fact that the data suffered from outlying cases, which made the distribution skewed. The plots of unstandardized residuals obtained from OLS versus predicted values, which are given in Annex D suggest that the variances of the error terms were not homoscedastic and also the linearity assumption was not strictly satisfied.

Additionally, the existence of outliers and leverage points were checked by the method of least trimmed squares (LTS) regression. The results based on LTS indicate that many outlying cases were available in each of the regions production data (see Annex E). Therefore, a further test for heteroscedasticity was made using the modified Goldfeld- Quandt (MGQ) test to check whether real heteroscedasticity was present or the errors seemed heteroscedadstic due to outliers.

H_0: The error variance is homogeneous

H_1: The error variance is heterogeneous

$\alpha = 0.05$

Table 4.5-MGQ Test Results:

	Region			
	Tigray	**Amhara**	**Oromia**	**SNNP**
Number of cases(n)	1676	6157	6371	8084
C	176	257	371	484
MSDR2	0.0917	0.1017	0.16813	0.3831
MSDR1	0.0868	0.0973	0.1660	0.4128
MGQ	**1.0565**	**1.0452**	**1.0127**	**0.9281**

x variable	Family size	Plowing oxen	Fertilizer	Family size
Fcrit (α = 0.05)	**1.1287**	**1.0626**	**1.0620**	**1.0549**

. Note that Fcrit(α = 0.05) is the tabulated value of F statistic with numerator and denominator degrees of freedom each of (n-c-2x11)/2.

The above MGQ test results of each region error components, with details given in Annex F, indicate that the error terms in each of the regions production functions were not heteroscedastic. Furthermore, the graphs of unstandardized robust residuals versus predicted values verify that the errors were not heteroscedastic (see Annex G). Above all, the graphs displayed in Annex G indicate that the linearity assumption was satisfied after the robust regression was used. Finally, the assumption of normality of error terms was assessed after the robust regression. The QQ plots given in Annex H indicate that the assumption of normality was achieved after robust regression was done. Therefore, the robust regression estimates, which were taken as final model estimates are presented as follows.

4.5: Robust Regression Estimates

The estimates for the parameters of production functions and associated statistics for each region using robust regression methods are given below. The existence of outliers was a central problem in the credibility of parameter estimates using the method of OLS regression. Therefore, the regression equations were re-estimated using robust methods in order to adjust the effects of outlier problem. In the vein of the results obtained from robust method, the R^2 values (summary measures for overall goodness of fit) were greater, the standard errors of the coefficients were smaller and the coefficient estimates in general were found to differ from that of OLS methods. Additionally, the confidence intervals for each of the parameter estimates have come to be narrower than those obtained from OLS. Thus, these confirm that the OLS estimates were misleading due to the occurrence of outliers and/ or leverage points.

Because of its robustness and efficiency properties, the reweighted least squares based on an analysis of least trimmed squares (LTS) residuals regression was applied for re-estimating the parameters in the production functions. The re-estimated production functions of each region adjusted for outlying cases are given hereunder.

Table 4.6: Parameter estimates and associated statistics for final WLS fit in Tigray region

Parameter	DF	Estimate	Standard Error	95% Confidence Limits		Chi-Square	Standardized Beta Coefficient
Intercept	1	0.9329	0.1055	0.726	1.1397	78.13*	0.0000
LnEduc	1	-0.0259	0.0203	-0.0656	0.0138	1.63	-0.0176
LnFer	1	0.0486	0.0135	0.0221	0.0751	12.9*	0.0556
LnWt	1	0.0718	0.029	0.015	0.1287	6.13**	0.0675
LnOx	1	0.1348	0.0444	0.0477	0.2219	9.2*	0.0437
LnAr	1	1.735	0.0845	1.5694	1.9006	421.5*	0.7418
LnHs	1	0.1273	0.0293	0.0699	0.1848	18.88*	0.0598
Ext	1	0.0357	0.0193	-0.0022	0.0736	3.40***	0.0222
Irr	1	-0.0125	0.0225	-0.0565	0.0315	0.31	-0.0077
Dam	1	-0.0687	0.0269	-0.1215	-0.0159	6.51**	-0.0404
Own	1	-0.006	0.024	-0.0531	0.0411	0.06	-0.0034
Scale	0	0.4446					

Note * = significant at $p < 0.01$; ** = significant at $p < 0.05$; and *** = significant at $p < 0.1$

$R^2 = 0.7109 \sum\hat{\beta} = 2.1175$

Table 4.7: Parameter estimates and associated statistics for final WLS fit in Amhara region

Parameter	DF	Estimate	Standard Error	95% Confidence Limits		Chi-Square	Standardized Beta Coefficient
Intercept	1	1.3456	0.0461	1.2553	1.436	852.07*	0.0000
LnOx	1	0.0373	0.0178	0.0204	0.0722	4.39**	0.0157
LnAr	1	1.3585	0.0247	1.31	1.407	3014.96*	0.5918
LnFer	1	0.0571	0.0037	0.0498	0.0645	232.66*	0.1454
LnHs	1	0.1056	0.0213	0.0638	0.1474	24.49*	0.0425
LnWt	1	0.0212	0.0081	0.0052	0.0371	6.79*	0.0259
LnEduc	1	-0.0231	0.015	-0.0525	0.0063	2.36	-0.0123
Ext	1	0.0254	0.0157	-0.0053	0.0561	2.63	0.0145
Irr	1	-0.0071	0.0143	-0.0351	0.0208	0.25	-0.0039
Dam	1	-0.0616	0.0283	-0.1172	-0.0061	4.74**	-0.0344
Own	1	0.0199	0.0158	-0.0111	0.0508	1.58	0.0100
Scale	0	0.5351					

Note * = significant at $p < 0.01$; ** = significant at $p < 0.05$; and *** = significant at $p < 0.1$

$R^2 = 0.5791 \sum\hat{\beta} = 1.5797$

Table 4.8: Parameter estimates and associated statistics for final WLS fit in Oromia region

Parameter	DF	Estimate	Standard Error	95% Confidence Limits		Chi-Square	Standardized Beta Coefficient
Intercept	1	2.4652	0.0536	2.3601	2.5703	2114.02*	0.0000
LnEduc	1	0.0181	0.0091	0.0003	0.0358	3.96**	0.0192
LnHs	1	0.0497	0.0201	0.0104	0.089	6.14**	0.0250
LnFer	1	0.0516	0.0084	0.0352	0.068	37.99*	0.0769
LnWt	1	-0.1283	0.01	-0.148	-0.1086	163.17*	-0.1735
LnAr	1	1.0251	0.0226	0.9809	1.0693	2064.28*	0.5986
LnOx	1	0.0479	0.0261	-0.0034	0.0991	3.35***	0.0203
Ext	1	-0.0934	0.0147	-0.1222	-0.0646	40.43*	-0.0624
Irr	1	0.004	0.0143	-0.024	0.032	0.08	0.0027
Dam	1	0.0034	0.0142	-0.0245	0.0313	0.06	0.0023
Own	1	0.0053	0.0152	-0.0245	0.0352	0.12	0.0033
Scale	0	0.5443					

Note * = significant at $p < 0.01$; ** = significant at $p < 0.05$; and *** = significant at $p < 0.1$

$R^2 = 0.5193 \sum\hat{\beta} = 1.0641$

Table 4.9: Parameter estimates and associated statistics for final WLS fit in SNNP region

Parameter	DF	Estimate	Standard Error	95% Confidence Limits		Chi-Square	Standardized Beta Coefficient
Intercept	1	1.3236	0.0599	1.2062	1.4411	487.66*	0.0000
LnHs	1	0.7804	0.025	0.7313	0.8294	972.67*	0.3088
LnFer	1	-0.0046	0.0117	-0.0275	0.0182	0.16	-0.0068
LnWt	1	-0.1194	0.0441	-0.2058	-0.0329	7.33*	-0.1344
LnAr	1	0.6102	0.0267	0.5578	0.6626	521.12*	0.2451
LnOx	1	0.1182	0.023	0.0731	0.1633	26.36*	0.0627
LnEduc	1	0.0842	0.013	0.0587	0.1096	41.89*	0.0602
Ext	1	0.0513	0.0204	0.0114	0.0913	6.33**	0.0233
Irr	1	-0.0303	0.024	-0.0774	0.0168	1.59	-0.0140
Dam	1	-0.0682	0.0259	-0.119	-0.0174	6.93*	-0.0308
Own	1	-0.0524	0.0251	-0.1017	-0.0031	4.34**	-0.0228
Scale	0	0.858					

Note * = significant at $p < 0.01$; ** = significant at $p < 0.05$; and *** = significant at $p < 0.1$

$R^2 = 0.5328 \sum\hat{\beta} = 1.4736$

where,

- Ln = the natural logarithm of a quantity;
- The dependent variable is Ln of production in quintals per peasant;
- Hs = the total number of individuals per household;
- Ox = the total number of plowing oxen a household own;
- Ar = Area of agricultural land in hectares;
- Fer = Amount of chemical fertilizer employed in kilogram;
- Wt = Weight of improved and/or non-improved seed employed in Kg;
- Educ = Education (highest grade) attained by the head of the household;
- Ext = Dummy variable scoring 1 for farmers having extension contact and 0 otherwise;
- Irr = Dummy variable scoring 1 for farmers who applied irrigation on their farms and 0 otherwise;
- Dam = Dummy variable scoring 1 if there was any crop damage and 0 otherwise;
- Own = Dummy variable scoring 1 if land ownership type is private and 0 if land ownership type is rent/leased;
- $\sum\hat{\beta}$ is the sum of coefficient estimates for continuous predictor variables (variable inputs).

4.6: Discussion and Interpretation of Regression Coefficients

As indicated earlier, the occurrence of outliers was the major problem in each of the regions production data. Hence, the iteratively reweighted least squares using LTS residuals regression (robust regression) was applied as a means of handling this problem. Regression analysis indicated that number of statistically significant parameters varied among the regions considered in this study. However, the F-test results showed that overall regression models for each region were statistically significant. The estimated production functions were able to fit the observed data reasonably well at each region as suggested by R^2 greater than 0.50 in that over 50% of the variation in crop production was explained by the included explanatory variables. This performance is good; however, model improvement can be achieved through incorporation of variables, such as farmer ability and precipitation. In most of the study regions, area of agricultural land, weight of seed employed, number of plowing oxen and number of individuals per household variables were statistically significant (these results are consistent with Addis et. al, 2001; Weir and Knight, 2000; Yao, 1996 and Yohannes and Coffin, 1993). The statistical significance of such coefficients in the estimated production functions indicates that these variables were the most significant inputs in crop productions and require special attention.

The production function estimation results in Tigray region showed that an increase of one percent in the average number of individuals per household would result in a 0.13 % increase in the average crop yield while all other variables were held constant. An increase in one percent of the average crop area in hectare had contributed for a 1.74% increase on the average crop yield keeping the remaining variables fixed. The output response to a 1% increase in input due to number of plowing oxen was 0.13%, indicating that the contribution of number of plowing oxen for yield maximization was not high. This is partly because the average number of plowing oxen owned by each private peasant was few (less than 2 on average). Also the coefficient estimates for the variables fertilizer and seed were statistically significant implying that crop yield increases with increasing the amount of these inputs. A 10% increase in the amount of fertilizer resulted in a 0.49% increase in crop yield. Moreover, the estimated coefficient of education variable in Tigray region was not statistically significant. The implication is that farmers in the region could no longer use their skills obtained from education for the crop production system. This was of course related to their low level of education (the mean grade level was less than three).

Furthermore, the standardized beta coefficients indicate that areas of crop land followed sequentially by weight of seed, labour force participation (number of individuals per household), fertilizer employed and plowing oxen had many contributions for the maximization of crop yield per household in Tigray region.

The estimated coefficients for the variables extension contact and crop damage entered as dummy variables were statistically significant. The coefficient estimate for the variable extension contact being positive, suggests that the amount of crop yield was higher for farmers who were included in the agricultural extension programme than for those who were not. However, the coefficient estimate for the variable crop damage, -0.0687 indicates that the expected percentage decrease in crop yield by farmers who faced crop damage was 6.87%. In other words, this result implies that farmers, who faced crop damage, produced about 6.87% less than those who did not provided the other variables were the same for each of the farmers in Tigray region.

Above all, a number of conclusions can be drawn based on the Reweighted Least Squares (RLS) regression estimation results of Tigray region crop farms. Firstly, crop production is mainly determined by three major factors: agricultural land, human labor and seed. Secondly, fertilizer and oxen power had statistically significant effects on crop production, an indication that crop production is also dependent on number of plowing oxen possessed per household and fertilizer in addition to the above three major factors. Finally, peasants who had no crop damage in their farms produced a greater amount of crop than those who had faced crop damage; this is in turn a condition that may lead to the minimal crop yield in the region due to the fact that crops are continuously damaged by some natural or artificial disasters.

All the coefficients in the robust regression estimation of Amhara region had the expected a priori signs except the variables education and irrigation. The standardized beta coefficient for area of agricultural land (0.5918) was the highest compared to others. This reveals that crop production was highly dependent upon the area of agricultural land in the region. Fixing the effect of other variables the same, the coefficient estimate of 1.3585 with respect to crop yield implied that a 1% increase in area of agricultural land would lead to an increase of 1.3585% in the crop yield. Similar results were obtained for the variables oxen power, fertilizer, seed and human labour to positively affect crop yield in the region. From among the coefficient estimates of these variables, the contribution by fertilizer was the highest (standardized beta coefficient = 0.1454) followed sequentially by human labour, seed and plowing oxen. Thus, these entail that these variables among others were playing a pivotal role for increasing crop yield. Particularly, the significance of inputs such as fertilizer is derived from the fact that fertilizer is the major land augmenting input that increases crop yield by improving the fertility and productivity of the agricultural land.

The estimated coefficient for education variable had no statistically significant effect on crop yield in Amhara region. The non significance of this variable was attributed to the proportion and level of education, i.e., only few farmers were educated and even from among those the level of education did not on average exceed grade 2. There was a statistically significant ($p < 0.05$) negative relation between the crop damage variable and crop yield in the Amhara region crop farms. The coefficient estimate of -0.0616 reveals that farmers who faced crop damages produced 6.16% less than those who did not keeping other factors constant. However, the estimated coefficients for variables irrigation and extension had no statistically significant contribution for crop yield. This is in fact in line with the descriptive results (less than 50% of the farmers used irrigation and had extension contacts).

The coefficient estimate for area of agricultural land in Oromia region was positive and statistically highly significant ($p < 0.01$), which suggests that area of agricultural land affected crop production positively. Holding other factors fixed, a 1% increase in the area of agricultural land would induce an increase of about 1.03% in crop yield. Similarly, the contributions of fertilizer, education and oxen power to crop production were statistically significant, though the sizes of the coefficients were very small, i.e. less than 6% of total crop production was explained by either of the three. The positive and significant effect of education variable indicate that farmers who had more year of formal education tend to produce more amount of crop per hectare, presumably due to their enhanced ability to acquire technical knowledge, which facilitated them focus on the best input output combination. Besides, farmers who had some level of education respond readily to the use of improved technology such as application of fertilizer, use of pesticides and improved planting materials.

The coefficient for weight of seed was negative and statistically significant at the 1% level in Oromia region, indicating that increasing in weight of seed decreased crop yield. This was perhaps due to the fact that yield of crop increased proportionately with increased in weight of improved/non improved seed up to certain level but started declining afterwards. The other possible reason is that farmers in the region may mostly applied non-improved seed in such a way that the crop yield may not increased with increasing seed quantity.

Though the sign for extension contact variable was negative, it had statistically significant ($p < 0.01$) contribution for crop production in Oromia region. This sign for extension contact variable was consistent with the descriptive result for which the mean crop yield by peasants who had extension contacts was smaller (30.69 quintals) than those who had no extension contacts (mean crop yield of 36.79 quintals). This was perhaps related to the fact that the agricultural extension service providers were not serving the farmers appropriately. On the other hand, the standardized beta coefficient for irrigation variable was small and not statistically significant. This was perhaps related to the generally well distributed and timely rainfall in Ethiopia during the surveyed year (see FAO/WFP, 2008) so that production differences were not observed due to water availability.

We found statistically significant effects of crop land, human labour, oxen power, education and seed on crop yield at SNNP region. All these coefficients were positive except the coefficient of seed variable. The standardized beta coefficient for human labor (0.3088) was found to be the highest compared to others. This implies that human labor was the major input in maximizing crop yield at the region. An increase of 1% in the average number of individuals per household led to an increase of 0.78% in crop yield provided the other variables were constant. This result was of course different from the other regions where the coefficient estimate for area of crop land was the highest. The possible reason behind the highest coefficient estimate for the human labor variable is that large proportions of individuals at the region were in the economically working age (15 - 64), that was 49.7% (FDREPCC, 2008). Thus, they were highly and effectively participating in the agricultural practices.

Though it is generally believed that application of fertilizer in crop farms increases the crop yield to the greater extent, the contribution was not statistically significant and received sign different from theoretical expectation in SNNP region. The non-significance of fertilizer variable may be attributed to the level of use in which the mean amount was the minimum (27.04 Kg) as compared to the remaining regions.

The coefficient estimates for all variables entered as dummy except irrigation were statistically significant and had signs expected a priori in SNNP region. Counter to theoretical expectations, the coefficient estimate for the irrigation variable was negative implying that farmers who used irrigation produced smaller amount of crops than those who did not. This was perhaps due to the inappropriate application of water on the field owing to lack of sufficient knowledge.

In summary, the change in output relative to a unit change in input, the elasticity of production for the Cobb- Douglas function is the same as the coefficients of the estimated model. Accordingly, the production elasticity for each of the inputs at each region was less than unity except for crop areas in Tigray, Amhara and Oromia regions which were 1.7350, 1.3585 and 1.0251 respectively. This reveals that the relation between inputs and output was inelastic except crop areas in the stated regions. That is, holding other factors constant, the marginal return to each factor will decrease as more of the factors are used. The production elasticities for crop areas in Tigray, Amhara and Oromia regions imply the maximum technical efficiency. Also the coefficients of returns to scale, which are obtained by summing the estimated coefficients of variable inputs (inputs which change with the volume of output over a specified time period (e.g. fertilizer, seeds, fuel, harvest labor etc, Ellis, 1994, pp. 42), were greater than unity for each of the regions indicating increasing returns to scale. From these results it follows that an increase in all factors of production by 1% will lead to an increment of crop yield by $\sum \hat{\beta}$ percent. This will in fact hold true only if the peasant can actually make a proportionate change in every input factor. To this end, the finding pertaining to increasing returns is in full agreement with the claim that cases of increasing returns to scale occur at relatively low levels of output, which are characteristics of small scale farming (Mbanasor and Obioha, 2003 citing Olayide and Heady, 1982).

CHAPTER FIVE

5. CONCLUSION AND RECOMMENDATIONS

5.1: Conclusion

This thesis applied robust regression method for production functions analysis of four major crop producing regions in Ethiopia namely, Tigray, Amhara, Oromia and SNNP. At first, three production functions (linear, exponential and Cobb-Douglas) were proposed, but the statistical model diagnostics and checking in addition to a priori theoretical expectations suggested that crop production function for each of the region was found to be appropriately represented by the Cobb-Douglas production function. This is in fact in conformity with the literature (e.g., Yao (1996); Addis, et.al (2001); Pender and Gebremedhin (2007)).

In general, the statistical findings from each region reveal that farm size, fertilizer, seed employed, oxen power and human labor were playing a pivotal role for the maximization of crop yield. From among these variables, the great contribution was found to be due to farm size (highest standardized beta coefficient) in each of the regions with SNNP an exception in which the great share was due to human labor. However, given that the possibility for increasing farm size is impracticable due to the prevailing population growth maximizing land productivity can only be achieved through effective involvement of labor and efficient use of modern agricultural practices. This comprises educating farmers (offering formal education or short term trainings which focus on the wise application of agricultural inputs), intensifying use of insecticides and pesticides, increasing application of chemical fertilizer, rising irrigated areas and expansion of agricultural extension services to the greater extent. The conclusion in this analysis is strongly consistent with the policy direction set in ADLI strategy. What the conclusion in this analysis really argues is that expansion of modern agricultural inputs and practices have to be performed to the maximum possible scale.

To this end, the contribution of fertilizer for significant crop yield response was observed to be good in many of the regions (see standardized beta coefficients in Tables 4.5 – 4.8). Surprisingly, the coefficient estimate was statistically insignificant for SNNP region. This implies that farmers in SNNP region were either applying chemical fertilizer to the smallest extent (the mean amount was of course 27.04 Kg) or not at all. The reason behind this was perhaps shortage of transportation facilities which provide fertilizer timely to the farms, monetary inability to afford the cost of fertilizer and unwillingness to apply chemical fertilizer relating to some unscientific justifications.

Education had no statistically significant contribution for the maximization of crop yield at Tigray and Amhara regions. This was because farmers were almost illiterate (with average year of formal schooling less than three) in such a way that they could no longer be easily volunteers at least for the acceptance of modern agronomical inputs such as application of improved seeds and chemical fertilizer. This eventually led to minimum crop yield per hectare in the regions.

From among those variables entered as dummy in the production functions, extension contact exerted statistically significant and positive effects on crop yield in Tigray and SNNP regions while it exerted statistically significant and negative effect in Oromia region. The negative sign for extension contact implies that policymakers and the Ministry of Agriculture and Rural Development should do their level best to mitigate extension's negative effect on crop productivity in Oromia region. Though not statistically significant, the coefficient estimate for irrigation variable at each of the studied regions except Oromia was negative. This non significance of this variable was perhaps related to proportions of peasants who used irrigation was small (see Table 4.2). Furthermore, the inappropriate application of water on the farm due to lack of good knowledge might led to the negative sign for irrigation variable. Crop damage, which may encompass vulnerability to droughts, flooding and crop viruses, was the basic problem which reduced crop productivity in each of the studied regions except Oromia.

By far, crop and livestock production are intertwined in Ethiopian agriculture and hence it was impossible to get the values of each variable incorporated in this research separately for the crop production. Thus, the data taken in the analysis were more aggregated. We therefore believe that this study opens a possible avenue for further investigation in the field provided it is possible to get/ collect measurable data separately for crop farming system.

5.2: Recommendations

Based on the findings in this thesis work, the following are recommended.

1. By and large, formal education as well as farm trainings to the farmers should be pursued with more vigor, because, education brings about greater awareness on the part of the farmers and adoption of better production techniques and use of improved inputs, and thus brings about higher output.
2. The most important factor explaining the crop yield in all of the regions except SNNP was farm size; as farm size increased, so did the crop yield. Since increasing farm size is difficult, use of irrigation as an alternative to produce crop is paramount in such a way that farmers can produce many times per year with the available land. Moreover, irrigation is probably the most important factor which can guarantee sustainable productivity and output growth not only because it can reduce the vulnerability of production to droughts, but also because it can enhance the effectiveness of fertilizer and improved seeds. As a result, the increase of irrigated areas to the maximum possible scale is mandatory.
3. Crop damage was found to be one of the most crucial components, which led to the minimum crop productivity. To decrease or possibly alleviate this problem, immense efforts should be made, such as introducing drought resistant seed varieties, effective pre or post damage application of insecticides and pesticides and disseminating information on valid whether forecasts so that appropriate measures can be taken such as harvesting yields before crop damage occurred.
4. Since productivity was found to increase with number of plowing oxen, policymakers and the Ministry of Agriculture and Rural Development should work at their level best for improving the provision of credit systems to the farmers so that the peasants are either able to increase their plowing oxen or purchase modern agricultural machineries such as tractor. Also, lower interest rate credit systems to the peasants are required so that farmers will be able to purchase fertilizer and improved seeds.

5. Since extension service was found to be one of the viable alternatives for the wise application of factor inputs, the government should allocate more funds to strengthen the extension department and expand network of extension services to reach each and every farmer.

REFERENCE

Addis, T., Teklu, T., Mwangi ,W. and Verkuijl, H. (2001). Gender Differentials in Agricultural Production and Decision-Making among Smallholders in Ada, Lume, and Gimbichu Woredas of the Central Highlands of Ethiopia. Mexico, D.F.: International Maize and Wheat Improvement Center (CIMMYT) and Ethiopian Agricultural Research Organization.

Bakhsh, K., Ahmad, B. and Hassan, S. (2006). Food Security through Increasing Technical Efficiency. Asian Journal of Plant Sciences 5 (6): pp. 970 – 976.

Chunkwuji ,C. O., Inoni, O. E., Ogisi, O. D. and Oyaide, W. J. (2006). A Quantitative Determination of Allocative Efficiency in Broiler Production in Delta State, Nigeria. Agriculturae Conspectus Scientificus, Vol. 71, No. 1, pp. 21-26.

Chatterjee, S. and Price, B. (1977). Regression Analysis by Example. John Wiley and Sons, Inc.

Central Statistical Agency (CSA), (2008). Agricultural Sample Survey 2007/08. Report on Area and Production of Crops (Private peasant holdings, Meher Season). Vol. 1, CSA, Addis Ababa.

Crown, W. H. (1998). Statistical Models for the Social and Behavioral Sciences: Multiple Regression and Limited-Dependent Variable Models. Praeger Publishers, Westport, CT .

Draper, N.R. and Smith, H. (1998). Applied Regression Analysis, John Wiley & Sons, Inc., New York. 3rd edition.

Ellis, F. (1994). Peasant Economics: Farm Households in Agrarian Development. Cambridge University Press. 2nd edition.

FAO/WFP (2008). Special Report: Crop and Food Security Assessment Mission to Ethiopia (Phase 1), Rome.

Federal Democratic Republic of Ethiopia Population Census Commission (FDREPCC), (2008). Summary and Statistical Report of the 2007 Population and Housing Census.

Finger, R. and Hediger, W. (2007). The Application of Robust Regression to a Production Function Comparison–the Example of Swiss Corn. http://mpra.ub.uni-muenchen.de/9097/.

Frank, D. and Beattie, R. (1979). The Economic Value of Irrigation Water in the Western United States: An Application of Ridge Regression. Technical Report No. 99. Texas Water Resources Institute. Texas A&M University

Ghebremariam, W.K., Ortmann, G.F. and Nsahlai, I.V., (2006). A Production Function Analysis of Commercial Dairy Farms in the Highlands of Eritrea Using Ridge Regression. Agrekon, Vol. 45, No. 2.

Griffin, C. R., Montgomery, M. J. and Rister, E. M. (1987). Selecting Functional Form in Production Function Analysis. Western Journal of Agricultural Economics, 12(2): 216-227.

Gujarati, N. D. (1995), Basic Econemtrics. McGraw-hill international editions, 3rd edition

Iqbal, M., Ahmad, M. and Abbas, K. (2003). The Impact of Institutional Credit on Agricultural Production in Pakistan. The Pakistan Development Review 42, pp. 469–485

Jason, O. and Waters, E. (2002). Four Assumptions of Multiple Regression that Researchers Should Always Test. Practical Assessment, Research & Evaluation, 8(2).

Lawrence, D. K. and Marsh, C. L. (1984). Robust Ridge Estimation Methods for Predicting U.S. Coal Mining Fatalities. Commun. Statist. – Theor. Meth. , 13(2), pp. 139-149.

Lopez, E. (1997). The Structure of Production of the Spanish Telecommunications Sector", Empirical Economics, Vol.22, pp. 321-330.

Madariaga B. and McConnell K. E. (1984). Value of Irrigation Water in the Middle Atlantic States: An Econometric Approach. Southern Journal of Agricultural Economics, pp. 91-98.

Maddala, G. S. (1992). Introduction to Econometrics. Macmillan, Inc. 2nd edition.

Mbanasor, J. A. and Obioha, L. O. (2003). Resource Productivity under Fadamas Cropping System in Umuahia North Local Government Area of Abia State, Nigeria. Tropical and Subtropical Agricultura, Vol. 2, pp. 81 – 86.

Ministry of Finance and Economic Development (MOFED). Federal Democratic Republic of Ethiopia (2007). Ethiopia: Building on Progress: A Plan for Accelerated and Sustained Development to End Poverty (PASDEP). Annual Progress Report 2006/07, Addis Ababa.

Olarinde, L. O., Manyong, V. M. and Akintola, J. O. (2007). Attitudes towards Risk among Maize Farmers in the Dry Savanna Zone of Nigeria: Some Prospective Policies for Improving Food Production. African Journal of Agricultural Research, Vol. 2 (8), pp. 399-408

Pender, J. and Gebremedhin, B. (2007). Determinants of Agricultural and Land Management Practices and Impacts on Crop Production and Household Income in the Highlands of Tigray, Ethiopia. Journal of African Economies Vol. 17, No. 3, pp. 395-450.

Qunaibet, M. H. Al., Limam, M.M.T. and Sofian, B. E. E. (1995). Testing for Returns to Scale in Dairy Farms in Saudi Arabia. J. King Saudi Univ. Vol. 7, Agric. Sci. (1).

Rana, S., Midi, H. and Imon, R. A.H.M. (2008). A Robust Modification of the Goldfeld-Quandt Test for the Detection of Heteroscedasticity in the Presence of Outliers, Journal of Mathematics and Statistics 4 (4), pp. 277-283.

SAS Institute (2008), SAS/STAT 9.2 User's Guide. Cary, NC: SAS Institute Inc.

United Nations Economic and Social Council (UNESC), 2007. The Agricultural Development Led Industrialization (ADLI) Strategy Ethiopia, 2007, available online at http://webapps01.un.org/nvp/frontend!policy.action?id=124

Verbeek, M. (2008). A Guide to Modern Econometrics. John Wiley and Sons, Inc. 3rd edition.

Weir, S. and Knight, J. (2000) Education Externalities in Rural Ethiopia: Evidence form Average and Stochastic Frontier Production Functions.

Yao, S. (1996). The Determinants of Cereal Crop Productivity of the Peasant Farm Sector in Ethiopia, 1981-87. Journal of International Development, Vol. 8, Issue 1, pp. 69-82.

Yilmaz, I. and Ozkan, B. (2004). Econometric Analysis of Land Tenure Systems in Cotton Production in Turkey. International Journal of Agriculture & Biology, 1560–8530/2004/06–6–1023–1025, http://www.ijab.org

Yohannes, M. and Coffin, G. (1993). Crop and Milk Production Structure of Smallholders in Ethiopia. MPRA paper No.405.

ANNEX A: Correlation Matrices between the Response and Predictor Variables

Pearson Correlation Coefficients, N = 1676 for Tigray region
Prob[5] > |r| under H0: Rho=0

	Prod	Educ	Hsize	Fert	Weight	Area	Ox
Prod	1.0000	-0.0072	0.6282	0.3506	0.62073	0.7162	0.261
		0.7679	<.0001	<.0001	<.0001	<.0001	<.0001
Educ	-0.0072	1.0000	0.0006	0.0371	-0.0008	0.00455	0.0226
	0.7679		0.9818	0.1289	0.9742	0.8524	0.3557
Hsize	0.6282	0.0006	1.0000	0.3711	0.81083	0.88705	0.2669
	<.0001	0.9818		<.0001	<.0001	<.0001	<.0001
Fert	0.3506	0.0371	0.3711	1.0000	0.41087	0.42512	0.1255
	<.0001	0.1289	<.0001		<.0001	<.0001	<.0001
Weight	0.6207	-0.0008	0.8108	0.4109	1.0000	0.88697	0.2674
	<.0001	0.9742	<.0001	<.0001		<.0001	<.0001
Area	0.7162	0.00455	0.88705	0.4251	0.88697	1.0000	0.2945
	<.0001	0.8524	<.0001	<.0001	<.0001		<.0001
Ox	0.2610	0.0226	0.2669	0.1255	0.2674	0.2945	1.0000
	<.0001	0.3557	<.0001	<.0001	<.0001	<.0001	

Pearson Correlation Coefficients, N = 6157 for Amhara region
Prob > |r| under H0: Rho=0

	Prod	Ox	Areah	Fert	Hsize	Wt	Educ
Prod	1.0000	0.2818	0.6235	0.3985	0.2616	0.307	0.0375
		<.0001	<.0001	<.0001	<.0001	<.0001	0.0033
Ox	0.2818	1.0000	0.4359	0.2652	0.2939	0.2077	0.0171
	<.0001		<.0001	<.0001	<.0001	<.0001	0.1793
Areah	0.6235	0.4359	1.0000	0.3752	0.3197	0.3748	0.0565
	<.0001	<.0001		<.0001	<.0001	<.0001	<.0001
Fert	0.3985	0.2652	0.3752	1.0000	0.1398	0.2815	0.0928
	<.0001	<.0001	<.0001		<.0001	<.0001	<.0001
Hsize	0.2616	0.2939	0.3197	0.1398	1.0000	0.1451	0.0336
	<.0001	<.0001	<.0001	<.0001		<.0001	0.0085
Wt	0.307	0.2077	0.3748	0.2815	0.1451	1.0000	0.0672
	<.0001	<.0001	<.0001	<.0001	<.0001		<.0001
Educ	0.0375	0.0171	0.0565	0.0928	0.0336	0.0672	1.0000
	0.0033	0.1793	<.0001	<.0001	0.0085	<.0001	

[5]Note that numbers at the bottom of each value of the correlation coefficient indicate the probability of rejecting the null hypothesis Rho=0

ANNEX A Continued...

Pearson Correlation Coefficients, N = 6371 for Oromia Region Prob > |r| under H0: Rho=0							
	Prdod	Educ	Hhsize	Fert	Weight	Area	Oxen
Prdod	1.0000	0.0473	0.2130	0.3129	0.2538	0.4787	0.2574
		0.0002	<.0001	<.0001	<.0001	<.0001	<.0001
Educ	0.0473	1.0000	0.0924	0.0468	0.0335	0.0329	0.0448
	0.0002		<.0001	0.0002	0.0074	0.0087	0.0003
Hhsize	0.2130	0.0924	1.0000	0.1657	0.1782	0.3066	0.2704
	<.0001	<.0001		<.0001	<.0001	<.0001	<.0001
Fert	0.3129	0.0468	0.1657	1.0000	0.5675	0.5014	0.3638
	<.0001	0.0002	<.0001		<.0001	<.0001	<.0001
Weight	0.2538	0.0335	0.1782	0.5675	1.0000	0.5792	0.3972
	<.0001	0.0074	<.0001	<.0001		<.0001	<.0001
Area	0.4787	0.0329	0.3066	0.5014	0.5792	1.0000	0.4792
	<.0001	0.0087	<.0001	<.0001	<.0001		<.0001
Oxen	0.2574	0.0448	0.2704	0.3638	0.3972	0.4792	1.0000
	<.0001	0.0003	<.0001	<.0001	<.0001	<.0001	

Pearson Correlation Coefficients, N = 8084 for SNNP region Prob > |r| under H0: Rho=0							
	PRR	Educ	Ox	Area	Weight	Fert	Hsize
Prdod	1.0000	0.1043	0.1446	0.1352	0.1025	0.1316	0.3347
		<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
Educ	0.1043	1.0000	0.0161	0.0421	-0.0431	0.0381	0.1448
	<.0001		0.1485	0.0002	0.0001	0.0006	<.0001
Ox	0.1446	0.0161	1.0000	0.0139	0.3053	0.3554	0.0812
	<.0001	0.1485		0.2093	<.0001	<.0001	<.0001
Area	0.1352	0.0421	0.0139	1.0000	0.1917	0.1108	0.1267
	<.0001	0.0002	0.2093		<.0001	<.0001	<.0001
Weight	0.1025	-0.0431	0.3053	0.1917	1.0000	0.4813	0.1384
	<.0001	0.0001	<.0001	<.0001		<.0001	<.0001
Fert	0.1316	0.0381	0.3554	0.1108	0.4813	1.0000	0.1156
	<.0001	0.0006	<.0001	<.0001	<.0001		<.0001
Hsize	0.3347	0.1448	0.0812	0.1267	0.1384	0.1156	1.0000
	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	

ANNEX B: OLS Regression Estimates

TableB1: ANOVA and model summary for Tigray region crop production function

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Squares**	**F Value**	**Pr > F**
Model	10	630.87823	63.08782	230.21	<.0001
Error	1665	456.28226	0.27404		
Corrected Total	1675	1087.16048			
Root MSE	0.5235	**R-Square** = 0.58034	**Durbin-Watson D = 1.81**		
Dependent Mean	2.5916	**Adj R-Sq** = 0.5778			
Coeff Var	20.1995				

Table B2: OLS regression estimates for Tigray region crop production function

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits		VIF
Intercept	Intercept	1	0.7347	0.1191	6.17	<.0001	0.5012	0	0
E	Log of Education	1	-0.0254	0.0235	-1.08	0.2791	-0.0714	1.0065	1.0065
FER	Log of Fertilizer	1	0.0569	0.0157	3.63	0.0003	0.0262	1.2746	1.2746
HS	Log of family Size	1	0.1059	0.0307	3.46	0.0006	0.0459	4.5480	4.5480
W	Log of seed weight	1	0.1253	0.0327	3.83	0.0001	0.0612	3.7491	3.7491
AR	Log of crop area	1	1.5438	0.0973	15.86	<.0001	1.3528	6.8714	6.8714
OX	Log of oxen	1	0.1415	0.0513	2.76	0.0058	0.0409	1.0944	1.0944
EXT2	Extension= 1	1	0.0615	0.0259	2.37	0.0180	0.0106	1.0283	1.0283
IRRIGA2	Irrigation=1	1	-0.0013	0.0260	-0.05	0.9618	-0.0523	0.0498	1.0126
DAMAGE2	Damage=1	1	-0.0112	0.0272	-0.41	0.6817	-0.0645	0.0422	1.0129
OWNT2	Ownership= 1	1	-0.0049	0.0279	-0.18	0.8610	-0.0596	0.0498	1.0048

TableB3: ANOVA and model summary for Amhara region crop production function

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Squares**	**F Value**	**Pr > F**
Model	10	2369.60107	236.96011	620.06	<.0001
Error	6146	2348.75377	0.38216		
Corrected Total	6156	4718.35484			
Root MSE	0.6182	**R-Square** = 0.5022	**Durbin-Watson D** = 1.78		
Dependent Mean	2.7963	**Adj R-Sq** = 0.5014			
Coeff Var	22.1072				

Table B4: OLS regression estimates for Amhara region crop production function

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits		VIF
Intercept	Intercept	1	1.0447	0.0515	20.27	<.0001	0.9437	0.0000	0.0000
LED	Log of Education	1	-0.0194	0.0171	-1.13	0.2575	-0.0529	1.0184	1.0184
LFER	Log of fertilizer	1	0.0602	0.0043	14.07	<.0001	0.0518	1.4677	1.4677
LHS	Log of family size	1	0.1402	0.0241	5.82	<.0001	0.0929	1.1608	1.1608
LWT	Log of seed	1	0.0726	0.0089	8.08	<.0001	0.0549	1.4863	1.4863
LAREA	Log of area	1	1.2966	0.0278	46.72	<.0001	1.2422	1.8042	1.8041
LOX	Log of oxen	1	0.0025	0.0239	0.10	0.9176	-0.0444	1.2592	1.2592
EXT2	Extension=1	1	0.0315	0.0178	1.77	0.0770	-0.0034	0.0664	1.2676
IRRIG2	Irrigation==1	1	-0.0129	0.0162	-0.80	0.4237	-0.0447	0.0188	1.0177
DAM2	Crop damage= 1	1	-0.0791	0.0393	-2.01	0.0443	-0.1561	-0.0021	1.0368
OWN2	Ownership= 1	1	0.0079	0.0179	0.44	0.6575	-0.0273	0.0432	1.0084

TableB5: ANOVA and model summary for SNNP region crop production function

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Squares**	**F Value**	**Pr > F**
Model	10	4046.12165	404.61217	635.33	<.0001
Error	8073	5141.25722	0.63685		
Corrected Total	8083	9187.37887			
Root MSE	0.7980	**R-Square** = 0.4404	**Durbin-Watson D** = 1.768		
Dependent Mean	2.9934	**Adj R-Sq** = 0.4397			
Coeff Var	26.6591				

Table B6: OLS regression estimates for SNNP region crop production function

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits		VIF
Intercept	Intercept	1	1.0232	0.0654	15.66	<.0001	0.8951	1.1513	0.0000
LEDUC	Log of education	1	0.0679	0.0143	4.74	<.0001	0.0398	0.0960	1.0462
LHS	Log of family size	1	0.8563	0.0263	32.53	<.0001	0.8047	0.9079	1.0818
LFERT	Log of fertilizer	1	0.0194	0.0129	1.50	0.1327	-0.0059	0.0447	3.6629
LWT	Log of seed weight	1	0.0560	0.0130	4.30	<.0001	0.0305	0.0816	2.1418
LAREA	Log of crop area	1	0.3395	0.0261	12.99	<.0001	0.2883	0.3907	1.0983
LOX	Log of oxen	1	0.1406	0.0256	5.49	<.0001	0.0904	0.1909	1.8421
ETEN2	Extension= 1	1	0.0079	0.0225	0.35	0.7228	-0.0362	0.0521	1.0429
IRRIG2	Irrigation= 1	1	0.0344	0.0262	1.31	0.1890	-0.0169	0.0858	1.4689
DAMAGE2	Crop damage= 1	1	-0.0638	0.0288	-2.21	0.0268	-0.1203	-0.0073	1.6827
OWN2	Land ownership= 1	1	-0.0353	0.0275	-1.28	0.2003	-0.0893	0.01871	1.4298

ANNEX C: The Normal Qunatile-Quantile Plots

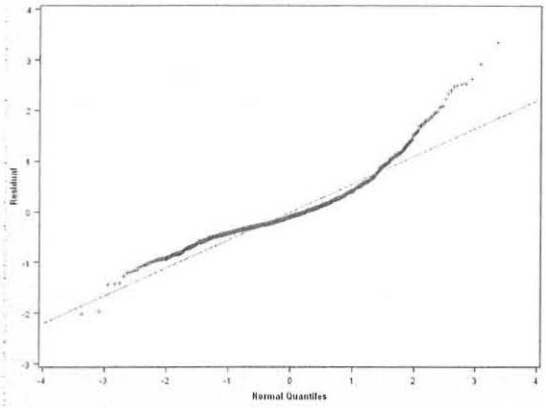


Figure C1: The normal Qunatile-Quantile plot of OLS regression for Tigray Region

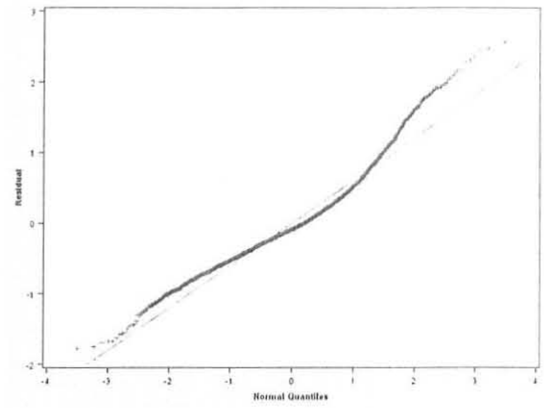


Figure C2: The normal Qunatile-Quantile plot of OLS regression for Amhara Region

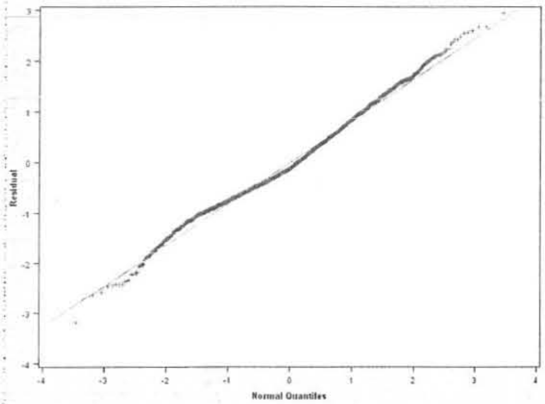


Figure C3: The normal Qunatile-Quantile plot of OLS regression for Oromia Region

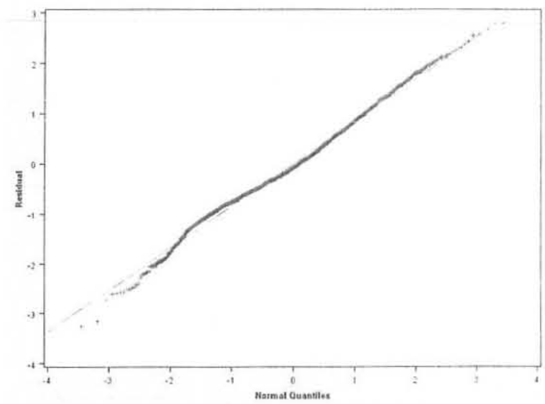
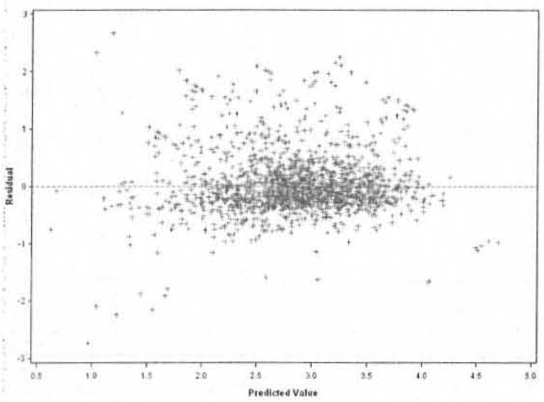


Figure C4: The Normal Qunatile-Quantile plot of OLS regression for SNNP Region

ANNEX D: Graphs for Checking Linearity and Heteroscedasticity



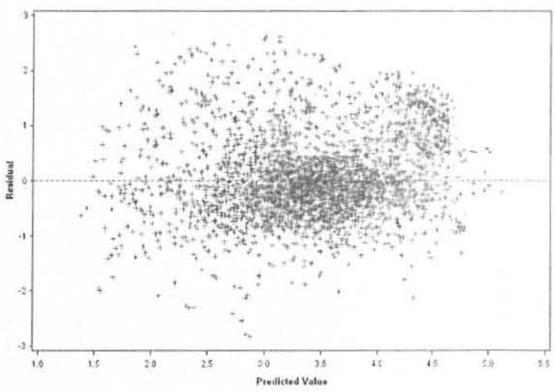
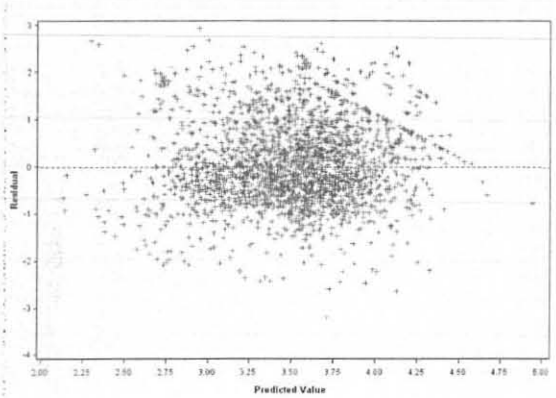


Figure D1: Graph of unstandardized OLS predicted values- Tigray region

Figure D2: Graph of unstandardized OLS residuals versus residuals versus predicted values- Amhara region



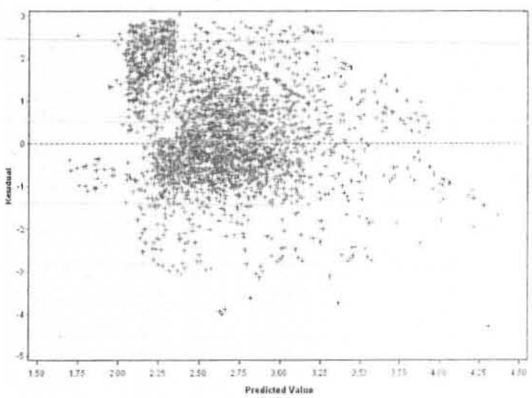


Figure D3: Graph of unstandardized OLS versus predicted values -Oromia region

Figure D4: Graph of unstandardized OLS residuals residuals versus predicted values- SNNP region

ANNEX E: Outlier Diagnostics Results

Table E1: Outlier diagnostics for Tigray region crop production function

Observation	Standardized Robust Residual	Observation	Standardized Robust Residual
148	3.4814	770	3.8977
177	-5.2942	773	3.504
182	3.8747	825	4.019
188	3.4731	841	3.3127
251	4.4537	845	-3.0142
285	4.1963	971	-3.5703
319	3.1189	977	3.0185
368	-4.184	1088	-3.0286
372	-5.2013	1091	3.6029
378	-4.3555	1101	-4.0804
571	-4.4555	1143	-3.1399
573	-4.1859	1167	4.1395
591	3.3445	1181	3.4512
592	4.4172	1182	4.2063
598	4.798	1237	3.0122
600	3.6611	1268	-3.4659
603	3.2652	1441	-6.926
604	4.0127	1454	3.1335
610	3.1649	1457	3.7919
645	4.2885	1462	3.9105
648	3.3415	1479	4.5379
650	3.3863	1486	-3.1447
701	3.4693	1515	3.1568
746	3.3751	1521	3.2396
751	3.5776	1582	4.3915
769	3.5726		

Diagnostics summary for Tigray region

Observation Type	Proportion	Cutoff
Outlier	0.0304	± 3.0000

Table E2: Outlier diagnostics for Amhara region crop production function

Observation	Standardized Robust Residual	Observation	Standardized Robust Residual	Observation	Standardized Robust Residual
5	-3.7085	1488	3.3614	1853	3.0216
33	-3.2601	1499	4.1536	1998	-4.1825
47	3.0775	1502	3.8217	2000	-3.182
56	3.4341	1513	-3.659	2039	3.0823
61	-3.5992	1553	-3.3147	2050	3.2758
90	3.1925	1569	-3.5154	2060	3.2259
93	3.4035	1570	-3.5763	2069	3.5358
95	3.2018	1572	-3.2107	2070	3.0425
96	3.0407	1583	-4.1895	2148	3.2066
135	-3.5821	1594	3.1797	2219	-3.7167
147	-3.6144	1598	3.0522	2233	-3.26
172	-3.6666	1607	3.1795	2236	-4.7201
225	4.0724	1624	-3.0035	2238	-3.891
333	-3.5943	1631	-3.8531	2240	-4.8239
399	-4.378	1635	-3.5943	2266	-3.5931
458	-3.1501	1636	-3.2227	2278	-3.4392
459	3.2685	1640	-3.4451	2299	3.1335
476	-4.0449	1641	-3.7311	2301	3.2847
478	-3.0821	1643	-3.3522	2447	-3.1031
490	-3.1955	1665	-3.0615	2562	-3.0314
595	-3.514	1678	3.6959	2563	-4.2929
615	-3.5821	1679	3.7186	2564	-3.4993
718	3.6134	1684	-3.7359	2565	-4.4349
811	-3.3319	1699	-3.4537	2567	-4.1767
888	3.2301	1714	-3.3423	2568	-3.911
893	3.4195	1734	-4.1319	2570	-5.2123
914	-3.7183	1740	-4.5281	2571	-4.2049
950	-3.538	1742	-5.1874	2572	-3.6059
983	3.5515	1744	-4.2858	2573	-5.6427
1024	-3.205	1745	-4.5525	2575	-4.5158
1056	4.0917	1746	-3.0698	2576	-3.8892
1163	-3.0846	1748	-3.9099	2672	-3.4518
1190	4.9001	1749	-4.2482	2678	-3.3798
1244	-3.4049	1768	3.0094	2681	-3.4269
1262	3.3135	1803	3.3228	2767	3.4455
1376	4.0193	1824	3.4148	2773	-3.069

Table E2 Continued...

Observation	Standardized Robust Residual	Observation	Standardized Robust Residual	Observation	Standardized Robust Residual
2883	-3.1659	4268	-3.2018	5258	-3.2291
2891	-3.7137	4297	3.4532	5260	-4.1487
2896	-3.0026	4347	-3.673	5280	3.1231
2906	3.4776	4363	-3.5858	5293	3.2305
3033	-3.6182	4369	-3.8375	5339	3.3764
3095	-3.6898	4380	-3.1225	5395	-3.5804
3281	3.2425	4515	-3.2239	5494	-3.1766
3374	3.4191	4545	-3.2574	5516	-3.3316
3377	-3.297	4596	-3.1493	5673	3.0043
3403	-3.3732	4603	-3.3093	5680	-3.4071
3448	3.109	4606	3.9554	5698	-3.4422
3449	3.0522	4614	3.9184	5773	-3.2852
3455	3.178	4626	-3.0182	5774	-3.4292
3457	3.0029	4640	-3.477	5781	-3.1637
3486	-3.5184	4744	-3.2733	5797	-3.4545
3507	-3.5452	4834	-3.5517	5800	-3.4012
3632	-3.0539	4845	-3.3059	5806	-4.0074
3657	-3.4432	4846	-3.2479	5816	-3.668
3661	-3.3482	4847	-3.2836	5873	-4.5593
3670	-3.4265	4848	-3.4215	5874	-3.3878
3904	-3.6192	4849	-3.1486	5896	-3.4457
3906	-3.4567	4873	-3.3306	5916	-3.4744
3954	3.2815	4950	-3.5075	5934	-3.5691
3956	3.3308	5200	-3.4323	5935	3.8459
4022	-3.2012	5211	-3.4272	5939	-3.5013
4109	-3.7983	5214	-3.0142	6027	-3.0853
4120	-3.9345	5250	-3.6039	6062	-3.3331
4121	-3.8648	5251	-3.1534	6080	-3.1504
4147	-3.0744			6119	-3.3536
4220	-3.7603			6154	-3.6519

Diagnostics summary for Amhara region

Observation Type	Proportion	Cutoff
Outlier	0.0318	± 3.0000

Table E3: Outlier diagnostics for Oromia region crop production function

Observation	Standardized Robust Residual	Observation	Standardized Robust Residual	Observation	Standardized Robust Residual
600	3.0418	2023	-3.7369	4853	-3.3476
605	3.0028	2031	-3.1624	4856	-4.0031
666	-3.6281	2065	3.1303	4864	-3.1331
724	3.1568	2104	3.2082	5068	-3.5318
884	-3.7667	2867	-4.2716	5149	-3.5339
930	-3.8693	2886	-3.0929	5159	-3.2007
948	-3.3329	2948	-3.5733	5284	3.1252
950	-3.0021	2953	-3.3461	5425	-3.5036
1107	-3.5465	2994	-3.1177	5428	-3.3534
1110	-3.6211	2996	-3.713	5431	-4.0576
1113	-4.3054	3005	-6.4921	5439	-3.1388
1121	-3.3866	3009	-3.5284	5643	-3.5118
1179	-3.1723	3010	-4.0926	5724	-3.5627
1325	-3.7367	3072	-3.4781	5734	-3.2007
1362	-3.1381	3074	-3.7083	5848	-4.0109
1406	-3.8018	3076	3.0743	5859	3.1053
1416	-3.2007	3077	-3.8047	6075	3.0222
1541	3.0265	3664	3.1199		
1756	-4.2336	3858	-3.5118		
1771	-3.9409	3934	-3.3816		
1776	-3.8249	3936	-3.0221		
1785	-3.1395	4093	-3.4837		
1924	-3.9803	4096	-3.3334		
1928	-4.5375	4099	-4.0376		
1938	-3.4943	4107	-3.1189		
2010	-3.9478	4567	3.1448		
2013	-3.1471	4579	-3.2746		
2022	-4.0372	4850	-3.4549		

Diagnostics summary for Oromia region

Observation Type	Proportion	Cutoff
Outlier	0.0115	± 3.0000

Table E4: Outlier diagnostics for SNNP region crop production function

Observation	Standardized Robust Residual	Observation	Standardized Robust Residual	Observation	Standardized Robust Residual
14	-3.4374	446	-4.0479	1164	-3.3835
38	-3.235	447	-3.1381	1212	-3.469
52	-3.6166	448	-3.5872	1216	-3.3835
71	-3.0068	551	-3.2773	1240	-3.1285
74	-3.4661	556	-3.2969	1250	-3.5719
112	-3.6534	561	-3.7947	1276	-3.5104
116	-4.4892	572	-3.4218	1282	-3.6152
118	-3.4589	601	-3.3835	1350	-3.5869
121	-3.2605	616	-3.6257	1360	-3.3811
122	-3.2685	646	-3.4442	1449	-3.0108
123	-3.2863	677	-3.4101	1564	-3.8204
124	-3.4607	713	-3.26	1565	-3.2452
126	-4.1084	721	-3.4555	1573	-3.6254
128	-3.4683	788	-3.2747	1576	-3.7472
135	-3.1309	792	-3.0046	1729	-3.9316
152	-3.5096	834	-3.4352	1740	-4.1721
167	-3.629	839	-3.0482	1760	-3.2164
168	-3.0808	882	-3.3179	1777	-4.3909
179	-3.548	983	-3.7801	1779	-3.554
184	-3.3639	986	-5.4083	1871	-3.8035
185	-3.5535	989	-3.6615	1874	-3.0412
264	-3.7538	995	-3.1704	1893	-4.2201
297	-3.4243	1009	-3.3742	1898	-3.0596
341	-3.0702	1031	-4.4457	1928	-3.2041
345	-3.7559	1054	-3.8072	1954	-3.8626
382	-3.2371	1068	-3.0008	1961	-3.4435
428	-3.1133	1071	-3.1576	1962	-3.4323
429	-3.1106	1131	-3.2537	1967	-3.2829
435	-3.0587	1132	-3.4233	1978	-4.0803
445	-3.4155	1161	-3.7784	2008	-3.2006

Table E4 Continued…

Observation	Standardized Robust Residual	Observation	Standardized Robust Residual	Observation	Standardized Robust Residual
2078	-4.0232	3125	-3.4977	4018	-3.2767
2107	-3.0255	3126	-3.1403	4025	-3.0051
2147	-3.196	3127	-3.301	4026	-3.0566
2157	-3.4177	3130	-3.349	4027	-3.7518
2230	-3.3356	3131	-3.3826	4029	-3.8708
2266	-3.4164	3132	-3.2882	4030	-3.8708
2295	-3.353	3133	-3.165	4034	-3.7974
2326	-3.4892	3136	-3.2224	4035	-3.471
2332	-3.4073	3196	-3.2726	4038	-3.4321
2342	-3.4435	3197	-3.2464	4052	-3.1425
2349	-3.1684	3198	-3.3226	4098	-3.1844
2363	-3.5844	3199	-3.426	4121	-3.4177
2429	-3.1294	3200	-4.2579	4124	-3.439
2445	-3.857	3207	-3.173	4126	-3.5063
2522	-3.3664	3209	-3.3174	4130	-3.2832
2527	-3.1146	3222	-4.2799	4146	-3.034
2887	-3.0503	3225	-3.3233	4151	-3.4192
2917	-3.026	3230	-3.3457	4337	-3.4607
2920	-3.3042	3236	-3.426	4355	-3.9264
2953	-3.4076	3248	-3.3181	4356	-3.4442
2964	-3.0875	3274	-3.1116	4480	-3.0194
2985	-3.2432	3596	-3.4192	4722	-3.0061
2988	-3.5949	3642	-3.3946	4771	-4.3105
3063	-3.5226	3643	-3.4595	4805	-3.1421
3072	-3.0754	3803	-3.1449	4911	-3.7273
3104	-3.433	3820	-3.0295	4946	-3.0121
3105	-3.3226	3864	-3.0546	4948	-3.8543
3106	-3.3304	3908	-3.4323	4956	-3.5156
3107	-3.388	3939	-3.7198	5014	-3.3533
3124	-3.4607	3997	-3.4535	5041	-4.1704

Table E4 Continued...

Observation	Standardized Robust Residual	Observation	Standardized Robust Residual
5052	-3.6307	5947	-3.6534
5057	-3.4122	5951	-4.5664
5062	-3.2947	5953	-3.5361
5073	-4.1661	5956	-3.1833
5074	-3.7004	5957	-3.3835
5075	-3.1978	5958	-3.4013
5076	-3.2511	5959	-3.3835
5078	-3.2453	5961	-4.1084
5079	-3.4128	5963	-3.4683
5082	-3.0712	5970	-3.1309
5086	-3.5839	5987	-3.5096
5099	-3.2924	6002	-3.7439
5117	-3.191	6003	-3.1581
5135	-3.541	6014	-3.548
5137	-3.9309	6019	-3.3639
5138	-3.172	6020	-3.5535
5140	-3.5188	6090	-3.2371
5146	-3.3551	6136	-3.1906
5148	-3.5333	6137	-3.2256
5192	-3.2497	6143	-3.1359
5306	-3.4506	6153	-3.4927
5575	-3.0598	6154	-4.0479
5727	-3.1311	6155	-3.2153
5824	-3.241	6156	-3.51
5851	-3.5146	6259	-3.3546
5875	-3.235	6264	-3.2969
5889	-3.6166	6269	-3.7947
5908	-3.0068	6280	-3.499
5910	-3.3419		

Diagnostics summary for SNNP region

Observation Type	Proportion	Cutoff
Outlier	0.0293	± 3.0000

ANNEX F: MGQ Test Results

Table F1.1: Estimation results for MGQ test (smaller group) in Tigray region

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Squares**	**F Value**	**Pr > F**
Model	10	260.37165	26.03717	112.17	<.0001
Error	703	163.17940	0.23212		
Corrected Total	713	423.55105			
Root MSE	0.48179	**R-Square** = 0.6147			
Dependent Mean	2.08806	**Adj R-Sq** = 0.6093			
Coeff Var	23.07347				

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	-0.54452	0.18040	-3.02	0.0026
E	Log of Education	1	-0.00052	0.03346	-0.02	0.9876
FER	Log of Fertilizer	1	0.06962	0.02372	2.94	0.0034
HS	Log of Household Size	1	0.43978	0.13866	3.17	0.0016
W	Log of seed weight	1	0.13910	0.04026	3.46	0.0006
AR	Log of area in hectares	1	2.29848	0.14932	15.39	<.0001
OX	Log of plowing oxen	1	0.24091	0.07260	3.32	0.0010
EXT2	Extension= 1	1	0.03877	0.03732	1.04	0.2992
IRRIGA2	Irrigation= 1	1	-0.00813	0.03662	-0.22	0.8245
DAM2	Crop damage=1	1	0.01618	0.03757	0.43	0.6667
OWN2	Land ownership=1	1	0.02957	0.03911	0.76	0.4499

Table F1.2: Estimation results for MGQ test (larger group) in Tigray region

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	10	98.05921	9.80592		<.0001
Error	726	79.61423	0.10966		
Corrected Total	736	177.67344			
Root MSE	0.33115	**R-Square** = 0.5519			
Dependent Mean	3.06887	**Adj R-Sq** = 0.5457			
Coeff Var	10.79067				

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	2.03252	0.14876	13.66	<.0001
E	Log of Education	1	-0.02510	0.02219	-1.13	0.2585
FER	Log of Fertilizer	1	0.01404	0.01453	0.97	0.3343
HS	Log of Household Size	1	-0.29971	0.07105	-4.22	<.0001
W	Log of seed weight	1	-0.02641	0.03775	-0.70	0.4844
AR	Log of area in hectares	1	1.48241	0.09047	16.39	<.0001
OX	Log of plowing oxen	1	0.12939	0.04892	2.65	0.0083
EXT2	Extension= 1	1	0.02523	0.02474	1.02	0.3082
IRRIGA2	Irrigation= 1	1	-0.03172	0.02505	-1.27	0.2058
DAM2	Crop damage= 1	1	0.01531	0.02668	0.57	0.5663
OWN2	Land ownership=1	1	0.00170	0.02656	0.06	0.9489

Table F2.1: Estimation results for MGQ test (smaller group) in Amhara region

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	10	1001.34978	100.13498	244.88	<.0001
Error	2888	1180.92317	0.40891		
Corrected Total	2898	2182.27296			
Root MSE	0.63946	**R-Square** = 0.4589			
Dependent Mean	2.70604	**Adj R-Sq** = 0.4570			
Coeff Var	23.6308				

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	0.87866	0.08154	10.78	<.0001
LED	Log of Education	1	-0.03367	0.02435	-1.38	0.1668
LFER	Log of fertilizer in Kg	1	0.02815	0.00701	4.01	<.0001
LHS	Log of family size	1	0.12471	0.03781	3.30	0.0010
LWT	Log of seed weight in Kg	1	0.11096	0.01458	7.61	<.0001
LAREA	Log of area in hectare	1	1.30614	0.04127	31.65	<.0001
LOX	Log of plowing oxen	1	0.08318	0.04074	2.04	0.0412
EXT2	Extension= 1	1	0.00896	0.02593	0.35	0.7297
IRRIG2	Irrigation=1	1	0.00147	0.02399	0.06	0.9511
DAM2	Crop damage= 1	1	-0.00847	0.02710	-0.31	0.7545
OWN2	Land ownership= 1	1	-0.02435	0.02736	-0.89	0.3735

Table F2.2: Estimation results for MGQ test (larger group) in Amhara region

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	10	725.91862	72.59186	349.82	<.0001
Error	2851	591.60997	0.20751		
Corrected Total	2861	1317.52859			
Root MSE	0.45553	**R-Square** = 0.5510			
Dependent Mean	3.08039	**Adj R-Sq** = 0.5490			
Coeff Var	14.78814				

Variable	**Label**	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	Intercept	1	2.08113	0.08532	24.39	<.0001
LED	Log of Education	1	0.01581	0.01836	0.86	0.3890
LFER	Log of fertilizer in Kg	1	0.05874	0.00439	13.39	<.0001
LHS	Log of family size	1	0.07169	0.02771	2.59	0.0097
LWT	Log of seed weight in Kg	1	-0.01910	0.01029	-1.86	0.0635
LAREA	Log of area in hectare	1	1.19082	0.02871	41.48	<.0001
LOX	Log of plowing oxen	1	-0.32176	0.05953	-5.40	<.0001
EXT2	Extension= 1	1	0.00358	0.01919	0.19	0.8520
IRRIG2	Irrigation= 1	1	-0.00751	0.01760	-0.43	0.6697
DAM2	Crop damage= 1	1	-0.02214	0.01783	-1.24	0.2143
OWN2	Land ownership= 1	1	-0.00337	0.01955	-0.17	0.8630

Table F3.1: Estimation results for MGQ test (smaller group) in Oromia region

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	10	781.25373	78.12537	224.02	<.0001
Error	2955	1030.55918	0.34875		
Corrected Total	2965	1811.81291			
Root MSE	0.59055	**R-Square** = 0.4312			
Dependent Mean	3.10212	**Adj R-Sq** = 0.4293			
Coeff Var	19.03698				

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	2.58497	0.09460	27.33	<.0001
LEDUC	Log of education	1	0.01422	0.01564	0.91	0.3634
LHS	Log of household size	1	0.08771	0.03534	2.48	0.0131
LFERT	Log of fertilizer	1	0.01964	0.01568	1.25	0.2104
LWT	Log of seed weight	1	-0.20317	0.01590	-12.78	<.0001
LAREA	Log of area in hectare	1	1.32506	0.03977	33.32	<.0001
LOX	Log of plowing oxen	1	-0.00224	0.04803	-0.05	0.9628
EXT2	Extension= 1	1	-0.09829	0.02631	-3.74	0.0002
IRRIG2	Irrigation= 1	1	-0.00007	0.02462	-0.00	0.9978
DAM2	Crop damage= 1	1	0.02117	0.02501	0.85	0.3973
OWN2	Land ownership= 1	1	0.01163	0.02611	0.45	0.6560

Table F3.2: Estimation results for MGQ test (larger group) in Oromia region

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	10	353.20472	35.32047	203.94	<.0001
Error	2961	512.81426	0.17319		
Corrected Total	2971	866.01898			
Root MSE	0.41616	**R-Square** = 0.4078			
Dependent Mean	3.48920	**Adj R-Sq** = 0.4058			
Coeff Var	11.9271				

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	1.73712	0.10282	16.89	<.0001
LEDUC	Log of education	1	-0.00046	0.01034	-0.04	0.9646
LHS	Log of household size	1	0.05061	0.02217	2.28	0.0225
LFERT	Log of fertilizer	1	0.15957	0.01837	8.69	<.0001
LWT	Log of seed weight	1	-0.02598	0.01296	-2.00	0.0451
LAREA	Log of area in hectare	1	0.72750	0.02574	28.26	<.0001
LOX	Log of plowing oxen	1	0.09953	0.02831	3.52	0.0004
EXT2	Extension= 1	1	-0.09686	0.01607	-6.03	<.0001
IRRIG2	Irrigation= 1	1	-0.01158	0.01613	-0.72	0.4729
DAM2	Crop damage= 1	1	-0.01211	0.01573	-0.77	0.4414
OWN2	Land ownership= 1	1	-0.02351	0.01727	-1.36	0.1734

Table F4.1: Estimation results for MGQ test (smaller group) in SNNP region

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	10	1871.77494	187.17749	334.58	<.0001
Error	3698	2068.80388	0.55944		
Corrected Total	3708	3940.57882			
Root MSE	0.74796	**R-Square** = 0.4750			
Dependent Mean	2.72824	**Adj R-Sq** = 0.4736			
Coeff Var	27.41548				

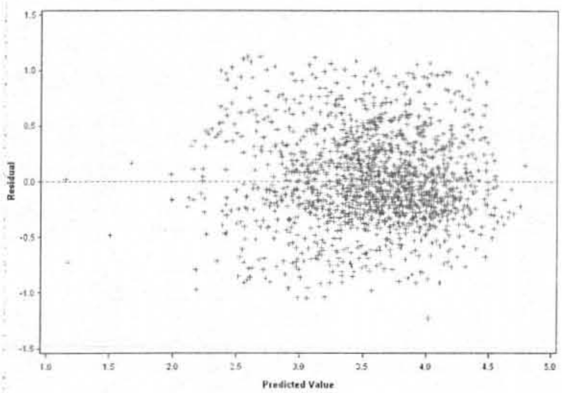
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	0.67277	0.10703	6.29	<.0001
LEDUC	Log of education	1	0.12141	0.02043	5.94	<.0001
LHS	Log of family size	1	1.07481	0.06110	17.59	<.0001
LFERT	Log of fertilizer in Kg	1	0.01270	0.01882	0.68	0.4996
LWT	Log of seed weight	1	0.00487	0.02002	0.24	0.8077
LAREA	Log of area in hectare	1	0.86194	0.04756	18.12	<.0001
LOX	Log of plowing oxen	1	-0.02231	0.04174	-0.53	0.5931
EXT2	Extension= 1	1	0.05733	0.03182	1.80	0.0716
IRRIG2	Irrigation=1	1	0.00403	0.03589	0.11	0.9106
DAM2	Crop damage=1	1	-0.03823	0.03937	-0.97	0.3315
OWN2	Land ownership= 1	1	-0.00921	0.03951	-0.23	0.8156

Table F4.2: Estimation results for MGQ test (larger group) in SNNP region

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	10	910.04095	91.00409	257.29	<.0001
Error	3578	1265.55438	0.35370		
Corrected Total	3588	2175.59533			
Root MSE	0.59473	**R-Square** = 0.4183			
Dependent Mean	3.45784	**Adj R-Sq** = 0.4167			
Coeff Var	17.19946				

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	2.75439	0.13880	19.84	<.0001
LEDUC	Log of education	1	0.02877	0.01658	1.74	0.0827
LHS	Log of family size	1	0.22172	0.06509	3.41	0.0007
LFERT	Log of fertilizer in Kg	1	-0.00111	0.01473	-0.08	0.9399
LWT	Log of seed weight	1	-0.03303	0.01403	-2.36	0.0186
LAREA	Log of area in hectare	1	0.42231	0.03131	13.49	<.0001
LOX	Log of plowing oxen	1	0.21248	0.02707	7.85	<.0001
EXT2	Extension= 1	1	-0.00083	0.02619	-0.03	0.9747
IRRIG2	Irrigation= 1	1	-0.06877	0.03240	-2.12	0.0338
DAM2	Crop damage= 1	1	-0.04474	0.03532	-1.27	0.2053
OWN2	Land ownership= 1	1	-0.04566	0.03099	-1.47	0.1408

ANNEX G: Graphs for Checking Linearity and Heteroscedasticity after Robust Regression



FigureG1: Graph of unstandardized robust residuals predicted values for Tigray region

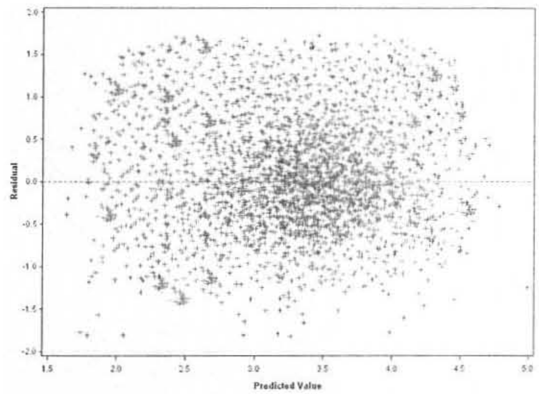


Figure G2: Graph of unstandardized robust residuals versus versus predicted values for Amhara region

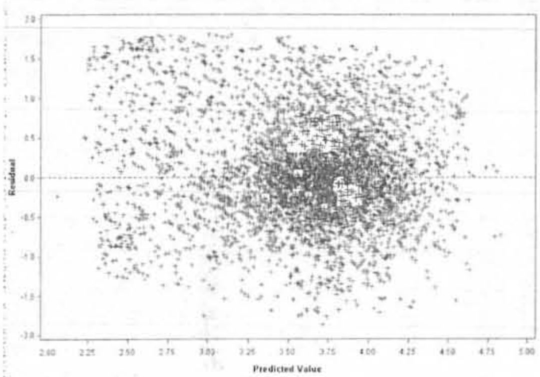


Figure G3: Graph of unstandardized robust predicted values-Oromia region

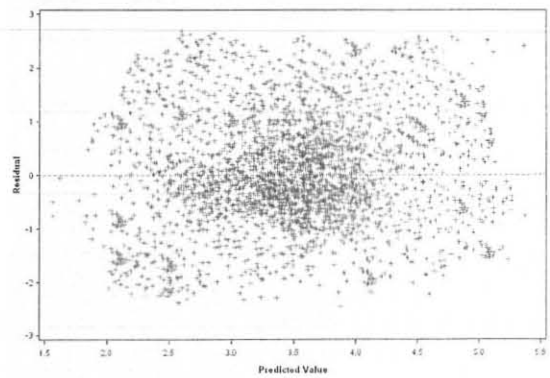


Figure G4: Graph of unstandardized robust residuals versus residuals versus predicted values-SNNP region

ANNEX H: The Normal Qunatile-Quantile Plots after Robust Regression

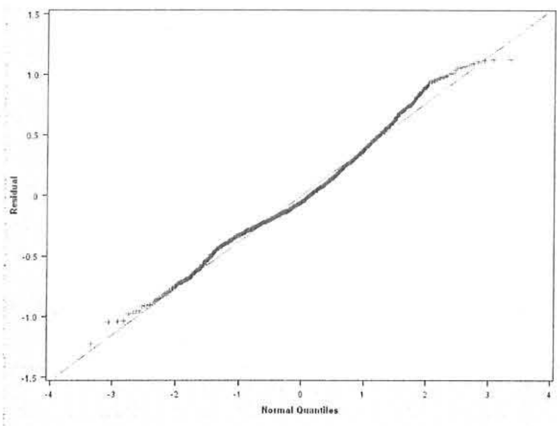


Figure H1: The normal Q-Q plot of robust regression for Tigray region

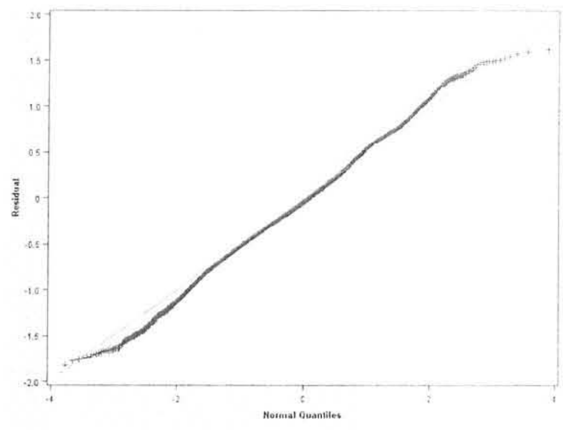


Figure H2: The normal Q-Q plot of robust regression for Amhara region

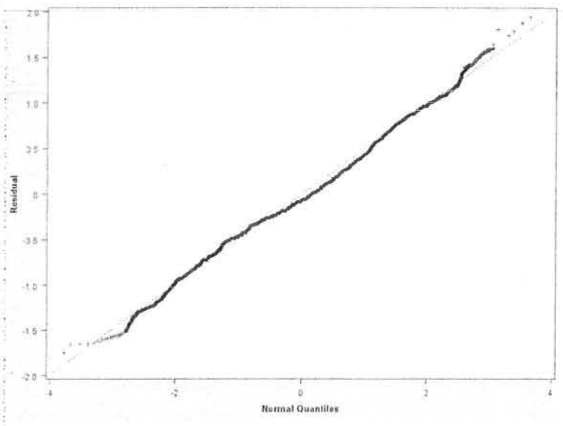


Figure H3: The normal Q-Q plot of robust regression for Oromia region

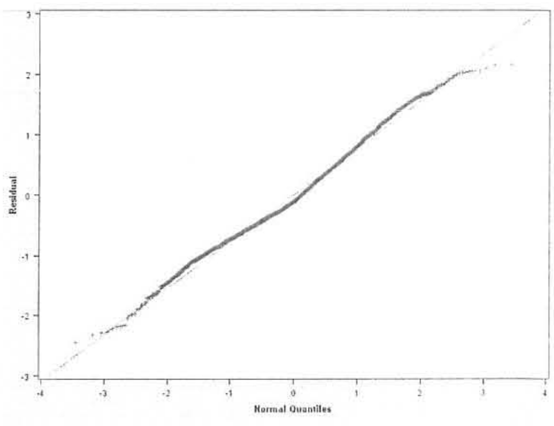


Figure H4: The normal Q-Q plot of robust regression for SNNP region

DECLARATION

I, the undersigned, declare that the thesis is my original work, has not been presented for degrees in any other university and all sources of material used for the thesis have been duly acknowledged.

Name: Taddesse Kassahun

Signature: Taddesse Kassahun

Place: Faculty of Science, Addis Ababa University

Date: 06/07/2009

This thesis has been submitted for examination with my approval as a university advisor.

06/07/2009

Fentaw Abegaz (Ph.D.)