ESTIMATION AND VALIDATION OF LINEAR AND NON-LINEAR PRODUCTION FUNCTIONS THROUGH ROBUST REGRESSION

by

Rizwan Yousuf

(J-17-D-24-BS)

Thesis submitted to
Faculty of Basic Sciences
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

IN

STATISTICS

Division of Statistics and Computer Science
Sher-e-Kashmir University of Agricultural Sciences & Technology of Jammu
Main Campus, Chatha, Jammu- 180009

2021

CERTIFICATE –I

This is to certify that the thesis entitled **"Estimation and Validation of Linear and Non-Linear Production Functions through Robust Regression"** submitted in partial fulfillment of the requirements for the degree of **Doctor of Philosophy in Statistics** to the Faculty of Post Graduate Studies, Sher-e-Kashmir University of Agricultural Sciences and Technology of Jammu, is original work and has similarities with the published work not than minor similarities as per UGC norms of 2018 adopted by the university. Further the level of minor similarities has been declared after checking the manuscript with URKUND plagiarism detection software provided by the University.

The work has been carried out by **Mr. Rizwan Yousuf,** Registration No. **J-17-D-24-BS** under my supervision and guidance. No part of the thesis has been submitted for any other degree or diploma. It is further certified that such help and assistance received during the course of investigation have been duly acknowledged.

10.09.2021
(Major Advisor)

Place: Jammu

Date: 10-09-2021

10.09.2021
Head of the Division

05/10/2021
Dean

CERTIFICATE – II

We, the members of the Advisory Committee of **Mr. Rizwan Yousuf,** Registration No. **J-17-D-24-BS**, a candidate for the degree of **Doctor of Philosophy in Statistics,** have gone through the manuscript of the thesis entitled **"Estimation and Validation of Linear and Non-Linear Production Functions through Robust Regression"** and recommend that it may be submitted by the student in partial fulfillment of the requirements for the degree.

10.09.2021

Major Advisor &
Chairman Advisory
Committee

Place: Jammu
Date: 10-09-2021

Advisory Committee Members

Dr. S.E.H. RIZVI
Professor (Statistics)
Div. of Statistics and Computer Science
(Member from Major Subject)

10/9/2021

Dr. Anil Bhat
Asst.Professor (Agril. Econ.)
Div. of Agril. Economics and ABM
(Member from Minor Subject)

10.9.2021

Dr.M.Iqbal Jeelani Bhat
Asstt. Professor (Statistics)
Div. of Statistics and Computer Science
(Member from Supporting Subject)

10.09.2021

Dr.Moni Gupta
Assoc. Professor (Biochemistry)
Div. of Biochemistry
(Dean's Nominee)

10/09/2021

CERTIFICATE–III

This is to certify that the thesis entitled **"Estimation and Validation of Linear and Non-Linear Production Functions through Robust Regression"** submitted by **Mr. Rizwan Yousuf,** Registration No. **J-17-D-24-BS** to the Faculty of Post Graduate Studies, Sher-e-Kashmir University of Agricultural Sciences and Technology of Jammu, in partial fulfillment of the requirements for the degree of **Doctor of Philosophy in Statistics** was examined and approved by the advisory committee and external examiner (s) on ....12/11/21..........

12/11/21
External Examiner

Dr. Imran Khan
Assoc. Prof. (Statistics)
SKUAST-K

24.11.2021
Dr. Manish Kumar
Major Advisor
Professor (Statistics) and Head
Division of Statistics and Computer Science

24.11.2021
Head of Division
Division of Statistics and Computer Science

24/11/2021
Dean, FBSc.

ACKNOWLEDGEMENTS

"This thesis is dedicated to my Parents"

I express gratitude to Allah The almighty for all of the blessings he has bestowed upon me throughout my career. He's been the driving force behind all I've done. It was an immense pleasure for me to get an opportunity to persue my PhD in the Division of Statistics and Computer Science Faculty of Basic Sciences, Sher-e-Kashmir University of Agricultural Sciences and Technology of Jammu

My profound gratitude and warm regards go out to my research supervisor **Dr. Manish Kumar Sharma** Professor and Head, Division of Statistics and Computer Science ,of Basic Sciences, Sher-e-Kashmir University of Agricultural Sciences and Technology of Jammu for his invaluable supervision, affection, encouragement and support throughout the period of my PhD work. He has given me all the freedom to pursue my research, while silently and non-obtrusively ensuring that I stay on course and do not deviate from the core of my research. Without his able guidance, this thesis would not have been possible and I shall eternally be grateful to his for his assistance.

All of my advisory members. Dr. S.E.H Rizvi Professor Division of Statistics and Computer Science, Dr. Anil Bhat, Assistant Professor, Division of Agricultural Economics and ABM, Dr. M.Iqbal Jeelani Bhat Assistant Professor, Division of Statistics and Computer Sciences, Dr. Moni Gupta Associate Professor Division of Biochemistry, have my heartfelt gratitude and respect for their noble thoughts during my research period.

I feel obligation to make a special mention of Dr. S.E.H Rizvi Professor Division of Statistics and Computer Science and Dr. M.Iqbal Jeelani Bhat Assistant Professor, Division of Statistics and Computer Science for their constant support and guidance throughout the course of study.

It gives me great pleasure to express my appreciation to everyone of my Seniors and juniors namely, Dr Banti, Dr Sunali, Dr Faizan Danish, Dr Archana, Mr Zakir, Mr Khalid, Mr Sajad, Mr.Arshid, Mr Younis, Ms. Afshan, Ms. Shavi Gupta, Mr. Imran, Mr. Aryan,Ms Ritika,Mr Saqib and Ms Chetna for helping me out in any possible way during the entire period of my PhD work.

I am also thankful to non-teaching staff of Division of Statistics and Computer Science Faculty of Basic Sciences, Sher-e-Kashmir University of Agricultural Sciences and Technology of

Jammu Mr.Sajad Ahmed Bhat and Mrs.Bandhana Devi, Shri Joginder Lal and Mr. Vipan Salaria for their cordial behavior during my study.

I genuinely convey my thanks and appreciation to my friends Dr zahid Nabi Sheikh, Mr. Zubair A Khan, Mr. Sheikh Sajad, Mr obaid Zaffar, Mr. Waqar ul Ashraf, Mr. Shabir A Magray, Sheikh Amjad for providing their generous support throughout the period of my PhD research work.

I convey my heartfelt gratitude to my parents and other family members for their unwavering support, love, and sacrifices during my study. Without their assistance, I would not have been able to complete the study. I admire the confidence and self-belief bestowed upon me by my father Mr. Mohd Yousuf war, My mother Mrs. Raja Bano, My sisters Ms Aarifa Yousuf, Ms Tabasum Yousuf and Ms Asiya Yousuf, My Brother Mr Nasir Yousuf and My cousin Brother Mr. Mushtaq Ahmad Rather. I wish for them blessings of Allah in this world and world here after.

The Merit Fellowship awarded to me to carry out research is gratefully acknowledged here. My Sincere thanks are extended to university for timely disbursal of the grant.

I wish to offer my thanks to all those names which are not included but will always be fondly remembered

Rizwan

Rizwan Yousuf

Place : Jammu

Date: 10|09|21

ABSTRACT

Title of the thesis	:	Estimation and Validation of Linear and Non-linear Production Functions through Robust Regression
Name of the Student	:	Rizwan Yousuf
Registration No.	:	J-17-D-24-BS
Major Subject	:	Statistics
Name and Designation of Major Advisor	:	Dr Manish Kumar Professor and Head (Statistics)
Degree to be Awarded	:	Ph.D. (Statistics)
Year of award of Degree	:	2021
Name of the University	:	Sher-e-Kashmir University of Agricultural Sciences and Technology of Jammu

As in the presence of High leverage points (HLP) i.e. outliers, the ordinary least square (OLS) method does not provide true estimates of production function. In this study, the impact of HLP have been checked in case of linear and non linear production functions viz Quadratic function (QF),Square root function (SF) and Mitscherlich–Baule (MB), Cobb Douglas(CD) respectively. The robust estimation techniques M, MM, S, LTS and OLS after replacing the HLP by robust values had provided the precise estimates of production functions as OLS method led to misleading conclusions. It has been observed that the input variables individually contribute significantly in case of quadratic function whereas the interaction is significant in case of square root production function and QF outperforms SF on the basis of high R^2 value, minimum AIC and BIC.In case of large variation, the S estimation technique outperformed than the other as observed in all types of data specifically in apple yield, followed by modified OLS after handling HLP, M and MM estimation techniques. In case of MB function, Marquardt method found best followed by Gauss Newton. The estimates of the production functions have been compared and observed that the influential observations have affected the size, sign and significance of the parameter(s).The marginal value of productivity was found to be positive in SF and MB thereby indicating that an increased used of these inputs could increase the output because these were sub optimally used. In QF and CD Marginal value of productivity was found to be negative hence indicating excess use and should be avoided to check the fall of returns in production. Lastly, Cross validation techniques consisting of validations set approach (data split) and k-folded method having training error not more than the validation error which indicated that the fitted model found to be reliable.

Key words: *Production Function Estimation, Robust Regression, High Leverage Points(HLP), Cross Validation ,MCD Distance*

Signature of the Major Advisor **Signature of the Student**

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Introduction

"Statistics are the triumph of the quantitative method and the quantitative method is the victory of sterility and death" Hilaire Belloc

Regression is the most common and useful statistical tool which can be used to quantify the relationship between a response variable (y) and explanatory variables (x). If we have dependent variable Y and independent variables X_1, $X_{2...,}X_n$ then the regression analysis is to estimate population regression function .The model generally can be expressed as:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + u_i \quad i = 1,2,\ldots,n$$

Where$\beta_0, \beta_1, \ldots., \beta_p$regression parameters and error u is normally distributed with mean zero and equal variance discussed as by Montgomery and Peck (2006). Problem that arises in regression analysis is to determine the best estimates for $\beta_0, \beta_1, \ldots., \beta_p$. If assumptions are violated.

Simple linear regression allows the mean function E(y) to depend on more than one explanatory variable and to have shapes other than straight lines although does not allow for arbitrary shapes. To this end, the seminal works of both Legendre (1805) and Gauss (1809) proposed the method of ordinary least squares (OLS) which has arguably become the most popular approach to conducting a regression analysis. This popularity is likely attributable to the fact that the OLS method can be expressed in closed form and achieve minimum variance among all unbiased estimators when the underlying error distribution follows normality. It has been observed that OLS may not be appropriate when the response variable differs from the regression function in an asymmetric manner which is commonly encountered usually in medical data, agricultural economic data and among other situations. Regression analysis is used to predict a continuous dependent variable from a number of independent variables. The independent variables used in regression can be either continuous or dichotomous. If the dependent variable is dichotomous, then logistic regression is advisable. Independent variables with more than two levels can also be used in regression analyses but they first must be converted into variables that have only two levels. This is called dummy coding. Usually, regression analysis is used with naturally occurring variables, as opposed to experimentally manipulated variables.

In lieu of these deficiencies, a general regression methodology is proposed which allows for the possibility of non-normal tail behaviour and asymmetry in the conditional distribution of y given x, but will still perform well for symmetric and/or normally distributions as per Yang *et al*., (1993).The goal of regression is to estimate the parameters of a model relating two sets of variables, given a training data set. However, the presence of outliers in the training dataset will make this estimate unreliable. Real world data are almost always corrupted with outliers and hence robust parameter estimation is of paramount importance Mitra *et al.* (2010).One of the most important statistical tools is a linear regression analysis for many fields. Nearly, all regression analysis relies on the method of least squares for estimation of the parameters in the model. A problem that we often encountered in the application of regression is the presence of an outlier or outliers in the data. Outliers can be generated by from a simple operational mistake to including a small sample from a different population they result in serious effects on statistical inference. Even one outlying observation can destroy least squares estimation resulting in parameter estimates that do not provide useful information for the majority of the data. Robust regression analyses have been developed as an improvement to least squares estimation in the presence of outliers and to provide us information about what a valid observation is and whether this should be thrown out. The primary purpose of robust regression analysis is to fit a model which represents the information in the majority of the data. The properties of efficiency, breakdown point and bounded influence are used to define the measure of robust technique performance in a theoretical sense. Efficiency can tell us how well a robust technique performs relative to least squares on clean data (without outliers). Linear regression models are widely used in many fields of science such as engineering, economics, sociology, health, etc. Due to the simplicity of the idea behind the OLS method, the minimization of the sum of squared residuals, and the interpretability of the final model parameter estimates, the OLS method is very popular among practitioners. However it is also well known that outliers considered here are heterogeneous observations in comparison with the majority of the data and might strongly affect these estimators. Not all outliers are influential and not all influential points will hurt our modelling, but we still need robust methods to prevent possible influential outliers and bad leverage points from leading us to the wrong choice of model. Statisticians who neglect to consider robust methods are ignoring the fact that the methods they are using can perform very poorly when their data does not strictly satisfy the assumptions of their chosen method as per S.Wesiburg (1985), Roncheeti *et al.* (1997) and William (1983). Then, robust estimators are regarded as more reliable and most commonly used statistical techniques. Robust methods may not be as efficient as other methods that require strict assumptions. When those assumptions are met, we can have better estimation with the help of the robust method. Out of many possible regression techniques, the ordinary least squares (OLS) method has been generally adopted because of being a conventional/standard method and its ease of computation. However, OLS estimation of regression weights in the multiple regression is affected by the occurrence of outliers, non-normality, multicollinearity and missing data. Robust estimation methods comprise those which down the weight observations with extreme residuals and eradicate the observations pointed out as an outlier detection method. Robust regression is a method of regression analysis considered to avoid some restrictions of traditional parametric and nonparametric methods. Certain widely used method of regression is ordinary least squares and has properties if their fundamental premises are true but can yield misleading results in case premises are not reliable. Robust regression methods are designed to be not overly affected by violations of assumptions.

Robust regression analysis provides an alternative to the least square regression model when fundamental assumptions are unfulfilled by the nature of the data Marona *et al. (*2006).The analyst estimates his statistical regression models and tests his assumptions and frequently finds that the assumptions are substantially violated. Sometimes, the analyst can transform his variables to conform to those assumptions. However, a transformation will not eliminate the leverage of influential outliers that bias the prediction and distort the significance of parameter estimates. Under these circumstances, robust regression is resistant to the influence of the outliers and may be the only reasonable recourse. The well –known methods of robust estimation are M estimation, S estimation and MM estimation. M estimation is an extension of the maximum likelihood method and is a robust estimation. According to Levin (1998) the use of least squares method will not be suitable in resolving problems having outlier or extreme observation, because the assumption of normality cannot be held. The M-estimation anticipates this by defining a function of $\rho(\varepsilon)$, which is called the Huber function. In the M-estimation, $b_0, b_1, \ldots b_p$ are estimates of $\beta_O, \beta_1, \ldots, \beta_P$ chosen such that $\sum \rho(\varepsilon)$ is minimum and Huber function is defined as.

$$\rho(\varepsilon) = \begin{cases} \varepsilon^2 & if -k \le \varepsilon \le k \\ 2k|\varepsilon| - k^2 & if \ \ \varepsilon < -k \ or \ \varepsilon > k \end{cases}$$

Where $k = 1.5\hat{\sigma}$. To estimate σ , if it is to be needed we use $\hat{\sigma}$ = 1.483.Median of Absolute Deviation(MAD) is the median of the remaining absolute M-estimators of a large class of estimators which are acquired from a given data set by minimizing the sum of certain functions of the data. M-estimation is derived from statistical function like finding the value of a parameter allowing the maximum likelihood function equal to zero. The S estimation was proposed by Rousseeuw and Yohai, (1984).It is based on residual scale of M estimation. The weakness of M estimation is the lack of consideration on the data distribution and not a function of the overall data because only using the median as the weighted value. This method uses the residual standard deviation to overcome the weaknesses of median as discussed by Susanti *et al.* (2014). MM estimation technique is to estimate the regression parameter using S estimation which minimize the scale of the residual from M estimation and then proceed with M estimation. MM estimation targets to obtain estimates that have a high breakdown value and more efficient. The Breakdown value is common measure of the proportion of outliers that can be addressed before these observations affect the model as discussed by C. Chen (2002). Further, MM estimation is the development of M estimation method. By using these robust methods, it is possible to eliminate some of the data which in some cases could not always be done additionally if that data are important, such as those which often be found on agricultural field. Robust regression methods attempt to fit regression model to data so that the fit is less sensitive to the behaviour on the tails of the errors and is reasonably stable when the errors come from a high –tailed distribution. M estimation a generalized version of maximum likelihood (ML) is a popular robust procedure used in regression analysis as per Lange and Sinsheimer (1993) for using adaptive robust regression methods that are based on ML estimation. Adaptive methods provide an attractive compromise between classical normal theory methods and modern robust methods discussed by

Lange and Sinsheimer (1993) and Hogg (1974).In adaptive estimation, the distribution of the errors depends on tuning parameters that are estimated or set to achieve robust inference for statistical modelling of data sets involving errors with longer than normal tails. The best studied adaptive methods are based on the distributional families of normal/independent type as proposed by Large and Sinsheimer (1993) and Dempster *et al.* (1977). Consider the regression model $y_i = f_i(\theta)$+e_i,Where y_i are the observed dependent variable, f_i are the theoretical response functions with a known form generally depending on a set of explanatory variable x_i and indexed by a set of parameters θ, and $e_i$denote the errors. In adaptive robust regression the classical normal theory assumptions of $e_i \sim^{iid}$N (0,σ^2) is commonly replaced by the assumption that e_i are identically and independently distributed (iid) with a distribution from a N/I family. A popular example from the N/I family, is the student‘s t distribution.

Robust regression is a regression method that is used when the distribution of residual is not normal or there are some outliers that affect the model. This method is an important tool for analyzing the data which is affected by outliers so that the resulting models are stout against outliers. The transformation appeared unlikely to eliminate or reduce the influence of outliers, which finally became inaccurate predictions, when researchers set up regression models and tested the common assumption that the regression assumptions are violated. Under these circumstances, robust regression is resistant to the influence of outliers is the best method .Robust regression is used to detect outliers and provide results that are resistant to outliers. The properties of efficiency, breakdown, and high leverage points are used to define robust technique performance in a theoretical sense. One aim of robust estimators is a high finite sample breakdown point defined by Donoho and Huber (1983).The breakdown point can be defined as the point ε_n or limiting percentage of contamination in the data at which any test statistics first becomes swamped as per Alma (2011). Robust regression methods provide an alternative to least squares regression by requiring less restrictive assumptions. These methods attempt to dampen the influence of outlying cases in order to provide a better fit to the majority of the data. Outliers have a tendency to pull the least squares fit too far in their direction by receiving much more "weight" than they deserve. Typically, you would expect that the weight attached to each observation would be on average 1/n in a data set with n observations. However, outliers may receive considerably more weight, leading to distorted estimates of the regression coefficients. This distortion results in outliers which are difficult to identify since their residuals are much smaller than they would otherwise be if the distortion wasn't present. As scatter plots may be used to assess outliers when a small number of predictors are present. However, the complexity added by additional predictor variables can hide the outliers from view in these scatter plots. Robust regression down-weights the influence of outliers, which makes their residuals larger and easier to identify. Least sum of squares may strongly be influenced by the presence of outliers. Robust regression has been developed to handle contaminated data, i.e., data containing outliers. Many of the robust methods are based on the median statistic. Among them are the single median, the repeated median and the least median of squares methods. The first two methods were developed for straight line models, least median of squares (LMS) were developed for models linear in the parameters P. Vankeerberghcn *et al.*(1995). The breakdown point as defined by Hampel (1985) i.e., the smallest fraction of contaminated data which leads to large deviating model parameters changes from 0 percent for LS to 50 percent for LMS. Practically, this means that up to 50 percent of the data set may be contaminated with outliers before LMS breaks down and produces aberrant parameter estimations P.Vankeerberghcn *et al.* (1995). In contrast with the LS principle which minimises the sum of squared residuals, residual $(\sum_{i=1}^{n}(res)^2)$ LMS minimises the median of squared residual $[med(res)^2]$the median is defined as the $\left[\frac{n}{2}+1\right]^{th}$ ranked value. Once the robust model is obtained, outlier detection can be applied. A robust estimator of the pure error is computed from the median of squared residuals and is compared with each residual to define whether the corresponding data point is outlying or not discussed by P. Vankeerberghcn *et al.*(1995).Robust regression techniques are critical to fitting data with noise in real-world applications. A robust estimation technique is essentially a method which tolerates the presence of data points not obeying the model whose parameters are to be estimated. Such techniques are widely used today in computer vision C. V. Stewart (1999). The theory of robust estimators was developed in the last 30 years in statistics and two main classes of methods, M-estimators and least median of squares (LMedS) were successfully applied to vision problems. However, some of the "most" robust techniques used in the vision community. Robust regression through the use of functions related to least squares has been the subject of intense research. Dempster *et al.* (1980) discussed statistical properties of the estimates under the assumption that the observation errors are independent normal. Dianne P (1990) developed a high quality set of routines to compute robust estimators for eight weight functions. Robust regression techniques are a class of estimators that are comparatively indifferent to the presence of one or more outliers in the data. They are specially well suited to data that require large numbers of statistical tests and may contain outliers due to factors not of experimental interest Wager *et al.*(2005).Some popular statistical packages have procedures for robust regression analysis. Among them are SAS, STATA, S-PLUS, LIMDEP, and E-Views. They need to know in which statistical package the type of robust regression appropriate for that particular application can be found. There is a family of robust regression analysis that replaces the sum of squared errors as the criterion to be minimized with one less influenced by outliers. Least absolute deviations, sometimes called L_1 regression, is alternative method that pursues to minimize the influenced of outliers. Weighted least squares can be employed to serve this purpose. There is a method called iteratively reweighted least squares using robust corrections for influence of the outliers Chatterjee and Mächler (1995). In many real-world applications, such linear regression models suffer from two drawbacks: lack of robustness to outliers/noises and the curse of dimensionality. A typical solution to the former is to estimate noises under an assumed parametric distribution such as Gaussian, whereas a solution to the latter is to select appropriate features by using dimensionality reduction such as Principal Component Analysis (PCA). However, if there exist a small number of gross outliers among the samples, the estimate of model parameters would drift obviously. Moreover, the linear regression models often do not work well in processing high dimensional data. For regression tasks on high-dimensional data like face images, we often cannot collect and label enough samples. Robust methods have been studied widely to overcome the impact of outliers/noise Zhang.*et al.* (2017).In order to develop a robust regression procedure one could consider two competing approaches; i.e., perform the regression analysis with respect to multiple loss functions or allow the characteristics of the data to dictate the selection of the loss function. In order to improve the efficiency of quantile regression. Zou and Yuan (2008) introduced composite quantile regression (CQR), which optimizes over a sum of multiple quantile loss functions. As a robust regression procedure, CQR combines the strength of multiple quantile regressions to estimate the same “slope” coefficients across different quantiles. Kai *et al.* (2010) adapted CQR to the local polynomial framework and established that for many common non-normal errors this extension provided for gains in estimation efficiency when compared to its local LS counterpart. Regretfully, when implementing CQR it is still unclear how many quantiles should be used and simply increasing the number of quantiles does not necessarily improve the efficiency of the estimator Kai *et al.* (2010). Alternately, one could consider a convex combination of loss functions Zheng *et al.* (2013) extended CQR by embedding the usage of an empirically weighted average of quantile loss functions and the LS loss function, so that the LS loss tends to be weighted heavier for normally distributed data. Rather than using several quantiles, tact would be to let the data select the quantile of interest in quantile regression Yang *et al.* (2006). We use simulation as an instructional tool in our study. Simulation studies are computer experiments which involve creating data by pseudo-random sampling from known probability distributions. They are an invaluable tool for statistical research, particularly for the evaluation of new methods and the comparison of competing methods. Simulation studies are much used in the pages of Statistics in Medicine, it has been seen that many statisticians lack the necessary understanding to execute a simulation study with confidence. Proper understanding of simulation studies would enable people to run simulation studies themselves and to critically appraise published simulation studies. Issues with design and reporting, lead to uncritical use or appraisal of simulation studies. In this context, better understanding of the rationale, design, execution, analysis and reporting of simulation studies is necessary to improve what researchers can learn from them. Simulation studies are used to obtain empirical results about the behaviour of statistical methods in certain scenarios, as opposed to analytic results. It is not always possible to obtain analytic results, or may be extremely difficult. Simulation studies come into their own when methods make wrong assumptions or data are messy; that is, they can assess the resilience of methods discussed by Morris *et al.* (2017). This is not always possible with analytic results, which may assume that data arise from a specific model. Simulation studies are empirical experiments, and statisticians – particularly those doing working in applications such as clinical trials – should be familiar with good practice regarding design, analysis, presentation and reporting as previously lamented by Hoaglin and Andrews (1975),Hauck and Anderson(1984), Ripley (1987),Burton *et al.* (2006) and Koehler *et al.*(2009). For example, few reports of simulation studies acknowledge that Monte Carlo procedures will give different results when based on a different set of random numbers; failing to report measures of uncertainty would be unacceptable in medical research. By keeping above facts, the study "Estimation and validation of linear and non linear production functions through robust regression" has been taken into consideration with the main objectives as

- Identification of the bad leverages in the data and to estimate the robust values.
- To estimate the parameters of linear production functions through various robust estimation techniques.
- To estimate the parameters of Non-linear production functions through various robust estimation techniques.
- To compare the estimated parameters of linear and non-linear production functions through different techniques with OLS.
- Validation of the proposed linear and non-linear production functions with respect to best robust technique.

Limitations of the Study

I. Due to unavailability of road transport because of COVID -19 restrictions unable to consider other districts for collection of the data therefore nearby district Budgam had been selected in case of survey.

II. The most common robust methods and production functions are used only in the study.

III. The more time for research discussion was online rather than in person discussion with respondents.

IV. In case of survey the numbers of respondents taken in the study are limited due to in person interview and due to pandemic condition.

Chapter -2

Review of Literature

The relevant literature to study were collected and presented in chronological order related to regression modelling, robust regression, leverages and production functions. The review in detailed as per the study are:

Holland (1973) described that when there are weights associated with each observation, the formulas for and derivation of ridge regression methods were developed. The motivation is Bayesian and many k values are discussed. It is suggested that ridge regression be combined with robust regression methods.

Pollak and Wachter (1975) discussed the home production function approach and its application to time allocation are criticised. Many applications of the model, particularly those involving implicit "commodity prices," are said to necessitate the household's technology having constant returns and no joint production; otherwise, implicit commodity prices are said to be dependent on the household's preferences as well as its technology and market goods prices. Furthermore, in situations involving the allocation of time, joint production is common. In cases where home production theory does not provide a sufficient analytical framework.

Aigner*et al.* (1977) explained that previous investigations of the so-called border production function have not used an appropriate characterisation of the disturbance term. They define the disturbance term as the sum of symmetric normal and half normal random variables, which is an adequate definition. The coefficient of a production function with an additive disturbance term of this type is then estimated using maximum-likelihood estimation.

Harrell *et al.*(1984) had employed a generic metric of predictive discrimination to assess a model's ability to predict survival in an independent test sample of patients suspected of having coronary artery disease based on training samples of varied sizes. After creating clinical indices from variable clusters, they examined three approaches of model fitting: traditional ‘step-up' variable selection, incomplete principle components regression and Cox model regression. In the test sample, we discovered that regression using principal components provided superior predictions whereas regression using indices provided easily interpretable models that were nearly as excellent as the principal components models.

Chaloner and Brant (1988) developed a method for finding outliers in a linear model. An outlier is an observation that has a substantial random error and was created by the linear model in question. The posterior distribution of random errors is examined to find outliers. As a graphical aid in locating outliers, an augmented residual plot is also suggested.

Wooldridge (1990) suggested for (potentially) dynamic econometric models, proposes a generic approach to robust, regression-based specification testing. The suggested tests are distinguished by the fact that, in addition to estimation under the null hypothesis, computation only necessitates a matrix linear least squares regression followed by an ordinary least squares regression comparable to those used in common non robust tests. The proposed statistics are robust to deviations from distributional assumptions that are not being evaluated in the leading cases of conditional mean and/or conditional variance tests. Furthermore, the statistics can be obtained using any vT-consistent estimator, simplifying some otherwise complex —contexts significantly. Conditional mean tests for models estimated by weighted nonlinear least squares under conditional variance misspecification, tests of jointly parameterized conditional means and variances estimated by quasi-maximum likelihood under nonnormality, and some new, computationally simple specification tests for the tobit model are among the examples covered.

Battese (1992) described that during the last two decades, econometric research has focused on the modelling and estimation of frontier production functions. He gave overviews of the concepts and models involved, as well as citations to some of the empirical applications that had been published at the time.

Huggins (1993) suggested that M-estimators that are resistant to contamination by outliers are constructed using the mentioned robust variant of the log likelihood for multivariate normal data. The robust estimators are found using a minimization technique that keeps the multivariate normal approach's flexible parameterizations. The estimators' asymptotic properties are determined, and the estimates' computation and usage in outlier identification tests are explained.

Fung (1993) defined that identifying several outliers and leverage points is challenging. For unmasking these facts, Rousseeuw and van Zomeren advised employing high breakdown robust estimating methods such as the least median of square and minimum volume ellipsoid. These methods have a tendency to classify too many observations as extreme, but a stepwise analysis is presented here for confirming outliers and leverage points found using robust methods.

Hadi and Smirnoff (1993) developed two test approaches for detecting numerous outliers that appear to be less prone to this issue. Both approaches try to split the data into a set of "clean" data points and a collection of data points with probable outliers. Using a suitably scaled version of the prediction error, the probable outliers are then examined to see how extreme they are in comparison to the clean subset. The data sets and Monte Carlo simulations show that both processes are effective in detecting multiple outliers in linear models and outperform alternative methods, such as robust fit methods (e.g., least median of squares residuals). The methods, in particular, do not necessitate a predetermined number of outliers to test for, do not necessitate an estimator's efficiency level, do not necessarily require Monte Carlo to determine cutoff values, are not computationally intensive, and are relatively resistant to both masking and swamping effects.

Battese and Coelli (1995) obtained that non-negative technical inefficiency effects are believed to be a function of firm-specific factors and time in a stochastic frontier production function for panel data on companies. The inefficiency effects are assumed to be distributed independently as truncations of normal distributions with constant variance and means that are a linear function of observable variables. This panel data model is an extension of recently proposed cross-sectional data models for inefficiency effects in stochastic frontiers.

Hubert and Rousseeuw (1996) developed a robust regression method for conditions in which the regressors are both continuous and binary. Encoding one or more category variables frequently results in the latter. In the first phase, they compute robust distances in the space of continuous regressors to reduce the weight of leverage points. Then, as a function of the continuous and binary regressors, they execute a weighted least absolute values fit. Finally, the error scale is calculated with confidence. They focus on the two-way model, in which the suggested estimator is compared to an algorithm that alternates between treating continuous and categorical data.

Koutsias and Karteris (1998) determined that using two Landsat-5 Thematic Mapper (TM) images to create a logistic regression model for burned area mapping. The spectral channels of the two photographs were used as explanatory variables in logistic regression models. The overall accuracy of the results, as well as other statistical indicators, suggest that logistic regression modelling may be used to map burned areas successfully. The model with the spectral channels TM4, TM7, and TM1 and an overall accuracy of 97.62 percent was found to be the most appropriate.

Croux and Haesbroeck (2000) used an influence function technique, they investigated robust estimators of the covariance or correlation matrix. If the robust M estimator is more efficient than the least squares estimator, then the corresponding ridge type M-estimator described here is better than the ordinary ridge estimator with k chosen appropriately in terms of mean squared error criterion.

Blundell and Bond (2000) analyzed the estimation of Cobb-Douglas production functions using panel data from a large number of enterprises over a short period of time. When utilizing the usual first differenced estimator, the generalized technique of moments (GGM) of estimators has been found to yield substantial finite sample biases. By using realistic stationarity constraints on the initial conditions procedure, these biases can be drastically eliminated. They discover that the extended GGM estimator's extra instruments produce substantially more accurate parameter estimations.

Zaman *et al.* (2001) showed that in applied econometrics, robust regression approaches are rarely used. They describe a method based on Rousseeuw and Van Zomeren that eliminates many of the challenges associated with applying such techniques to economic data.

Chen *et al.* (2001) highlighted that many structures can appear in the data in many vision issues (e.g., stereo, motion), in which case several instances of the same model must be recovered from a single data set. The robust statistical methods typically utilised in the vision community, on the other hand, tend to fail when the measurement noise gets very great in comparison to the gap between the structures. All of these techniques are particular examples of the general class of M-estimators with auxiliary scale, and their failure in the presence of noisy multiple structures is explained in this study. To deal with data with different structures, vision-based techniques (Hough and RANSAC) should be supplemented with statistically robust methods. The implications of our study are demonstrated by presenting a simple process for 2D multi structured data that is incompatible with all currently techniques available.

Salibian and Zamar (2002) presented a new computer-intensive method for estimating the robust regression estimates distribution. Our strategy is based on bootstrapping a reweighted representation of the estimates. They include the auxiliary scale estimate in our reweighted representation of the estimations to obtain an asymptotically valid bootstrap approach. Because we only have to solve a linear system of equations for each bootstrap sample, the procedure is computationally simple. Outlying observations are given tiny weights because the weights employed are diminishing functions of the absolute value of the residuals. As a result, the bootstrap method is immune to the presence of outliers in the data. The quantile estimates obtained with this approach have greater breakdown points than those obtained with the bootstrap. They demonstrated their methodology on two datasets and presented the results of a Monte Carlo experiment on confidence intervals for linear model parameters.

Hubert and Branden (2003) explained the partial least squares regression technique (PLSR) was created to deal with high-dimensional regressors and one or more response variables. Because of its speed and efficiency, they introduced robustified versions of the SIMPLS method, which is the leading PLSR algorithm. SIMPLS' conclusions are influenced by anomalous observations in the data set since it is based on the empirical cross-covariance matrix between the response variables and the regressors and linear least squares regression. A robust covariance matrix for high-dimensional data and robust linear regression are used to create two robust algorithms, RSIMCD and RSIMPLS.The effectiveness and resilience of the novel methodologies are demonstrated by several simulation results and the analysis of real data sets. RSIMPLS is the overall best approach because it is nearly twice as quick as RSIMCD.

Levinsohn and Petrin (2003) examined approaches for modelling the production function for cognitive success in a way that captures theoretical conceptions that child development is a cumulative process influenced by family and school inputs as well as innate ability. It analyses various approaches dealing with data limitations and clarifies the defining assumptions required to justify alternative approaches. It has been demonstrated that commonly used specifications impose restrictive assumptions on production technology.

Battese*et al.* (2004) proposed a model of a metafrontier production function for enterprises in various groups with varying technologies. The metafrontier model allows for the calculation of similar technological efficiencies for businesses using various technologies. The model also allows for the estimation of technology gaps for enterprises under various technologies in comparison to the prospective technology available to the industry as a whole.

Bloom *et al.*(2004) suggested that macroeconomists have identified two characteristics as basic components of human capital: work experience and health. They calculated a production function model of aggregate economic growth that included both variables. They claim that the life expectancy effect in growth regressions appears to be a true labour productivity effect, rather than a proxy for worker experience.

Petrin *et al.* (2004) recommended that we should consider for the association between input levels and productivity when estimating production functions. Increases in productivity prompt profit-maximizing enterprises to increase their use of factor inputs. Methods that ignore endogeneity, such as OLS and the fixed-effects estimator, will produce erratic estimates of the production function's parameters. Levinsohn and Petrin (2003) present an estimate that uses intermediate inputs to proxy for the unobservable productivity factor, building on the work of Olley and Pakes (1996). The majority of plant-level databases include information on the consumption of intermediate inputs like energy and materials.

Wager *et al.*(2005)used both simulations and experimental data and found that robust estimating methods can provide significant advantages in neuroimaging analyses. Robust approaches are ideally suited for situations in which data may contain artifactual outliers, and automated robust techniques, such as IRLS, provide significant gains when each regression analysis cannot be reviewed individually for assumptions violations. Neuroimaging data falls into each of these categories.

Nakaya *et al.* (2005) introduced GWPR and its semi-parametric form as a novel statistical tool for analysing illness maps emerging from spatially non-stationary processes. The method is a sort of conditional kernel regression that estimates spatial changes in Poisson regression parameters using a spatial weighting function. It allows us to create surfaces of local parameter estimates that represent regional differences in illness rates and socioeconomic factors.As a result, the method can be used to evaluate the widely held assumption that the processes being represented are stable in space, which is commonly taken for granted in global spatial data modelling. It can also be used to highlight areas of the research region where 'interesting' linkages are present and where more inquiry is needed. Traditional global modelling can often miss such exceptions; hence GWPR provides disease analyzers with a valuable new set of statistical tools.

In analytical chemistry, outliers of various types are frequently seen in experimental data Rousseeuw *et al.* (2006).Outliers can impact the outcomes of the most commonly used chemometrical/statistical approaches, which are sensitive to them. They give an overview of reliable chemometrical/statistical methods that look for the model that fits the majority of the data and are thus less affected by outliers. As an added bonus, outliers can be identified by their considerable divergence from the robust fit. They also talked about how to perform multiple linear regression, PCA, PCR, PLS, and classification using robust approaches for estimating position and scatter.

Motulsky and Brown (2006) described a novel strategy for detecting outliers while using nonlinear regression to fit data. They used a robust type of nonlinear regression to fit the data, assuming that the dispersion follows a Lorentzian distribution. They came up with a new adaptive technique that becomes more resilient as time goes on. Adapted the false discovery rate technique to handle multiple comparisons to define outliers. They next eliminate the outliers and use ordinary least-squares regression to assess the data. The ROUT method is named for the fact that it combines robust regression and outlier removal. Our approach discovers (falsely) one or more outliers in only around 1–3% of experiments when evaluating simulated data with Gaussian scatter. The ROUT approach works effectively at outlier identification when examining data contaminated with one or more outliers, with an average False Discovery Rate of less than 1%.

Preminger *et al.*(2007) explained least squares estimation approach, like other ordinary estimation methods for regression models, can be adversely influenced by a small number of outliers, resulting in poor out-of-sample projections. They proposed using a robust regression strategy based on the S-estimation method to build forecasting models that are less susceptible to outlier data contamination. To investigate the predictability of two exchange rates over one, three, and six months, a robust linear autoregressive (RAR) and a robust neural network (RNN) model are estimated. In terms of prediction accuracy and sign predictability measures, they compare the resilient models to the random walk (RW), standard linear autoregressive (AR), and neural network (NN) models. At all time horizons, we find that robust models tend to increase the AR and NN predicting accuracy. At all projection horizons, robust models have been found to have strong market timing abilities.

Verardi and Croux (2008) suggested that the existence of outliers in a dataset can cause the classical least-squares estimator to be highly distorted, resulting in incorrect conclusions. The rreg and qreg commands in Stata provide access to several of these methods. Unfortunately, these strategies are only successful against certain sorts of outliers and are ineffective in other cases. They also presented more powerful robust estimators that can be used in Stata. They also showed a graphical tool that distinguishes the types of outliers that have been found.

Yasar *et al.* (2008) developed a semi parametric approach for reliably estimating production function parameters and obtaining reliable productivity measurements by correcting for such biases. This research examines the method before introducing a Stata command to apply it. When simultaneity and selection biases are not adjusted for, the coefficients for variable inputs are biased upward while the coefficients for fixed inputs are biased downward, according to the researchers.

Xu *et al.* (2008) presented a feature-wise disturbance resilient least-squares regression They show that this formulation leads to tractable convex optimization problems, and we show that the robust problem is equal to regularised regression for a specific uncertainty set (Lasso). This is how Lasso is interpreted from the standpoint of robust optimization. They extend this robust formulation to include a wider range of uncertainty sets, resulting in tractable convex optimization problems. As a result, they offer a new approach to developing regression algorithms that generalise existing formulations. The benefit is that robustness to disturbance is a physical quality that can be exploited: they utilise it directly to illustrate Lasso's sparsity qualities, as well as to prove a general consistency result for robust regression issues, including Lasso, from a unified robustness approach.

Yuan and Zhong (2008) defined that outliers are data points that stray from the factor model rather than the data cloud's centre. Through study, the impacts of each type of outlying observation on the maximum likelihood estimator based on the normal distribution and the accompanying likelihood ratio statistic are investigated. The distinction between outliers and leverage observations also helps to clarify the responsibilities of three robust Mahalanobis distance-based techniques. All of the robust processes are built to reduce the impact of outlier observations.

Corsi and Reno (2009) suggested a dynamic model for financial market volatility with three components: continuous volatility, leverage, and jumps, each having a heterogeneous structure. They discovered that each of the three components plays an important role in volatility forecasting, and that ignoring one of them might lead to poor predicting results. Importantly, the leverage effect has extraordinary predictive power for negative historical returns at all frequencies investigated, revealing a novel heterogeneous structure in the leverage effect. They further shown that the occurrence of leaps is significant for two reasons, utilising simulation studies: Explicitly modelling jumps has two effects: first, it trims the dynamics of the persistent volatility component, and second, it has a favorable and significant impact on future volatility.

Ferrari and Neto (2010) proposed that using a parameterization of the beta law that is indexed by mean and dispersion parameters, presents a regression model where the response is beta distributed. The proposed model is useful in situations where the variable of interest is continuous and restricted to the range (0,1), and is related to other variables via a regression structure. Unlike the parameters of a linear regression that uses a transformed response, the beta regression model's regression parameters can be interpreted in terms of the response mean and, when the logit link is utilised, an odds ratio. Maximum likelihood is used to estimate the data. They gave closed-form equations for the scoring function, Fisher's information matrix, and the inverse of Fisher's information matrix. Hypothesis testing is done with approximations derived from the maximum likelihood estimator's asymptotic normality.

The breakdown of a generalised leverage matrix useful for identifying important units and observations in linear mixed models is shown to identify large leverages for both the marginal fitted values and the random effect component of the conditional fitted values Nobre and Singer (2010). They use a simulated scenario as well as real data to demonstrate the distinct uses of the two components of the decomption.

Nguyen and Welsch (2010) suggested that the outlier detection was used to investigate the robust covariance estimation problem. They vary from others in that they introduced a weight vector on the observations and used that weight vector to build an optimization problem where outliers are given lower probabilities at the optimal solution. It was discovered that solving a system of nonlinear equations is comparable to solving this optimization issue. To solve the system, they employed the Newton–Raphson method. Because the optimization model is analogous to a semi-definite programming problem, the technique performs well in terms of computing.

Epple *et al.* (2010) has addressed how to estimate the housing production function. The fundamental issue with estimating is that housing prices and quantity are never observed independently. They created a new technique that allows us to identify and estimate the underlying production functions without depending on strong functional form assumptions by treating prices and quantities as latent variables. The technique's main finding is that change in land prices and housing values per unit of land is enough to identify the housing supply function per unit of land. If there is a continuous return to scale in the production functions.

Wang *et al.* (2011) studied a bivariate Student-t distribution which was utilised to simulate the error innovations of the return and volatility equations in a heavy-tailed stochastic volatility (SV) model with leverage impact. The bivariate Student-t distribution was expressed as a scale mixing of bivariate normal distributions by Choy et al. (2009). They suggest a new formulation in which they first derive a conditional Student-t distribution for the return and a marginal Student-t distribution for the log-volatility, and then express these two Student-t distributions as a scale mixture of normal (SMN) distribution. Their approach distinguishes the sources of outliers and enables for differentiation between outliers generated by the return process and outliers generated by the volatility process, which is an advance over Choy et al approach. (2009). Additionally, using the WinBUGS programme, it enables for efficient model implementation.

Alma (2011) proposed that when evaluating data, the ordinary least squares estimate is the optimal method for obtaining regression weights in classical multiple regressions assuming assumptions are met. However, sample estimates and outcomes can be misleading if the data does not meet some of these assumptions. Outliers, in particular, defy the least squares regression assumption of normally distributed residuals. Outlying data, both in the direction of the dependent and explanatory variables, pose a hazard to least squares regression since they might have a significant negative impact on the estimate and go unreported. As a result, statistical strategies for dealing with or detecting outlier findings have been created. Robust regression is a useful technique for assessing data that has been tainted by outliers. It can be used to detect outliers and produce resistant outcomes when outliers are present. In all of the linear regression cases, the simulation research is utilised to determine which methods work best.

Fox and Weisberg (2011) proposed a anticipated estimator or statistical process which is robust if it offers meaningful information even if some of the assumptions used to justify the estimating approach are not relevant. The majority of this appendix is about robust regression which are estimate approaches for linear regression models that are unaffected by outliers and have potentially large leverage points.

Ramalho and Silva (2011) described that in regression analysis of capital structure choices and suggested linear models used. Models like the tobit, the fractional regression model, and its two-part form are a superior choice because leverage ratios are proportionate and bounded. They go over the basic econometric assumptions and features of those models, present a theoretical framework for their application in leverage ratio regression analysis, and go over various statistical tests that can be used to evaluate their specification. Carry out a comprehensive comparison of the alternative models using a dataset previously considered in the literature, finding that the most relevant functional form issue in this framework is the choice between a single model for all capital structure decisions and a two-part model that explains the decisions to issue debt and, conditional on the first decision, the amount of debt.

Huang *et al.*(2012)proposed discriminative approaches (e.g., kernel regression, SVM) have been widely employed to handle problems like object detection, image alignment, and pose estimation from pictures, according to the researchers. Image features (X) are commonly mapped to continuous (e.g., pose) or discrete (e.g., object category) values using these approaches. Existing discriminative approaches have a key flaw in that samples are immediately projected into a subspace, failing to account for outliers that occur often in practical training sets due to occlusion, specular reflections, or noise. It's vital to note that current discriminative techniques assume noise-free input variables X.Thus; discriminative algorithms face severe performance deterioration when gross outliers are present. Despite its obvious importance, the subject of robust discriminative learning has been relatively neglected in computer vision. They provide an effective convex technique that exploits recent breakthroughs in rank minimization to create the theory of Robust Regression (RR). Robust linear discriminant analysis, regression with missing data, and multi-label classification are among the challenges that the framework can solve in computer vision. To demonstrate the benefits of RR, several synthetic and real-world examples are presented, including head pose prediction from photos, image and video classification, and facial attribute classification with missing data.

Bayes *et al.* (2012) proposed Markov Chain Monte Carlo (MCMC) method, a Bayesian inference approach is proposed. Simulation studies on the impact of outliers were conducted using contaminated data and four perturbation patterns to generate outliers. The results show that the Beta rectangular regression model appears to be a new robust alternative for modelling proportion data, and that the Beta regression model is sensitive to regression coefficient estimation, to the posterior distribution, and to the posterior distribution. Two applications are also provided to demonstrate the Beta rectangular model's robustness.

Ostertagova (2012) suggested that polynomial regression model were chosen because it is useful when there is reason to believe that the connection between two variables is curvilinear. The description of the relationship between strains and drilling depth was used to use the polynomial regression model. The model's parameters were calculated using the least square method. Following the fitting, the model was tested using some of the most common metrics for evaluating regression model accuracy.

Imon and Hadi (2013) proposed leverages values which are employed in regression diagnostics as measurements of uncommon data in the x-space. Due to their role in hiding outliers, detecting high leverage observations or points is critical. High leverage points (HLP) in linear regression are those that are far from the data's centre (mean), and hence the most extreme locations in the covariate space have the largest leverages. But Hoseemer and Lemeshow Applied logistic regression, Wiley, New York, (1980) pointed out that in logistic regression , The leverage measure incorporates a component that might make the leverage values of real HLP appear to be very small, creating concern in case identification. Attempts to identify the HLP using median distances from the mean have been made, but because they are built for identifying a single high leverage point, they may not be very useful in the presence of several HLP due to their making (false-negative) and swamping (false – positive)effects.

Griffin and Brown (2013) established a strong Bayesian prior for regression that combines aspects of both ridge and g-prior regressions while allowing for flexible weighting between them The prior spanning the entire parameter space is singular in the p > n 1 scenario, which shrinks away components with no information. This is necessary if we want to build priors that are dependent on the design matrix and so have a structure comparable to the g-prior. Because the prior is spherical and can be extended to unnamed components, a simple non-singular extension of the Ridge-CNG is simple to construct, even if it has no influence on prediction mean square error.

El Karoui (2013) studied regression M-estimates A scalar random variable whose deterministic limit $r_\rho k$ can be studied when p=n→κ>0 plays a central role in this representation. We discover a nonlinear system of two deterministic equations that characterizes $r_\rho k$ interestingly, the system shows that $r_\rho k$ depends on ρ through proximal mappings of ρ as well as various aspects of the statistical model underlying our study. Several surprising results emerge. In particular, they show that, when p=n is large enough, least squares becomes preferable to least absolute deviations for double-exponential errors.

Lewis and Ward (2013) recommended regression modelling which is one of the most extensively used methodologies in epidemiological research. It provides a tool for detecting statistical connections, which may then be used to study potential causal associations relevant to disease control. Multivariable regression has long been the conventional model, having a single dependent variable (typically disease) and numerous independent variables (predictors). Multivariable regression can be generalised to multivariate regression, with all variables possibly statistically dependent. This provides a much richer modelling framework. We examine and contrast these approaches using a number of basic illustrative cases. While a relative newcomer to the epidemiological literature, the technological methodology utilised to accomplish multivariate regression is well established – Bayesian network structure discovery – and has a long history in computing science. Multivariate analysis in epidemiological studies can help researchers better understand disease processes at the population level, which can lead to the development of more effective disease management and prevention strategies.

Marubini and Orenti (2014) studied ordinary least square (OLS) regression analysis, claimed that identifying and assessing outliers plays a key role. In regression analysis, they provide a robust two-stage technique for identifying outlying observations. Through a strong distance estimator based on Minimum Covariance Determinant, the exploratory stages identify leverage points and vertical outliers (MCD). Following the deletion of these points, the confirmatory stage runs an OLS analysis on the remaining sample of data to see how adding back in the previously removed observations affects the results. Bootstrapping is used to establish cut-off points for various diagnostics, and the cases are clearly labelled as good-leverages, bad-leverages, vertical outliers, and usual cases. Through the use of jack-knife following bootstrap robust cut-off points, this approach is able to find and correctly categorize vertical outliers, good and bad leverage points.

Alguraibawi *et al.* (2015) proposed new strategy for identifying poor leverage points has been proposed. The conventional ordinary least square (OLS) versus Mahalanobis distance (MD) graphic fails to correctly identify the problematic leverage points the majority of the time. The robust least median of square (LMS) versus robust mahalanobis distance (RMD) plot successfully classifies observations into four categories. They suggest novel modified generalised studentized residuals versus diagnostic robust generalised potentials (DRGP) in this regard, which is highly effective in classifying observations into regular observations, vertical outliers, and good and bad leverage points. The MGt-DRGP plot can correctly detect problematic leverage points with a low rate of masking and swamping, according to a Monte Carlo simulation research. It's important to notice that the OLS-MD has a masking problem, whereas the LMS-RMD has a swamping problem.

Midi and Mohammed (2015) presented a novel method for determining which leverage points are excellent and which are bad. They used certain well-known data sets to test the performance of the suggested technique. They proposed a new diagnostic plot-based method for identifying poor leverage points. The traditional diagnostic plot misses the poor leverage points. They offer MGti-DRGP, which is particularly effective in classifying observations as regular, vertical outliers, good, and bad leverage points.

The impact of financial crises on two key aspects of stock returns, the risk-return trade-off and the leverage effect, was investigated Christensen *et al.* (2015). For daily stock return data, they use the fractionally integrated exponential GARCH-in-mean (FIEGARCH-M) model, which incorporates both properties and allows for the coexistence of long memory in volatility and short memory in returns. They adapt this model to accommodate for especially in the economic parameters that determine the volatility-in-mean effect and the leverage effect during financial crises. An analysis of the daily U.S. stock index return data from 1926 to 2010 reveals that during crises, both financial effects grow dramatically. Surprisingly, the risk-return trade-off is considerably favorable only during financial crises, whereas non-crisis times are minor. Throughout, the leverage effect is negative, but during financial crises, it increases by around 50% in magnitude. During NBER recessions, no such changes are found hence financial crises are unique in this regard. The increase in the leverage effect is confirmed by applications to a number of major developed and emerging international stock markets, but international evidence on the risk-return relationship is mixed.

Nurunnabi *et al.* (2016) suggested a well-known masking and swamping effects, detecting multiple atypical observations such as outliers, high leverage points, and influential observations (IOs) in regression remains a difficult task for statisticians. They proposed a six-fold charting strategy based on the well-known group deletion approach to classify regular observations, outliers, high leverage points, and IOs simultaneously in linear regression, and they presented a robust influence distance that can identify numerous IOs.

Imon and Apu (2016) proposed that previous cut-off point logically appears to be erroneous; a new robust Mahalanobis distance cut-off point was proposed to discover several high leverage locations. The proposed method is empirical and based on a non-parametric approach. As a result, there is no need for a table and it is very simple to calculate. A number of well-known data sets clearly demonstrate the benefits of employing the proposed cut-off point rather than the existing one. When they used the existing cut-off values for leverages measures to compute robust Mahalanobis distances, it swamps a substantial amount of observations, regardless of whether we use the MVE or the MCD. When they used the proposed cut-off point, however, it successfully identifies all high-leverage situations.

In a very broad global stock universe Guerard (2016) used robust regression approaches to predict stock returns and build stock selection models. To a worldwide stock universe, he applies Markowitz portfolio creation and optimization approaches. Using a given stock selection model, he evaluated expected return models in the worldwide market and generated statistically significant active returns from various portfolio creation strategies. He claimed that in a global stock market, mean-variance procedures with strong regression-created expected returns continue to construct portfolios capable of generating excess returns above transaction costs and statistically meaningful asset selection. The stock selection model utilised in this analysis is the McKinley Capital Management proprietary model in its "public form." Our proprietary model produces very statistically significant asset selection and factor risk exposures, but the public form model does not.

Yu and Yao (2017) proposed that Ordinary least-squares (OLS) estimators for a linear model are extremely sensitive to outliers or unexpected values in the design space. Even a single typical value can have a significant impact on parameter estimates. This article will evaluate and detail some widely used robust approaches, as well as some newly invented ones, as well as compare them in terms of breakdown point and efficiency. Furthermore, a simulation study and a real-data application were utilised to compare the performance of existing robust approaches in various scenarios.

Diskin *et al.* (2017) suggested deep neural networks in the context of robust regression was considered with outlier-corrupted observations in the usual linear model. They propose a deep neural network that generalises Huber's robust regression and the classical least trimmed squares estimator and achieves good accuracy with low processing complexity. A single training phase is used to train the network for arbitrary linear models. The network can manage a wide range of Signal-to-Noise Ratio (SNR) and is robust to different sorts of outliers, according to numerical testing with synthetic data.

Chaudhuri and Basu (2017) explained the Self Updating Process (SUP) is a clustering technique that starts with a set of data points and replicates the process as the data points move about and cluster themselves. Using this Self Updating Process, they looked at the challenge of robust regression in a simple linear configuration. Chen (2014) has previously used the algorithm in certain real-world scenarios. However, the technique has some unresolved issues with the scale parameter in the weight function in question, and it can be highly unstable in the face of significant leverage points. The issues can be solved, at least in part, by selecting a weight function that is a decreasing function of the data points' robust Mahalanobis Distance and leverage values. An comprehensive simulation study was conducted to investigate the empirical efficiency of the improved SUP algorithm, demonstrating that it is a good compromise between efficient approaches such as the ordinary least squares (OLS) and robust methods such as the least median of squares (LMS).

Cadila Jan *et al.*(2017) presented the aggregate production function, particularly in its Cobb-Douglas and more general CES forms, is still quite popular. However, there is a long-standing and widespread criticism of the function. The core of the critique that has persisted until now is that neither the Cobb-Douglas nor the CES aggregate production functions are anything more than disguised income identities. Such a truth would have significant ramifications for modern macroeconomics, both in theory and in practise. They used the standard method to estimate Cobb-Douglas and CES production functions. Then, instead of capital stock and labour, they employed capital and labour services. They came to the conclusion that when elasticities equal factor shares, it is required to adopt the latter strategy to deal with the factor's utilisation in order to produce statistically sound findings. However, as predicted by the critique, doing so revealed that the modified estimate is really just a disguised estimate of income identity.

Bhatia *et al.* (2017) studied the robust regression problem, offered the first efficient and provably consistent estimator. Because of its relevance in settings with corrupted data and addressing model mis-specifications, the field of robust learning and optimization has sparked a lot of attention in the learning and statistics fields in recent years. The fundamental topic of robust linear regression, where estimators that can withstand corruption in up to a certain fraction of the response variables are frequently investigated, has gotten a lot of attention. Surprisingly, no polynomial time estimator that provides a consistent estimate in the presence of dense, unbounded corruptions has been discovered too far. The consistency studies a novel two-stage proof technique that includes a detailed examination of the stability of ordered lists, which may be of interest in and of itself. They demonstrated that CRR is not only consistent in its estimations, but also outperforms various other newly published algorithms for the robust regression problem, such as extended Lasso and the TORRENT method, empirically. In comparison, CRR offers comparable or better model recovery but with runtimes that are faster by an order of magnitude.

Deyoreo and Kottas (2018) presented quantile regression modelling, a Bayesian nonparametric methodology. They suggested nonparametric prior probability models, in particular, allow the data to influence the error density, resulting in more reliable predictive inference than models based on parametric error distributions. They examine quantile regression extensions for data sets with censored observations. They also use dependent Dirichlet processes to create quantile regression models that allow the error distribution to shift non-parametrically when the covariates change. Markov chain Monte Carlo algorithms are used to implement posterior inference.

Schulz *et al.*(2018)introduced Gaussian process regression as a powerful technique for modelling, exploring, and exploiting unknown functions. Gaussian process regression (GPR) is a non-parametric Bayesian approach to regression issues that can be used in both exploration and exploitation scenarios. This tutorial attempts to provide you a basic understanding of these techniques. They introduced Gaussian processes, which generate distributions over functions used in Bayesian non-parametric regression, and demonstrated their use in applications and didactic examples such as simple regression problems, a demonstration of kernel-encoded prior assumptions and compositions, a pure exploration scenario within an optimal design framework, and a bandit-like exploration–exploitation scenario.

Al Sayed *et al.* (2018) demonstrated that the presence of outliers in panel data may violate the standard OLS estimator's premise. They may cause the relationship between variables to be overestimated. The goal of the study was to use several diagnostic tools to find outliers, leverages, and influencing points. The goal of the study was to compare the results of utilising a sample with outliers and samples after eliminating the outliers to see how outliers affected the OLS estimator. The studies are performed on environmental science data in order to estimate the relationship between CO_2 emissions, energy consumption, and economic growth for two sets of countries: developed and developing.Because it employs robust mean and robust variance estimation, the Robust Mahalanobis distance approach surpasses the other methods in finding outliers. Furthermore, after removing the outliers from the analysis, the estimated model reflected a more accurate estimate of yield.

Chapter-3

Material and Methods

In order to achieve the objectives the following methodology has been used.

3.1 Identification of bad leverages:

There are several types of outliers in regression problems which include residual outliers, vertical outliers, and high leverage points. The term "residual outlier" refers to any observation with a big residual. Vertical outliers (VO), also known as y-outliers, are observations that are extreme or outlying in y-axis. High leverage points (HLPs)on the other hand are those observations which are extreme or outlying in X-Coordinate. HLP can be classified into good leverage points (GLP) and bad leverage points (BLP). GLP are outliers in the explanatory variables that follow the majority of the data's pattern, whereas BLP are the absolute opposite. The computed values of various estimates are more affected by BLP. The bad and good Leverages have been identified by using Residual analysis, Standardized Residuals, Studentized Residuals and Deleted residuals. Because the bad leverages suffer from a masking effect, these approaches may fail to detect high leverage points. As a result, in addition to the methods mentioned above, the effective steps were used to identify the bad leverages. The following techniques have been used for identification of HLP.

3.1.1 Weighted sum of squared distance

The weighted sum of squared distance of i^{th} point from the center of data (WSSD) say$WSSD_I$=$\sum_{j=1}^{K}\frac{\widehat{\beta}_j(\bar{X}_{ij-x_j})^2}{\sqrt{MSE}}$. It was proposed by Wood (1980**)** used to locate the points that are remote in X-space. The general procedure to rank the points in increasing order $WWSD_I$ and concentrate on points for which the statistic is large.

3.1.2 Hat matrix

It was proposed by Hoaglin and Welsch (1978).The role of hat matrix H=X$(X'X)^{-1}X'$in identifying influential observation. The diagonal elements of H matrix are called the Hat values denoted by h_{ii} given by: h_{ii}=$x_i{}^T(X^TX)^{-1}x_i$, $i = 1,2,3 \ldots \ldots n$ Since H determines the variance and covariance of $\acute{y}$and e, since v(y) $=\acute{}\sigma^2 H$ and v(e)=$\sigma^2(1-H)$.The elements $h_{ij}$of H may be interpreted as the amount of leverage exerted by y_j on $\acute{y}_i$.

3.1.3 Cook's distance

Cook's distance was proposed by Cook (1979) useful for identifying outliers in the observations for predictor variables. It also shows that the influence of each observation on the fitted response values. An observation with Cook's distance larger than three times the mean Cook's distance might be an outlier. Cook's distance is the scaled change in fitted values. Each element in Cook's distance is the normalized change in the vector of coefficients due to the omission of an observation.A conventional cut-off point is 4/n, where n is the number of observations in the data set. Cook gave the method for evaluating the influential observations $D_i = \frac{f_i^2}{p}\frac{h_{ii}}{(1-h_{ii})}$, i=1,2….n where f_i^2 is the i^{th} studentized residual and h_{ii} is the i^{th} diagonal element of H.In statistics,cook's distance is a commonly used estimate of the influence of the data point when doing least square regression analysis. Cook's distance measures the effect of deleting a given observation

3.1.4 Potential Matrix

A single case deleted measure called potential matrix proposed by Hadi (1992). The diagonal elements of potential matrix denoted by p_{ii} are given by $p_{ii} = x_i^T(XX_i^TX_i)^{-1}$ i=1, 2…, n, Where $X_i$is the matrix X excluding the i^{th} row.We can rewrite p_{ii}as a function of h_{ii}as $p_{ii} = \frac{h_{ii}}{1-h_{ii}}$.The potential matrix is efficient to identify a single outlier.

3.1.5 Robust Mahalanobis distance

It is proposed by Rousseeuw and Leroy (1987)**.**In multivariate space MD is the distance between two locations in multivariate space. The independent variables in a regression equation constitute a multidimensional space in which each observation can be plotted. Construct a point that represents the means of all independent variables. The term "centroid" refers to the mean point in multidimensional space. The MD is the distance between an observation and the centroid as specified by the independent variables that are correlated. The putative HLPs are identified using this method, which is based on Rousseeuw and Leroy's minimal volume ellipsoid (MVE).The means for all independent variables. In multidimensional space this mean point is known as 'centroid'.The MD is the distance of a observation from the centroid defined by the correlated independent variables

In this method the suspected HLP are determined on the base of the minimum volume ellipsoid (MVE) developed by Rousseeuw and Leroy as

$$RMD_i = \sqrt{[X - T(X)]^T [C(X)]^{-1} [X - T(X)]}$$

Where, T(X) and C(X) are robust locations and shape estimates of the MVE respectively. Habshah *et.al (2009*) suggested using the following cut-off value for the robust Mahalanobis distance $Median(RMD_i) + 3MAD(RMD_i)$where, MAD is the median absolute deviation.

3.2 Estimation of Robust values

The selected data incorporated with the identified bad leverage and the high leverage points have been replaced by robust values using the linear interpolation method. Linear interpolation is a curve fitting technique that uses linear polynomials to produce new data points within the range of a discrete set of known data points. Interpolation or predicting the value of a function for an intermediate value of the independent variable is frequently necessary. Linear interpolation necessitates the knowledge of two points as well as their constant rate of change. Interpolate values between those two points using this data. The assumption underlying linear interpolation is that the rate of change between known values is constant.

3.3 Estimation of the Linear and non-linear productions functions

The yield response curve also known as the production function is a non-negative mathematical function that connects the quantity of inputs used to the quantity of output produced. It assumes technical efficiency and calculates the maximum output for various input combinations, establishing a boundary between attainable and unattainable outputs. In most cases, a production function is supposed to specify the highest output possible from a given set of inputs. As a result, the production function depicts a boundary or frontier that represents the maximum amount of output that can be obtained from each possible combination of input. A production function can also be defined as the specification of the minimum input requirements required to produce specified output quantities. The assumption that maximum output can be obtained from given inputs allows economists to ignore the technological and managerial issues involved in achieving such a technical maximum, allowing them to focus solely on the problem of allocative efficiency, which is concerned with the economic decision of how much of a factor input to use, or the degree to which one factor can be substituted for another. The output-to-input relationship in a production function is non-monetary; that is, a production function connects physical inputs to physical outputs without accounting for prices or costs. The maximum output that may be produced with a particular input combination is determined by the production function. Technology is the most fundamental component. Its utilization determines the level of detail and precision of the production function. In a general theoretical context, it is given in more generalised terms than in specific empirical applications. Production function parameters computed using profit maximisation assumptions are "as if" production function parameters. There is a distinction to be made between technological and behavioral relationships. Behavioral relationships are those in which technology and behavioral assumptions are both present. The behavioral approach of production function parameters evaluated under duality is significant. Even production theory acknowledges that strict profit maximisation is impossible; behaviour must adjust to uncertainty and risk, and new models of production behaviour under uncertainty are developing. Bounded rationality is a concept introduced by Simon (2005).He claimed that humans' ability to process and interpret data is limited. As a result, selections are imperfect and reflect this limitation. One of the difficulties is determining the factors that influence production decisions and understanding what motivates producers to make judgments. Under risk aversion and restricted analytic capacities, decisions may differ from those made under rigorous profit maximisation. Under differing behavioral assumptions, the same technology relationships may produce distinct consequences. However, separating the behavioral and technological components to the reported outcome is extremely difficult. The linear and non-linear production functions that are widely used in practice are Quadratic, Square root Mitscherlich – Baule function and Cobb Douglas. Their Mathematical forms are as:

3.3.1 Quadratic production function

It consists of an additive composition of the input factors, their squared values, and an additional interaction term. The latter elucidates whether the input factors are independent of each other or not. The quadratic production function is formally defined as follows:

$$Y = \alpha_0 + \alpha_1 * x_1 + \alpha_2 * x_2 + \alpha_3 * {x_1}^2 + \alpha_4 * {x_2}^2 + \alpha_5 * x_1 * x_2$$

Where, Y denotes study variable $x_1$and x_2 are the causes used to study the Y.

3.3.2 Square root production function

It is very similar to the quadratic form but produces different shapes of the curves. The square root form is defined as follows:

$$Y = \alpha_0 + \alpha_1 * {x_1}^{1/2} + \alpha_2 * {x_2}^{1/2} + \alpha_3 * x_1 + \alpha_4 * x_2 + \alpha_5 * (x_1 * x_2)^{1/2}$$

To ensure decreasing marginal productivity of each input factor, the parameters must satisfy the same conditions as for the quadratic form, and their interpretation is identical

3.3.3 Mitscherlich-Baule production function

Mitscherlich-Baule production function allows for a growth plateau, which follows from the Von Liebig approach to production functions. Moreover, this functional form is characterized by continuously positive marginal productivities of the input factors. It does not exhibit negative marginal productivities, as the above polynomial forms. Formally, the Mitscherlich-Baule function is given by

$$Y = a * \left(1 - \exp\left(-b * (c + x_1)\right)\right) * (1 - \exp\left(-d * (e + x_2)\right))$$

Where, a representing the growth, and $c\ and\ e$ the natural input endowments that include $x_1$and x_2.The coefficients c and $e$describes the influence of the corresponding input factors on the study variable.

3.3.4 Cobb Douglas production function

The Cobb-Douglas production function is based on the empirical study of the American manufacturing industry made by Paul H. Douglas and C.W. Cobb (1928). It is a non linear homogeneous production function of degree one which takes into account two inputs, labour and capital, for the entire output. One of the most popular functions used of recent years is the Cobb-Douglas.

$$P(L, K) = bL^{\alpha}K^{\beta}$$

Where, P =total production (the monetary value of all goods produced in a year), L is the labor input (the total number of person-hours worked in a year), K is the capital input (the monetary worth of all machinery, equipment and buildings), b is the total factor productivity and α and β are output elasticities of labour and capital respectively. The researchers created a model economy in which neoclassical theory would provide a poor causal explanation for the data. Yet the Cobb-Douglas still produced a remarkably good fit. These findings demonstrate that the Cobb-Douglas functional form is adaptable enough to fit data effectively even when it lacks a relevant economic interpretation. If several theories can explain a set of empirical data, we should choose the one that has the fewest assumptions, according to Occam's razor. Indeed, the presence of an accounting identity that occurs in any actual economy, regardless of the underlying production technology, is the simplest explanation for the Cobb-Douglas strong empirical fit in many real world studies.

3.4 Estimation of the parameters of linear and non-linear models has been done through the following ways in order to see efficiency of the parameters.

3.4.1 Ordinary Least Square (OLS)

$$Y = x\beta + \epsilon \qquad\qquad Y = \begin{bmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ y_n \end{bmatrix}$$

$$X = \begin{bmatrix} 1 & X_{11} & X_{21} & \dots & X_{k1} \\ 1 & X_{12} & X_{22} & \dots & X_{K2} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ 1 & X_{1n} & X_{2n} & \dots & X_{kn} \end{bmatrix} \qquad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \cdot \\ \cdot \\ \beta_K \end{bmatrix}$$

Where, $\acute{\beta} = (X'X)^{-1}X'Y$

The "goodness of fit" of the model are described by the quantity R^2 and adjusted R^2.In SLR model, R^2 was plainly the square of the correlation coefficient between the Independent and the dependent variable. This, however, will not work with the MLR model. R^2 is usually defined as the proportion of variation in the dependent variable that can be explained by regression on all of the model's independent variables. In an MLR model, adding more independent variables always raises the R^2 value. Using more independent variables, on the other hand, is not recommended. The modified R^2 can be used to determine the proportion of variance explained by the number of independent variables.

3.4.2 Robust Estimation Techniques

As the data are contaminated through high leverage values so the robust estimation techniques as per the similarity of the models have been used.

3.4.2.1 M Estimation:

One of the robust regression estimation approaches is M estimation. Huber (1973) introduced it, and it is the most basic technique both computationally and theoretically. The letter M denotes maximum likelihood type estimation. If estimator at M estimation is $\acute{\beta}=\beta_n(x_1, x_2, \dots, x_n)$then, $E[\beta_n(x_1, x_2, \dots, x_n)] = \beta$ shows that the estimator $\acute{\beta} = \beta_n(x_1, x_2, \dots, x_n)$ is unbiased andhas minimum variance, so M-estimator has the smallest variance estimator compared to other estimators of variance:

$$Var(\hat{\hat{\beta}}) \geq \frac{[\bar{\beta}']^2}{nE\left(\frac{d}{d\beta} lnf(x_i; \beta)\right)^2}$$

Where$\hat{\hat{\beta}}$is an extension of the maximum likelihood estimate method and a robust estimation see other linear and unbiased estimator for estimating (Susanti 2008). It is feasible to remove certain data using this method, which is not always suitable to do, especially if the data or seed being eliminated is valuable, as is often the case in agriculture (Susanti *et al* 2009).

3.4.2.2 MM Estimation

Yohai (1987) introduced the MM Estimation, which combines high breakdown value estimation and M estimation. When the errors have a normal distribution and the breakdown point is 0.5, MM Estimation has high efficiency properties. The robust MM regression method produces a model with a structure that is nearly equal to that of a regular linear regression model. The MM estimation technique involves employing S estimation to estimate the regression parameter, which reduces the scale of the residual from M estimation. The goal of MM estimation is to produce estimates with a high breakdown value that are more efficient.

Breakdown value is a common measure of the proportion of outliers that can be addressed before these observations affect the model.MM-estimator is the solution of

$$\sum_{i=1}^{n} \rho_1' \,(ui) X_{ij} = 0$$

$$\sum_{i=1}^{n} \rho_1' \left(\frac{Y_i - \sum_{j=0}^{k} X_{ij} \,\widehat{\beta_J}}{s\, MM} \right) X_{ij}$$

Where $s\, MM$ the standard deviation is obtained from the residual of S estimationandρ is a Turkey's biweight function.

3.4.2.3 Least trimmed squares (LTS) Estimation

Least trimmed squares (LTS) is a statistical technique for estimating unknown parameters in a linear regression model that offers a more reliable alternative to the traditional regression method of minimizing the sum of squared residuals. Rousseeuw introduced the LTS Estimation, which is a high breakdown value method. The breakdown value is a measurement of how much contamination a technique can handle while still remaining stable. LTS is a reliable statistical strategy for fitting a function to a set of data while minimizing the impact of outliers. It is one of a number of robust regression algorithms. There is no closed-form solution for this method, as points are either included or excluded. As a result, methods for finding the LTS solution sift through various data combinations in an effort to analyze the k subset with the lowest sum of squared residuals. For small n, there are methods that will find the exact solution; but, as n goes larger, the number of possible combinations expands rapidly, leading to methods that seek to find approximation (but generally sufficient) solutions.

3.4.2.4 S Estimation

Outlying observations can have a significant impact on the estimation result in least squares, least absolute deviations, and even generalised M-estimation, disguising an essential and interesting relationship that exists in the majority of observations. In fitting regression functions to data, the S-estimators are a class of estimators that overcome this challenge by smoothly down weighting outliers. Rousseeuw and Yohai (1984) proposed the S-estimator, which minimises a scale estimator.

$$\tilde{\beta}_n = \text{Argmin } \tilde{\sigma}_n(\beta)$$

and the estimator of scale $\tilde{\sigma}_n(\beta)$ solves the equation

$$\frac{1}{n}\sum_{i=1}^{n} \rho\left(\frac{Y_i - X_i'\beta}{\tilde{\sigma}_n(\beta)}\right) = b$$

For each fixed β. where ρ (x) is a symmetric, continuous function, non decreasing in |x|, and $b = \int \rho(x) d\phi(x)$ with Φ being the standard normal distribution function.

3.5 Estimation of the parameters through robust values

The estimation of the linear and non-linear production functions has been done through OLS using the estimated robust values. Finally, the comparison of all the estimated models w.r.t different techniques have been done.

3.6 Validation of the proposed linear and non-linear production functions: The following techniques have been used.

3.6.1 Cross validation

It is a method for determining how well the results of a statistical analysis will generalise to a different data set. Cross validation is a method of model evaluation that is preferable to residuals. The difficulty with residual evaluations is that they don't show how well the learner will perform when asked to generate new predictions for data it hasn't seen before. When training a learner, one method to avoid this problem is to not use the complete data set. Before the training begins, some of the data is eliminated. After training, the data that was eliminated can be used to test the learned model's performance on new data.

3.6.1.1 Validation set approach (or data split)

The process works as follow:

- Build (train) the model on the training data set.
- Apply the model to the test data set to predict the outcome of new unseen observations.

- In this study 80% is taken as training data set and 20% as validation set
- Quantify the prediction error as the mean squared difference between the observed and the predicted outcome values. Note that, the validation set method is only useful when you have a large data set that can be partitioned.

3.6.1.2 K-Fold cross validation

It is a method that attempts to maximize the use of the available data for training and then testing a model. The algorithm is as follows:

- Randomly split the data set into k-subsets (or k-fold) (for example 5 subsets).
- Reserve one subset and train the model on all other subsets.
- Test the model on the reserved subset and record the prediction error.
- Repeat this process until each of the k subsets has served as the test set.
- Compute the average of the k recorded errors. This is called the cross-validation error serving as the performance metric for the model.

K-fold cross-validation (CV) is a robust method for estimating the accuracy of a model.

3.7 Fitness of the models will be checked through the following statistical and economical techniques.

3.7.1 Akaike Information Criterion (AIC)

Compares the quality of a set of statistical models to each other. Akaike Information Criterion is usually calculated with software. The basic formula is defined as: AIC=-2log $-liklelihood + 2K$.Where K is the number of model parameters (the number of variables in the model plus the intercept).Log-likelihood is a measure of model fit. The higher the number, the better the fit. This is usually obtained from statistical output. For small sample size (n/K$\leq\approx 40$) use the second order AIC. $AIC_C = -2\,(\log likelihood) + 2K + (2\,K(K + 1/n - k - 1))$Where n = sample size, K= number of model parameters, Log-likelihood is a measure of model fit. Akaike weights are a little more cumbersome to calculate but have the advantage that they are easier to interpret they give the probability that the model is the best from the set.$w_i = \frac{\exp\,(-\Delta_i/2)}{\exp_{i=1}^{R}\,(-\Delta_i/2)}$.AIC scores are reported as ΔAIC scores or Akaike weights. The ΔAIC Scores are the easiest to calculate and interpret. TheΔAIC is the relative difference between the best model (which has anΔAIC of zero) and each other model in the set the formula is $\Delta AIC = AIC_I - minAIC$.Where AIC_i is the score for the particular model i.Min AIC is the score for the "best" model.

3.7.2 Bayesian information criterion (BIC):

Bayesian information criterion (BIC) is derived to serve as an asymptotic approximation to a transformation of a candidate model's Bayesian posterior probability. The model preferred by BIC in large sample scenarios ideally corresponds to the candidate model that is a posteriori most probable, i.e., the model that is rendered most credible by the facts at hand. BIC is calculated using the empirical log-likelihood method, which does not require the identification of priors. As a result, BIC provides utility in many Bayesian modelling issues where accurate priors are difficult to obtain.BIC values are interpreted in terms of bayes factors, posterior model probabilities, and model averaging weights using the Bayesian rationale. These applications demonstrate that BIC is a useful scientific measure for describing the outcomes of a model selection challenge. Where K is the number of estimable parameters and n is the sample size, BIC = 2L + K log n. (for now we assume that n can be approximated by the total number of characters in the alignment). The difference between two BIC estimates may be a decent approximation to the natural log of the Bayes factor because the BIC was designed as an approximation to the log marginal likelihood of a model Kass and Wasserman(1995).

3.7.3 Root mean square error (RMSE)

It is the residuals' standard deviation (prediction errors). The root-mean-square error (RMSE) is a commonly used metric for comparing predicted and observed values (sample or population values) by a model or estimator. The normal distribution that underpins the application of the RMSE well describes the existence of outliers and their probability of occurrence. When calculating the RMSE, it may be justified to exclude outliers that are several orders greater than the other samples, especially if the number of samples is limited. If the model biases are strong, systematic errors may need to be removed before calculating the RMSE's. RMSE's have the distinct advantage of avoiding the usage of absolute value, which is quite undesirable in many mathematical calculations. It represents the model's absolute fit to the data. It gives you the average model prediction error in units of the variable you want to know about. They are negative scores, which means that the lower the value, the better.

3.7.4 Coefficient of determination (R^2)

R-squared is a statistical measure that is used to assess the goodness of fit of our regression model. In R-squared there is a baseline model. This baseline model doesn't make use of any independent variables to predict the value of dependent variable Y. Instead it uses the mean of the observed responses of dependent variable Y and always predicts this mean as the value of Y. Any regression model which is fitted is compared to this baseline model to understand its goodness of fit. The coefficient of determination or R squared method is the proportion of the variance in the dependent variable that is predicted from the independent variable. It indicates the level of variation in the given data set.

R^2 can be calculated by using the following formula:

$$R^2 = 1 - \frac{\text{ESS}}{\text{TSS}}$$

- The coefficient of determination is the square of the correlation(r), thus it ranges from 0 to 1.
- With linear regression, the coefficient of determination is equal to the square of the correlation between the x and y variables.
- If R^2 is equal to 0, then the dependent variable cannot be predicted from the independent variable.
- If R^2 is equal to 1, then the dependent variable can be predicted from the independent variable without any error For instance, in a few fields like rocket science, R^2 is expected to be nearer to 100 %.
- If R^2 is between 0 and 1, then it indicates the extent that the dependent variable can be predictable. If R^2 of 0.10 means, it is 10 percent of the variance in the y variable is predicted from the x variable. If 0.20 means, 20 percent of the variance in the y variable is predicted from the x variable, and so on.

The value of R^2 shows whether the model would be a good fit for the given data set. In the context of analysis, for any given per cent of the variation, it (good fit) would be different. But R^2 = 0(minimum theoretical value), which might not be true as R^2 is always greater than 0(by Linear Regression).

3.8 Elasticity of production function

The elasticity of production, also called output elasticity, is the percentage change in the production of a good by a firm, divided the percentage change in an input used for the production of that good, for example, labour or capital. The elasticity of production shows the responsiveness of the output when there is a change in one input. It is defined as the proportional change in the product, divided the proportional change in the quantity of an input. If a production function, for example: Q=f (K, L), is used to calculate the input, and the function is differentiable, the elasticity of production can be calculated using derivatives. It is denoted by:

$$e_p = \frac{dQ/Q}{dL/L}$$

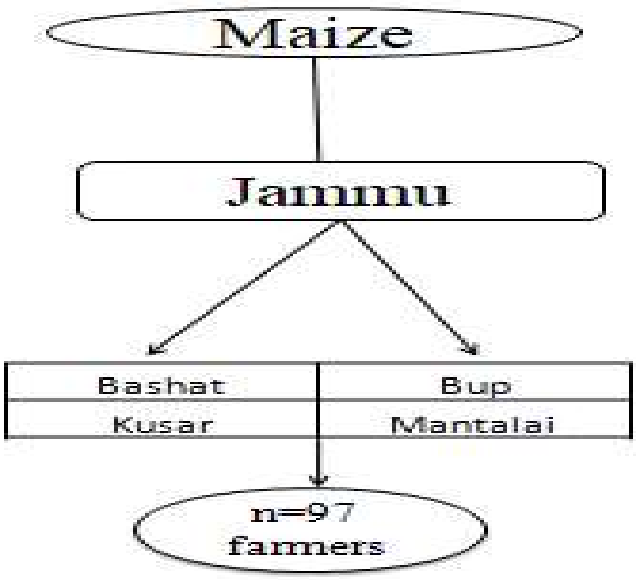
Elasticity is an economic metric that indicates how sensitive one economic dimension is to another, such as changes in supply or demand in response to price changes, or changes in demand in response to changes in income. When the elasticity value is greater than 1.0, it means that a change in the price of a good or service affects demand more than proportionally. When the elasticity number is less than 1.0, it means that demand for the good or service remains generally stable despite price fluctuations. A value less than 1.0 indicates that demand is relatively price insensitive, or inelastic. When prices rise, consumers' buying habits remain relatively unchanged, and when prices fall, consumers' buying habits remain relatively unchanged. It is said to be 'perfectly' inelastic if elasticity = 0, which means its demand will remain unchanged at any price. Elasticity is an economic word that describes how the aggregate quantity demanded of an item or service changes as the price of that good or service changes. When a product's quantity demand varies more than proportionally in response to price changes, the product is said to be elastic. A product, on the other hand, is said to be inelastic if its quantity demand changes very little when its price changes.

3.9 Marginal Value Product (MVP)

The extra revenue generated by raising the quantity of an input used by one unit while keeping all other input quantities fixed is referred to as the marginal value product. The marginal value product is calculated by multiplying the input's marginal product by the product's price. When a firm's output is sold at the same price regardless of quantity sold, i.e. when the market is in perfect competition, marginal revenue equals price, because the income generated by selling an extra unit of output is always equal to the price. As a result, the marginal revenue product and the marginal value product are equal in this example, and the latter accurately represents the additional revenue generated by increasing the quantity of an input used by one unit. MVP was determined using the geometric mean value of the output and inputs, as well as the regression coefficients of each input.

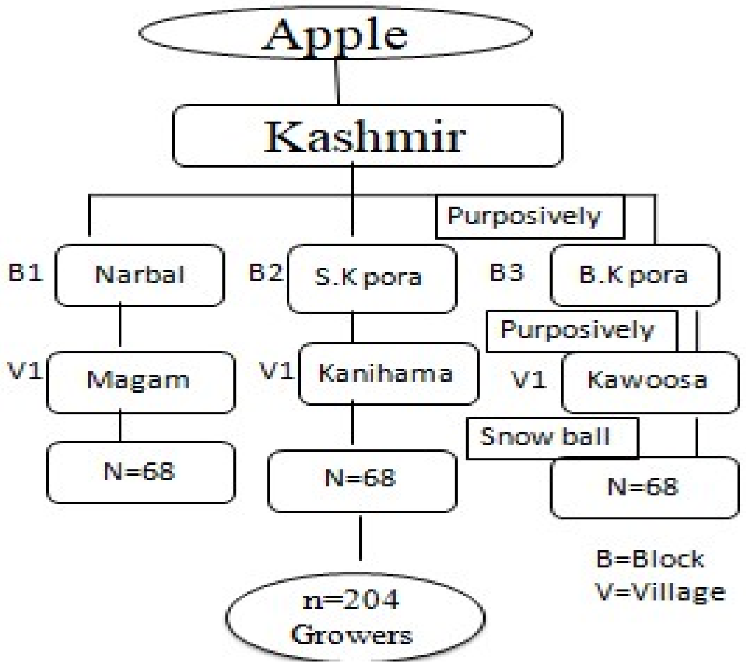
3.10 Simulation technique

Simulation is the use of purposely generated data to test a hypothesis or statistical procedure in statistics. When a new statistical approach is developed or utilised, certain assumptions must be validated and confirmed. Statisticians use simulated data to test their interpretations. When there is a lack of theoretical basis, simulation is employed in statistics to test the performance of a method. Simulation has been employed in a wide range of physical and social sciences and technical sectors from nuclear fusion to economic forecasting to space shuttle design. For varied contexts and systems, many sorts of models are utilised. When classifying simulations, it's vital to distinguish between the sorts of models that are simulated and the types of programme structures that are used to carry out the simulation. Deterministic and stochastic simulation models are the two types of simulation models (meaning probabilistic). In a deterministic simulation, all occurrences and relationships between variables are governed only by a set of well-known, albeit tough, rules. Deterministic simulations are commonly used to study physical systems. The advantage of simulation is that you can still answer the issue even if the model is too complex to solve analytically. In a stochastic simulation, random variables are used to reflect the influence of aspects that are unpredictable, unknown, or outside the scope of the model. Another criterion for categorising simulation models is the temporal dimension. The data had been analyzed through statistical software's SPSS 14.0, SAS 9.2, R and NCSS Software. The two types of data sets as per the study have been used first Simulation Data: Simulated data of 1000 observations and 500 observations generated by SAS 9.2 and second real data sets one of which is secondary and the other is primary. The secondary data is the survey done by Division of Agricultural Economics and ABM during the year 2018 to study the yield of maize w.r.t to the labour cost and capital involves. The multi stage sampling procedure was adopted for the selection of unit block as the first, village's as second and farmers as third stage. The Chenani block of Udhampur district of Jammu was selected purposively whereas Bashat, Bupp, Kosar and Mantalai villages were selected randomly with sample sizes of 25, 24, 24 and 24 respectively comprising of 97 respondents. The chart for this is as follows:



Source: Division of Agricultural Economics and ABM SKUAST JAMMU

The other real data is primary survey pertains to study the effect on yield of Apple w.r.t the cost involved in labour and infrastructure. The survey is conducted in district Budgam of central Kashmir purposively as the maximum growers are in this area and moreover due to Covid-19 restrictions during the month of April and May 2020.At first stage three blocks have been selected viz Narbal, S.K.Pora and S.K.Pora, at stage second three villages viz Magam, Kanihama and Kawoosa were selected purposively due to the maximum growers are in this area and due to COVID restrictions. At final stage 68 growers equally selected from all the villages on the basis of snowball sampling which comprises total sample size of 204 growers. The chart for this is as follows:


Apple
Kashmir
Purposively
B1
Narbal
B2
S.K pora
B3
B.K pora
Purposively
V1
Magam
V1
Kanihama
V1
Kawoosa
Snow ball
N=68
N=68
N=68
B=Block
V=Village
n=204
Growers

Chapter-4

Results

The results pertaining to the present investigation "**Estimation and validation of linear and non linear production functions through Robust Regression** "presented in this chapter through tables and figures. The chapter has been classified under following two sections

(I) Estimation of linear production function of Quadratic and Square root function.

(II) Estimation of non-linear production function of Cobb-Douglas and Mitscherlich Baule.

The estimation of each production functions under the section have been done analysed through Ordinary Least Square, M,MM,S, LTS and estimates of production function after handling HLP.

4.1 Estimation of linear production functions of quadratic and square root function:

The procedure used for the estimation of the same is discussed in detail in the following sections w.r.t to all the data sets.

4.1.1 Estimation of quadratic production function through robust method(s)

In case of summary statistics of simulation data N=500 (P_1) see table 4.1, it has been observed that the mean and standard deviation of the study variable w.r.t simulated data were found to 24.31unit and 38.47 unit which clearly indicates variation is more in the data as the mean is smaller than the standard deviation that is due to the presence of influential observations. Further, the mean and standard deviation of auxiliary variables of $x_1$and$x_2$in unit were found to found to be 0.05 & 0.98; 0.03 & 1.03 respectively which also depicts the variation is more in the simulated data set.

Table 4.1: Summary Statistics of Simulation data (P_1) w.r.t. endogenous and exogenous variables for quadratic production function.

Variable	Q_1	Median	Q_3	Mean	Standard Deviation	MAD	C V (%)
x_1	-0.59	0.02	0.75	0.05	0.98	0.99	1650.50
x_2	-0.76	-0.07	0.69	0.03	1.03	1.08	3433.33
y	1.81	8.12	15.68	24.31	38.47	9.82	158.24

Among the Mean Absolute Deviation, the minimum was for x_1 it was 0.99 units. Q_1 for x_1 and x_2 were found to be negative 0.59 unit and 0.76 unit whereas Q_1 of y found to be positive 1.81 unit. Low standard deviation indicates that the data points tend to be very close to the mean and high standard deviation indicates that the data points are spread out over a large range of value. Coefficient of variation of the Y is found to be low 158.24.whereas for x_1 and x_2 coefficient of variation is very high indicating inconsistency in the data set which may be due to HLP.

Table 4.2 : Detection of outliers of simulation data (P_1) and values of distance/residual through various techniques for quadratic production function.

Observations (HLP)	Mahalanobis	Robust MCD	Standardized Robust	Cook's	Studentized	WSSDI	Hat Diagonal (E-04)	Deleted
6	2.81	6.96	161.39	0.01	2.14	0.17	2.00	78.04
11	2.54	5.03	150.09	0.01	1.97	0.02	2.00	76.01
16	2.18	5.75	155.29	0.01	2.09	0.16	1.00	79.25
21	1.19	1.57	145.21	0.04	1.99	0.01	1.00	79.06
26	1.02	1.47	136.60	0.05	1.93	0.01	1.00	76.89
31	2.13	5.64	143.96	0.01	2.06	0.10	1.00	81.59
36	3.33	7.06	145.49	0.01	1.91	0.01	3.00	74.52
41	1.12	1.47	140.05	0.03	2.02	0.01	1.00	79.94
46	1.67	2.84	129.85	0.01	1.83	0.05	1.00	71.60
51	1.05	1.54	141.67	0.01	1.95	0.01	1.00	78.16
56	1.51	2.33	145.73	0.01	1.97	0.00	1.00	78.16
61	3.33	8.04	164.39	0.02	2.16	0.18	3.00	77.21
66	1.06	1.46	141.80	0.02	1.92	0.01	1.00	76.67
71	1.78	3.06	134.24	0.01	1.87	0.04	1.00	73.83
76	1.05	1.42	146.55	0.04	2.00	0.01	1.00	79.61
81	6.83	18.09	175.99	0.04	2.30	0.39	1.00	76.87
86	1.04	1.54	137.85	0.03	1.92	0.01	1.00	76.81
91	2.75	5.06	122.64	0.01	1.86	0.02	3.00	66.81
96	2.52	7.27	154.65	0.01	2.11	0.19	2.00	80.29
101	3.69	7.67	163.21	0.02	2.11	0.07	4.00	74.31
106	1.13	1.48	133.98	0.03	1.92	0.01	1.00	75.24
111	2.25	4.20	122.80	0.01	1.83	0.04	2.00	66.73
116	1.12	1.47	148.75	0.01	2.02	0.04	1.00	79.64
121	1.43	2.04	141.96	0.01	1.94	0.01	1.00	77.30
126	1.73	2.90	125.62	0.01	1.85	0.02	1.00	69.47
131	1.17	1.59	153.91	0.01	2.10	0.01	1.00	82.63
136	1.11	1.49	141.78	0.02	2.03	0.02	1.00	80.71
141	1.07	1.47	144.53	0.02	1.97	0.01	1.00	78.63

146	2.81	8.25	149.17	0.01	2.12	0.17	2.00	83.48
151	1.57	2.47	133.97	0.01	1.87	0.03	1.00	74.00
156	1.18	1.76	151.05	0.01	2.06	0.05	1.00	80.79
161	3.42	7.93	121.07	0.01	1.79	0.14	4.00	64.46
166	1.71	2.91	149.40	0.01	2.00	0.03	1.00	78.35
171	1.18	1.94	143.30	0.01	2.03	0.04	1.00	80.65
176	2.24	4.30	153.29	0.01	2.02	0.01	2.00	77.54
181	1.58	2.62	127.47	0.01	1.83	0.04	1.00	70.36
186	1.07	1.35	149.52	0.01	2.05	0.02	1.00	81.09
191	1.86	4.62	141.76	0.01	2.03	0.08	1.00	80.30
196	1.56	2.78	154.93	0.01	2.07	0.06	1.00	78.67
201	1.16	1.52	152.21	0.01	2.06	0.03	1.00	80.46
206	2.38	6.10	137.67	0.01	2.00	0.09	2.00	78.61
211	2.35	6.60	149.00	0.01	2.13	0.13	2.00	83.84
216	4.05	8.58	159.54	0.02	2.07	0.01	4.00	76.53
221	1.20	1.50	134.98	0.02	1.92	0.02	1.00	75.43
226	8.45	17.16	110.48	0.03	1.83	0.03	2.70	55.40
231	1.53	3.33	152.18	0.01	2.06	0.09	1.00	79.59
236	3.17	8.89	141.85	0.01	2.06	0.14	3.00	81.38
241	1.03	1.34	141.01	0.02	2.01	0.01	1.00	80.02
246	1.07	1.38	147.32	0.02	2.05	0.03	1.00	81.61
251	1.52	2.63	153.42	0.01	2.04	0.06	1.00	77.84
256	5.58	13.36	142.28	0.02	1.91	0.11	8.00	77.93
261	3.70	7.64	160.66	0.02	2.07	0.02	4.00	74.85
266	2.48	6.10	136.55	0.01	2.00	0.07	2.00	78.39
271	1.31	1.94	131.58	0.01	1.92	0.01	1.00	73.78
276	1.05	1.37	143.30	0.04	2.00	0.02	1.00	79.88
281	1.06	1.42	145.68	0.05	1.98	0.03	1.00	78.60
286	1.45	2.30	151.63	0.01	2.04	0.01	1.00	79.51
291	1.57	2.59	127.70	0.01	1.85	0.04	1.00	70.59
296	2.62	5.78	128.55	0.01	1.86	0.10	2.00	70.41
301	1.39	2.09	131.70	0.01	1.94	0.02	1.00	74.58
306	4.30	12.64	161.88	0.02	2.15	0.31	5.00	78.09
311	1.68	2.89	132.49	0.01	1.89	0.05	1.00	73.48
316	1.26	2.37	144.36	0.01	2.03	0.06	1.00	80.72
321	2.04	5.64	153.88	0.01	2.12	0.15	1.00	82.33
326	3.38	7.03	162.46	0.02	2.10	0.08	3.00	74.70
331	4.41	13.15	146.92	0.02	2.13	0.22	5.00	84.82
336	1.12	1.36	138.84	0.02	1.95	0.01	1.00	77.49
341	1.14	1.48	151.01	0.03	2.05	0.03	1.00	80.09
346	1.48	2.30	130.05	0.01	1.91	0.01	1.00	73.54
351	2.69	7.42	157.47	0.01	2.11	0.19	2.00	78.68
356	2.62	5.38	122.78	0.01	1.82	0.08	2.00	66.15
361	1.85	3.32	126.47	0.01	1.86	0.04	1.00	69.55
366	1.04	1.58	140.85	0.02	1.96	0.15	1.00	78.32
371	1.67	2.73	127.81	0.01	1.89	0.21	1.00	71.31
376	2.69	5.15	125.43	0.01	1.90	0.01	2.00	70.34
381	1.36	2.01	130.78	0.01	1.89	0.02	1.00	72.94
386	1.11	1.57	144.07	0.02	2.01	0.04	1.00	80.17
391	1.14	1.52	146.50	0.04	2.00	0.03	1.00	79.32
396	1.12	1.61	147.12	0.05	2.06	0.04	1.00	81.86
401	1.90	3.53	130.02	0.01	1.83	0.06	1.00	71.50
406	1.92	3.43	147.31	0.01	1.98	0.01	1.00	77.97
411	1.72	2.82	136.50	0.01	1.87	0.02	1.00	74.52
416	1.83	3.26	156.10	0.01	2.06	0.03	1.00	77.90
421	1.17	1.88	142.32	0.01	2.02	0.04	1.00	80.23
426	1.21	1.67	150.44	0.01	2.05	0.01	1.00	80.83
431	3.81	8.04	124.57	0.01	1.91	0.01	5.00	70.28
436	1.05	1.39	146.54	0.04	2.03	0.02	1.00	81.00
441	3.20	7.50	128.77	0.01	1.86	0.13	3.00	70.47
446	1.77	2.93	125.63	0.01	1.86	0.02	1.00	69.85
451	1.32	1.88	133.80	0.01	1.93	0.02	1.00	74.89
456	1.28	2.00	138.31	0.01	2.00	0.02	1.00	78.90
461	2.14	3.84	123.78	0.01	1.84	0.03	2.00	67.72
466	2.29	4.38	159.20	0.01	2.10	0.01	2.00	79.26
471	2.07	4.20	159.40	0.01	2.11	0.08	1.00	78.57
476	1.10	1.39	149.10	0.03	2.02	0.02	1.00	79.51

481	1.59	2.76	155.11	0.01	2.06	0.05	1.00	78.34
486	2.87	5.76	162.53	0.01	2.12	0.05	2.00	77.14
491	3.70	7.83	154.48	0.02	2.00	0.06	4.00	75.33
496	1.17	1.60	129.57	0.02	1.87	0.06	1.00	72.89
500	1.02	1.45	137.10	0.03	1.95	0.07	1.00	77.59

An outlier elucidation is obtained in Table 4.2, and estimation results are presented and discussed. About Twenty percent of the 500 observations were found as outliers. The different methods used for identify the outliers were Mahanablios distance, Robust MCD Distance, and Standardized Robust Residual, cook's distance, and studentized residual, WSSDI, Hat Diagonal and Deleted Residual. In Mahanablios distance, it has been observed the value of outliers through Mahanablios distance lies between 1.02 to 8.45 whereas through Robust MCD1.34 to 18.09, For standardized robust residuals, cooks distance, WSSDI, Hat diagonal and deleted the range of the observations are between 110.48 to 175.99, 0.01 to 0.05, 1.79 to 2.30, 0.01 to 0.39, 1.0E-04 to 8.0E-04, 55.39 to 84.82 respectively.

The Fig 4.1 shows outlier and leverage diagnostics for dependent variable for simulated data for quadratic production function. The figure shows out of 500 observations we have 100 outliers and 160 leverage points. A scatter plot of the standardized robust residuals against the robust distances is an especially useful plot for enlightening outliers and leverage points (RD plot). The robust distances are plotted against the classical Mahalanobis distances in Fig 4.2. Outliers as well as high leverage points can be identified using the plot.

Table 4.3: Estimates of quadratic production function (P_1)through Ordinary Least Square (OLS).

Variable	Regression coefficient (S.E)	Parameters
Intercept	21.43* (2.48)	F-value 1.52(P<0.01) R^2=0.01 AIC=5074.24 BIC=5103.74
x_1	1.09 ^{NS} (1.70)	
x_2	3.107* (1.16)	
${x_1}^2$	1.62^{NS} (1.31)	
${x_2}^2$	1.16 ^{NS} (1.11)	
$x_1 * x_2$	1.90^{NS} (1.83)	

*** 5 % level of significance   NS= non significant**

The estimates results for the Quadratic production function with standard error have been shown in the the table 4.3. It depicts that all the coefficients were found to be positive and non significant the as the F-value was found to be significant indicates the model is adequate w.r.t to study variable. It further shows that each estimation coefficient may be having the correct sign due to the presence HLP. The estimates of the regression coefficient have been done through OLS.The R^2 value found to minimum (0.015) and having AIC (5074.24) and SBIC (5103.74), this may be due to the presence of HLP.Therfore robust methods have been applied to study the effect on the estimates of the parameter of quadratic function.

In Fig 4.3 the panel of diagnostics of ordinary least square for quadratic production function model observed to be poor. These diagnostics reveal that this model is a resounding failure. The residuals and studentized residuals plots versus expected values show poor trends. The point on the plot of the dependent variable(y) versus the predicted values predicts the behaviour of the dependent variable. The plot of studentized residual versus leverage showed that leverage points are lying outside the interval -2 to +2 and are HLP. Further, the graph of Cook's indicates the maximum number of observations are outside limit and having high jump the cut-off clearly indicates that data are influenced by outliers.

Table 4.4: Estimation of quadratic functions through robust method for simulation data (P_1) and comparison with OLS thereof.

Variable	Estimates of the regression coefficients					
	OLS (S.E)	**M Estimation (S.E)**	**LTS Estimation (S.E)**	**MM Estimation (S.E)**	**S Estimation (S.E)**	**Modified estimate after handling HLP (S.E)**
Intercept	21.43* (2.48)	5.01* (0.03)	5.01* (0.03)	4.99* (0.05)	4.99* (0.05)	5.37* (0.21)
x_1	1.09 NS (1.70)	5.02* (0.02)	5.01* (0.02)	5.04* (0.03)	5.04* (0.03)	5.128* (0 .21)
x_2	3.10* (1.16)	5.04* (0.02)	5.04* (0.02)	3.06* (0.06)	3.06* (0.06)	5.05 NS (0 .19)
${x_1}^2$	1.62 NS (1.32)	-0.02 NS (0.02)	-0.03 NS (0.02)	-0.01 NS (0.01)	-0.05 NS (0.01)	-0.10NS (0 .10)
${x_2}^2$	1.16 NS (1.112)	0.01 NS (0.07)	0.09 NS (0.06)	-0.04 NS (0.01)	-0.04 NS (0.01)	-0.08 NS (0 .15)
$x_1 * x_2$	1.90NS (1.83)	-0.03 NS (0.02)	0.044 NS (0.02)	-0.06 NS (0.01)	-0.06 NS (0.02)	-0.03 NS (0 .20)
R^2	0.01	0.70	0.99	0.99	0.70	0.73
AIC	5074.24	956.61	-1257.15	323.12	412.25	714.78
BIC	5103.74	987.85	-1231.86	350.12	432.21	744.27

*** 5%level of significance NS= non significant**

The table 4.4 presents the estimates of the regression coefficient which have been done through OLS, M Estimation, MM, S, LTS and modified estimates after dealing with HLP. It has been observed that the OLS estimate of ${x_1}^2$ was positive 1.62 and non significant, but through M,MM, S Estimation and LTS were found to be negative and non-significant 0.02, 0.01, 0.05 and 0.03.Modified estimate of production function also after handling HLP was found to be negative and non-significant 0.10.Moreover, the coefficient of determination is remarkably higher in case LTS estimation but this method is not advisable as it is based on trimming of data Chen and Cary (1987) followed by MM estimation i.e. 0.99 which indicates that 99 percent variation of study variable is explained through this method whereas the least was found to be in case of OLS method i.e.0.01.The AIC and BIC in case of MM estimation was found to be low. If exceptional observations are omitted in the analysis, the linear pattern formed by the remaining observations explains more of the variation in simulated data. It has also been observed that the OLS estimate of ${x_2}^2$ was positive 1.16 and non-significant and in case of M and LTS estimation found to be positive 0.01, 0.09 and also non-significant. The coefficient of ${x_2}^2$ in case of MM and S estimation were found to be same i.e. negative and non-significant 0.04. The estimate of modified production function after handling HLP was also found to be negative and non-significant having value 0.08. Further, the simulated data of 1000(P_2) observation have been considered to see the effect of outliers in case of big data. The summary statistics of the data has been reflected in table 4.5. The table depicts that the mean and standard deviation of study variable

Table 4.5: Summary Statistics of simulation data (P_2)w.r.t. endogenous and exogenous variables for quadratic production function.

Variable	Q_1	Median	Q_3	Mean	Standard Deviation	MAD	C V(%)
x_1	-0.69	-0.03	0.63	0.02	1.00	0.97	4996.86
x_2	0.31	0.72	1.44	0.97	1.03	0.73	9443.68
y	4.62	8.47	13.62	17.19	28.15	6.52	163.72

of simulated data were 17.19 unit and 28.15 unit respectively, indicating that there is wider diversity in the data because the mean is smaller than the standard deviation, which is due to presence of influenced observations. Furthermore, the mean and standard deviation of auxiliary variables $x_1$and $x_2$were determined to be 0.02and 1.00

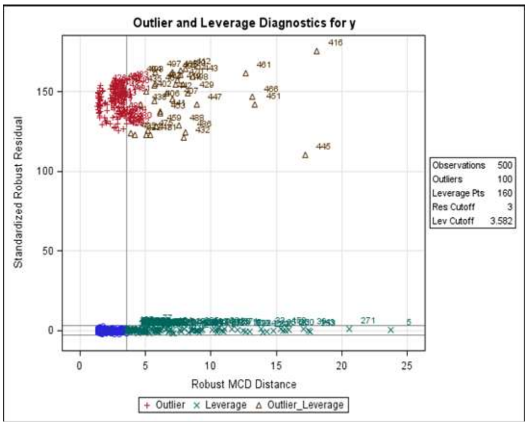


Fig 4.1 : Outlier and Leverage diagnostics for dependent variable for Simulation data (P_1) for quadratic production function..

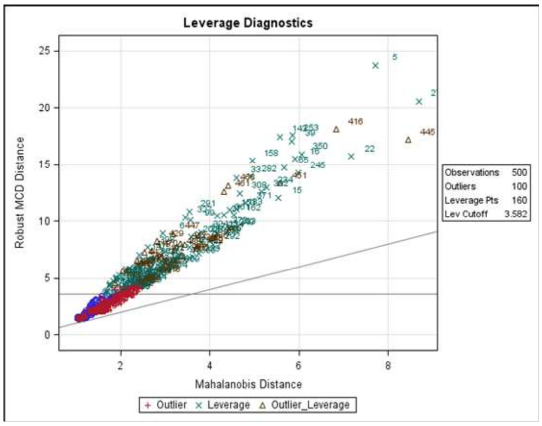


Fig 4.2 : Leverage diagnostics for dependent variable for Simulation data (P_1) for quadratic production function.

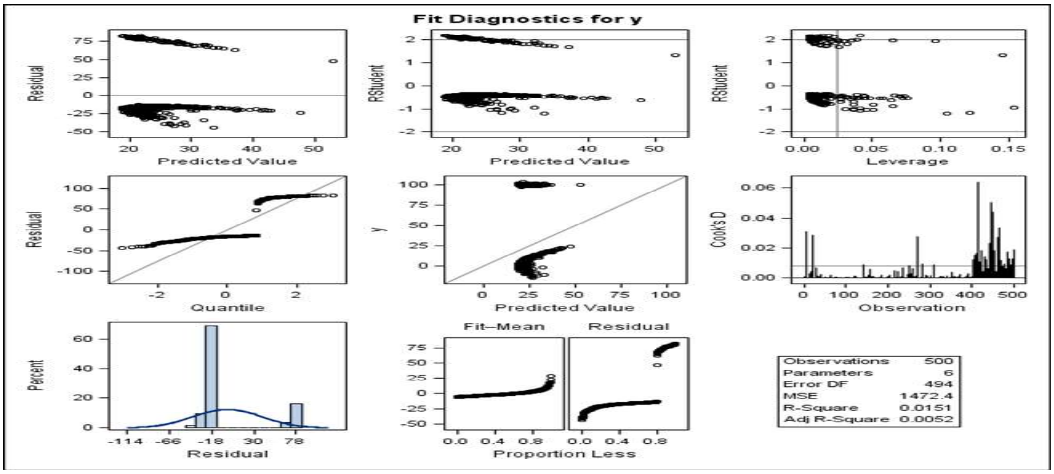


Fig 4.3: Diagnostics Panel of quadratic production function through OLS.

and 0.97and 1.03, respectively, indicating that the simulated data set has higher variability. The lowest mean absolute deviation was 0.73 forx_2. The Q_1 for $x_1$was found to be negative 0.69 and for x_2 positive 0.31 respectively, the Q_1value for study variable was positive i.e. 1.81. Low standard deviation means the data points are closed to the mean while high standard deviation means the data points are spread out over a wide range of values. The study variables coefficient of variation is found to be low 163.72. However the coefficient of variation for $x_1$and x_2 is quite high, indicating inconsistency in the data set, which could be attributable to HLP.

Table 4.6: Detection of outliers of simulation data $(P_2)$and the values of distance/residual through various techniques for quadratic production function.

Observations (HLP)	Mahalanobis	Robust MCD	Standardized Robust	Cook's	Studentized	WSSDI	Hat Diagonal (E-04)	Deleted
10	1.34	4.18	144.19	0.04	2.87	0.07	1.00	80.63
20	1.75	7.24	151.01	0.06	2.95	0.16	1.00	83.96
30	0.95	1.49	143.42	0.04	2.89	0.03	1.00	80.88
40	2.07	3.16	142.17	0.06	2.91	0.11	1.00	81.27
50	0.85	1.43	146.95	0.04	2.95	0.03	6.00	82.75
60	1.50	1.83	155.32	0.06	3.12	0.22	4.00	86.86
70	2.75	42.36	136.31	0.07	2.62	1.58	1.00	73.58
80	3.55	6.27	136.49	0.06	2.82	0.13	1.00	78.74
90	1.27	1.26	159.39	0.06	3.17	0.12	3.00	87.37
100	0.79	1.49	154.71	0.03	3.09	0.01	2.00	86.79
110	0.80	1.49	150.22	0.03	3.02	0.01	3.00	84.14
120	0.83	1.46	152.21	0.03	3.03	0.03	1.00	85.08
130	1.90	8.92	140.11	0.07	2.80	0.19	1.00	77.65
140	6.49	95.28	132.14	0.05	2.50	3.82	5.00	70.01
150	1.91	9.92	136.82	0.06	2.72	0.23	1.00	75.50
160	2.70	4.11	168.19	0.07	3.30	0.05	1.00	90.04
170	1.49	3.73	138.12	0.06	2.78	0.01	5.00	77.38
180	6.47	51.53	123.00	0.03	2.44	1.72	6.00	61.64
190	0.96	1.57	159.59	0.06	3.16	0.02	4.00	87.88
200	1.71	1.91	161.65	0.09	3.21	0.16	4.00	87.70
210	2.07	2.81	166.02	0.03	3.27	0.06	1.00	89.43
220	1.90	2.31	165.47	0.09	3.29	0.18	5.00	89.71
230	7.38	68.52	119.40	0.02	2.34	2.46	8.00	57.97
240	1.42	5.11	143.63	0.04	2.86	0.10	2.00	80.20
250	2.41	30.67	136.34	0.03	2.67	1.07	1.00	73.43
260	2.55	3.65	165.45	0.05	3.25	0.07	1.00	88.38
270	2.38	6.45	137.71	0.01	2.79	0.02	1.00	76.97
280	0.82	1.47	158.13	0.04	3.14	0.02	2.00	87.85
290	5.71	49.52	151.08	0.08	2.87	1.77	3.00	86.21
300	0.88	1.33	155.63	0.04	3.09	0.05	4.00	86.19
310	1.51	1.78	151.61	0.06	3.06	0.21	1.00	85.51
320	2.28	3.37	166.01	0.04	3.26	0.05	1.00	89.21
330	1.45	1.70	155.16	0.06	3.11	0.20	2.00	86.49
340	4.95	73.73	126.32	0.05	2.42	2.86	3.00	64.36
350	1.42	3.92	156.33	0.06	3.07	0.05	1.00	86.90
360	1.72	7.52	151.81	0.06	2.97	0.18	1.00	84.56
370	2.11	3.16	140.11	0.02	2.86	0.15	1.00	80.07
380	2.34	3.26	168.14	0.08	3.34	0.21	1.00	90.51
390	0.80	1.46	148.87	0.03	2.97	0.03	2.00	83.31
400	0.79	1.50	153.41	0.03	3.05	0.01	2.00	85.65
410	1.07	1.16	149.59	0.04	3.00	0.11	1.00	83.91

420	4.92	23.37	158.72	0.03	3.04	0.59	2.00	89.56
430	1.16	2.20	141.97	0.05	2.87	0.08	1.00	80.14
440	2.32	31.28	139.47	0.03	2.69	1.12	1.00	76.28
450	1.42	1.62	153.00	0.05	3.08	0.20	1.00	85.61
460	1.05	2.02	152.28	0.04	2.99	0.03	3.00	84.29
470	1.56	4.62	156.05	0.06	3.06	0.06	3.00	86.59
480	1.54	1.81	150.70	0.07	3.05	0.20	4.00	85.41
490	1.22	1.25	146.77	0.05	2.96	0.12	5.00	82.98
500	1.27	1.33	154.80	0.05	3.10	0.16	2.00	86.13
510	0.79	1.50	152.11	0.03	3.03	0.01	1.00	84.81
520	2.88	28.24	131.79	0.02	2.59	0.93	1.00	70.34
530	1.10	1.48	144.47	0.05	2.92	0.06	2.00	81.74
540	2.14	28.93	140.24	0.03	2.72	1.03	1.00	76.65
550	1.98	5.38	159.64	0.03	3.12	0.03	4.00	87.84
560	1.56	9.21	143.20	0.05	2.82	0.27	2.00	79.58
570	1.96	2.65	163.57	0.02	3.22	0.06	3.00	88.08
580	1.70	2.47	164.64	0.01	3.25	0.03	2.00	89.27
590	1.61	8.34	147.07	0.05	2.88	0.23	3.00	81.89
600	0.96	1.25	154.99	0.04	3.09	0.08	4.00	85.88
610	4.21	6.75	172.29	0.06	3.37	0.07	2.00	91.11
620	2.32	31.19	136.53	0.03	2.66	1.10	1.00	73.69
630	1.40	1.52	158.49	0.06	3.17	0.18	2.00	87.57
640	0.80	1.50	152.87	0.03	3.03	0.01	4.00	84.98
650	1.83	8.22	139.63	0.06	2.79	0.17	2.00	77.45
660	0.94	1.76	149.87	0.04	3.02	0.03	1.00	84.66
670	2.84	6.06	164.93	0.06	3.22	0.03	1.00	89.30
680	2.41	3.75	168.50	0.05	3.31	0.09	1.00	90.70
690	1.21	2.62	137.83	0.05	2.77	0.04	5.00	77.35
700	2.41	3.38	166.34	0.03	3.30	0. 68	1.00	89.27
710	1.97	23.84	143.07	0.04	2.79	0. 89	1.00	78.66
720	1.63	2.94	141.21	0.04	2.87	0.02	7.00	80.06
730	1.25	1.26	148.94	0.06	3.01	0.13	1.00	84.24
740	1.39	1.63	146.83	0.07	2.98	0.11	5.00	83.46
750	1.03	1.20	155.80	0.04	3.11	0.10	5.00	86.44
760	3.05	4.83	167.41	0.07	3.31	0.26	1.00	88.93
770	1.16	1.25	159.90	0.06	3.17	0.08	1.00	87.71
780	1.55	1.88	151.63	0.07	3.07	0.21	1.00	85.77
790	2.36	13.80	135.40	0.08	2.69	0.35	1.00	74.09
800	0.88	1.53	149.22	0.04	3.01	0.02	2.00	84.29
810	0.89	1.49	151.64	0.03	3.00	0.05	3.00	84.44
820	0.81	1.51	153.58	0.03	3.05	0.07	4.00	85.45
830	1.06	1.72	149.07	0.03	2.96	0.03	2.00	83.51
840	0.98	1.61	144.88	0.05	2.92	0.07	4.00	81.85
850	1.58	2.03	152.54	0.06	3.06	0.07	2.00	85.16
860	0.91	1.48	153.42	0.03	3.04	0.05	4.00	85.61
870	1.47	1.65	146.94	0.06	2.97	0.17	1.00	83.26
880	1.84	2.44	161.68	0.09	3.22	0.23	1.00	88.14
890	1.12	1.21	160.65	0.06	3.19	0.08	3.00	88.39
900	2.08	3.14	145.40	0.01	2.97	0.22	1.00	83.17
910	0.98	1.82	143.09	0.03	2.86	2.23	1.00	80.40
920	1.30	3.81	146.17	0.03	2.89	0.08	2.00	81.69
930	3.49	9.77	165.30	0.09	3.21	0.07	1.00	90.79
940	5.66	85.90	130.96	0.08	2.50	3.40	4.00	67.51
950	2.08	2.86	162.03	0.09	3.22	0.23	1.00	87.54
960	1.17	1.32	143.47	0.05	2.90	0.09	4.00	81.11
970	1.90	3.62	138.55	0.09	2.82	0.01	1.00	78.45
980	1.22	1.22	151.72	0.05	3.06	0.15	2.00	85.36
990	4.29	7.70	167.97	0.02	3.27	0.01	1.00	89.28
1000	0.87	1.61	156.86	0.04	3.10	0.06	2.00	86.84

An outlier interpretation and estimation results are shown and discussed in Table 4.6. About 10 percent of Outliers were found among 1000 observations as per the Mahanablios distance, robust MCD distance, standardized robust residual, cook's distance, studentized residual, WSSDI, Hat Diagonal, and Deleted Residual were some of the approaches used to find outliers. The value of outliers in Mahanablios distance is 0.79 to 6.49, whereas in Robust MCD 1.16 to 95.28, respectively. For standardised robust residuals, cooks distance, WSSDI, Hat diagonal, and deleted, the range of observations is 119.40 to 172.29, 0.01 to 0.09, 2.42 to 3.37, 0.01 to 3.82, 1.00 to 8.00 and 57.97 to 91.11

Outlier and leverage diagnostics for the dependent variable for simulated data for the quadratic production function are shown in Fig 4.4 Out of 1000 observations; there are 100 outliers and 160 leverage points, as seen in the Fig 4.4. A scatter plot of the standardised robust residuals vs. robust distances is very useful for identifying outliers and leverage points (RD plot). In Fig 4.5 the robust distances are plotted against the conventional Mahalanobis distances. The plot can be used to identify outliers, as well high-leverage points.

Table 4.7: Estimates of quadratic production function(P_2) through ordinary least square (OLS).

Variable	Regression coefficient (S.E)	Parameters
Intercept	13.95* (1.750)	F-value 7.11 (p<0.0001) R^2=0.03 AIC=9491.00 BIC=9525.35
x_1	2.332^{NS} (1.252)	
x_2	1.72^{NS} (2.254)	
$x_1{}^2$	0.51^{NS} (0.598)	
$x_2{}^2$	0.51^{NS} (0.565)	
$x_1 * x_2$	1.31^{NS} (0.937)	

*** 5%level of significance NS= non significant**

Table 4.7 displays the standard error estimates for the Quadratic production function. The fact that all of the coefficients were found to be positive and non-significant, and that the F-value was found to be significant, suggests that the model is appropriate in relation to the study variable. It demonstrates that due to the presence of HLP, each estimated coefficient may have the proper sign. OLS was used to calculate the regression coefficient estimations. Furthermore, the R^2 value is 0.03, with AIC (9491.00) and BIC (9525.35), which may be related to the presence of HLP. As a result, robust approaches were used to investigate the effect on the estimations of the parameter of the quadratic function.

Fig 4.6 shows panel of diagnostics of ordinary least square for quadratic production function model have been studied and is poor. These diagnostics reveal that this model is a resounding failure. The residuals and studentized residuals plots versus expected values show poor trends. The points on the plot of the dependent variable(y) versus the predicted values predict the behaviour of the dependent variable. The plot of studentized residual versus leverage showed that leverage points are lying outside the interval -2 to +2 are HLP. Further, the graph of Cook's indicates the maximum numbers of observations are outside and having high jump the cut-off clearly indicates that data are influenced by outliers.

Table 4.8: Estimation of quadratic functions through robust methods for simulation data $(P_2)$and comparison with OLS thereof.

Variable	Estimates of the regression coefficients					
	OLS (S.E)	**M estimation (S.E)**	**LTS estimation (S.E)**	**MM estimation (S.E)**	**S estimation (S.E)**	**Modified Estimate after handling HLP (S.E)**
Intercept	13.95* (1.75)	5.02* (0.05)	4.99* (0.03)	5.05* (0.05)	5.08* (0.03)	4.94* (0.18)
x_1	2.33 NS (1.25)	5.05* (0.05)	5.04* (0.04)	5.04* (0.05)	5.03* (0.02)	5.09* (0.09)
x_2	1.72 NS (2.25)	3.04* (0.04)	3.07* (0.03)	3.02* (0.05)	3.09* (0.06)	3.298* (0.23)
${x_1}^2$	0.51 NS (0.59)	0.04 NS (0.02)	0.04 NS (0.01)	0.03 NS (0.02)	0.02 NS (0.01)	-0.02NS (0.06)
${x_2}^2$	0.51NS (0.56)	-0.05 NS (0.01)	-0.06 NS (0.01)	-0.04 NS (0.01)	-0.03 NS (0.01)	-0.07 NS (0.05)
$x_1 * x_2$	1.33 NS (0.93)	0.06 NS (0.08)	0.07 NS (0.08)	0.06 NS (0.09)	0.05 NS (0.02)	-0.03 NS (0.06)
R^2	0.03	0.78	0.99	0.75	0.98	0.79
AIC	9491.00	1311.50	-2767.00	844.00	376.00	1455.63
BIC	9525.35	1345.73	-2737.56	879.92	346.56	1489.98

*** 5%level of significance NS= non significant**

The estimation results for the Quadratic production function are shown in Table 4.8, which illustrates that each estimation coefficient may be correct in terms of the expected sign. After dealing with HLP, the regression coefficient was estimated using OLS, M Estimation, MM, S, LTS, and updated estimates. The OLS estimate of ${x_1}^2$was found to be positive 0.51 and non significant, however the M,MM, S

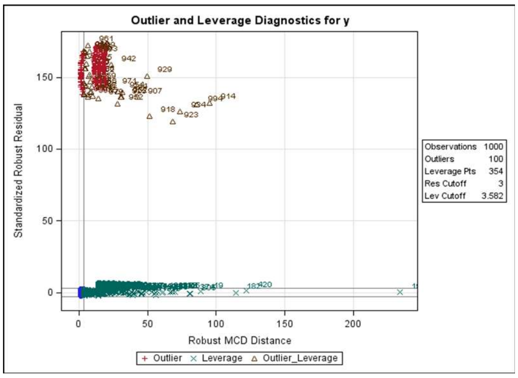


Fig 4.4: Outlier and Leverage diagnostics for dependent variable for Simulation data (P_2)for quadratic production function.

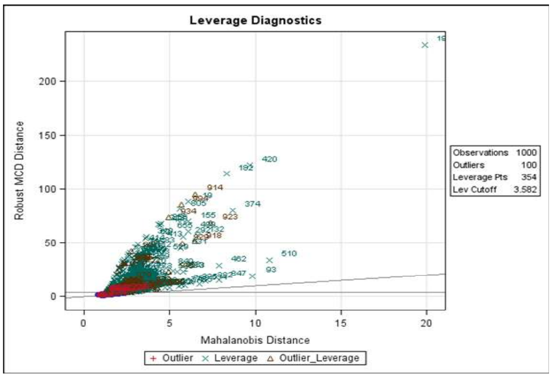


Fig 4.5 : Leverage diagnostics for dependent variable for Simulation data (P_2) for quadratic production function.

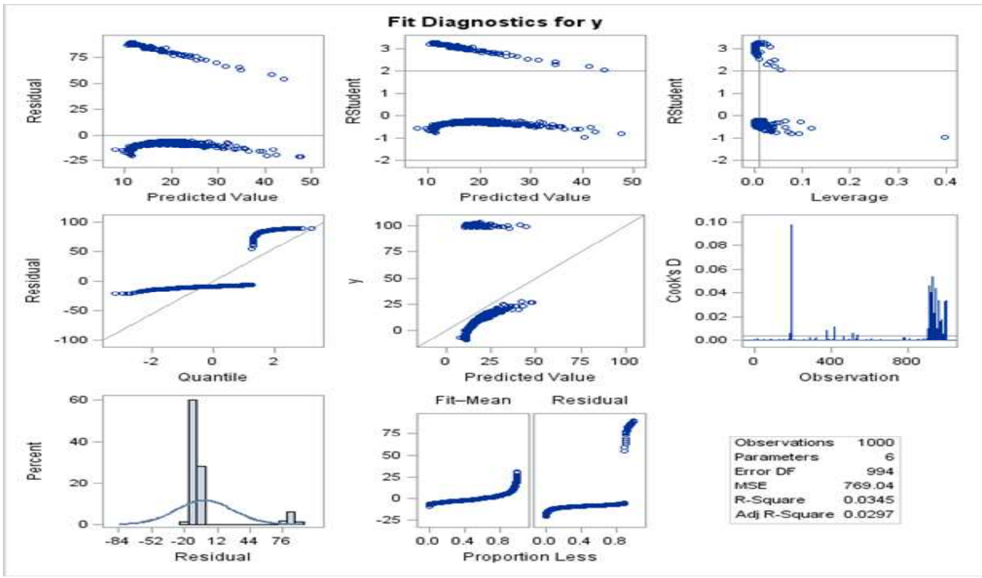


Fig 4.6: Diagnostics Panel of quadratic production function through OLS.

Estimation, and LTS were found to be positive and non significant 0.04, 0.04, 0.03 and 0.02. After handling HLP, modified estimate of the production function were found to be negative and non-significant (0.02). Furthermore, the coefficient of determination for LTS estimation 0.99 is significantly greater than for S estimation 0.98, indicating that this method can explain 99 percent of the fluctuation in the study variable but this method is not advisable Chen, C. (2002). S estimation is based on variance. The least was found in case of OLS i.e. 0.03. If exceptional observations are excluded from the study, the remaining observations create a linear pattern that explains more of the variation in the simulated data. It was also found that the OLS estimates of ${x_2}^2$was positive 0.51 and non significant, and M,MM,S and LTS estimates were negative and non-significant0.05,0.04, 0.03 and 0.06. After handling HLP, modified estimate of the production function was determined to be negative and non-significant (0.07). After observing the impact of outliers in case of simulation data, the real data sets have been taken to observe the effect of outliers in case of linear and nonlinear production functions. The first real data in this series is pertained to the survey done at Kashmir pertained to yield of apple w.r.t cost involved in labour and capital. The demographic profile of the real dataset is reflected in table4.9.

Table 4.9:Socio-economic characteristics of growers of Kashmir involved in Apple yield.

Age (Years)	Frequency	Percentage	Test value	P value
21-30	44	21.56		
31-40	5	25.49	χ^2=38.87*	0.007
41-50	25	12.25		
51-60	66	32.35		
Above	60	17	8.33	
Sex(M/F)	Frequency	Percentage		
Male	141	72.16		
Female	63	27.83	Z= 6.33*	0.025
Marital Status(Married/Unmarried)				
Married	160	82.47		
Unmarried	44	17.52	Z=9.33*	0.012
Occupation(farming/other)				
Farming	165	87.62		
Other than farming	39	12.37	Z=26.6*	0.006

*** Significant at 5% los.**

It illustrates that age has a significant impact on farming activities since it impacts the efficiency and competency of labour available for apple production. According to the figures, the majority of respondents 91.66 percent were between the ages of 21 and 60. This implies that, apple cultivation is done by young adult growers within this age bracket. This is because these categories of growers are still strong and have the ability to supply the required labour in agricultural activities to boost production as well to increase resource use efficiency. The chi square value (38.87) was found to be significant which indicates the maximum number of growers involved in the production of apple in the age group of 51-60 followed by growers in age group of 31-40 and minimum in above 60 because of the age factor. The result also revealed that about 87.62 percent growers have opted farming as main occupation due to years of farming experience, low-land apple production, better farm management methods and resource utilisation efficiency, lowering costs and improving output. Further, the participation of male (72 percent) is more in apple production and married (82 percent) as their Z values were significant. Thus males and married are more involved in the production of apple in Kashmir region having farming as main occupation.

Table 4.10:Summary Statistics of Apple yield w.r.t labour and capital for quadratic production function.

Variable (unit)	Q_1	Median	Q_3	Mean	Standard Deviation	MAD	C V (%)
x_1 labour (Rs/day)	11200.00	13500.00	19200.00	14455.40	5267.06	6671.00	36.43
x_2 Capital(Rs/kanal)	11600.00	14000.00	19600.00	15183.80	6058.04	6523.04	39.89
Y Yield(q/hectare)	0.80	1.18	3.14	2.05	2.18	0.02	5.31

According to Table 4.10, the mean and standard deviation of the study variable based on real data were 2.05 q/ha and 2.18q/ha, respectively, having more variation in the data as the mean is smaller than the standard deviation, which is presence of influential observations. The smallest among the mean absolute deviation was 0.02 q/ha for study variable y. The $Q_1$for $x_1$and $x_2$were found to be positive 11200 Rs/dayand 11600 Rs/day, respectively, whereas the $Q_1$for y was determined to be positive 0.80q/ha. A low standard deviation implies that the data points are fairly close to the mean, whereas a high standard deviation suggests that the data points are spread out throughout a wide range of values. The CV of x_2 (36.43) was more than x_1(39.89) respectively.

Table 4.11: Detection of outliers in case of Apple yield data and values of distance/residual through various techniques in case of quadratic production function.

Observations (HLP)	Mahalanobis	Robust MCD	Standardized Robust	Cook's	Studentized	WSSDI	Hat Diagonal	Deleted
4	3.98	17.73	-6.53	0.11	-0.86	834.52	0.05	-15.45
8	3.98	17.73	-6.53	0.11	-0.86	834.52	0.05	-15.45
9	3.98	17.73	-6.53	0.11	-0.86	834.52	0.05	-15.45
15	3.98	17.73	-6.53	0.11	-0.86	834.52	0.05	-15.45
31	3.98	17.73	-6.53	0.11	-0.86	589.31	0.05	-15.45
68	2.51	105.43	22.85	0.02	0.61	741.89	0.08	10.75
70	1.50	20.36	3.89	0.04	-0.39	741.89	0.04	-6.75
73	1.50	20.36	3.89	0.04	-0.39	741.89	0.04	-6.75
81	1.50	20.36	3.89	0.04	-0.39	741.89	0.04	-6.75
87	1.50	20.36	3.89	0.04	-0.39	381.88	0.04	-6.75
104	2.96	29.92	28.49	0.05	-0.24	589.31	0.02	-4.35
109	2.51	105.43	22.85	0.02	0.61	477.05	0.08	10.75
111	5.09	453.40	12.63	0.03	0.38	214.43	0.08	7.11
119	3.99	279.05	50.34	0.01	0.81	352.58	0.07	14.56
121	2.75	28.96	26.80	0.04	-0.23	328.98	0.02	-4.17
122	1.74	90.42	81.08	0.78	4.70	834.52	0.01	81.62
133	3.98	17.73	-6.53	0.12	-0.86	381.88	0.03	-15.49
135	2.96	29.92	28.49	0.05	-0.24	328.98	0.02	-4.37
137	1.74	90.42	81.08	0.78	4.70	214.43	0.01	81.62
144	3.99	279.05	50.34	0.10	0.81	589.31	0.04	14.57
145	2.51	105.43	23.41	0.06	0.65	741.89	0.07	11.48
149	1.50	20.36	3.89	0.04	-0.39	381.88	0.08	-6.72
153	2.96	29.92	28.49	0.05	-0.24	214.43	0.04	-4.37
154	3.99	279.05	50.34	0.10	0.81	589.31	0.04	14.57
157	2.51	105.43	23.41	0.06	0.65	214.43	0.02	11.48
158	3.99	279.05	50.34	0.10	0.81	477.05	0.07	14.57
160	5.09	453.40	14.03	0.06	0.48	932.01	0.08	8.96
161	2.07	150.21	99.42	0.39	5.58	871.66	0.07	97.12
163	2.75	28.96	58.18	0.26	1.90	352.58	0.02	33.45
164	5.09	453.40	48.48	0.22	2.95	477.05	0.01	54.52
165	2.75	28.96	58.18	0.26	1.90	352.58	0.17	33.45
166	2.07	150.21	99.42	0.39	5.58	932.01	0.07	97.12
167	2.07	150.21	99.42	0.39	5.58	932.01	0.04	97.12
172	2.96	29.92	28.49	0.05	-0.24	381.88	0.07	-4.37
195	2.96	29.92	28.49	0.05	-0.24	381.88	0.01	-4.37
196	1.50	20.36	3.89	0.04	-0.39	741.89	0.01	-6.72

An outlier explication as well as estimation findings are given and discussed in table 4.11. Outliers were detected in about 17 percent of the 204 observations.

Mahanablios distance, Robust MCD Distance, and Standardized Robust Residual, cook's distance and studentized residual, WSSDI, Hat Diagonal, and Deleted Residual were the approaches used to find outliers. The value of outliers in Mahanablios distance has been observed to be between 1.50 to 5.09, whereas in Robust MCD 17.73 to 453.40, respectively. The range of observations in standardised robust residuals, cooks distance, WSSDI, Hat diagonal, and deleted are between -6.53 to 99.42, 0.01 to 0.78, -0.86 to 5.58, 214.43 to 932.01, 0.01 to 0.17, -15.45 to 97.12 respectively.

Outlier and leverage diagnostics for the dependent variable for real data for the quadratic production function are shown in Fig 4.7 and 4.8. Out of 204 observations, there are 36 outliers and 47 leverage points, as observed in the Fig. A scatter plot of the standardised robust residuals vs. robust distances is very useful for identifying outliers and leverage points (RD plot). In Fig 4.8, the robust distances are plotted against the conventional Mahalanobis distances. The plot can be used to identify outliers, as well as high leverage points.

Table 4.12: Estimates of Apple yield w.r.t labour and capital for quadratic production function through Ordinary Least Square (OLS).

Variable	**Regression coefficient (S.E)**	**Parameters**
Intercept	45.81* (9.64)	F-value 192.70(p<0.0001) R^2=0.82 AIC=1746.82 BIC=1770.05
x_1	-0.03 ^{NS} (0.08)	
x_2	0.02 ^{NS} (4.56)	
${x_1}^2$	0.05 ^{NS} (0.07)	
${x_2}^2$	0.03* (0.01)	
$x_1 * x_2$	-0.02^{NS} (0.05)	

*** 5% level of significance NS= non significant**

The estimates results for the Quadratic production function with standard error have been shown in the the Table 4.12. The F-value was found to be significant indicates the model is adequate w.r.t to study variable. It further shows that each estimation coefficient may be having the correct sign due to the presence HLP. The estimates of the regression coefficient have been done through OLS.The R^2 value found to be (0.82) and having AIC (1746.82) and BIC (1770.05), this may be due to the presence of HLP.Therfore robust methods have been applied to study the effect on the estimates of the parameter of quadratic function.

The panel of diagnostics for the quadratic production function model using ordinary least squares found to be inadequate. This shows that this model is a resounding failure. In Fig 4.9poor trends can be seen in the residuals and studentized residuals plots versus predicted values. The points on the plot of the dependent variable (y) versus the anticipated values predict the dependent variable's behaviour. The plot of studentized residual versus leverage reveals that leverage points outside of the range -2 to +2 are HLP. Furthermore, the cook's graph showed that the majority of observations are outside, and the huge jump in the cut-off plainly suggests that data are influenced by outliers.

Table 4.13:Estimation of quadratic functions through robust method for Apple yield w.r.t labour and capital and comparison with OLS thereof.

Variable	Estimates of the regression coefficients					
	OLS (S.E)	M Estimation (S.E)	LTS Estimation (S.E)	MM Estimation (S.E)	S Estimation (S.E)	Modified Estimate after handling HLP (S.E)
Intercept	45.81* (9.64)	28.50 NS (1.67)	25.449 NS (1.599)	3.29 NS (0.78)	5.21 NS (0.14)	38.25 NS (12.07)
x_1	-0.03 NS (0.08)	-0.01* (0.01)	-0.084* (0.012)	0.07 NS (0.09)	0.09* (0.06)	0.02* (0.01)
x_2	0.02 NS (4.56)	0.08 NS (0.01)	0.045* (0.009)	-0.08* (0.08)	0.01* (0.01)	0.01* (0.01)
${x_1}^2$	0.02 NS (0.07)	0.01* (0.01)	0.02* (0.01)	0.01* (0.03)	0.01* (0.31)	0.06 NS (0.06)
${x_2}^2$	0.03* (0.01)	0.07* (0.01)	0.06* (0.01)	0.05* (0.01)	0.03* (0.02)	0.03* (0.08)
$x_1 * x_2$	-0.02 NS (0.05)	-0.01* (0.04)	-0.02* (0.04)	-0.02* (0.02)	-0.01* (0.01)	-0.01* (0.06)
R^2	0.82	0.75	0.99	0.77	0.72	0.83
AIC	1837.94	345.33	-683.84	166.30	384.21	313.21
BIC	1861.17	371.66	-663.93	197.13	391.12	360.35

*** 5%level of significance NS= non significant**

The table 4.13 displays the estimation results for the Quadratic production function, demonstrating that each estimation coefficient may be correct in terms of the expected sign. Following the treatment of HLP, the regression coefficient was estimated using OLS, M Estimation, MM, S, LTS, and modified OLS. The OLS estimate of ${x_1}^2$was found to be positive 0.02 and non-significant, further the M, MM, S Estimation, and LTS were found to be positive and significant i.e. 0.01, 0.01, 0.01, and 0.02 respectively. Modified estimate of the production function was also found to be positive and non significant after handling HLP found to be 0.06. Furthermore, the coefficient of determination for LTS estimation 0.99 is much greater than the coefficient of determination for MM estimation 0.77, indicating that this method can explain 99 percent of the variation in the study variable but is not advisable Chen. C (2002).Modified OLS after handling HLP has $R^2$0.83 which also shows better results. The AIC and BIC for Modified OLS estimation is also low. When unusual observations are omitted from the study, the remaining observations form a linear pattern that explains a greater proportion of the variation in the simulated data. It was also observed that the OLS estimate of ${x_2}^2$was positive 0.03and significant, whereas the M, MM, S, and LTS estimates were positive and significant 0.07, 0.05, 0.03, and 0.06 respectively. After handling HLP modified estimate of the production function was found to be positive and significant 0.03.Overall it has been observed that the estimate of x_1 labour cost is non significant as per OLS method whereas found to be significant in case of robust method. Since LTS and S estimation are usually applied in trimming of data respectively. Therefore among the rest three the modified OLS after handling HLP having maximum R^2 as compared to M and MM estimation. Thus if one unit increases in the cost of labour will decrease overall yield of apple. It may be due to reduction of labour as the total cost deviates in terms of labour and will be decreased by two percent of the total. The estimate of $x_2$capital cost is also non significant as per OLS method whereas found to be significant in case of robust method. If there is one unit increase in the capital cost will increase the yield of apple and it increases by one percent. Now the case of second real data set is considered to observe the behaviour of the linear production function in presence of HLP.

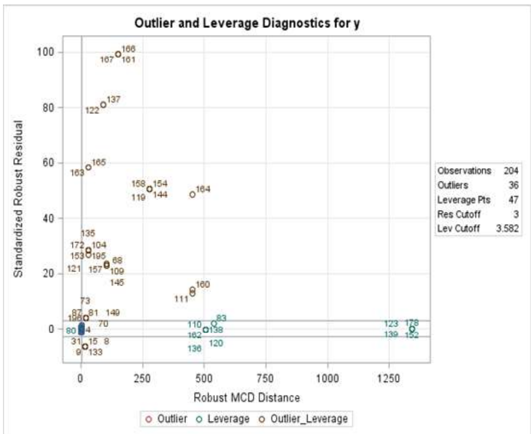


Fig 4.7 : Outlier and Leverage diagnostics for dependent variable for Apple data for quadratic production function.

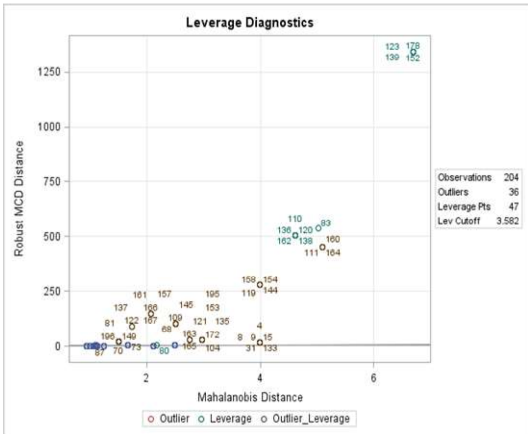


Fig 4.8 : Leverage diagnostics for dependent variable for Apple data for quadratic production function.

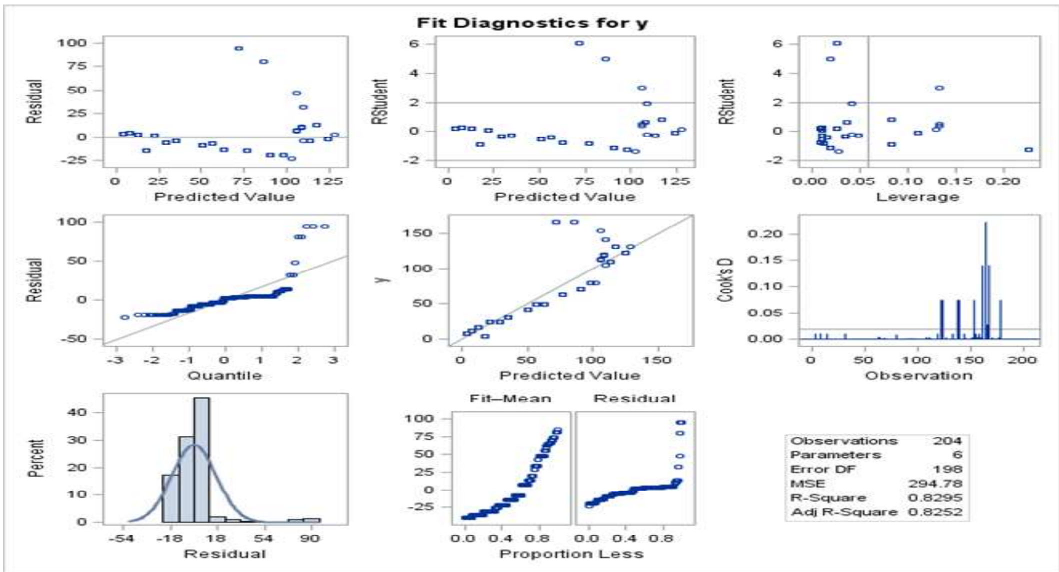


Fig 4.9: Diagnostics Panel of quadratic production function through OLS.

Table 4.14: Socio-economic characteristics of farmers involved in maize yield.

Age (Years)	Frequency	Percentage	Test value	P value
21-30	03	3.09		
31-40	22	22.68	$\chi^{2=}$21.89*	0.013
41-50	25	25.77		
51-60	30	30.92		
Above 60	17	17.52		
Sex(M/F) Frequency Percentage				
Male	70	72.16		
Female	27	27.83	Z=4.4*	0.0016
Marital Status(Married/Unmarried)				
Married	80	82.47		
Unmarried	17	17.52	Z=7.0 *	0.002
Occupation(farming/other)				
Farming	85	87.62		
Other than farming	12	12.37	Z=7.4*	0.001

*** Significant at 5% los.**

Table 4.14 illustrates that age has a significant impact on farming activities since it impacts the efficiency and competency of labour available for maize production. According to the figures, the majority of respondents 82.47 percent were between the ages of 21 and 60. This implies that, maize cultivation is also done by young adult farmers mostly within this age bracket. This is because these categories of farmers are still strong and have the ability to supply the required labour in agricultural activities to boost production as well to increase resource use efficiency. The chi square value (21.89) was found to be significant which indicates the maximum number of farmers involved in the production of maize were in the age group of 51-60 followed by farmers in age group of 41-50 and least in 21-30 in case of maize production. The result also revealed that about 87.62 percent farmers have opted farming as occupation. Years of farming experience in low-land maize production are expected to improve individuals' expertise of better farm management methods and resource utilisation efficiency, lowering costs and improving output. Further the participation of male (70 percent) is more in maize production and married (80 percent) as their Z values were significant. Thus males and married are more involved in the production of maize in Jammu region having farming as main occupation.

Table 4.15: Summary Statistics of maize w.r.t labour and capital for quadratic production function.

Variable	Q_1	Median	Q_3	Mean	Standard Deviation	MAD	CV (%)
x_1 labour (Rs/day)	1800.00	3200.0	6000.0	4750.51	4402.61	2965.21	92.67
x_2 Capital(Rs/kanal)	500.00	800.0	200.0	1308.00	1154.21	8153.10	88.23
Y Yield(q/hectare)	0.55	0.87	1.5	0.77	1.11	0.55	3.47

The mean and standard deviation of the variable in this study based on maize data were 0.77 q/ha and 1.11q/ha, respectively, in table 4.15, demonstrating that the data has more variability because the mean is less than the standard deviation, which is attributable to the presence of influential observations. The study variable y had the smallest mean absolute deviation of 0.55 q/ha. The Q_1 for x_1 and $x_2$were positive 1800.00 Rs/day and 500.00 Rs. /kanal, respectively, while the $Q_1$for y was positive 0.55 unit. A low standard deviation indicates that the data points are close to the mean, whereas a high standard deviation indicates that they are spaced.

Fig 4.10 shows the outlier and leverage diagnostics for the dependent variable for real data for the quadratic production function. It is observed that there are 3 outliers and 33 leverage points among the 97 observations. Outliers and leverage points can be easily identified using a scatter plot of standardised robust residuals vs. robust distances (RD plot). The robust distances are plotted against the traditional Mahalanobis distances in Fig 4.11. Outliers, as well as high leverage points can be identified using the plot.

Table 4.16: Detection of outliers in case of maize data and values of distance/residual through various techniques for quadratic production function.

Observations (HLP)	Mahalanobis	Robust MCD	Standardized Robust	Cook's	Studentized	WSSDI	Hat Diagonal	Deleted
19	1.61	3.34	3.59	0.05	2.99	3761	0.01	17.92
75	1.75	1.99	3.47	0.05	2.75	1116	0.01	16.75
91	3.91	21.15	3.95	0.30	2.98	1008	0.07	19.75

Table 4.16 provides a description of outliers as well as estimation results that are shown and discussed. Outliers were found in around 3 percent of 97 observations. The methodologies employed to find out outliers were Mahanablios distance, Robust MCD Distance, and Standardized Robust Residual, cook's distance and studentized residual, WSSDI, Hat Diagonal, and Deleted Residual. Outliers in Mahanablios distance range from 1.61 to 3.91, while in Robust MCD they range from 1.99 to 3.34, respectively. Standardised robust residuals, cooks distance, WSSDI, Hat diagonal, and deleted observations ranged from 3.47 to 3.95, 0.05 to 0.30, 2.75 to 2.99, 1008 to 3761, 0.01 to 0.07, 16.75 to 19.75 respectively.

Table 4.17: Estimates of Maize w.r.t labour and capital for quadratic production function through Ordinary Least Square (OLS).

Variable	Regression coefficient (S.E)	Parameters
Intercept	7.98* (1.810)	F-value 107.63 (p<0.0001) R^2=0.85 AIC=632.28 BIC=650.30
x_1	0.03* (0.01)	
x_2	-0.05^{NS} (0.01)	
${x_1}^2$	-5.29^{NS} (2.56)	
${x_2}^2$	1.17^{NS} (4.47)	
$x_1 * x_2$	2.16^{NS} (1.46)	

*** 5%level of significance NS= non significant**

The standard error estimates for the Quadratic production function are shown in Table 4.17. The coefficient of $x_1$was found to be positive and significant, while as $x_2$was found to be negative and non-significant. F-value was significant, implies that the model is appropriate for the study variable. It indicates that each estimated coefficient may have the correct sign due to the presence of HLP. The regression coefficient estimations were calculated using OLS. Furthermore, with AIC (632.28) and BIC (650.30), the R^2 value is 0.85, which could be related to the presence of HLP. As a result, robust procedures were employed to explore the impact on quadratic function -parameter estimations.

Fig 4.12 shows the diagnostic panel for the quadratic production function model based on ordinary least squares was found to be insufficient. These results revealed that this model is a total failure. The residuals and studentized residuals graphs vs projected values show poor trends. The behaviour of the dependent variable (y) is predicted by the points on the plot of the dependent variable (y) versus the expected values. The plot of studentized residual vs. leverage showed that leverage points outside of the -2 to +2 range are HLP. Furthermore, the Cook's graph demonstrates that the majority of observations are made outside, and the large leap in the cut-off clearly indicates that data are influencing the decisions.

Table 4.18: Estimation of quadratic functions through robust methods for maize yield w.r.t labour and capital and comparison with OLS thereof.

Variable	Estimates of the regression coefficients					
	OLS (S.E)	**M Estimation (S.E)**	**LTS Estimation (S.E)**	**MM Estimation (S.E)**	**S Estimation (S.E)**	**Modified Estimate after handling HLP (S.E)**
Intercept	7.98* (1.80)	7.36* (1.51)	3.32* (1.45)	7.05* (1.76)	6.57* (1.79)	7.56* (1.58)
x_1	0.03* (0.01)	0.03^{NS} (0.03)	0.05* (0.02)	0.03^{NS} (0.04)	0.03^{NS} (0.04)	0.05* (0.01)
x_2	-0.05^{NS} (0.01)	-0.04^{NS} (0.06)	-0.03^{NS} (0.01)	0.02* (0.01)	0.01^{NS} (0.05)	0.04* (0.01)
${x_1}^2$	-5.29^{NS} (2.50)	-5.67^{NS} (0.01)	-1.31^{NS} (2.16)	-2.30^{NS} (0.05)	-3.12^{NS} (0.31)	-6.21^{NS} (2.27)
${x_2}^2$	1.17^{NS} (4.41)	9.73* (0.01)	5.37^{NS} (3.62)	6.67^{NS} (6.24)	5.54* (2.21)	9.31* (3.90)
$x_1 * x_2$	2.16^{NS} (1.40)	2.71* (0.01)	3.20* (1.27)	2.30* (0.11)	3.10* (0.01)	3.18* (1.30)
R^2	0.85	0.88	0.92	0.63	0.82	0.87
AIC	632.28	123.21	-121.46	168.77	181.75	605.27
BIC	650.30	139.54	-106.01	190.03	197.17	630.21

*** 5%level of significance NS= non significant**

The estimation results for the Quadratic production function are shown in Table 4.18, indicating that each estimation coefficient may be right in terms of expected sign. The regression coefficient was estimated using OLS, M Estimation,

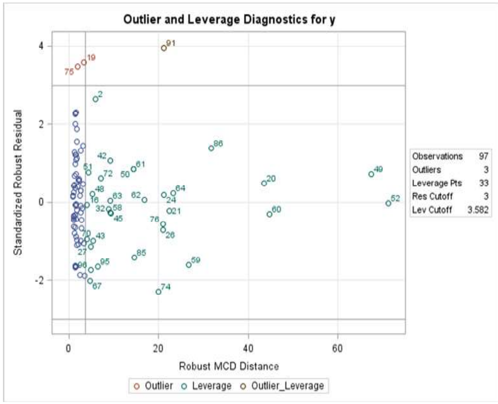


Fig 4.10 : Outlier and Leverage diagnostics for dependent variable for maize data for quadratic production function

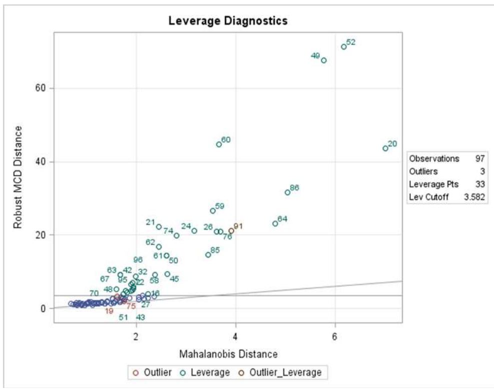


Fig 4.11 : Leverage diagnostics for dependent variable for maize data for quadratic production function

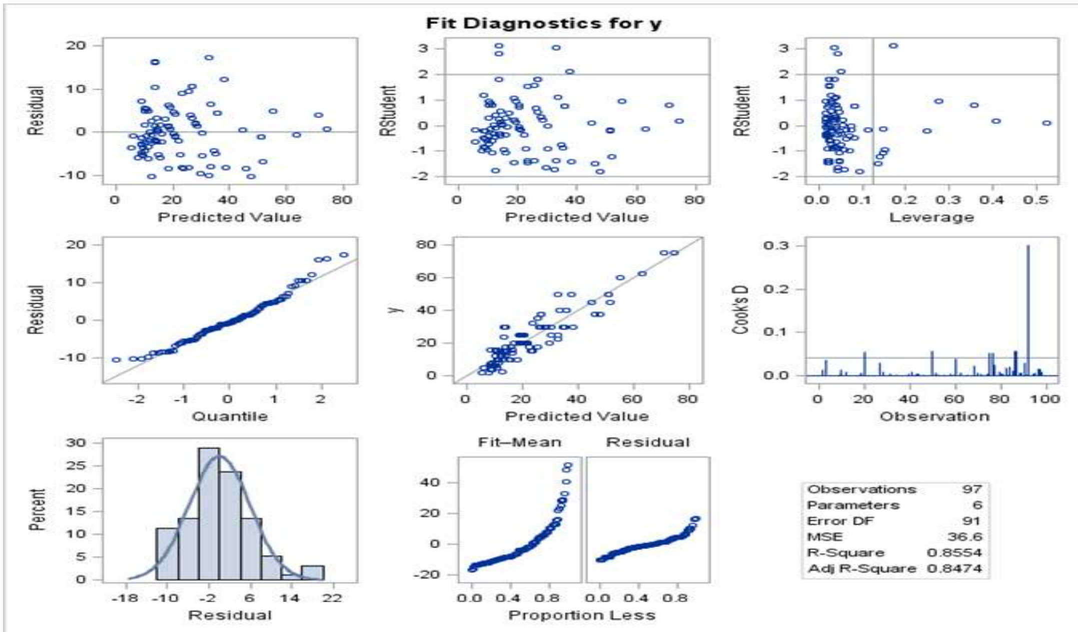


Fig 4.12 : Diagnostics Panel of quadratic production function through OLS.

MM, S, LTS, and modified OLS, in accordance with HLP's behaviour. The OLS estimate of ${x_1}^2$was found to be negative 5.29 and non-significant. Furthermore the M, MM, S Estimation, and LTS estimates i.e. negative and non-significant 5.67, 2.30, 3.12, and 1.31 respectively. After handling HLP, modified estimates of the production function were also found to be negative and non-significant (-6.21). Furthermore, the coefficient of determination for LTS estimate 0.92 is significantly higher but is not recommended Chen and Cary (1987) M estimation has R^2value 0.88, indicating that this method can explain 88 percent of the variation in the study variable, the modified OLS has an R^2 of 0.87, indicating that it provides positive results. For M estimation, the AIC and BIC are also low. When unusual observations are excluded from the study, the remaining observations form a linear pattern that explains a larger percentage of the variation in the simulated data. The OLS estimate of ${x_2}^2$ was also positive 1.17 but non-significant, whereas M and S estimates were positive and significant, MM and LTS estimates were positive but non-significant 6.67 and 5.37 respectively. The improved estimates of the production function were found to be positive and significant 9.31 after handling HLP. In general, the estimate ofx_1 labour cost was determined to be significant using the OLS approach, but non significant using the M, MM, and S estimation methods of the robust method. As a result, a one-unit decrease in labour costs will result in a higher overall maize yield. Because the overall cost of labour will rise by 5 percent. According to the OLS approach, the estimate of $x_1$capital cost is non-significant however it is significant when using the robust method. The yield of maize will decline by four percent if the capital cost is reduced by one unit.

4.1.2 Estimation of Square root production function through Robust method(s).

Table 4.19: Summary Statistics of simulation data $(\mathbf{P_1})$w.r.t. endogenous and exogenous variables for Square root production function.

Variable	Q_1	Median	Q_3	Mean	Standard Deviation	MAD	C V (%)
x_1	-0.59	0.02	0.75	0.05	0.98	0.99	1650.50
x_2	-0.76	-0.07	0.69	0.03	1.03	1.08	2700.52
y	2.80	7.58	13.77	24.35	38.27	8.30	157.18

Table 4.19 shows that the study variable's mean and standard deviation based on simulated data were 24.35 unit and 38.27unit respectively having more variation in

the data because the mean is found to be smaller than the standard deviation, which is due to presence of influencing observations. Furthermore, the mean and standard deviation of auxiliary variables $x_1$andx_2 were determined to be 0.05 unit and 0.98 unit; 0.03 unit and 1.03 unit respectively, indicating that the simulated data set has higher variability. The smallest among the mean absolute deviation was 0.99 units forx_1. The $Q_1$for $x_1 and\ x_2$ were negative 0.59 and 0.76, respectively, whereas the Q_1 for y was positive 2.80. Low standard deviation means the data points are close to the mean, while high standard deviation means the data points are spread out over a wide range of values. The coefficient of variation for study variable is found to be low as 157.18 percent but the coefficient of variation for x_1 and x_2 is quite high 1650.50 and 2700.52 percent respectively, showing inconsistency in the data set, which could be due to HLP.

Outlier and leverage diagnostics for the dependent variable for simulated data for the square production function are shown in Fig 4.13. Out of 500 observations, there are 100 outliers and 121 leverage points, as seen in the diagram. A scatter plot of the standardized robust residuals vs. robust distances is very useful for identifying outliers and leverage points (RD plot). In Fig 4.14, the robust distances are plotted against the conventional Mahalanobis distances. The plot can be used to identify outliers, as well as high leverage points.

Table 4.20: Detection of outliers of simulation data ($\mathbf{P_1}$) and values of distance/residual through various techniques for square root production function.

Observations	Mahalanobis	Robust MCD	Standardized Robust	Cook's	Studentized	WSS DI	Hat Diagonal	Deleted
5	2.46	3.23	169.93	0.01	2.07	0.04	0.02	75.44
10	2.49	3.63	151.97	0.01	1.99	0.07	0.02	74.72
15	2.40	3.27	163.50	0.01	2.03	0.02	0.02	77.13
20	1.09	1.34	146.86	0.04	2.00	0.01	0.01	77.91
25	2.26	2.85	139.41	0.02	1.93	0.01	0.02	79.17
30	1.96	2.61	150.96	0.09	2.01	0.03	0.01	74.05
35	2.86	3.87	145.83	0.01	1.95	0.09	0.02	69.88
40	1.19	1.51	143.78	0.04	2.01	0.01	0.01	78.63
45	2.57	3.61	128.18	0.07	1.89	0.01	0.02	74.56
50	2.29	3.12	143.99	0.02	1.96	0.01	0.02	80.23
55	1.41	1.87	146.95	0.06	2.00	0.03	0.01	75.77
60	2.90	4.13	173.11	0.06	2.09	0.06	0.02	75.07
65	2.03	2.42	144.78	0.03	1.92	0.01	0.01	78.37
70	1.63	2.04	132.99	0.07	1.92	0.02	0.01	72.31
75	1.50	1.74	150.08	0.03	2.00	0.01	0.01	80.01
80	4.79	7.64	187.95	0.04	2.19	0.12	0.06	75.80
85	2.46	3.46	139.43	0.02	1.93	0.01	0.02	79.86

90	2.32	3.13	122.05	0.02	1.90	0.06	0.02	68.89
95	3.96	6.48	163.46	0.08	2.03	0.02	0.04	80.81
100	3.05	4.15	169.66	0.07	2.06	0.10	0.03	73.86
105	1.10	1.23	133.99	0.04	1.96	0.08	0.01	76.09
110	2.11	3.01	120.98	0.08	1.88	0.03	0.02	68.41
115	1.03	1.16	153.84	0.04	2.00	0.08	0.01	78.17
120	1.20	1.56	142.54	0.06	1.97	0.02	0.01	74.81
125	1.59	2.12	125.01	0.06	1.89	0.03	0.01	70.22
130	1.93	2.70	157.13	0.05	2.10	0.02	0.01	84.11
135	0.94	1.10	146.06	0.04	2.02	0.01	0.02	78.35
140	2.03	2.38	147.15	0.03	1.98	0.08	0.01	80.44
145	2.38	3.12	157.64	0.01	2.05	0.04	0.02	75.02
150	1.67	2.03	132.89	0.06	1.92	0.01	0.01	73.78
155	1.23	1.51	156.68	0.05	2.03	0.09	0.01	79.18
160	2.90	3.49	116.84	0.07	1.88	0.04	0.03	68.53
165	1.84	2.69	151.33	0.08	2.02	0.04	0.01	77.01
170	0.95	1.13	148.51	0.04	2.00	0.01	0.04	77.03
175	2.85	4.97	155.96	0.10	2.03	0.06	0.02	78.81
180	1.81	2.20	125.89	0.06	1.89	0.01	0.01	72.15
185	1.08	1.17	154.05	0.03	2.03	0.06	0.01	80.27
190	1.72	2.23	148.20	0.08	1.98	0.03	0.01	73.54
195	1.44	1.81	160.98	0.07	2.03	0.02	0.01	76.08
200	0.94	1.12	156.96	0.05	2.04	0.01	0.02	78.74
205	2.20	3.01	144.15	0.05	1.95	0.05	0.02	70.96
210	2.09	2.71	156.78	0.12	2.06	0.03	0.01	76.10
215	4.33	7.94	162.51	0.09	2.07	0.13	0.05	80.38
220	2.05	2.53	134.60	0.05	1.96	0.08	0.01	78.52
225	4.72	6.97	108.89	0.02	1.87	0.23	0.07	63.87
230	1.82	2.49	159.09	0.07	2.01	0.06	0.01	77.60
235	2.81	4.20	149.62	0.01	1.99	0.05	0.02	71.52
240	1.29	1.47	144.47	0.03	2.00	0.06	0.01	79.87
245	2.02	2.38	151.88	0.03	2.03	0.02	0.01	82.85
250	1.39	1.76	159.26	0.07	2.01	0.08	0.01	75.33
255	4.60	7.74	139.44	0.02	2.00	0.18	0.05	68.25
260	3.04	4.31	165.22	0.01	2.06	0.18	0.03	76.07
265	2.26	2.98	142.55	0.01	1.96	0.06	0.02	71.44
270	1.14	1.41	131.42	0.05	1.96	0.01	0.01	74.15
275	1.47	1.75	147.53	0.03	1.98	0.02	0.01	79.73
280	1.49	1.70	149.09	0.03	1.97	0.08	0.01	79.04
285	2.28	3.81	154.42	0.07	2.04	0.01	0.02	80.93
290	1.63	1.94	126.25	0.06	1.90	0.08	0.01	71.93
295	3.64	5.43	125.29	0.01	1.94	0.03	0.04	75.74
300	1.71	2.48	133.58	0.05	1.94	0.02	0.01	75.87
305	3.32	4.19	172.68	0.02	2.06	0.05	0.03	75.87
310	3.48	5.28	130.77	0.07	1.95	0.01	0.04	79.24
315	1.12	1.39	150.07	0.05	1.99	0.01	0.01	76.78
320	3.14	4.90	161.94	0.09	2.06	0.02	0.03	81.14
325	2.92	4.05	169.15	0.05	2.06	0.09	0.02	73.77
330	3.54	5.55	156.38	0.09	2.04	0.07	0.04	72.22
335	1.39	1.51	139.29	0.04	1.98	0.06	0.01	78.57
340	0.90	1.04	155.89	0.04	2.02	0.01	0.02	78.32
345	2.17	3.47	132.21	0.06	1.92	0.03	0.01	75.68
350	2.52	3.35	166.41	0.02	2.04	0.03	0.02	76.37
355	2.37	3.18	119.88	0.07	1.89	0.03	0.02	68.70
360	1.78	2.28	124.77	0.07	1.91	0.02	0.01	70.88
365	4.77	7.89	143.63	0.02	1.95	0.02	0.06	84.18
370	1.49	1.84	128.03	0.06	1.92	0.03	0.01	71.95
375	3.96	7.24	127.82	0.01	1.90	0.08	0.05	75.88
380	1.33	1.57	129.99	0.05	1.94	0.01	0.01	73.80
385	1.13	1.32	149.10	0.04	1.98	0.05	0.01	78.27
390	1.42	1.74	148.78	0.04	2.01	0.01	0.01	79.42
395	1.27	1.51	152.29	0.04	2.03	0.05	0.01	80.28
400	2.26	2.96	128.07	0.07	1.89	0.02	0.02	72.63
405	1.82	2.42	148.36	0.08	2.01	0.49	0.01	74.92
410	1.51	1.97	135.90	0.07	1.92	0.26	0.01	71.74
415	1.79	2.36	160.96	0.09	2.04	0.04	0.01	76.59
420	0.92	1.09	147.38	0.04	1.99	0.01	0.01	76.64

425	1.68	2.35	153.01	0.05	2.05	0.02	0.01	81.35
430	3.42	5.17	127.92	0.04	1.90	0.12	0.04	71.98
435	2.17	2.76	150.69	0.03	2.02	0.17	0.01	82.60
440	3.91	5.81	124.92	0.13	1.94	0.84	0.05	75.31
445	1.58	1.92	125.95	0.06	1.89	0.03	0.01	70.64
450	1.46	1.73	133.01	0.06	1.98	0.01	0.01	76.18
455	1.23	1.50	142.60	0.05	1.98	0.02	0.01	75.97
460	1.98	2.80	122.60	0.08	1.89	0.03	0.01	69.07
465	2.55	4.15	163.14	0.01	2.10	0.07	0.02	80.11
470	1.96	2.59	166.15	0.03	2.07	0.04	0.01	76.05
475	0.97	1.09	153.35	0.04	2.00	0.01	0.02	78.38
480	1.49	1.89	160.80	0.07	2.03	0.03	0.01	75.99
485	2.51	3.26	168.41	0.04	2.09	0.08	0.02	76.14
490	3.31	5.15	156.72	0.06	2.02	0.19	0.03	75.5
495	1.03	1.21	130.71	0.04	1.89	0.06	0.01	73.24
500	2.07	2.46	139.59	0.03	1.96	0.05	0.01	79.96

An outlier explanation and estimation results are shown and discussed in Table 4.20. Outliers were found in about 20 percent of the 500 observations. Mahanablios distance, robust MCD Distance, standardized robust residual, cook's distance, studentized residual; WSSDI, Hat diagonal, and deleted residual were among the techniques used to find outliers. The value of outliers using Mahanablios distance is between 0.90 and 4.77 unit and through Robust MCD lying between 1.04 to 7.94 units. For standardised robust residuals, cooks distance, WSSDI, Hat diagonal, and deleted residual the range of observations is between 108.89 to 187.95, 0.01 to 0.13, 1.87 to 2.19, 0.01 to 0.84, 0.01 to 0.07, 63.87 to 84.18 unit respectively.

Table 4.21: Estimates of Square root production function(P_1)through Ordinary Least Square (OLS).

Variable	Regression coefficient (S.E)	Parameters
Intercept	15.74* (12.40)	F-value 0.93 (p<0.01) R^2=0.09 AIC=5072.04 BIC=5101.54
x_1	0.82^{NS} (1.75)	
x_2	1.81^{NS} (1.66)	
${x_1}^{1/2}$	3.78^{NS} (13.10)	
${x_2}^{1/2}$	3.53^{NS} (13.44)	
$(x_1 * x_2)^{1/2}$	3.74^{NS} (14.53)	

*** 5%level of significance NS= non significant**

The estimated results for the square root production function with standard error are presented in Table 4.21. All of the coefficients were positive and non-significant, and the F-value(0.93*) was significant, indicating that the model is appropriate in terms of the study variable. Because of the presence of HLP, each estimated coefficient may have the proper sign. The OLS method was used to calculate the regression coefficient estimations. The R^2 value is also the lowest (0.09), with AIC (5072.04) and BIC (5101.54), which could be attributed to the existence of HLP. As a result, robust approaches have been used to examine the effect on the estimates of the square root function parameter.

In Fig 4.15 the diagnostic panel for the square root production function model based on ordinary least squares was assessed and found to be inadequate. These findings demonstrated that this model is a complete failure. Poor trends may be seen in the residuals and studentized residuals graphs vs projected values. The points on the plot of the dependent variable (y) versus the expected values predict the behaviour of the dependent variable (y). Leverage points outside of the -2 to +2 range are HLP, as seen by the plot of studentized residual vs. leverage. In addition, Cook's graph showed that the majority of observations are taken outside, and the big jump in the cut-off plainly implies that data are being collected outside.Further

Table 4.22: Estimation of square root functions through robust method for simulation data (P_1) and comparison with OLS thereof.

Variable	Estimates of the regression coefficients					
	OLS (S.E)	**M Estimation (S.E)**	**LTS Estimation (S.E)**	**MM Estimation (S.E)**	**S Estimation (S.E)**	**Modified Estimate after handling HLP (S.E)**
Intercept	15.74 ^{NS} (12.40)	5.13* (0.18)	5.16* (0.18)	5.13* (0.19)	5.13* (0.19)	4.94* (1.63)
x_1	0.82 ^{NS} (1.75)	5.02* (0.02)	5.02* (0.02)	5.03* (0.07)	5.03* (0.02)	4.78* (0.23)
x_2	1.81 ^{NS} (1.66)	3.04* (0.02)	3.05* (0.02)	3.06* (0.02)	3.06* (0.02)	2.66* (0.21)
${x_1}^{1/2}$	3.78 ^{NS} (13.10)	-0.12 ^{NS} (0.20)	-0.16 ^{NS} (0.19)	-0.12 ^{NS} (0.20)	-0.12 ^{NS} (0.20)	-0.60 ^{NS} (1.72)
${x_2}^{1/2}$	3.53 ^{NS} (13.44)	-0.05 ^{NS} (0.20)	-0.10 ^{NS} (0.20)	-0.04 ^{NS} (0.20)	-0.04 ^{NS} (0.20)	0.09 ^{NS} (1.79)
$(x_1 * x_2)^{1/2}$	3.74 ^{NS} (14.53)	0.05 ^{NS} (0.20)	0.11 ^{NS} (0.21)	0.04 ^{NS} (0.22)	0.05 ^{NS} (0.22)	1.06 ^{NS} (1.93)
R^2	0.09	0.69	0.98	0.99	0.70	0.54
AIC	5072.04	965.28	-1250.86	107.21	464.64	2974.91
BIC	5101.54	996.30	-1225.57	132.26	697.66	3004.41

*** 5%level of significance   NS= non significant**

the estimation results for the square root production function are shown in Table 4.22, indicating that each estimation coefficient may be correct in terms of expected sign. The regression coefficient was estimated using OLS, M Estimation, MM, S, LTS, and modified OLS, in accordance with HLP's behaviour. The OLS estimate of $x_1$was found to be positive (0.82) and non-significant, and M, MM, S and LTS Estimation were positive and significant (5.02 5.02 5.03 4.78). After handling HLP modified estimate of the production function was also found to be positive and significant (4.78).Furthermore, the coefficient of determination for MM estimate is 0.99 is significantly higher than that for LTS estimation 0.98, indicating that this method can explain 99 percent of the variation in the study variable. When unusual observations are excluded from the study the remaining observations form a linear pattern that explains a larger percentage of the variation in the simulated data. The OLS estimate of x_2 was also positive (1.81) but non-significant, whereas M, MM, S and LTS estimates were positive and significant 3.04, 3.05, 3.06, 3.06 respectively. The improved estimates of the production function were found to be positive and significant 2.66 after handling HLP.

Table 4.23: Summary Statistics of simulation data (P_2)w.r.t. endogenous and exogenous variables for square root production function.

Variable	Q_1	Median	Q_3	Mean	Standard Deviation	MAD	C V (%)
x_1	-0.65	0.02	0.70	0.02	0.99	1.00	4474.32
x_2	-0.78	-0.07	0.68	0.04	1.03	1.08	2592.01
y	-0.49	6.57	13.82	14.44	29.75	10.66	206.00

Table 4.23 indicated that the mean and standard deviation of the study variable based on simulated data were 14.44 and 29.75 unit respectively, showing that there is greater variation in the data because the mean is less than the standard deviation, which is due to the presence of influencing observations. Furthermore, the mean and standard deviation of auxiliary variables x_1 and x_2 were calculated and found to be 0.02 and 0.99 unit; 0.04 and 1.03unit, showing that the simulated data set has larger variability. For x_1 the smallest mean absolute deviation was 1.00 units. $Q_1$for $x_1$and x_2 were both negative 0.65 and 0.78 unit respectively, also $Q_1$for y was negative, it was 0.49.Low standards means that the data points are close to the average, whereas high standards means that the data points are distributed among a given set of values. The coefficient of variation Y is low and was 206.00 percent. But for x_1 (4474.32 percent)

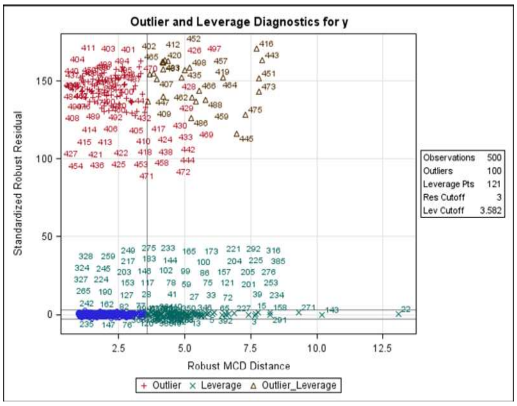


Fig 4.13 : Outlier and Leverage diagnostics for dependent variable for Simulation data (P_1) for square root production function.

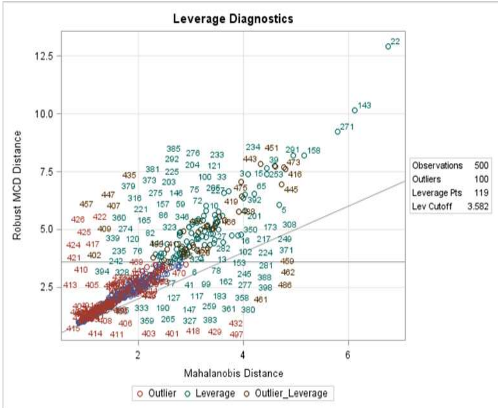


Fig 4.14 : Leverage diagnostics for dependent variable for Simulated data (P_1) for square root production function.

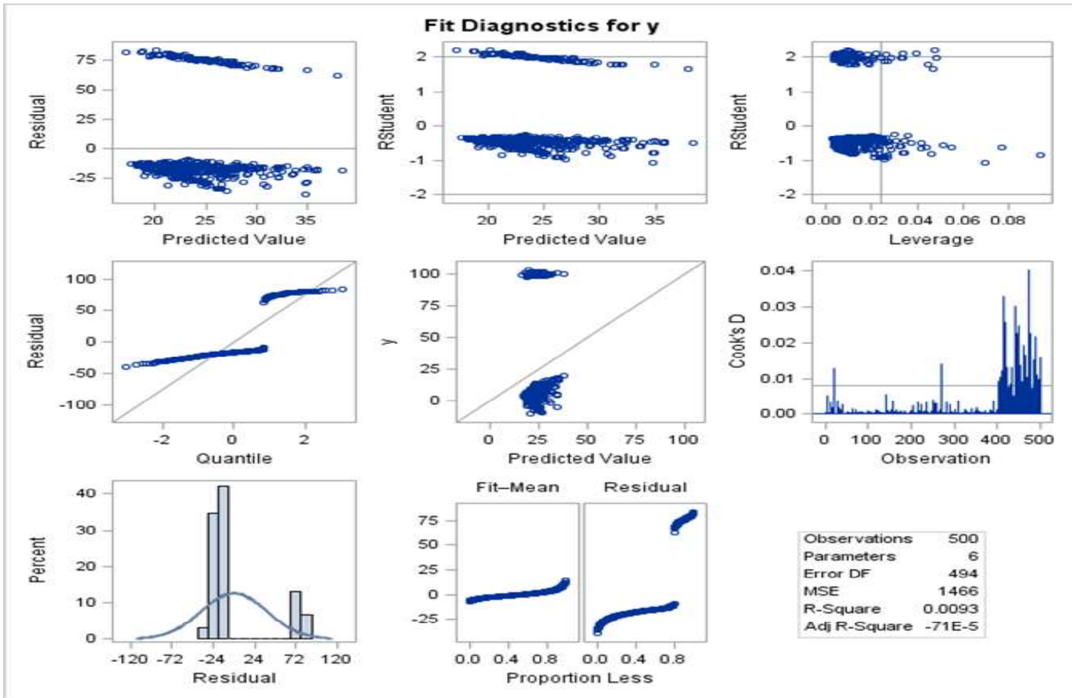


Fig 4.15: Diagnostics Panel of square production function through OLS.

and x_2(2592.01 percent) are relatively high with inconsistencies in the group of data that could be caused by the HLP.

Fig 4.16 shows the outlier and leverage diagnostics for the dependent variable for simulated data for the square root production function. As seen in the diagram, there are 100 outliers and 257leverage points among 1000 observations. Outliers and leverage points can be easily identified using a scatter plot of standardised robust residuals vs. robust distances (RD plot). The robust distances are plotted against the traditional Mahalanobis distances in Fig 4.17. Outliers, as well as high leverage points, can be identified using this plot whereas the

Table 4.24: Detection of outliers of simulation data (P_2)and values of distance/residual through various techniques for square root production function.

Observations (HLP)	Mahalanobis	Robust MCD	Standardized Robust	Cook's	Studentized	WSSDI	Hat Diagonal	Deleted
10	2.87	3.94	3.18	0.04	-0.17	0.70	0.07	-4.39
20	1.96	2.48	3.31	0.06	-0.31	0.32	0.01	-8.54
30	2.45	3.58	176.26	0.08	3.12	0.18	0.01	92.62
40	1.81	2.06	168.08	0.03	2.99	0.03	0.01	86.27
50	1.47	1.70	157.95	0.04	2.85	0.06	0.01	83.04
60	2.02	2.33	188.72	0.04	3.33	0.36	0.02	97.58
70	1.21	1.49	156.01	0.05	2.80	0.20	0.01	81.18
77	3.64	5.32	154.91	0.07	2.75	0.91	0.01	79.95
80	1.57	1.99	153.59	0.06	2.76	0.32	0.01	80.36
90	1.97	2.74	177.87	0.07	3.16	0.22	0.02	91.86
100	1.81	2.11	157.11	0.03	2.83	0.04	0.01	80.65
110	1.58	1.84	186.37	0.07	3.29	0.30	0. 01	95.75
120	2.15	2.67	160.72	0.03	2.90	0.01	0.03	82.93
130	0.89	0.99	177.29	0.05	3.14	0.12	0. 01	91.13
140	4.16	6.04	205.15	0.03	3.59	1.06	0. 01	104.48
150	3.25	4.49	179.23	0.02	3.22	1.02	0.04	93.25
160	2.43	3.55	173.99	0.08	3.08	0.18	0.05	91.49
170	2.47	3.14	134.71	0.09	2.48	0.76	0.03	72.19
180	2.96	3.92	167.03	0.03	2.99	0.01	0.01	85.22
190	4.90	7.59	143.30	0.01	2.66	0.68	0.01	83.32
200	2.96	3.66	188.31	0.02	3.36	0.97	0.02	98.36
210	2.29	2.55	159.58	0.03	2.88	0.08	0.01	82.18
220	1.16	1.26	157.54	0.03	2.85	0.04	0.01	81.98
230	1.66	2.09	142.79	0.06	2.61	0.40	0. 01	75.87
240	2.44	3.12	151.64	0.01	2.73	0.72	0.02	79.51
250	1.93	2.51	135.75	0.07	2.48	0.49	0.03	72.16
260	1.15	1.41	160.09	0.05	2.88	0.20	0.02	83.43
270	1.27	1.60	176.20	0.04	3.12	0.19	0.01	90.44
280	0.92	1.04	161.24	0.07	2.89	0.11	0.04	83.51
290	2.91	3.86	146.30	0.03	2.63	0.92	0.05	76.46
300	2.31	2.90	168.74	0.03	3.02	0.01	0.03	86.03
310	2.29	2.73	191.65	0.06	3.37	0.47	0.04	98.51
320	1.32	1.46	157.78	0.03	2.85	0.04	0.05	82.30
330	3.25	4.97	186.58	0.07	3.32	0.40	0.01	97.98
340	1.43	1.76	175.56	0.05	3.12	0.12	0.01	90.25
350	4.94	7.78	192.34	0.02	3.41	1.03	0.02	102.99
360	2.85	4.26	179.52	0.08	3.19	0.32	0.01	93.66
370	1.16	1.30	155.40	0.04	2.82	0.07	0.01	81.90

380	0.87	0.93	175.17	0.05	3.11	0.09	0.02	90.01
390	1.61	1.93	167.27	0.08	2.97	0.26	0.01	86.94
400	1.74	1.95	170.75	0.03	3.05	0.04	0.02	87.18
410	2.01	2.46	156.96	0.09	2.86	0.32	0.03	84.23
420	1.82	2.37	180.90	0.09	3.20	0.36	0.01	93.08
426	1.89	2.13	173.92	0.04	3.11	0.07	0.01	89.14
430	1.19	1.29	161.83	0.03	2.91	0.04	0.02	84.31
440	1.77	2.07	169.59	0.03	3.03	0.02	0.01	86.50
450	2.83	4.01	136.84	0.01	2.54	0.62	0.04	74.16
460	3.98	5.89	128.22	0.01	2.41	0.98	0.01	70.48
470	3.02	4.15	168.35	0.03	3.01	0.03	0.01	85.47
480	1.09	1.26	175.87	0.06	3.11	0.11	0.04	90.65
490	2.27	3.10	179.25	0.01	3.16	0.26	0.01	93.56
500	2.01	2.60	175.37	0.05	3.13	0.14	0.01	90.51
510	4.02	6.27	145.55	0.01	2.67	0.40	0.03	82.32
520	2.81	4.22	176.55	0.01	3.11	0.27	0.01	93.06
530	1.39	1.52	155.10	0.03	2.80	0.09	0.01	80.48
540	2.35	2.98	167.16	0.03	2.98	0.01	0.02	84.85
550	1.84	2.41	171.51	0.05	3.05	0.15	0.03	88.39
560	1.42	1.73	147.13	0.06	2.69	0.21	0.04	78.61
570	3.31	4.62	201.04	0.05	3.53	0.88	0.02	102.41
580	1.67	2.02	163.14	0.08	2.90	0.29	0.01	84.54
590	4.69	7.15	139.43	0.03	2.55	1.09	0.02	77.50
600	1.81	2.30	152.53	0.04	2.76	0.18	0.01	79.81
610	1.62	2.06	167.48	0.09	3.02	0.31	0.01	87.93
620	2.04	2.57	148.89	0.07	2.68	0.52	0.02	78.26
630	0.94	1.01	174.18	0.04	3.09	0.08	0.03	89.37
640	0.96	1.14	149.51	0.04	2.72	0.16	0.01	78.70
650	1.29	1.61	145.86	0.05	2.66	0.22	0.01	77.57
660	0.98	1.15	150.44	0.04	2.73	0.12	0.02	79.42
670	1.05	1.13	160.80	0.04	2.88	0.07	0.01	82.94
680	2.51	3.45	167.67	0.01	3.04	0.53	0.02	88.49
690	2.20	2.47	172.60	0.04	3.09	0.05	0.01	88.25
700	0.87	0.94	176.74	0.05	3.14	0.10	0.01	90.88
710	3.12	4.64	142.53	0.07	2.58	0.45	0.01	76.20
720	1.77	2.21	149.34	0.07	2.74	0.23	0.01	80.57
730	0.97	1.09	175.94	0.05	3.12	0.10	0.03	90.61
740	2.60	3.43	144.76	0.09	2.66	0.38	0.02	79.64
750	2.67	3.47	193.09	0.08	3.39	0.60	0.02	98.58
760	1.57	1.73	172.44	0.04	3.08	0.05	0.04	88.24
770	3.07	4.39	175.01	0.01	3.16	0.79	0.02	91.58
780	3.35	4.71	179.57	0.01	3.16	0.54	0.01	95.21
790	2.41	2.97	184.83	0.01	3.25	0.41	0.01	96.10
800	1.07	1.13	180.03	0.06	3.19	0.16	0.01	92.60
810	1.52	1.67	160.75	0.03	2.89	0.06	0.01	82.84
820	1.97	2.48	181.96	0.01	3.26	0.45	0.05	94.89
830	2.11	2.55	158.66	0.03	2.85	0.03	0.01	81.20
840	3.03	3.72	175.42	0.02	3.11	0.69	0.01	91.36
850	1.32	1.48	156.02	0.03	2.81	0.03	0.20	80.88
860	2.24	2.82	189.93	0.05	3.34	0.55	0.02	96.90
870	3.35	4.72	161.30	0.09	2.95	0.77	0.01	86.06
880	0.90	0.96	155.42	0.04	2.82	0.08	0.02	81.53
890	1.31	1.59	161.67	0.07	2.88	0.22	0.04	83.79
900	1.31	1.61	167.32	0.05	2.98	0.12	0.02	86.84
910	1.78	2.12	187.32	0.01	3.30	0.36	0.05	96.03
920	0.91	1.07	165.78	0.05	2.98	0.12	0.01	86.64
930	1.56	1.99	152.78	0.05	2.77	0.14	0.01	81.70
940	2.69	3.50	167.53	0.02	2.99	0.09	0.01	85.46
950	2.63	3.37	195.28	0.05	3.43	0.94	0.02	99.63
960	2.28	2.86	158.01	0.01	2.87	0.50	0.04	84.18
970	1.18	1.36	176.34	0.05	3.14	0.45	0.01	90.63
980	1.93	2.38	178.91	0.06	3.20	0.72	0.01	93.15
990	2.76	3.67	160.09	0.03	2.87	0.05	0.02	81.36
1000	2.38	3.02	166.84	0.03	2.96	0.49	0.01	86.60

Table 4.24 shows and discusses an outlier explanation as well as estimation results. Outliers were revealed in roughly 10 percent of the total observations. The values of the outliers were estimated using different techniques as already discussed. The Outliers have a range of 0.87 to 4.94 unit when utilising Mahanablios distance and 0.93 to 7.78 unit when using Robust MCD. The range of observations for standardised robust residuals, cook's distance, WSSDI, Hat diagonal, and deleted residual is 3.18 to 205.15, 0.01 to 0.09, -0.31 to 3.59, 0.01 to 1.09, 0.01 to 0.04, -8.54 to 104.48 unit respectively. The coefficients of the variables in case of square root production function through OLS method is represented in

Table 4.25: Estimates of Square root production function data (P_2) through Ordinary Least Square (OLS).

Variable	Regression coefficient (S.E)	Parameters
Intercept	21.33* (6.48)	F-value 2683.2 (p<0.0001) R^2=(0.05) AIC=1013.09 BIC=1016.45
x_1	3.49* (0.92)	
x_2	5.70* (0.88)	
${x_1}^{1/2}$	-8.09 ^{NS} (7.00)	
${x_2}^{1/2}$	-6.20 ^{NS} (7.16)	
$(x_1 * x_2)^{1/2}$	7.51^{NS} (7.86)	

*** 5%level of significance NS= non significant**

table 4.25 with standard error. The F-value was found to be significant and the coefficients ofx_1 and x_2 were positive and significant i.e 3.49 and 5.70 unit respectivly indicating that the model is appropriate in terms of the study variable. Each estimated coefficient may have the proper sign due to the presence of HLP. The regression coefficient estimations were calculated using OLS. The AIC and SBIC values were1013.09 and 1016.45 unit respectively and the R^2 value is very low 0.05 which could be due to the presence of HLP. As a result, therfore robust methodologies have been employed to investigate the effect on square root function parameter estimates.

In Fig 4.18 the diagnostic panel for the ordinary least squares-based square root production function model was found to be insufficient. These results proved that the model is a complete failure. The residuals and studentized residuals graphs vs projected values show poor trends. The points on the dependent variable (y) against predicted values plot predict the dependent variable's behaviour (y). The plot of studentized residual versus. leverage showed that leverage points outside of the -2 to +2 range are HLP. The estimates of the square root function throgh robust methods have been reflected in

Table 4.26: Estimation of square root functions through robust methods for simulation data (P_2) and comparison with OLS thereof.

Variable	Estimates of the regression coefficients					
	OLS (S.E)	M Estimation (S.E)	LTS Estimation (S.E)	MM Estimation (S.E)	S Estimation (S.E)	Modified Estimate after handling HLP (S.E)
Intercept	21.33* (6.48)	4.99* (0.12)	5.023* (0.12)	5.02* (0.12)	5.02* (0.12)	4.95* (0.11)
x_1	3.49* (0.92)	5.00* (0.01)	5.05* (0.01)	5.05* (0.01)	5.08* (0.01)	5.07* (0.01)
x_2	5.70* (0.88)	7.01* (0.01)	7.01* (0.01)	3.01* (0.01)	3.02* (0.01)	7.01* (0.01)
${x_1}^{1/2}$	-8.09 NS (7.02)	0.03* (0.13)	-0.01 NS (0.13)	-0.02 NS (0.13)	-0.01 NS (0.13)	0.06 NS (0.12)
${x_2}^{1/2}$	-6.20 NS (7.16)	0.08* (0.13)	-0.03 NS (0.13)	-0.04 NS (0.13)	-0.03 NS (0.14)	0.02 NS (0.12)
$(x_1 * x_2)^{1/2}$	7.51 NS (7.86)	-0.03* (0.14)	0.01 NS (0.14)	0.04 NS (0.14)	0.03 NS (0.15)	-0.04 NS (0.14)
R^2	0.05	0.79	0.99	0.77	0.97	0.99
AIC	1013.09	1378.33	-2775.09	784.71	665.23	615.21
BIC	1016.45	1412.06	-2803.72	817.92	635.90	631.25

*** 5%level of significance NS= non significant**

Table 4.26 which shows the results of the estimation for the square root production function, suggesting that each estimation coefficient may be true in terms of predicted sign. According to HLP's behaviour, the regression coefficient was estimated using OLS, M Estimation, MM, S, LTS, and modified OLS. The OLS estimate of ${x_1}^{1/2}$was found to be negative 8.09 unit and non-significant and through M Estimation it was positive 0.03 unit and significant while as in case of MM, S and LTS Estimation were negative and non-significant 0.02,0.01 and 0.01 unit respectively. After handling HLP, modified estimates of the production function were also found to be positive 0.06 and non-significant. Moreover, the OLS estimate of

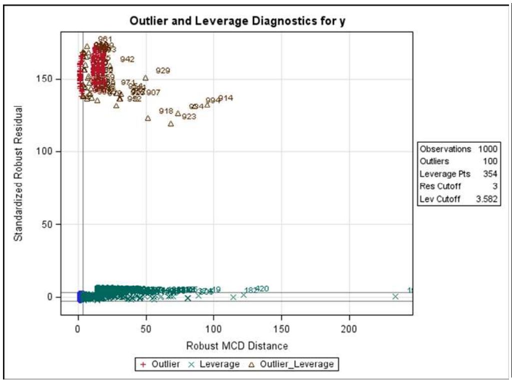


Fig : 4.16 Outlier and Leverage diagnostics for dependent variable for Simulation data (P_2) for square root production function.

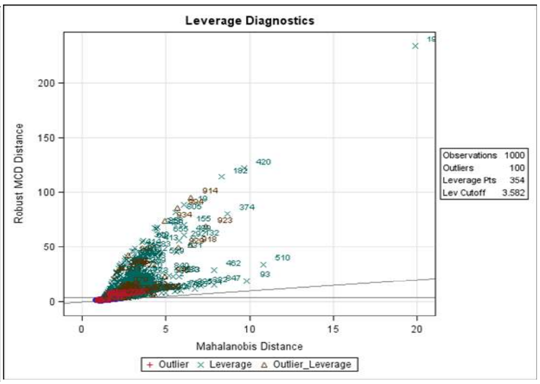


Fig 4.17 : Leverage diagnostics for dependent variable for Simulated data (P_2) for square root production function.

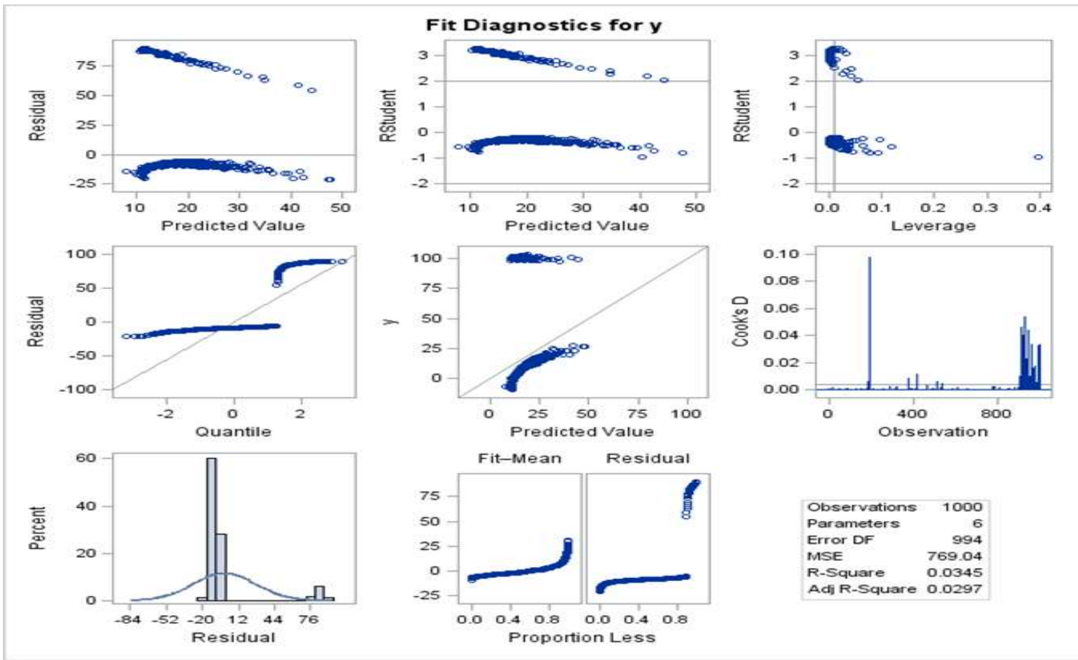


Fig 4.18: Diagnostics Panel of square production function through OLS.

$x_2{}^{1/2}$was also negative 6.20 unit but non-significant whereas M Estimation was positive 0.08 and significant. The estimate through MM, S and LTS estimates was negative and non-significant 0.03, 0.04 and 0.03unit respectively. The estimates of the production function after handling HLP through OLS were found to be positive and non-significant0.02 unit. Now to check the performance of square root production function in case of data pertained to the apple crop is done under the following section. The same procedure has been used.

Table 4.27: Summary Statistics of Apple yield w.r.t labour and capital for square root production function.

Variable	Q_1	Median	Q_3	Mean	Standard Deviation	MAD	C.V (%)
x_1 labour (Rs/day)	11200.0	13500.0	19200.0	14455.4	5267.6	6671.7	36.44
x_2 Capital(Rs/kanal)	11600.0	14000.0	19600.0	15183.8	6058.4	6523.4	39.90
y Yield(q/hectare)	0.80	1.18	3.14	2.05	2.18	0.87	1.76

Table 4.27 indicated that the mean and standard deviation of the study variable based on real data were 2.05 q/ha and 2.18q/ha respectively, showing that there is greater variation in the data because the mean is less than the standard deviation, which is due to the presence of influencing observations. Furthermore, the mean and standard deviation of auxiliary variables $x_1$andx_2 were calculated to be 14455.4 Rs/day and 5267.6Rs/day; 15183.8 Rs/Kanal and 6058.4Rs/Kanal showing that the data set has larger variability. For y the smallest Mean Absolute Deviation were 0.87q/ha. The Q_1 for $x_1$and x_2 were both positive 11200.0Rs/day and 11600.0 Rs/Kanal also $Q_1$for y was positive 0.80 q/ha. Low standards means that the data points are close to the average, whereas high standards means that the data points are distributed among a given set of values. The variation coefficient of Y is low at 35.21q/ha. But for x_1 (36.44)Rs/day and x_2 (39.90)Rs/kanal are relatively high, with inconsistencies in the group of data that could be caused by the HLP.

Fig 4.19 shows the outlier and leverage diagnostics for the dependent variable for real data for the square root production function. As seen in the Fig, there are 24 outliers and 46 leverage points among 204 observations. Outliers and leverage points can be easily identified using a scatter plot of standardised robust residuals vs. robust distances (RD plot). The robust distances are plotted against the traditional

Mahalanobis distances in Fig 4.20. Outliers, as well as high leverage points can be identified using the plot.

Table 4.28: Detection of outliers in case of Apple yield data and values of distance/residual through various techniques for square root production function.

Observations (HLP)	Mahalanobis	Robust MCD	Standardized Robust	Cook's	Studentized	WSSDI	Hat Diagonal	Deleted
68	2.3	36.88	10.07	0.08	1.01	593.40	0.01	23.21
104	2.6	21.87	16.18	0.02	0.58	383.42	0.01	13.34
109	2.3	36.88	10.07	0.08	1.01	593.45	0.01	23.21
111	4.49	163.47	3.80	0.03	1.45	480.44	0.09	33.74
119	3.61	110.24	27.60	0.02	1.76	216.37	0.01	40.82
121	2.51	20.93	15.11	0.01	0.47	355.58	0.01	10.79
122	1.70	35.45	33.45	0.08	3.79	331.01	0.02	86.82
135	2.67	21.80	16.18	0.02	0.58	383.40	0.01	13.34
137	1.70	35.42	33.45	0.05	3.79	331.02	0.02	86.82
144	3.61	110.23	27.60	0.02	1.67	216.31	0.16	40.82
145	2.32	36.81	10.31	0.08	1.04	593.45	0.01	23.88
153	2.67	21.82	16.18	0.02	0.58	383.42	0.01	13.34
154	3.61	110.21	27.60	0.02	1.76	216.37	0.06	40.82
157	2.32	36.85	10.31	0.08	1.04	593.45	0.01	23.88
158	3.61	110.28	27.60	0.04	1.76	216.37	0.01	40.82
160	4.49	163.44	4.39	0.03	1.52	480.44	0.08	35.42
161	2.11	65.49	43.96	0.12	4.04	944.74	0.02	92.71
163	2.52	20.92	28.34	0.02	2.07	355.58	0.08	47.49
164	4.49	163.44	18.92	0.17	3.31	480.44	0.01	76.79
165	2.52	20.93	28.34	0.02	2.07	355.58	0.08	47.49
166	2.11	65.45	43.96	0.12	4.04	944.74	0.02	92.71
167	2.11	65.45	43.96	0.12	4.04	944.74	0.09	92.71
172	2.67	21.81	16.18	0.02	0.582	383.42	0.01	13.34
195	2.67	21.81	16.18	0.02	0.582	383.42	0.01	13.34

Table 4.28 depicts and discusses a possible outlier explanation as well as estimation findings. Outliers were found in around 11 percent of the 204 observations. Mahanablios distance, Robust MCD Distance, and Standardized Robust Residual, as well as cook's distance and studentized residual; WSSDI, Hat Diagonal, and Deleted Residual, were used to find outliers. When using Mahanablios distance, outliers have a value of 1.7 to 4.49, and when using Robust MCD, they have a value of 20.92 to 163.47. The range of observations for standardised robust residuals, cook's distance, WSSDI, Hat diagonal, and deleted residual is 3.8 to 43.96, 0.01 to 0.17, 0.47 to 4.04, 216.31 to 944.74, 0.01 to 0.16, and 10.79 to 92.71 respectively.

Table 4.29:Estimates of apple yield w.r.t labour and capital for square root production function through Ordinary Least Square (OLS).

Variable	Regression coefficient (S.E)	Parameters
Intercept	210.91* (34.61)	F-value 161.37(p<0.001) R^2=0.80 AIC=1776.38 BIC=1799.61
x_1	0.10* (0.03)	
x_2	0.04^{NS} (0.09)	
${x_1}^{1/2}$	-10.25^{NS} (4.01)	
${x_2}^{1/2}$	5.52^{NS} (3.54)	
$(x_1 * x_2)^{1/2}$	-0.11^{NS} (0.03)	

*** 5%level of significance NS= non significant**

The estimated results for the square root production function with standard error are shown in Table 4.29. The coefficients of ${x_1}^{1/2}$ (10.25)unit are negative and non significant, whereas of ${x_2}^{1/2}$ (5.52) unit is positive and non significant. The F-value is significant, indicating that the model is adequate for the study variable. Due to the presence of HLP, any estimated coefficient may have the correct sign. OLS was used to get the regression coefficient estimates. The R^2 value is also the lowest (0.80) with AIC (1776.38) and BIC (1799.61), which could be attributed to the existence of HLP.

Fig 4.21 shows panel of diagnostics for the square root production function model using ordinary least squares has been evaluated and found to be inadequate. These tests show that this model is a resounding failure. Poor trends can be seen in the residuals and studentized residuals plots versus predicted values. The points on the plot of the dependent variable (y) versus the anticipated values predict the dependent variable's behaviour. The plot of studentized residual versus leverage reveals that leverage points outside of the range -2 to +2 are HLP. Furthermore, the Cook's graph showed that the majority of observations are outside, and the huge jump in the cut-off plainly suggests that data are influenced by outliers in case of SPF.

Table 4.30: Estimation of square root functions through robust methods for Apple yield w.r.t labour and capital and comparison with OLS thereof.

Variable	Estimates of the regression coefficients					
	OLS (S.E)	M Estimation (S.E)	LTS Estimation (S.E)	MM Estimation (S.E)	S Estimation (S.E)	Modified Estimate after handling HLP (S.E)
Intercept	210.91* (34.61)	46.89* (4.53)	55.66* (5.86)	77.79* (10.41)	43.21* (3.15)	47.51* (4.48)
x_1	0.10* (0.03)	0.03* (0.03)	-0.65 NS (0.02)	-0.55 NS (0.03)	-0.44 NS (0.04)	0.03* (0.03)
x_2	0.04NS (0.09)	0.06* (0.01)	-0.55NS (0.02)	-0.58 NS (0.03)	0.51* (0.02)	0.05* (0.01)
${x_1}^{1/2}$	-10.25 NS (4.01)	3.43* (0.52)	-1.57 NS (0.51)	-2.87 NS (0.90)	1.38* (0.07)	3.49* (0.49)
${x_2}^{1/2}$	5.52 NS (3.54)	-4.68NS (0.46)	-1.69NS (0.40)	-0.58 NS (0.71)	-1.21 NS (0.21)	-4.76 NS (0.43)
$(x_1 * x_2)^{1/2}$	-0.11 NS (0.03)	-0.09 NS (0.04)	1.17* (0.04)	1.09* (0.07)	1.05* (0.02)	-0.09 NS (0.04)
R^2	0.80	0.78	0.99	0.73	0.99	0.98
AIC	1776.38	284.65	1195.22	192.21	231.21	884.13
BIC	1799.61	309.87	1659.01	223.47	280.37	907.36

*** 5%level of significance NS= non significant**

The estimation results for the square root production function are shown in Table 4.30, indicating that each estimation coefficient may be as per the expected sign. The regression coefficient was estimated using OLS, M Estimation, MM, S, LTS, and modified OLS, in accordance with HLP's behaviour. The OLS estimate of $x_1$was found to be positive (0.10) unit and significant, whereas for M estimation it was found to be 0.03unit and significant and for MM, S LTS Estimation were negative and non-significant 0.55, 0.44 and 0.65 unit respectively. After handling HLP, modified estimates of the production function was found to be positive and significant i.e. 0.03 unit. Furthermore, the coefficient of determination for S estimation 0.99 is significantly higher. Generally it has been observed that the estimate of x_1 labour cost is significant as per OLS method whereas found to be non significant in case of LTS, MM and S estimation of robust method. Thus if one unit increase in the cost of labour will decrease overall yield of apple as the total cost of labour will be increased by three percent of the total. The estimate of $x_2$capital cost is non significant as per OLS method whereas found to be significant in case of LTS and MM robust method. If there is one unit increase in the capital cost it will increase the yield of apple and it increases by five percent.

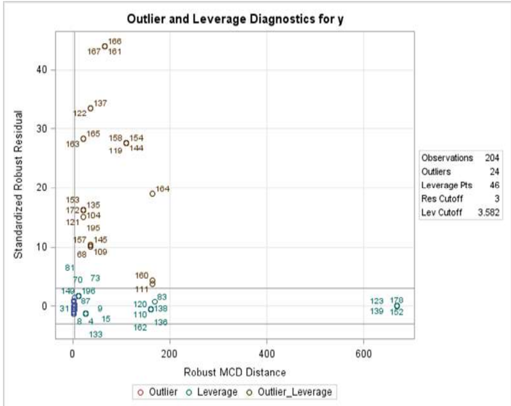


Fig 4.19 : Outlier and leverage diagnostics for dependent variable for Apple data for square root production function.

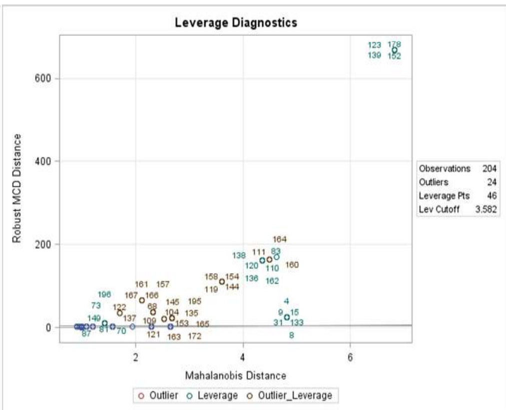


Fig 4.20 : Leverage diagnostics for dependent variable for Apple data for square root production function.

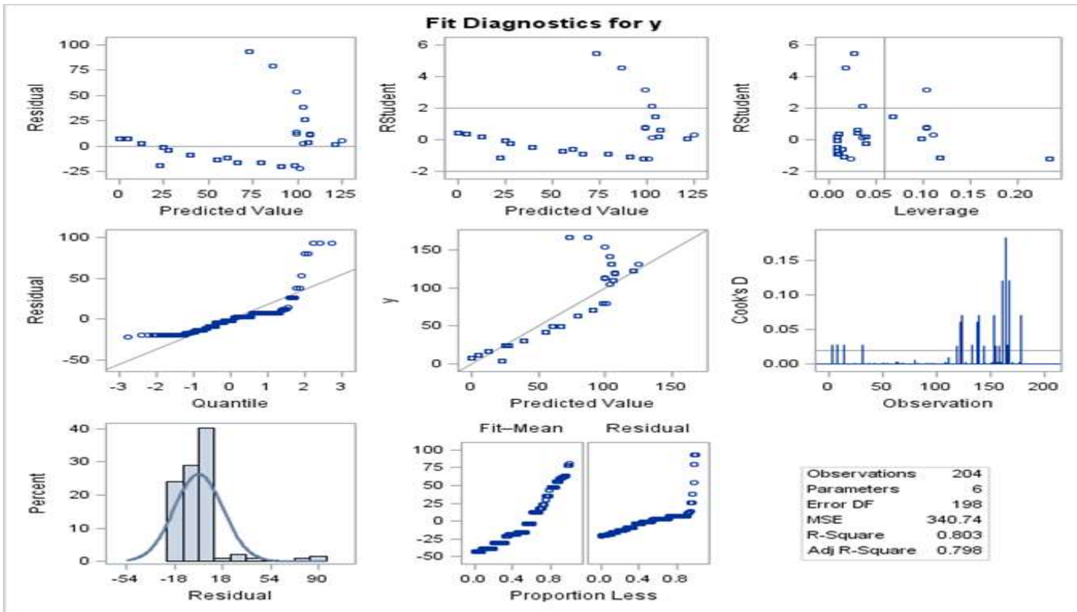


Fig 4.21: Diagnostics Panel of square root production function through OLS.

Fig 4.22 shows the outlier and leverage diagnostics for the dependent variable using real data for the square root production function. As seen in the diagram, there are 4 outliers and 29 leverage points among 97 data points. Outliers and leverage points can be easily identified using a scatter plot of standardised robust residuals vs. robust distances (RD plot). The robust distances vs the standard Mahalanobis distances are shown in Fig 4.23. Outliers, as well as high leverage points can be identified using the plot.

Table 4.31: Detection of outliers in case of maize yield w.r.t labour and capital and values of distances/residuals through various techniques for square root production function.

Observations (HLP)	Mahalanobis	Robust MCD	Standardized Robust	Cook's Distance	Studentized	WSSDI	Hat diagonal	Deleted
2	1.75	3.52	3.30	0.04	2.07	2310.89	0.07	13.76
19	1.52	2.25	4.16	0.04	2.79	908.40	0.03	17.85
75	2.18	3.47	3.77	0.06	2.83	2315.16	0.02	16.10
91	3.73	7.04	5.00	0.22	2.76	9774.77	0.05	21.92

Table 4.31 shows and analyses an outlier explanation, as well as estimation results. Outliers were detected in about 4 percent of the 97 observations. The values of the outliers were estimated using various techniques as already discussed. The outliers have a value of 1.52 to 3.73 units when utilizing Mahanablios distance, and a value of 2.25 to 7.04 units when using Robust MCD. Standardized robust residuals, cook's distance, WSSDI, Hat diagonal, and deleted residual observations vary from 3.3 to 5.00, 0.04 to 0.22, 2.07 to 2.83, 908.4 to 9774.77, 0.02 to 0.07, and 13.76to 21.92 units respectively.

Table 4.32: Estimates of maize yield w.r.t labour and capital for square root production function through OLS.

Variable	Regression coefficient (S.E)	Parameters
Intercept	8.32 NS (4.62)	F-value 107.88(p<0.0001) R^2=0.85 AIC=623.49 BIC=641.44
x_1	0.01 NS (0.06)	
x_2	0.05* (0.02)	
${x_1}^{1/2}$	0.26* (0.09)	
${x_2}^{1/2}$	-0.05 NS (0.02)	
$(x_1 * x_2)^{1/2}$	0.02* (0.01)	

*** 5%level of significance NS= non significant**

As from the table 4.32 it is observed that the estimated results for the square root production function with standard error${x_1}^{1/2}$ is 0.26 unit which found to be positive and significant coefficient but ${x_2}^{1/2}$is negative and non significant 0.05 unit. The model is suitable for the study variable because the F-value is substantial. Any predicted coefficient may have the proper sign due to the presence of HLP. The regression coefficient estimates were calculated using OLS. The existence of HLP could be attributed to the R^2 value of (0.85) with AIC (623.49) and SBIC (641.44) unit.

In Fig 4.24 the diagnostic panel for the square root production function model based on ordinary least squares was tested and found to be insufficient. These experiments demonstrate that this model is a complete failure. The residuals and studentized residuals graphs vs projected values showed poor trends. The behaviour of the dependent variable (y) is predicted by the points on the plot of the dependent variable (y) versus the expected values. The plot of studentized residual vs. leverage showed that leverage points outside of the -2 to +2 range are HLP. Furthermore, the Cook's graph demonstrates that the majority of observations are lying outside and the large leap in the cut-off clearly indicates that data are influencing the decision.

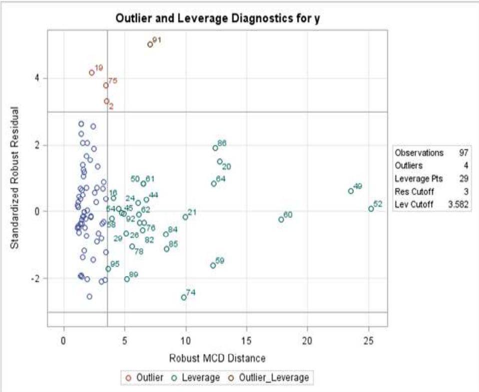


Fig 4.22 : Outlier and Leverage diagnostics for dependent variable for maize data for square root production function.

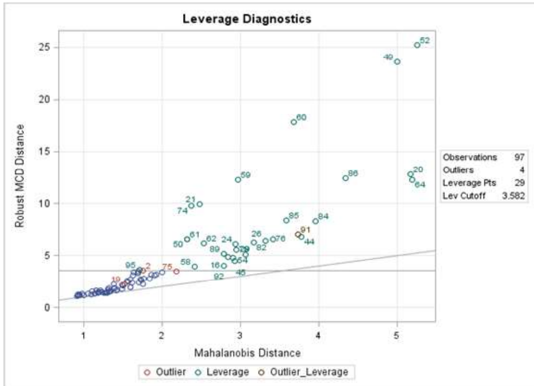


Fig 4.23 : Leverage diagnostics for dependent variable for maize data for Square root production function

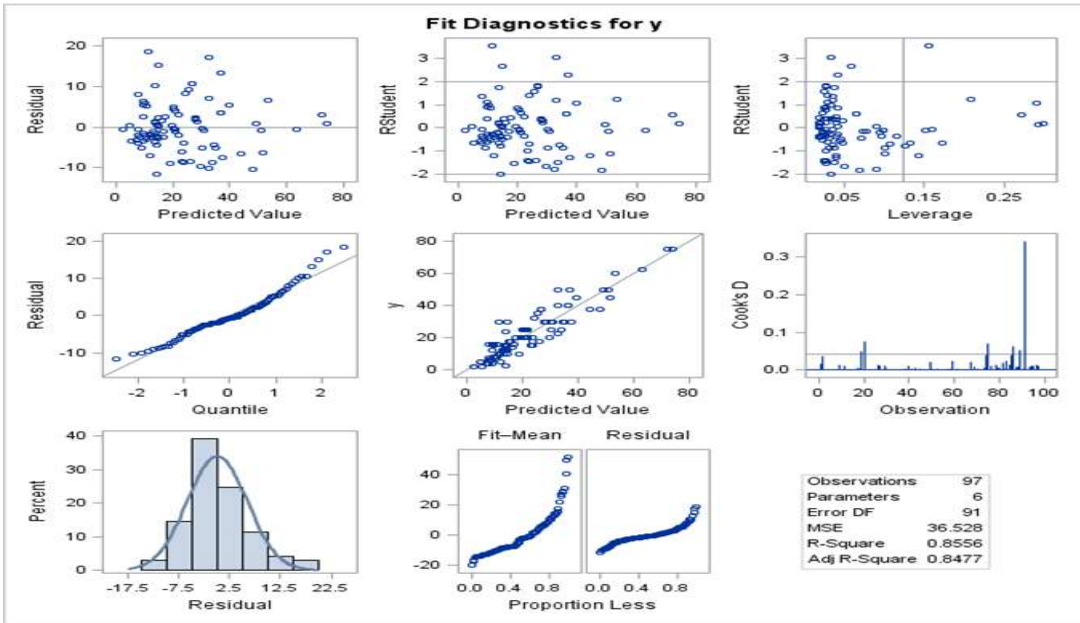


Fig 4.24: Diagnostics Panel of square production function through OLS.

Table 4.33: Estimation of square root production functions through robust methods for maize yield w.r.t labour and capital and comparison with OLS thereof.

Variable	Estimates of the regression coefficients					
	OLS (S.E)	**M Estimation (S.E)**	**LTS Estimation (S.E)**	**MM Estimation (S.E)**	**S Estimation (S.E)**	**Modified Estimate after handling HLP (S.E)**
Intercept	8.32 NS (4.62)	7.93* (3.73)	8.01* (3.96)	7.50NS (4.37)	7.53NS (4.32)	8.34 NS (3.91)
x_1	0.01 NS (0.06)	0.010* (0.05)	0.01 NS (0.06)	0.08* (0.03)	0.01 NS (0.06)	0.01* (0.01)
x_2	0.05* (0.02)	4.67* (2.03)	3.49 NS (2.50)	0.002* (0.001)	0.02* (0.01)	0.03* (0.01)
${x_1}^{1/2}$	0.26* (0.01)	0.20* (0.07)	0.18* (0.07)	0.20* (0.08)	0.19* (0.084)	0.19 NS (0.07)
${x_2}^{1/2}$	-0.05 NS (0.21)	-0.03 NS (0.02)	-0.05NS (0.01)	-0.04 NS (0.020)	-0.03 NS (0.02)	-0.05NS (0.01)
$(x_1 * x_2)^{1/2}$	0.09* (0.01)	0.04* (0.01)	0.03* (0.01)	0.05* (0.011)	0.03* (0.01)	0.02* (0.01)
R^2	0.85	0.64	0.86	0.61	0.83	0.86
AIC	623.49	145.12	-257.66	176.89	585.34	119.21
BIC	641.44	160.34	-242.2	197.84	603.29	123.21

*** 5%level of significance NS= non significant**

The table 4.33 shows the results of the estimation for the square root production function, suggesting that each estimation coefficient may be correct in terms of predicted sign. According to HLP's behaviour, the regression coefficient was estimated using OLS, M Estimation, MM, S, LTS, and modified OLS. The OLS estimate of ${x_1}^{1/2}$was found to be positive and significant having value 0.26 unit while as the M, MM, S, LTS estimates were generally positive and significant with coefficients 0.20, 0.03, 0.18, 0.20, 0.19 unit respectively. The estimate of the production function were also found to be positive and non-significant after handling HLP i.e. 0.19.Furthermore modified OLS has a much higher coefficient of determination i.e. 0.86 indicating that this method can explain 86 percent of the variation in the study variable. The OLS estimates of ${x_2}^{1/2}$ was negative 0.05 unit and non-significant, while as the M, MM, S, LTS estimate were also negative and non-significant 0.03, 0.05, 0.04, 0.03 unit respectively. The estimates of the${x_2}^{1/2}$ production function was found to be negative and non-significant after handling HLP 0.05. Thus the method OLS after handling HLP have same behaviour and more precise in many cases of Linear production function and can be considered as robust method. The estimate of x_1 labour cost is non significant as per OLS method whereas found to be non significant in case of LTS and S estimation. Thus if one unit increase in the cost of labour it will decrease overall yield of maize. The estimate of $x_2$capital cost is non significant as per OLS method whereas found to be significant in case of M, MM and S estimation. Modified OLS is also significant. If one unit increase in the capital cost it will increase the yield of maize and it increases by three percent.

4.2 Estimation of Non-linear production function of Cobb- Douglas and Mitscherlich-Baule.

The Cobb Douglas (CD) function is used widely in empirical studies of manufacturing industries. It is used to determine the relative shares of labour & capital in output. Its parameters represent elasticity coefficients that are used for inter sectoral comparisons. The Mitscherlich-Baule function is characterized by continuously positive marginal productivities of the input factors. It does not exhibit negative marginal productivities the function has been found to be preferable for use in an economic model because it allows for factor substitution following von Liebig's law of minimum.

4.2.1 Estimation of Cobb- Douglas production function through Robust method(s).

Table 4.34: Summary Statistics of simulation data (P_1)w.r.t. endogenous and exogenous variables for Cobb Douglas production function.

Variable	Q_1	Median	Q_3	Mean	Standard Deviation	MAD	C.V (%)
x_1	-0.48	-0.15	0.07	0.25	0.43	0.39	173.26
x_2	-0.47	-0.13	0.15	0.21	0.55	0.47	257.13
y	0.55	0.84	0.07	0.50	0.74	0.37	147.88

Based on simulated data, the mean and standard deviation of the study variable were 0.50 and 0.74 unit respectively in Table 4.34.There is greater variation in the data since the mean is smaller than the standard deviation which is due to the presence of influencing observations. The mean and standard deviation of auxiliary variables x_1 and$x_2$were found to be 0.25 and 0.43and 0.21 and 0.55 unit respectively indicating that the data set has more variability. The variable y had the least mean absolute deviation 0.37 unit. Both x_1 andx_2 had negative Q_1 of 0.48 and 0.47unit respectively, and y had a positive $Q_1$0.55 unit.The low standard deviation indicates that the data points tend to be very close to the mean and high standard deviation indicates that the data points are spread out over a large range of value. Coefficient of variation of the Y is found to be low 147.88 unit whereas for x_1 and x_2 coefficient of variation is very high 173.26 and 257.13 unit respectively indicating inconsistency in the data set which may be due to HLP.

The Fig 4.25 shows that the outlier and leverage diagnostics for the dependent variable for simulated data for the CD function. As observed there are 21 outliers and 51 leverage points among the 500 observations. Outliers and leverage points can be easily identified using a scatter plot of standardized robust residuals vs. robust distances (RD plot). The robust distances are plotted against the traditional Mahalanobis distances see Fig 4.26. Outliers as well as high leverage points can be identified using this plot.

Table 4.35: Detection of outliers of simulation data (P_1) and values of distance/residual through various techniques for Cobb Douglas production function.

Observations (HLP)	Mahalanobis	Robust MCD	Standardized Robust	Cook's	Studentized	WSSDI	Hat diagonal	Deleted
2	1.7	1.91	6.33	0.06	5.06	0.05	0.07	2.57
6	0.39	0.32	-3.06	0.04	-2.2	0.03	0.01	-1.12
26	0.57	0.56	-4.03	0.08	-2.95	0.19	0.09	-1.5
83	1.47	1.61	-4.58	0.02	-3.36	0.01	0.02	-1.7
86	0.41	0.36	-3.62	0.05	-2.64	0.03	0.09	-1.34
104	0.74	0.82	-4.3	0.01	-3.17	0.01	0.01	-1.61
150	0.88	0.87	-6.04	0.02	-4.49	0.02	0.08	-2.28
168	2.44	3.17	-3.42	0.03	-2.54	0.02	0.05	-1.29
211	0.5	0.49	-3.14	0.04	-2.26	0.05	0.01	-1.15
224	1.02	1.13	-4.05	0.01	-2.95	0.01	0.04	-1.49
236	0.83	1.31	-3.36	0.07	-2.46	0.09	0.04	-1.24
246	0.56	0.61	-5.89	0.07	-4.39	0.02	0.06	-2.23
250	1.49	2.08	-3.64	0.06	-2.67	0.06	0.05	-1.36
258	1.74	1.97	-3.07	0.03	-2.19	0.07	0.08	-1.11
272	1.19	1.34	-3.9	0.03	-2.84	0.02	0.08	-1.44
345	1.11	1.31	-4.34	0.06	-3.21	0.03	0.04	-1.63
363	0.7	0.82	-6.05	0.02	-4.5	0.01	0.07	-2.28
380	1.07	1.14	-3.28	0.08	-2.36	0.17	0.01	-1.2
412	0.6	0.65	-3.29	0.05	-2.38	0.08	0.01	-1.21
414	1.63	1.94	-8.43	0.1	-6.33	0.08	0.02	-3.22

474	1.35	1.59	-3.22	0.01	-2.32	0.01	0.05	-1.17

An outlier explanation and estimation results are shown and discussed in Table 4.35. 4 percent of the 500 observations found to be outliers. Further to examine HLP the techniques viz .mahalanobis distance, Robust MCD Distance, Standardized Robust Residual, cook's distance, studentized residual, WSSDI, Hat Diagonal, and Deleted Residual were used. The value of outliers coming very high using Mahanablios distance, the range is from 0.39 to 2.44, whereas in case of Robust MCD ranges from 0.32 to 3.17. For standardized robust residuals, cooks distance, WSSDI, Hat diagonal, and deleted the range of observations is from -8.43 to 6.33, 0.01 to 0.1, and 6.33 to 5.06, 0.01 to 0.2, 0.01 to 0.09, and 3.22 to 2.57 units respectively.

Table 4.36: Estimates of Cobb Douglas production function (P_1) through Ordinary Least Square (OLS).

Variable	Regression coefficient (S.E)	Parameters
Intercept	0.76* (0.07)	F-value 1.28(p<0.001) R^2=0.05 AIC=744.18 BIC=761.03
x_1	0.06 ^{NS} (0.05)	
x_2	0.05 ^{NS} (0.04)	

*** 5%level of significance   NS= non significant**

The estimated results for the CD function with standard error are shown in Table 4.36. in case of the coefficients x_1 (0.06) having value positive and non-significant whereas x_2 with size (0.05) also has a positive and non-significant. The F-value is significant, therfore the model is adequate for the study. Due to the presence of HLP coefficient may have the correct sign. The R^2 value is very low 0.05 with AIC (744.18) and BIC (761.03) respectively could be related to the occurance of HLP.

In Fig 4.27 the diagnostic panel for the CD function model utilizing ordinary least squares was tested and found to be insufficient. These tests demonstrate that this model is a huge failure. Poor trends can be seen in the residuals and studentized residuals plots versus predicted values. The points on the plot of the dependent variable (y) versus the anticipated values predict the dependent variable's behaviour. The plot of studentized residual versus leverage reveals that leverage points outside of the range -2 to +2 are HLP. Furthermore, the Cook's graph showed that the majority

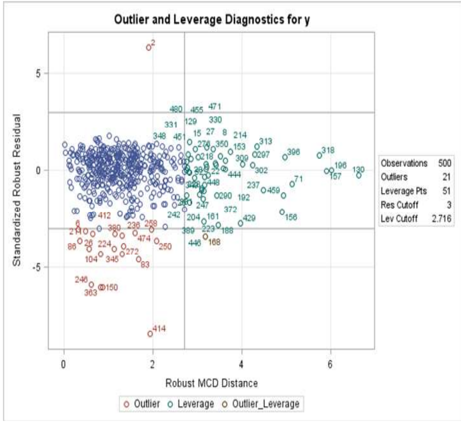


Fig 4.25 : Outlier and Leverage diagnostics for dependent variable for Simulation data (P_1) for Cobb Douglas production function.

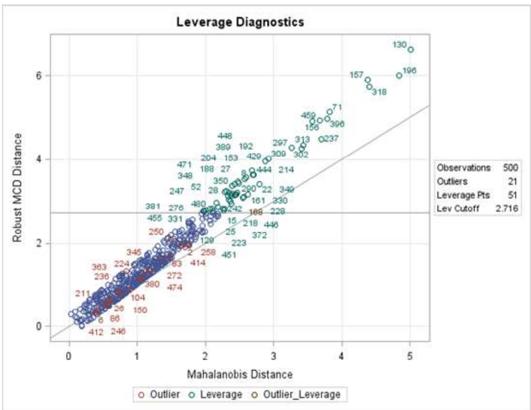


Fig 4.26 : Leverage diagnostics for dependent variable for simulated data (P_1) for Cobb Douglas production function

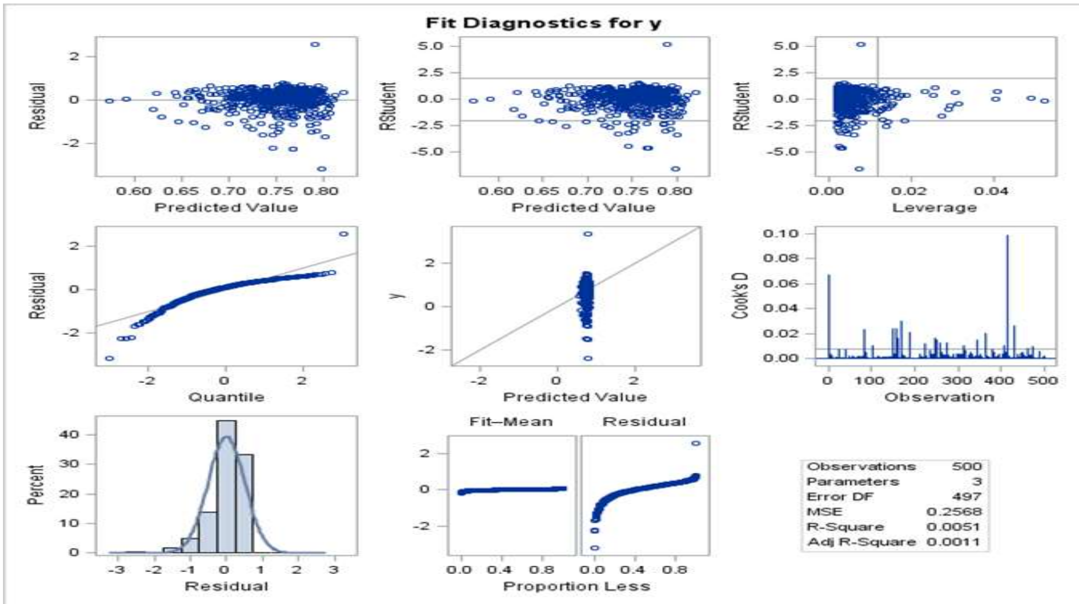


Fig 4.27: Diagnostics Panel of Cobb Douglas function through OLS.

of observations are outside and the huge jump in the cut-off plainly suggests that data are influenced by outliers in case of CD function.

Table 4.37: Estimation of Cobb Douglas functions through robust methods for simulation data (P_1) and comparison with OLS thereof.

	Estimates of the regression coefficients					
Variable	**OLS (S.E)**	**S Estimation (S.E)**	**MM Estimation (S.E)**	**M Estimation (S.E)**	**LTS Estimation (S.E)**	**Modified Estimates after handling HLP (S.E)**
Intercept	0.76* (0.02)	0.892* (0.02)	0.873* (0.02)	0.845* (0.021)	0.836* (0.01)	0.83* (0.02)
x_1	0.06 NS (0.05)	0.09 NS (0.03)	0.01 NS (0.03)	0.09 NS (0.04)	0.01 NS (0.03)	0.02 NS (0.03)
x_2	0.06 NS (0.04)	0.08* (0.03)	0.09* (0.03)	0.07* (0.03)	0.09* (0.02)	0.10* (0.02)
R^2	0.05	0.14	0.09	0.01	0.01	0.02
AIC	744.18	272.24	397.20	518.46	-2627.01	410.38
BIC	761.03	287.32	412.50	532.72	-2821.01	427.24

*** 5%level of significance NS= non significant**

Table 4.37 shows the estimation results for the Cobb Douglas production function, showing that each estimation coefficient may have correct in terms of expected sign. The regression coefficient was estimated using OLS, M Estimation, MM, S, LTS, and modified OLS after dealing with HLP. The OLS estimate of x_1 was positive 0.06 and non-significant, whereas the M, MM, S Estimation, and LTS was positive and non-significant 0.09, 0.01, 0.09 and 0.01 unit respectively. Modified estimate of the production function was also found to be positive and non-significant after handling HLP i.e. 0.02. For S estimation, AIC and BIC values are 272.24 and 287.32 respectively which are lowest among all the robust methods. If exceptional observations are excluded from the study, the remaining observations will create a linear pattern. It was also found that the OLS estimate of x_2 was positive 0.06 and non significant whereas M, MM, S and LTS estimate were positive and significant 0.08, 0.09, 0.07 and 0.09 unit respectively. After handling HLP, modified estimate of the x_2 was determined to be positive and significant 0.10 unit.

Table 4.38 : Summary Statistics of simulation data (P_2) w.r.t. endogenous and exogenous variables for Cobb Douglas production function.

Variable	**Q_1**	**Median**	**Q_3**	**Mean**	**Standard Deviation**	**MAD**	**C.V (%)**
x_1	-0.48	-0.17	0.06	0.27	0.46	0.39	171.43

x_2	-0.50	-0.14	0.15	0.22	0.53	0.48	241.89
Y	0.52	0.84	1.07	0.48	0.74	0.38	154.45

As per table 4.38,the study variable has mean and standard deviation as 0.48 and 0.74unit respectively. The mean is smaller than the standard deviation the data is asymmetrical which is due to the presence of influencing observations. The mean and standard deviation of auxiliary variables $x_1$were 0.27 and 0.46and forx_2 0.22 and 0.53 unit respectively showing that the data are influenced i.e. having more variability. The variable y had the smallest MAD as compared to others it is (0.38). The $Q_1$of x_1 and $x_2$are negative, 0.48 and 0.50 unit respectively, whereas the $Q_1$of y was positive which is 0.52 unit. Low standard deviation indicates that the data points tend to be very close to the mean and high standard deviation indicates that the data points are spread out over a large range of value. Coefficient of variation of the Y is found to be 154.45whereas for x_1 and $x_2$are very high 171.43 and 241.89 unit respectively indicating inconsistency in the data set which may be due to HLP.

The Fig 4.28 and 4.29 depicts that the outlier and leverage diagnostics for the CD function. Among the 1000 observations, there are 30 outliers and 114 leverage points, as shown in the diagram. The graph can be used to identify outliers as well as high leverage points.

Table 4.39: Detection of outliers of simulation data (P_2) and values of distance/residual through various techniques for Cobb Douglas production function.

Observations (HLP)	Mahalanobis	Robust MCD	Standardized Robust	Cook's	Studentized	WSSDI	Hat Diagonal	Deleted
2	1.70	1.97	6.14	0.04	5.27	0.22	0.09	2.54
26	0.58	0.56	-3.92	0.01	-3.14	0.03	0.05	-1.51
83	1.54	1.72	-4.41	0.01	-3.53	0.16	0.01	-1.70
86	0.43	0.35	-3.52	0.01	-2.80	0.02	0.05	-1.35
104	0.75	0.82	-4.21	0.01	-3.39	0.05	0.05	-1.63
150	0.89	0.91	-5.87	0.01	-4.77	0.06	0.04	-2.30
168	2.52	3.17	-3.44	0.02	-2.79	0.46	0.09	-1.35
211	0.53	0.54	-3.01	0.01	-2.37	0.02	0.06	-1.14
224	1.08	1.19	-3.87	0.01	-3.08	0.08	0.08	-1.48
236	0.86	1.29	-3.26	0.01	-2.60	0.05	0.07	-1.25
246	0.57	0.6	-5.74	0.01	-4.66	0.03	0.03	-2.24
250	1.39	2.15	-3.44	0.01	-2.74	0.15	0.02	-1.32
258	1.75	2.04	-3.00	0.01	-2.37	0.23	0.09	-1.14

272	1.26	1.39	-3.73	0.01	-2.97	0.11	0. 01	-1.43
345	1.11	1.32	-4.27	0.01	-3.45	0.10	0.07	-1.66
363	0.69	0.89	-5.81	0.01	-4.70	0.04	0.04	-2.26
380	1.12	1.18	-3.16	0.01	-2.50	0.09	0.01	-1.20
412	0.61	0.64	-3.21	0.01	-2.55	0.03	0.06	-1.23
414	1.66	2.02	-8.07	0.05	-6.58	0.20	0.06	-3.17
474	1.35	1.68	-3.01	0.01	-2.36	0.14	0.013	-1.14
525	0.64	0.58	-3.16	0.01	-2.51	0.03	0.06	-1.21
538	1.19	1.29	-3.00	0.01	-2.36	0.10	0.01	-1.14
565	0.99	1.02	-3.27	0.01	-2.59	0.07	0.08	-1.24
625	1.56	2.09	-4.65	0.02	-3.72	0.19	0.01	-1.79
636	0.95	1.04	-3.72	0.01	-2.95	0.06	0.07	-1.42
734	1.06	1.11	-3.29	0.06	-3.66	0.08	0.09	-1.77
743	3.43	4.54	-4.47	0.01	-3.06	0.81	0.04	-1.48
856	1.62	1.88	-3.86	0.01	-2.67	0.18	0.03	-1.29
878	1.15	1.52	-3.34	0.01	-3.60	0.09	0.09	-1.73
932	1.09	1.23	-4.46	0.01	-3.26	0.10	0.07	-1.22

Table 4.39 shows an outlier explanation as well as estimation results. Outliers were found in approximately three percent of the 1000 observations. To identify outliers as good/bad the methods used are Mahanablios distance, Robust MCD Distance, Standardized Robust Residual, cook's distance, studentized residual, WSSDI, Hat Diagonal and Deleted Residual. The values range from 0.43 to 3.43 unit when using Mahanablios distance and from 0.35 to 4.54 unit while using Robust MCD and are very high. The range of observations for standardized robust residuals, cooks distance, WSSDI, Hat diagonal, and deleted is -8.07 to 6.14, 0.01 to 0.06, and 6.58 to 5.27, 0.02 to 0.81, 0.01 to 0.06, and 3.17 to 2.54 units respectively indicating the leverages are bad leverage point.

Table 4.40 : Estimates of Cobb Douglas production function (P_2) through Ordinary Least Square (OLS).

Variable	Regression coefficient (S.E)	Parameters
Intercept	0.77* (0.01)	F-value 4.11(p<0.005) R^2=0.08 AIC=1378.78 BIC=1398.41
x_1	0.07* (0.03)	
x_2	0.04^{NS} (0.02)	

*** 5%level of significance   NS= non significant**

The estimated results for the CD function with standard error have been shown in table 4.40. The F-value is significant indicating that the model is acceptable w.r.t.x_1 and x_2.Further the coefficients of $x_1$and x_2 having value 0.07 and 0.04 unit are positive and significant but x_2(0.04) unit has a positive and non-significant coefficient. Any anticipated coefficient may have the correct sign due to the presence of HLP. The regression coefficient estimations were obtained using OLS. The R^2 value is 0.08 with AIC (1378.78) and BIC(1398.41) unit could be related to the occurrence of HLP.

Fig 4.30 shows the diagnostic panel for the CD function model was evaluated and found to be insufficient. This shows that the model is a complete failure. The plots of residuals and studentized residuals vs. projected values reveal dismal trends. The points on the plot of the dependent variable (y) versus the predicted values forecast the dependent variable's behaviour. The studentized residual vs leverage plot reveals that leverage locations outside of the -2 to +2 range are HLP. Furthermore, the Cook's graph showed that the majority of observations were outside and the big spike in the cut-off clearly indicating the data are influenced.

Table 4.41: Estimation of Cobb Douglas function through robust methods for simulation data (P_2) and comparison with OLS thereof.

	Estimates of the regression coefficients					
Variable	**OLS (S.E)**	**M Estimation (S.E)**	**LTS Estimation (S.E)**	**MM Estimation (S.E)**	**S Estimation (S.E)**	**Modified Estimate after handling HLP (S.E)**
Intercept	0.776* (0.01)	0.84* (0.01)	0.83* (0.01)	0.86* (0.01)	0.88* (0.01)	0.82* (0.01)
x_1	0.07* (0.03)	0.06* (0.02)	0.06* (0.02)	0.05* (0.02)	0.03^{NS} (0.02)	0.09* (0.02)
x_2	0.04^{NS} (0.02)	0.06* (0.02)	0.05* (0.02)	0.06* (0.02)	0.06* (0.02)	0.06* (0.02)
R^2	0.08	0.08	0.02	0.09	0.06	0.09
AIC	1378.78	1010.37	-2653.84	788.10	821.21	965.52
BIC	1398.41	1026.46	-2638.38	885.33	875.28	985.15

*** 5%level of significance NS= non significant**

Table 4.41 shows the estimation results for the Cobb Douglas production function. The regression coefficient were estimated using OLS, M Estimation, MM, S, LTS, and modified OLS after dealing with HLP. The OLS estimate of x_1 was positive 0.07 and significant, whereas the M, MM, Estimation, LTS were positive and significant 0.06, 0.06, 0.05 unit and S estimation was positive 0.03 and non

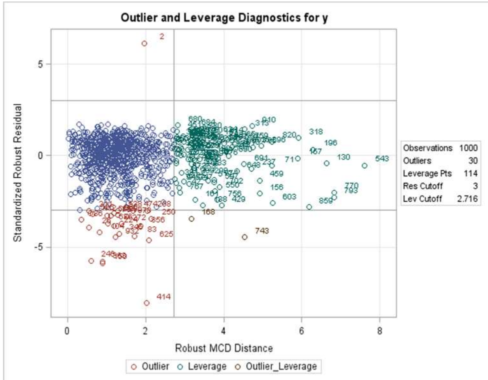


Fig 4.28 : Outlier and Leverage diagnostics for dependent variable for Simulation data (P_2) for Cobb Douglas production function.

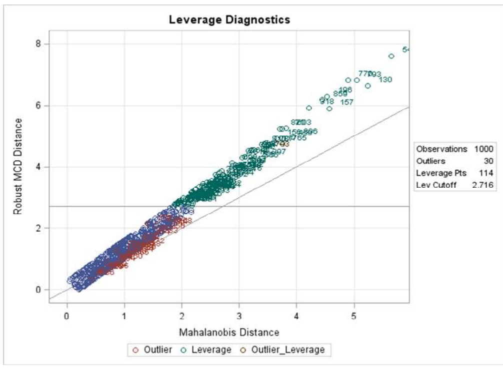


Fig 4.29 : Leverage diagnostics for dependent variable for Simulation data (P_2) for Cobb Douglas production function.

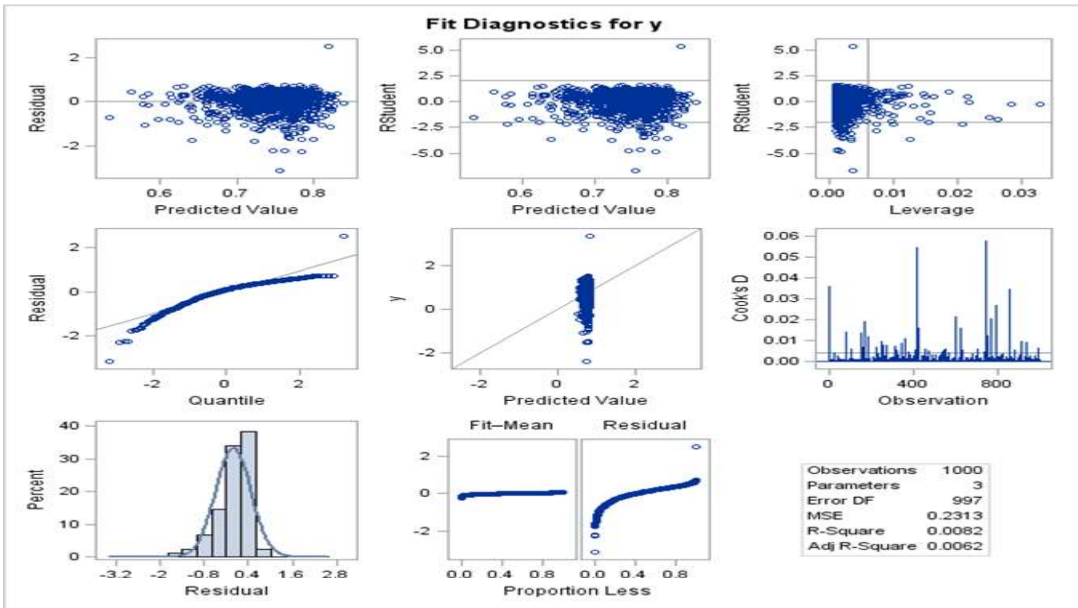


Fig 4.30: Diagnostics Panel of Cobb Douglas production function through OLS.

significant. Modified estimates of the production function was also found to be positive (0.09) and non-significant after handling HLP. If exceptional observations are excluded from the study, the remaining observations create a linear pattern that explains more of the variation in the simulated data. It was also found that the OLS estimate of x_2 was positive 0.04 and non significant whereas in case of M, MM, S and LTS estimates were positive and significant 0.06, 0.05, 0.06 and 0.06 unit. After handling HLP, modified estimate of the production function was found to be positive and significant 0.06 unit thus in case of HLP the estimates of CD function affected in case of small as well as big data. Therefore, it is matter of fact to propose the best estimates of the real data in case of CD method in the presence of HLP. The first data in series is the survey done at Kashmir.

Fig 4.31and Fig 4.32 depicts outlier and leverage diagnostics for the dependent variable for real data for the CD function. There are 14 outliers and 63 leverage points among the 204 observations. Their estimated values of HLP related to techniques used are mentioned in table 4.42

Table 4.42 : Detection of outliers in case of Apple yield data and values of distance/residual through various techniques for Cobb Douglas production function.

Observations (HLP)	Mahalanobis	Robust MCD	Standardized Robust	Cook's	Studentized	WSSDI	Hat Diagonal	Deleted
4	3.13	3.01	3.02	0.10	2.37	3.59	0.02	0.31
8	3.13	3.01	3.02	0.10	2.37	3.59	0.02	0.31
9	3.13	3.01	3.02	0.10	2.37	3.59	0.02	0.31
15	3.13	3.01	3.02	0.10	2.37	3.59	0.02	0.31
31	3.13	3.01	3.02	0.10	2.37	3.59	0.02	0.31
111	2.44	11.91	3.16	0.05	2.19	0.01	0.01	0.28
122	1.43	5.50	4.38	0.05	3.28	0.18	0.05	0.42
133	3.13	3.01	3.02	0.10	2.37	3.59	0.02	0.31
137	1.43	5.53	4.38	0.05	3.28	0.18	0.05	0.42
160	2.44	11.90	3.22	0.05	2.24	0.01	0.01	0.28
161	1.72	8.42	4.02	0.06	3.22	0.41	0.06	0.41
164	2.44	11.90	4.47	0.12	3.27	0.41	0.01	0.42
166	1.72	8.42	4.02	0.06	3.22	0.41	0.06	0.41
167	1.72	8.42	4.02	0.06	3.22	0.01	0.06	0.41

The first column represents the observations which are outliers where as the rest columns are the values of HLP w.r.t different techniques as discussed .The range of HLP through Mahanablios distance is 1.43 to 3.13unitwhereas for Robust MCD it was from 3.01 to 11.91 unit. The other ranges for standardized robust residuals, cooks distance, WSSDI, Hat diagonal, and deleted is 3.02 to 4.47, 0.05 to 0.12 and 2.19 to 3.28, 0.01 to 3.59, 0.01 to 0.06, and 0.28 to 0.42 unit respectively.

Table 4.43: Estimates of Apple yield w.r.t labour and capital for Cobb Douglas production function through Ordinary Least Square (OLS).

Variable	Regression coefficient (S.E)	Parameters
Intercept	-6.90 NS (0.19)	F-value 920.42(p<0.001) R^2=0.90 AIC= -257.54 BIC= -244.27
x_1	2.10* (0.23)	
x_2	-0.07 NS (0.21)	

*** 5%level of significance   NS= non significant**

Table 4.43 displays the Cobb Douglas production function's estimated results with standard error. The F-value is significant, indicating that the model for the variable in this study is adequate. The coefficient of $x_1$is 2.10 unitfound to be positive and significant, however the coefficient of x_2 is0.07 unit is negative and non-significant. Due to the presence of HLP, any expected coefficient may have the correct sign. OLS was used to estimate the regression coefficients. The R^2 value (0.90) with AIC (-257.54) and BIC (-244.27) unit could be related to the occurrence of HLP.

The panel of diagnostics for the CD function model using ordinary least squares has been evaluated and found to be inadequate like earlier due to HLP in Fig 4.33. And found that this model is a resounding failure. Poor trends can be seen in the residuals and studentized residuals plots versus predicted values. Further, the estimates of regression coefficients have been estimated through the OLS as well as various robust techniques .As per the comparisons purposes see table 4.44

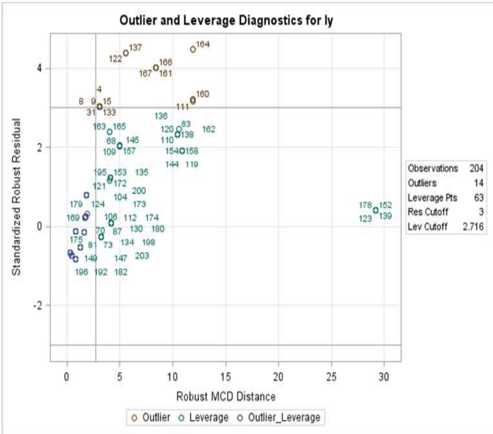


Fig 4.31: Outlier and Leverage diagnostics for dependent variable for Apple data for Cobb Douglas production function.

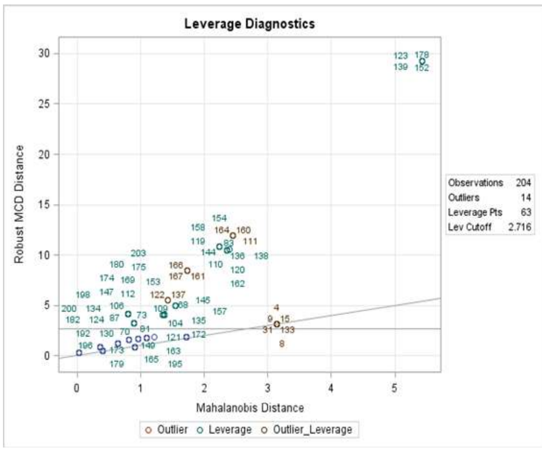


Fig 4.32 : Leverage diagnostics for dependent variable for Apple data for Cobb Douglas production function.

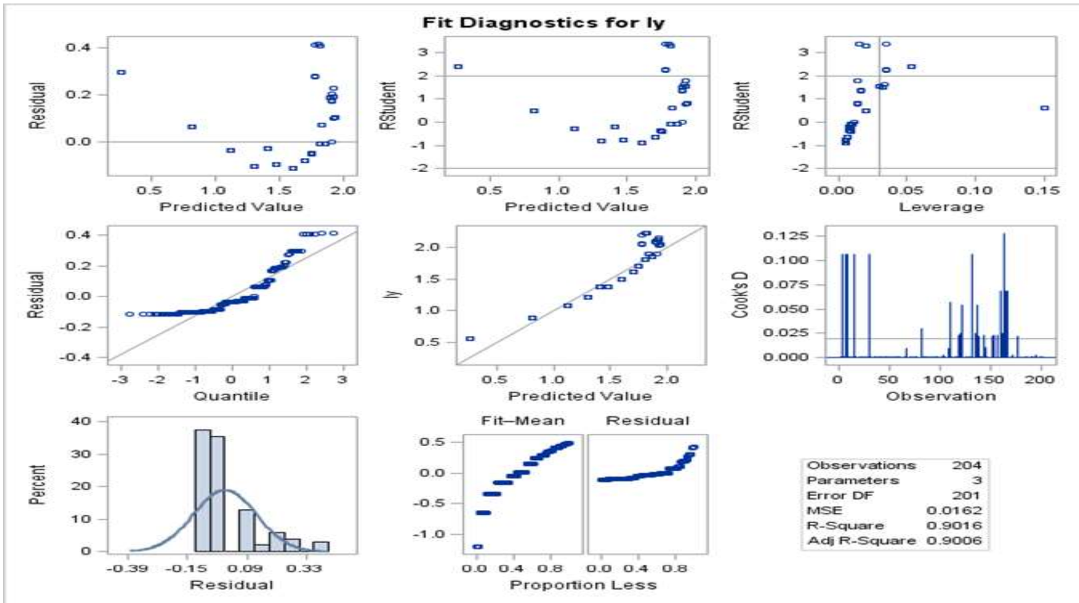


Fig 4.33: Diagnostics Panel of Cobb Douglas production function through OLS.

Table 4.44: Estimation of Cobb Douglas function through robust methods for apple yield data and comparison with OLS thereof.

	Estimates of the regression coefficients					
Variable	**OLS (S.E)**	**M Estimation (S.E)**	**LTS Estimation (S.E)**	**MM Estimation (S.E)**	**S Estimation (S.E)**	**Modified OLS after handling HLP (S.E)**
Intercept	-6.90 NS (0.17)	-6.93NS (0.19)	-6.832NS (0.10)	-6.03 NS (0.167)	-6.71 NS (0.128)	-7.55* (0.15)
x_1	2.10* (0.23)	2.35* (0.16)	-0.54 NS (0.13)	-0.66NS (0.26)	2.60* (0.12)	2.36* (0.15)
x_2	-0.07 NS (0.21)	-0.32 NS (0.15)	2.52* (0.12)	2.46* (0.23)	-0.63 NS (0.11)	-0.18NS (0.14)
R^2	0.90	0.77	0.98	0.70	0.95	0.94
AIC	-257.54	212.22	-521.05	485.02	171.45	-411.31
BIC	-244.27	224.60	-513.02	460.21	185.26	-398.04

*** 5%level of significance NS= non significant**

The OLS estimate of x_1 was found positive2.10 and significant whereas the coefficients of M and S Estimation were positive 2.35and 2.60 and significant. The MM and LTS estimates are negative 0.66 and 0.54 unit and non-significant. Modified estimate of the production function was also found to be positive 2.36 unit and significant after handling HLP. Moreover the coefficient of determination was remarkably higher in case of LTS estimation 0.98 which indicates that 98 percent variation of study variable is explained through this method but this method is not recommended Chen, C (2002). Further it explains if one unit decrease in the cost of labour will increase overall yield of apple. It will be increased by two percent of the total. The estimate of coefficient of x_2 capital cost is also non significant as per OLS method whereas found to be significant in case of robust method. Thus one unit decrease in the capital cost will decrease the yield of apple and it decreases by one percent. Further the data pertains to maize is used to see the impact of HLP in the estimation of parameters in case of CD function. The Fig 4.34 and 4.35 clearly depicted that the data are influenced by outliers. The figure showed that there are 3 outliers and 7 leverage points out of the 97 observations.

Table 4.45:Detection of outliers in case of maize yield data and values of distance/residual through various techniques for Cobb Douglas production function.

Observations	Mahalanobis	Robust MCD	Standardized Robust	Cook's	Studentized	WSSDI	Hat Diagonal	Deleted
67	1.13	1.18	-5.81	0.16	-4.48	0.01	0.05	-4.48
75	1.62	2.08	3.36	0.08	2.49	0.04	0.01	2.49
91	1.81	1.96	4.12	0.16	3.26	0.08	0.01	3.26

Further the Table 4.45 shows estimation results of outliers w.r.t different techniques used. Outliers were found in approximately three percent of the 97 observations. The outlier values range from 1.13 to 1.81 unit when using Mahanablios distance and from 1.18 to 2.08 unit while using Robust MCD. The range of observations for standardized robust residuals, cooks distance, WSSDI, Hat diagonal, and deleted is -5.81 to 4.12, 0.08 to 0.16, and -4.48 to 3.26, 0.01 to 0.08, and 0.01 to 0.05unit respectively.

Table 4.46:Estimates of maize yield w.r.t labour and capital for Cobb Douglas production function through Ordinary Least Square (OLS).

Variable	Regression coefficient (S.E)	Parameters
Intercept	-1.15^{NS} (0.21)	F-value 156.92(p<0.0001) R^2=0.85 AIC=632.28 BIC=650.30
x_1	0.74* (0.02)	
x_2	-0.04^{NS} (0.03)	

*** 5 % level of significance NS= non significant**

The CD function results predicted with standard error are shown in table 4.46. The F-value in this study is also significant indicating that the model for the variable is appropriate. The coefficient of x_1 was 0.74and is positive and significant whereas in case of $x_2$it was 0.04 and found to be negative and non-significant. Any expected coefficient may have the correct sign due to the presence of HLP. The regression coefficients were estimated using OLS. The R^2 value (0.85) in combination with the AIC (632.28) and BIC (650.30) could indicate the presence of HLP.Further, the comparision of different robust techniques w.r.t OLS has shown in table 4.57 to see which is the best method to tackle the problem of HLP in CD function.

Fig 4.36 shows the panel of diagnostics for the CD function model using ordinary least squares has been evaluated and found to be inadequate due to HLP. And found that this model is a resounding failure. Poor trends can be seen in the residuals and studentized residuals plots versus predicted values. Further, the estimates of regression coefficients have been estimated through the OLS as well as various robust techniques.

Table 4.47: Estimation of Cobb Douglas functions through robust methods for maize yield w.r.t labour and capital and comparison with OLS thereof.

	Estimates of the regression coefficients					
Variable	**OLS (S.E)**	**M Estimation (S.E)**	**LTS Estimation (S.E)**	**MM Estimation (S.E)**	**S Estimation (S.E)**	**Modified OLS after handling HLP (S.E)**
Intercept	-1.15 NS (0.21)	-1.31NS (0.17)	-1.49 NS (0.15)	-1.38 NS (0.17)	-1.50 NS (0.179)	-1.28NS (0.18)
x_1	0.74* (0.04)	0.76* (0.03)	0.79* (0.02)	0.76* (0.03)	0.789* (0.035)	0.74* (0.03)
x_2	-0.04 NS (0.03)	-0.02NS (0.03)	-0.01 NS (0.02)	-0.01 NS (0.02)	-0.08 NS (0.02)	-0.01NS (0.03)
R^2	0.82	0.65	0.86	0.62	0.82	0.85
AIC	632.28	109.82	-265.21	174.6	-188.19	104.52
BIC	650.30	118.76	-257.50	185.17	-180.46	94.22

*** 5%level of significance NS= non significant**

Table 4.47 shows the estimation results for the CD function, showing that each estimation coefficient may be correct in terms of expected sign. The regression coefficient was estimated using OLS, M Estimation, MM, S, LTS, and modified OLS after dealing with HLP. The OLS estimate of x_1 was positive 0.74 and significant whereas M, MM, S and LTS Estimation were positive 0.76, 0.79, 0.76 and 0.78 unit and significant respectively. Modified estimates of the production function were also found to be positive 0.74 unit and significant after handling HLP. Moreover, the coefficient of determination is remarkably higher in case of LTS estimation 0.86 which indicates that 86 percent variation of study variable is explained through this method but this method is not recommended because it is based on trimmed mean Chen and Cary (1987). The least were found to be in case of MM method it is0.62

unit. If the cost of labour is reduced by one unit, the overall output of apples will increase. It will be increased by seventy four percent of the total amount. The OLS technique finds the estimate of capital cost to be not significant whereas the robust method finds it to be significant. The yield of an apple will reduce by one percent if the capital cost is reduced by one unit.

Table 4.48:Estimates of Mitscherlich-Baule production function through robust methods for simulation data (P_2) and comparison with OLS thereof.

Variable	Least square method	Iteratively Reweighted Least Squares	
	Gauss Newton Method	Marquardt Method	Gradient Method
a	512.41^{NS} (262.20)	504.30^{NS} (241.01)	100.01^{NS} (763.01)
b	1.21^{NS} (3.01)	0.07* (0.33)	1.110^{NS} (0.43)
c	4.13^{NS} (0.35)	4.86^{NS} (3.99)	3.02^{NS} (0.36)
d	0.06* (2.97)	0.05* (0.01)	0.04* (0.40)
e	2.37^{NS} (3.10)	4.41^{NS} (0.10)	3.99^{NS} (6.07)
R^2	0.65	0.52	0.51
AIC	1836.83	6667.11	6663.21
BIC	1851.53	6681.22	6677.71

*** 5%level of significance NS= non significant**

Mitscherlich-Baule production function gives great flexibility in modelling the relationship between the response variable and independent (regressors) variables. It uses additional coding compared to model specifications in linear modelling procedures. It has sufficient flexibility to accommodate limited factor substitution. The Mitscherlich-Baule production function estimates of simulated data have been done through two methods least square method (LS) and Iteratively Reweighted Least Squares (IRLS).In LS method, the gauss newton method and in IRLS the marquardt and gradient method have been applied. It is observed from the table 4.48 that the coefficient estimates of parameters through Gauss newton method are a(512.41), b (1.21), c (4.13) and e (2.37) unit and non-significant while as the coefficient estimates of d (0.06)unit is positive as well as significant. The R^2 (0.65) value in case of gauss newton method is greater than other two methods which indicates that 65 percent variation of study variable is explained through this method. Further, the AIC (1836.83) and BIC (1851.53) values are low among other two methods. In case of marquardt method which is usually middle most between the both having the coefficient estimates positive and non significant a (504.30), c (4.86), e (4.41) unit

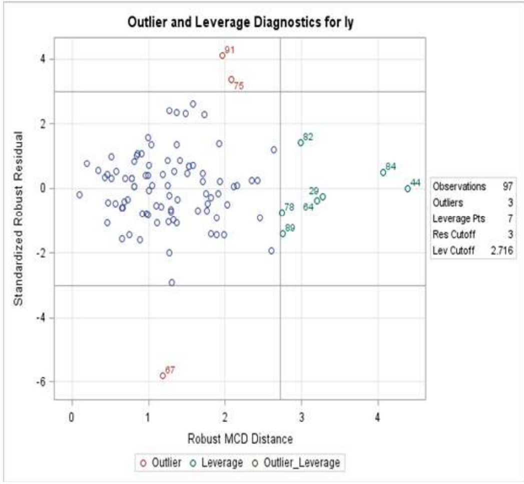


Fig 4.34 : Outlier and Leverage diagnostics for dependent variable for maize data for Cobb Douglas production function.

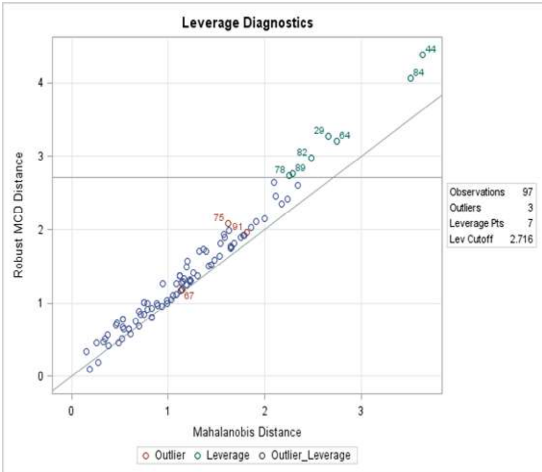


Fig 4.35 : Leverage diagnostics for dependent variable for maize data for Cobb Douglas production function.

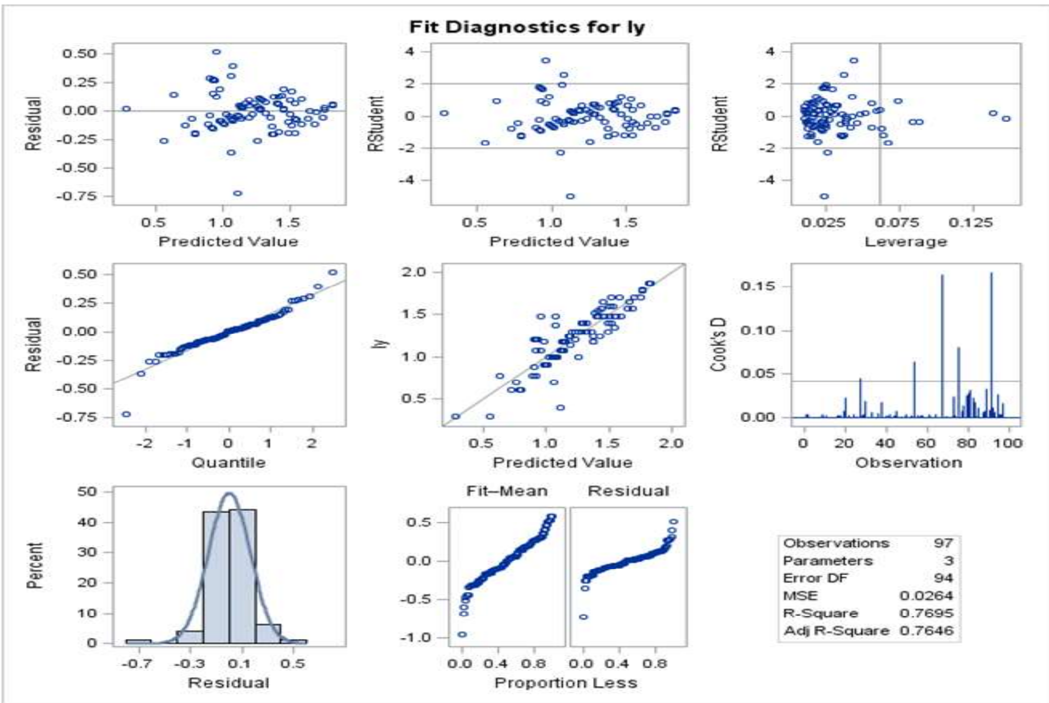


Fig 4.36: Diagnostics Panel of Cobb Douglas production function through OLS.

whereas b (0.07) and d (0.05) are significant. The R^2 is 0.52 with AIC and BIC value are 6667.11 and 6681.22 respectively. In case of Gradient method, the results of coefficient estimates are same as Guass Newton methods in terms of significant and non significant but different w.r.t sizes of the estimates i.e. a (100.01), b (1.11), c (3.02), d(0.04) and e (3.99) unit having R^2 value 0.51 with AIC and BIC values 6663.21 and 6677.71 respectively. The coefficients b and d describes the influence of the corresponding input factors on the study variable. c and $e$are the input variables that include $x_1$and x_2.The decision on the most appropriate estimation technique can exclusively be based on statistical measures. For instance, the goodness of fit with AIC and BIC. Thus in case of simulation data with large observation Gauss newton method outperforms Marquardt and Gradient method on the basis of coefficient of determination. But the parameters of marquardt method are more significant as compared to the other methods. In order to see the performance of MB, the simulation data of 500 observations is considered.

Table 4.49: Estimates of Mitscherlich-Baule production function through robust methods for simulation data (P_1) and comparison with OLS thereof.

	Least square method	Iteratively Reweighted Least Squares	
Variable	**Gauss Newton Method**	**Marquardt Method**	**Gradient Method**
a	131.00^{NS} (292.90)	442.22^{NS} (357.00)	100.01^{NS} (751.31)
b	4.18^{NS} (1.12)	0.07* (0.60)	2.01^{NS} (18.59)
c	1.94^{NS} (0.08)	7.58* (0.20)	4.00^{NS} (12.10)
d	0.01* (1.10)	14.60^{NS} (7.22)	0.05* (0.51)
e	90.17^{NS} (36.20)	72.11^{NS} (05.20)	5.01^{NS} (11.30)
R^2	0.99	0.99	0.47
AIC	3830.84	3611.00	3618.99
BIC	3842.63	3623.64	3631.64

*** 5%level of significance NS= non significant**

According to table 4.49 the coefficient estimates of gauss newton method area (131.00), b (4.18), c (1.94) and e (90.17) units and found non significant while as the coefficient estimate of d (0.01) was significant. The R^2 (0.99) value in case of Gauss newton and Marquardt method were same but AIC (3611.00) and BIC (3623.64) values in case of Marquardt is low among two methods. In case of Marquardt method the coefficient estimates of a (442.22), d (14.60) and e (72.11) unit were also non significant while as c (7.58) and b (0.07) is significant with R^2(0.99). In case of gradient method, the coefficient estimate of a (100.01), b (2.01), c (4.00) and e (5.01) unit are not significant while as d (0.05) is significant having R^2 value 0.47 with AIC and BIC 3618.99 and 3631.64 unit respectively. The coefficients b and $d$describes the influence of the corresponding input factors on the study variable. c and e are the input variables that include $x_1$and x_2. Here marquardt outperforms gauss newton method and gradient method which is robust than other two methods Cuthbert *et al.*(1980) as this method interpolates between the Gauss Newton and gradient method. In simulation data set the Gauss Newton and Marquardt method performs well. Now in order to see the performance of MB in case of real data sets the discussions are as follows.

Table 4.50: Estimates of Mitscherlich-Baule production function through robust methods for apple yield data w.r.t labour and capital and comparison with OLS thereof.

	Least square method	Iteratively Reweighted Least Squares	
Variable	**Gauss Newton Method**	**Marquardt Method**	**Gradient Method**
a	869.321* (382.27)	609.12^{NS} (338.21)	43.61* (2.91)
b	0.03 ^{NS} (0.02)	6.75* (0.03)	0.01 ^{NS} (6.04)
c	-10.60 ^{NS} (9.71)	-7.40 ^{NS} (14.61)	0.01 ^{NS} (2.31)
d	3.75* (0.09)	7.64* (0.03)	0.04 ^{NS} (0.21)
e	13.11^{NS} (106.21)	15.82* (3.58)	4.10* (1.24)
R^2	0.63	0.87	0.53
AIC	1411.56	1227.96	1990.92
BIC	1421.51	1237.91	2000.85

*** 5%level of significance NS= non significant**

This function is characterized by continuously positive marginal productivities of the input factors. It does not exhibit negative marginal productivities. As per table 4.50the coefficient estimate of b (0.03) c (-10.60), and e (13.11) unit are non significant while as the coefficient estimate of a (869.321) and d (3.75) unit are significant as per Gauss Newton Nethod. In Marquardt Method the coefficient estimate of a (609.12) and c (-7.40)are not significant while as b (6.75), d (7.64) and e (15.82) unit are significant. In Gradient method, the coefficient estimate of b (0.01), c (0.01), and d (0.04) unit are non significant while as a (43.61) and e (4.10)are significant. The R^2 value in case of marquardt method is remarkably high 0.87 among other two methods which indicates that 87 percent variation of study variable is explained through this method. The AIC (1227.96) and BIC (1237.91) values in case of Marquardt method is low among two methods and is best fitted as per adequacy criteria. The coefficients b and d describes the influence of the corresponding input factors on the study variable. In this case Marquardt outperforms Gauss Newton method and Gradient method. Thus in this case it is interesting to observe the efficiency of these methods in case of another real data set pertains to study the yield of maize w.r.t input variables as labour cost and capital cost involves in infrastructure. In the series now the data pertains to 97 observations had been used to examine the performance of MB.

Table 4.51: Estimates of Mitscherlich-Baule production function through robust methods for maize yield data w.r.t labour and capital and comparison with OLS thereof.

	Least square method	**Iteratively Reweighted Least Squares**	
Variable	**Gauss Newton Method**	**Marquardt Method**	**Gradient Method**
a	142.20* (30.21)	150.40* (62.28)	99.99* (18.71)
b	1.01* (0.02)	0.02 NS (1.60)	0.05* (0.01)
c	0.01 NS (0.02)	12.6* (0.01)	0.01 NS (2.31)
d	4.04 NS (5.49)	0.04 NS (1.09)	0.04 NS (7.75)
e	44.00* (20.01)	55.80* (12.35)	4.00* (0.01)
R^2	0.75	0.94	0.77
AIC	381.39	362.96	374.6
BIC	389.10	370.67	382.31

*** 5%level of significance NS= non significant**

The Mitscherlich-Baule production functions estimates of real data are presented in Table 4.51. In Gauss Newton method coefficient estimate of c (0.01) and d (4.04) are non significant while as the coefficient estimate of a (142.20), b (1.01) and e (44.00) are significant, R^2 value in case of Gauss Newton method is 0.75 while as values of AIC and BIC are 381.39 and 389.10 units. In marquardt method, the coefficient estimate of b (0.02) and d (0.04) are non significant while as a (150.40), c (12.6) and e (7.64) is significant. The R^2 value in this method is 0.94 which indicates that 94 percent variation of study variable is explained through this method and AIC and BIC have values are 362.96 and 370.67 unit respectively. In Gradient method the coefficient estimate of c (0.01) and d (0.04) are not significant while as a (99.99), b (0.05) and e (4.00) are significant. The coefficients b and d describes the influence of the corresponding input factors on the study variable. TheR^2 value in this method is 0.77 with AIC and BIC are 374.6 and 382.31unit respectively. The primary application of Marquardt algorithm is best among the others which is on the basis of IRLS.

Table 4.52: Estimated MVP and Elasticity of real data of linear and non-linear production functions.

	Linear production functions							
Production function	**Quadratic function**				**Square root function**			
Data / Variable	Apple (204)		Maize (97)		Apple (204)		Maize (97)	
	MVP	Elasticity	MVP	Elasticity	MVP	Elasticity	MVP	Elasticity
x_1	0.03	-33.40	0.01	-6.61	0.01	-101.20	0.01	-4.48
x_2	-0.03	24.09	-0.01	29.85	0.04	34.41	0.01	29.38
	Non Linear production functions							
Production function	**Cobb Douglas function**				**Mitscherlich Baule function**			
Data / Variable	Apple (204)		Maize (97)		Apple (204)		Maize (97)	
	MVP	Elasticity	MVP	Elasticity	MVP	Elasticity	MVP	Elasticity
x_1	0.65	-6.48	0.24	-2.11	0.03	-21.21	0.01	-5.58
x_2	-0.02	1.21	-0.01	1.54	0.05	2.08	0.02	6.71

The marginal value product is a measure of the additional profits raised during increasing the quantity of an input used by one unit while maintaining all other input quantities constant. The table 4.55 shows the MVP and elasticity of linear and non-linear production functions of real data. The MVP was calculated by using regression coefficients of each input of both the functions and geometric mean value of input values. The MVP values in case of quadratic production function for the apple data w.r.t the variables Labour cost and Capital i.e$x_1$and $x_2$were found to be 0.03 and -0.03 respectively. The elasticity for x_1 (-33.40) is less than zero showed a negative decreasing function to the factors indicating the over-utilization of the input implying that its allocation and utilization were in irrational stage of production (stage III) of the production process. Further, the elasticity values for x_2 (24.09) in quadratic production function was found to be greater than unity indicating that the output responds strongly to increases in the use of these inputs indicating that the allocation and utilization of the variable was in the Stage I of production process. In case of maize data for the QF the MVP for $x_1$and $x_2$were found to be 0.01 and -0.01 and the elasticity for x_1 (6.61) and x_2 (29.85) respectively. For the SF, the MVP values for $x_1$and x_2 ,in case of apple data, were found 0.01 and 0.04. The Elasticity for$x_1$is (-101.20) the negative sign indicates that the over-use of the input and the allocation of use were irrational in the production process (stage III), and the elasticity values for x_2 (34.41) was found to be larger than unity in square root production, showing that the result responds strongly indicating that the allocation and utilization of the variable was in the Stage I of production process. In case of maize data the MVP for $x_1$and $x_2$were found to be same 0.01 and the elasticity forx_1 (-4.48) was less than zero, suggesting an over-utilization of the input, meaning that its allocation and utilization were at an irrational stage of production (stage III) of the production process. Furthermore, the elasticity values for x_2 (29.38) in SF were found to be greater than unity, showing that the output responds strongly to increases in the use of these inputs, indicating that the variable was allocated and employed in Stage I of the production process. In case of CD function, the MVP value for x_1 and x_2 were found to be 0.65 and -0.02, respectively, while the elasticity for x_1 (-6.48) is less than zero showed a negative decreasing function to the factors, indicating over-utilization of the input implying that its allocation and utilization were in irrational stage of production (stage III) of the production process. The elasticity values forx_2 (1.21) was found to be greater than unity, indicating that the output responds strongly to increases in the use. For maize data the MVP for $x_1$and $x_2$were found to be 0.24 and -0.01 and the elasticity for x_1 and x_2 -2.11 and 1.54 respectively. The same trend is found in apple as in the case of maize. In case of Mitscherlich-Baule production function c and $e$are the inputvariables that include $x_1$and x_2 .For apple data the MVP for$x_1$and x_2 were found to be 0.03 and 0.05. x_2 is influencing the yield more than x_1. The elasticity for x_1 (-21.21) was less than zero, suggesting a negative decreasing function to the factors, indicating over-utilization of the input and implying that its allocation and utilization were in an irrational stage of the manufacturing process (stage III). The elasticity values for x_2 (2.08) are more than unity, indicating that the output responds substantially to increases in consumption. For maize data the MVP for $x_1$and $x_2$were found to be 0.01 and 0.02. The elasticity for x_1 (-5.58) was less than zero, signalling a negative decreasing function for the factors, indicating over-utilization of the input and implying that its allocation and use were in an irrational stage of the manufacturing process (stage III).The elasticity values for x_2 (6.71) are greater than one, showing that production responds to increases in consumption significantly. The x_1 and $x_2$describes the influence of the corresponding input factors on the study variable. Theoverall the trend of MVP and elasticity for apple and maize data were found to be similar in all the four functions.

Table 4.53 Performance criteria of production functions for simulation data using cross validation technique.

		N= 500										
Model Type	**Production Function**	**Root MSE**	**Dependent Mean**	**R-Square**	**Adj R-Sq**	**AIC**	**AICC**	**PRESS**	**SBC**	**CV PRESS**	**ASE (Train)**	**ASE (Validate)**
Linear	**Quadratic**	36.77	26.67	0.99	0.96	4109.75	4109.83	680021	3620.39	731078	1102.52	1620.31
	Square root	38.27	24.35	0.98	0.94	4147.78	4147.81	733942	3650.00	731099	1181.22	1753.13
Non-Linear	**Cobb Douglas**	0.50	0.74	0.05	0.30	-176.74	-176.69	128.67	-670.31	128.32	0.211	0.303
	Mitscherlich-Baule	36.82	26.68	0.02	0.01	4111.10	4111.18	682236	3621.74	684372	1121.87	1631.02
		N=1000										
Linear	**Quadratic**	27.73	17.19	0.97	0.96	7650.34	7650.38	772503	6663.06	779186	725.90	811.09
	Square root	28.21	4.98	0.96	0.99	-334.27	-334.23	262.85	-1321.55	262.39	0.27	0.25
Non-Linear	**Cobb Douglas**	0.48	0.74	0.08	0.06	-459.08	-459.04	232.09	-1446.36	231.28	0.200	0.267
	Mitscherlich-Baule	27.74	17.20	0.03	0.03	7650.43	7650.47	771936	6663.15	775720	725.902	811.09

The cross validation technique is a data resampling method by partitioning a data set into two training and testing data set. The training data set is used to fit a model and testing data set for evaluate the prediction performance of the fitted model through prediction error .As per Table 4.56, the performance for simulated data of production functions using cross validation techniques. In linear production functions the root mean square error (RMSE) value of QF and SF are 36.77 and 38.27 respectively. The former model is best to study being the lowest values of RMSE. The R^2value for QD and SF are 0.99 and 0.98 respectively. Further, AIC, AICC, predicted residual error sum of squares (PRESS), Schwarz criterion (SBC) and CVPRESS values for both the production function are as 4109.75and 4147.78; 4109.83and 4147.81; 680021and 733942; 3620.39 and 3650.00; 731078 and 731099 respectively. The overall R^2 value in case of QF was found to be greater with low AIC and BIC values as compared to SF. The lowest values of PRESS indicating the best model is QF. Further,both the models are acting normally, because training error is not higher than validation error.In non-linear production functions, the CD function has RMSE value 0.50 while as the value for MitscherlichBaule function is 36.82. Further, AIC, AICC, PRESS, SBC and CVPRESS values for both the production function are as -176.74 and 4111.10; -176.74and4111.18; 128.67 and 682236; -670.31and 3621.74; 128.32and684372 respectively.PRESS is the analogue of coefficient of determination. The lowest PRESS values indicate the best model.Generally, the lower the PRESS value, the better the prediction power of the model. Furthermore, both models are behaving appropriately because the training error is not greater than the validation error. Further, in case of simulation data pertains to big data the results of cross validation are depicted in Table 4.58. The linear production functions undertaken in study QF and SF their having RMSE value as 27.73 and 28.21 respectively. The former model is best to study being the lowest values of RMSE. The R^2value for QF and SF are 0.97 and 0.96 respectively. Further, AIC, AICC, PRESS, SBC and CVPRESS values for both the production function are as 7650.34 and -334.27; 7650.34 and -334.27; 772503 and 262.85; 6663.06 and -1321.55; 779186 and 262.39 respectively.The overall R^2 value in case of QF was found to be greater than the SF i.e. 0.97which indicates that 97 percent variation of variable is explained by this model. Also in QF the AIC and BIC values are low as compared to square root production function. Thus QF is the best in case of linear production functions. In general the smaller the PRESS values the better the models predictive ability which is QF. In non-linear production functions P_2CD function has RMSE value 0.48 while as the value for MitscherlichBaule function is 27.74. AIC, AICC, PRESS, SBC and CVPRESS values for both the production function are as -459.08 and 7650.43; -459.04 and 7650.47; 232.09 and 771936; -1446 and 6663.15; and 231.28 and 775720 respectively. Thusconcluded that both the models are acting normally, because training error is not higher than validation error.

Table 4.54 Performance criteria of production functions for real data using cross validation technique.

		Apple (204)										
Model Type	Production Function	Root MSE	Dependent Mean	R-Square	Adj R-Sq	AIC	AICC	PRESS	SBC	CV PRESS	ASE (Train)	ASE (Validate)
Linear	Quadratic	17.16	43.68	0.82	0.82	1371.90	1372.47	63021	1185.80	63253	246.95	526.97
	Square root	19.25	43.68	0.78	0.78	1415.70	1415.90	76613	1219.66	75967	228.03	530.37
Non-Linear	Cobb Douglas	0.12	1.46	0.35	0.31	-634.33	-634.21	3.36	-833.70	3.29	0.01	0.02
	Mitscherlich-Baule	22.70	43.68	0.40	0.41	1481.88	1482.00	106478	1282.52	105003	338.07	772.35

		Maize (97)									
Linear	**Quadratic**	6.09	22.26	0.85	0.85	454.31	455.24	3797.25	368.18	3712.95	31.64	47.22
	Square root	6.19	22.26	0.84	0.83	455.86	456.29	3796.72	364.58	3871.02	29.73	46.80
Non-Linear	**Cobb Douglas**	0.64	22.26	0.29	0.21	454.30	455.24	3797.24	368.18	3712.95	0.04	0.16
	Mitscherlich-Baule	6.32	22.26	0.32	0.33	458.73	458.99	3924.92	364.88	4001.1	31.64	47.22

The performance of production functions in case of real data sets using cross validation techniques is shown in Table 4.54. The quadratic production function has an RMSE of 17.16 in linear production functions of apple data while the square root function has an RMSE of 19.25. Because it has the lowest RMSE values, the former model is the best to investigate. QF and SF have R^2 values of 0.82 and 0.78, respectively. Furthermore, for both production functions, the AIC, AICC, PRESS, SBC, and CVPRESS values are 1371.90 and 1415.70; 1372.47and1415.90; 63021 and 76613; 1185.80 and1219.66; 63253and75967 respectively .The R^2 value for the QF was found to be greater than the SF, i.e. 0.82, indicating that this model can explain 82 percent of the variation in the variable. The PRESS values with the lowest values indicate the best model, which is QF. It has been concluded that both models are performing properly because the training error is not more than the validation error. In appledata Cobb Douglas production function has RMSE of 0.12 while Mitscherlich Baule function has RMSE of 22.70. Moreover, the AIC, AICC, PRESS, SBC, and CVPRESS values for both the production functions are -634.33 and 1481.88; -634.21 and 1482.00; 3.36 and 10647; -833.70 and1282.52; 3.29 and 10500, respectively. We also came to the conclusion that both models are functioning normally because the training error is not greater than the validation error. Moreoverin linear production functions of maize data QF has RMSE value 6.09 while as the value for SF is 6.19.The former model is best to study being the lowest values of RMSE.The R^2value for QF and SF is 0.85 and 0.84 respectively. Further, AIC, AICC, PRESS, SBC and CVPRESS values for both the production function are as 454.31 and 455.86; 455.24 and 456.29; 3797.25 and 3796.72; 368.18 and 364.58; 3712.95 and 3871.02. The lowest values of PRESS indicating the best model is QF. In case Cobb Douglas production function has RMSE value 0.64 while as the value for Mitscherlich Baule function is 6.32. The AIC, AICC, PRESS, SBC and CVPRESS values for both the production functions are as 454.30 and 458.73; 455.24 and 458.99; 3797.24 and

3924.92; 368.18 and 364.88; 3712.95 and4001.1 respectively.The lowest PRESS value indicates model has highest predictive power.The training error is not more than the validation error thus both models are operating normally.

Table 4.55 Performance criteria of production functions for simulation data using k-fold cross validation technique.

		N=500									
		K=5					K=10				
Model type	**Production Function**	**Root MSE**	**R-Square**	**Dependent Mean**	**Adj R-Sq**	**C V (%)**	**Root MSE**	**R-Square**	**Dependent Mean**	**Adj R-Sq**	**C V (%)**
Linear	**Quadratic**	36.86	0.01	26.88	0.05	139.67	36.22	0.01	25.93	0.05	137.11
	Square root	38.36	0.07	24.44	-0.06	161.72	37.65	0.09	23.37	-0.02	156.95
Non-Linear	**Cobb Douglas**	3.23	0.08	0.76	0.04	67.51	3.16	0.04	0.76	-0.05	66.16
	Mitscherlich-Baule	36.74	0.01	26.88	0.01	139.42	36.16	0.01	25.93	0.08	136.68
	N=1000										
		K=5					K=10				
Model type	**Production Function**	**Root MSE**	**R-Square**	**Dependent Mean**	**Adj R-Sq**	**C V (%)**	**Root MSE**	**R-Square**	**Dependent Mean**	**Adj R-Sq**	**C V (%)**
Linear	**Quadratic**	27.96	0.03	17.54	0.02	159.42	28.41	0.03	17.55	0.02	161.91
	Square root	23.21	0.99	4.48	0.99	10.21	21.12	0.99	4.91	0.99	10.12
Non-Linear	**Cobb Douglas**	3.01	0.01	0.74	0.01	65.20	2.95	0.08	0.75	0.06	63.07
	Mitscherlich-Baule	28.02	0.02	17.54	0.02	161.77	28.39	0.02	17.55	0.02	159.75

Further the K fold cross validation takes care of large portion of the data set. It is a good ratio of testing data points-folded cross validation is a procedure used to estimate the skill of the model on new data. There are common tactics that can use to select the value of k for your dataset. Table 4.55 shows performance criteria of production functions for both the simulated data using k-fold cross validation technique. The real error rate is small with a large number of folds (K), but the variation is large. Depending on the complexity of the models under consideration-fold validation has been replicated samples 10 times. We have reflected the value of K for 5 fold and 10 fold i.e. K=5 and K=10 for the linear and non linear production functions. Out of these two cases 10 folded comes out to be best as R^2 value is more. The Coefficient of variation in case of 5 fold was found be larger than 10 fold cross validation. Therefore 10 folded validation comes out to be best.

Table4.56 Performance criteria of production functions for real data using k-fold cross validation technique.

		N=204									
		K=5					K=10				
Model Type	**Production Function**	**Root MSE**	**R-Square**	**Dependent Mean**	**Adj R-Sq**	**C V(%)**	**Root MSE**	**R-Square**	**Dependent Mean**	**Adj R-Sq**	**C V (%)**
Linear	**Quadratic**	16.95	0.83	44.60	0.83	38.00	17.71	0.83	46.87	0.82	37.80
	Square root	18.34	0.80	44.60	0.80	41.12	18.99	0.80	46.87	0.79	40.52
Non-Linear	**Cobb Douglas**	1.34	0.89	1.47	0.89	8.77	1.34	0.90	1.49	0.90	8.69
	Mitscherlich-Baule	22.98	0.69	44.60	0.68	51.51	23.10	0.70	46.07	0.70	49.42
	N=97										
		K=5					K=10				
Model Type	**Production Function**	**Root MSE**	**R-Square**	**Dependent Mean**	**Adj R-Sq**	**C.V (%)**	**Root MSE**	**R-Square**	**Dependent Mean**	**Adj R-Sq**	**C V (%)**
Linear	**Quadratic**	5.81	0.87	21.98	0.86	26.01	5.75	0.86	22.12	0.85	26.03
	Square root	5.50	0.88	22.64	0.87	25.61	5.85	0.84	21.32	0.83	27.46
Non-Linear	**Cobb Douglas**	1.44	0.78	1.23	0.78	13.05	1.45	0.74	1.22	0.73	13.26
	Mitscherlich-Baule	6.04	0.86	22.64	0.86	26.70	6.06	0.82	21.32	0.82	28.45

Table 4.56 provides k-fold cross validation performance criteria for production functions for real data sets. With a large number of folds (K), the real error rate is minimal, but the variation is substantial. The computational time may be very long depending on the complexity of the models under consideration. We replicated samples ten times in K-fold validation. For the linear and nonlinear production functions, the K values used are 5 and 10, i.e. K=5 and K=10. The R^2 value in the case of 10 folded cross validation techniques is greater than the R^2 value in the case of 5 folded cross validation technique. Coefficient of variation in case of 5 fold was found be larger than 10 fold cross validation is .Therefore 10 folded validation comes out to be best. While as in case of maize data 5 folded comes out to be best as R^2 value is greater in case of 5 folded, also coefficient of variation in case of 5 folded is smaller than 10 folded cross validation technique. It is also observed that the RMSE value at K=10 for the QF in linear is least and found to be the best model in case of both data whereas CD and MB are usually used under different conditions so are not compared.

Chapter-5

Discussion

In any statistical data analysis the quality of the observations are very important. Generally, the real world datasets contaminated with the outliers. In this situation, the identification of outlier is the preliminary step in the process of data analysis. If the existing outliers are not properly treated, it will lead into inappropriate results. Hence, the analysts need to be very careful in dealing with these outliers. Datasets often used to estimate the parameters. Outliers can bias the estimates and therefore, may produce misleading results. It is often difficult to detect outliers through visual inspection without the aid of analytic or graphical tools. The experiment of Collett and Lewis (1976) indicates that visual inspection alone is not a reliable way to identify potential outliers. A superior approach uses a specific function of the observations. Such a measure aims not to identify observations as definite outliers, but only to label them as sufficiently suspect to merit further investigation. In this way, an outlier-labeling rule is an improved alternative to visual inspection. According to Iglewicz and Hoaglin (1993) the issues with regard to outliers can be categorized as outlier labeling method, outlier accommodation and outlier identification. Each method serves the need of different situations. The outlier identification methods formally test whether observations are outliers so that they can be set aside. As a basis for making such tests at a specified significance level, adopt the customary convention that the good observations adequately described as a sample from a normal distribution. Outlier accommodation emphasizes the techniques that suffer little adverse effect when outliers are present, but that do not attempt to remove the outliers. The outlier accommodation procedure can use robust statistical techniques that will not be excessively affect by outliers. There are many diagnostics available for identifying high leverage points that can distinguish between the good and bad ones. High leverage points contain both the good leverage points and bad leverage points. a leverage point is called ‘good leverage point’ if it does not deviate from the typical relationship. Good leverage points are no outliers and even improve the regression inference as these points reduce standard errors of coefficient estimates. The bad points are harmful to the fit while the good points improve it. The analysis of observations indicated as outliers is of particular relevance to exhibit the data structure. Outlier detection methods explicitly determine if observations are outliers, allowing them to be discarded. Adopt the standard practice that excellent observations are properly characterized as a sample from a normal distribution as a basis for running such tests at a defined significance level. When outliers are present, outlier accommodation emphasizes strategies that have little negative impact but do not attempt to remove the outliers. Outlier accommodation can make advantage of robust statistical procedures that aren't affected too much by outliers. There are a variety of diagnostics for identifying high leverage points that can tell the difference between good and bad ones. Both positive and harmful leverage points can be found in high leverage positions. If a leverage point does not stray from the conventional connection, it is referred to as a "good leverage point." There are no outliers at good leverage points, and they even help regression inference by lowering standard errors of coefficient estimations. The results obtain during the present investigation entitled "Estimation and Validation of linear and non linear production functions through Robust Regression" in preceding chapters have been discussed under the following sub-heads:

5.1 Robust Methods for Quadratic Production Function in Case of Simulated/Real Data in Presence of HLP.

5.1.1 Robust methods for Quadratic Production Function for Simulated Data (P_1).

The results of summary statistics Table (4.1) for independent variable and dependent variables indicted that the mean and standard deviation of the study variable w.r.t simulated data were found to 24.31 and 38.47 clearly indicates variation is more than mean that is due to the presence of influential observations. In multiple regression models, reaching a significant F statistics for a model in linear regression problem is not always an indicator of best fit, especially in determining the regression coefficient every observation requires a detailed examination. Even a single observation on a parameter can be very influential and removal of the set can completely change the regression equation. Both the influential observations and outliers violate the assumptions of multiple regression models. In the table (4.2) the detection of outliers has been done through Mahanablios Distance, Robust MCD Distance, and Standardized Robust Residual, cooks distance, studentized residual, WSSDI, Hat Diagonal and Deleted Residual. Table (4.3) provides Estimates of quadratic production function P_1 through ordinary least square (OLS).The F-value was found to be significant indicates the model is adequate w.r.t to study variable. Table (4.4) presents the estimation results for the Quadratic production function through M Estimation, MM, LTS, S and modified OLS after handling HLP through linear interpolation technique. The coefficient of determination is clearly higher in the case of MM estimation 0.99which indicates that 99 percent variation of study variable is explained through this method also AIC and BIC values for MM estimation are low as compared to other robust techniques which makes this estimation method best among other robust methods.

5.1.2 Robust methods for Quadratic Production Function for Simulated Data (P_2).

The examination of outlier observations is very important for displaying the data structure. The mean and standard deviation of the study variable w.r.t simulated data were found to be 17.19 unit and 28.15 unit respectively, indicating that variation is greater than mean due to the presence of influential observations, according to the results of summary statistics Table (4.5) for independent and dependent variables. Reaching a substantial F statistic for a model in a linear regression problem is not always an indicator of best fit in multiple regression models, especially when determining the regression coefficient for each observation. Outliers were detected using Mahanablios Distance, Robust MCD Distance, and Standardized Robust Residual, cooks distance, studentized residual, WSSDI, Hat Diagonal, and Deleted Residual in Table (4.6). Table 4.7 shows the results of ordinary least squares to estimate the quadratic production function. The fact that the F-value was significant shows that the model is adequate in terms of the study variable. Table (4.8) shows the results of M Estimation, MM, LTS, S, and modified OLS after handling HLP using the linear interpolation methodology for the Quadratic production function. The coefficient of determination is higher in the case of LTS estimation 0.991, but this method is not usually recommended because it is based on trimmed mean SAS/STAT 13.1 User's Guide. Cary, NC (2013). This is followed by S estimation 0.99 and AIC and BIC values are also low in S estimation method which makes it better than other robust methods.

5.1.3 Robust methods for Quadratic Production Function for Apple Data.

For exhibiting the data structure, it's critical to look at outlier observations. According to the results of summary statistics Table (4.10) for independent and dependent variables, the study variable's mean and standard deviation were 2.05 unit and 2.18 unit respectively, showing that variation is greater than mean due to the presence of influencing observations. In multiple regression models, achieving a significant F statistic for a model in a linear regression issue is not always an indicator of optimal fit, especially when determining the regression coefficient for each observation, which requires a thorough investigation. Outliers were found utilizing Mahanablios Distance, Robust MCD Distance, and Standardized Robust Residual in table 4.11, as well as cooks distance, studentized residual, WSSDI, Hat Diagonal, and Deleted Residual. The results of ordinary least squares estimation of the quadratic production function of apple data are shown in Table 4.12. The presence of a significant F-value indicates that the model is sufficient in terms of the study variable. After processing HLP employing the linear interpolation methodology for the Quadratic production function of apple data, Table (4.13) displays the outputs of M Estimation, MM, LTS, S, and modified OLS. Modified OLS after handling HLP has a higher coefficient of determination (0.83), which outperforms other robust approaches, followed by MM estimation (0.77).The AIC and BIC values for Modified OLS after handling HLP was lowest among other methods, lower the value of AIC ,BIC better the model is considered.

5.1.4 Robust methods for Quadratic Production Function for Maize Data.

A single outlier can lead the coefficient estimates to deviate arbitrarily far from the real underlying values, which the least squares estimator cannot control. Outliers lead to erroneous coefficient estimates as a result. In Table 4.15the outliers have been identified by Mahanablios distance, Robust MCD Distance, Standardized Robust Residual, Cooks distance, studentized residual, WSSDI, Hat Diagonal, and Deleted Residual, as well as HLP. Throughout the situation of three observations, outliers 19, 75, and 91 were discovered, and HLP was used. The maximum was determined to be 3.91 in Mahanablios distance and 21.15 in Robust MCD. The maximum was found to be 3.95, 0.30, 2.99, 3761, 0.07, 19.75 in standardized robust residuals, cook's distance, studentized residual, WSSDI, Hat diagonal, and deleted residual, respectively. Table 4.18 also shows the regression coefficients for estimating maize yield using exogenous factors such as labour and capital, utilizing several robust approaches. M estimate has a coefficient of determination of 0.881, which is higher than other robust approaches. The least was found in the case of MM estimate, which has 0.631. After dealing with HLP, the value of R^2 in the case of OLS was 0.88 and explained 88 percent of the predictor variables variation. The AIC and BIC values in case of M estimation are very low as compared to other robust methods. Using robust estimation, the spread of ideal input values is narrowed across the various functional forms. As a result, when robust regression is used to estimate production functions, the differences between different functional forms are minimized.

5.1.5 The Best Model of Linear Production Function for Simulated and Real Data for Quadratic Production Function

The Best Model of Linear Production Function for Simulated Data $P_1$for quadratic production function is as:

$$\hat{Y} = 4.99 + 504x_1{}^{*} + 3.06x_2{}^{*} - 0.01x_1{}^{2^{*}} - 0.04x_2{}^{2^{NS}} - 0.06x_1x_2{}^{NS}$$

In this function among different robust methods.MM estimations comes out to be best on the basis of Maximum R^2 value and Minimum AIC and BIC values. The parameters obtained from the OLS and robust regression estimation revealed the composition of the input factors, their squared values, and an additional interaction term.

The Best Model of Linear Production Function for Simulated Data (N=1000) for quadratic production function is as:

$$\hat{Y} = 5.08 + 5.03x_1{}^{*} + 3.09x_2{}^{*} + 0.02x_1{}^{2^{NS}} - 0.03x_2{}^{2^{NS}} + 0.05x_1x_2{}^{NS}$$

In this S estimation outperforms the other robust approaches due to big data. The input factor composition, their squared values, and an additional interaction term were revealed by the variables obtained from the OLS and robust regression estimate. In contrast to OLS, robust regression was selected because of the sign, size, and significance of the variable.

The Best Model of Linear Production Function for Apple Data for quadratic production function is as:

$$\hat{Y} = 38.25 - 0.02x_1{}^{*} + 0.01x_2{}^{*} + 0.06x_1{}^{2^{NS}} + 0.03x_2{}^{2^{*}} - 0.01x_1x_2{}^{*}.$$

In apple data for QF function modified OLS after handling HLP comes out to be best among different robust techniques on the basis of sign, size and significance of the variable. The parameters obtained from the robust regression estimation revealed that for yield of apple the variable x_2(capital) will maximize the yield whereas the increase in cost of labour will decrease. Further the interactions of both the variables are decreasing the yield.

The Best Model of Linear Production Function for Maize Data for quadratic production function is as:

$$\hat{Y} = 7.36 + 0.03{x_1}^{NS} - 0.04{x_2}^{NS} - 5.67{x_1}^{2^{NS}} + 9.73{x_2}^{2^{*}} + 2.71x_1{x_2}^{*}.$$

In maize data for QF function M estimation outperforms other robust techniques due to maximum R^2 with sign, size, and significance of the parameters. The parameters estimated by robust regression estimation revealed that for maize yield variables x_1 (labour), x_2 (capital) are non-significant whereas the square of the cost of infrastructure and interaction of both the variables i.e. labour and capital will increase the yield of maize.

5.2 Robust Methods for Square Root Production Function in Case of Simulated/Real Data in Presence of HLP.

5.2.1 Robust methods for Square root Production Function for Simulated DataP$_1$.

In summary statistics table (4.19) the study variable's mean and standard deviation were 24.35 and 38.27 for independent and dependent variables, respectively, indicating that variation is greater than mean due to the presence of influencing observations. A significant F statistic for a model in a linear regression problem is not always an indicator of optimal fit in multiple regression models, especially when determining the regression coefficient for each observation, which necessitates a comprehensive analysis. Outliers were identified using Mahanablios Distance, Robust MCD Distance, and Standardized Robust Residual, as well as cooks distance, studentized residual, WSSDI, Hat Diagonal, and Deleted Residual, as shown in Table 4.20. Table (4.22) demonstrates the outputs of M Estimation, MM, LTS, S, and modified OLS after handling HLP using the linear interpolation methodology for the square root production function of simulated data (500). MM estimation excels other robust techniques with a higher coefficient of determination (0.99), followed by

LTS estimation (0.98) .The least was found Modified OLS after handling high leverage points 0.54.in this case of simulation data of 500 observations MM estimation method comes out to be best among all robust methods, MM estimation has also Minimum AIC and BIC which also makes it best among all methods.

5.2.2 Robust methods for Square root Production Function for Simulated DataP_2.

Outlying observations deviate from the (quasi) linear relationship described by the majority of the data. They can have a large influence on estimation results. The results of summary statistics Table (4.23) for independent variable and dependent variables indicted that the mean and standard deviation of the study variable w.r.t simulated data were found to 14.44 and 29.75 unit clearly indicates variation is more than mean that is due to the presence of influential observations. In multiple regression models, reaching a significant F statistics for a model in linear regression problem is not always an indicator of best fit, especially in determining the regression coefficient. Every observation requires a detailed examination even a single observation on a parameter can be very influential and removal of the set can completely change the regression equation. Both the influential observations and outliers violate the assumptions of multiple regression models. In the Table (4.24) the detection of outliers has been done through Mahanablios Distance, Robust MCD Distance, and Standardized Robust Residual, cooks distance, studentized residual, WSSDI, Hat Diagonal and Deleted Residual. Table (4.26) presents the outputs of M Estimation, MM, LTS, S, and modified OLS after handling HLP using the linear interpolation methodology for the square root production function of simulated data. Modified OLS after handling high leverage points excels other robust techniques with a higher coefficient of determination (0.99) as compared to other robust approaches .AIC and BIC values in case of Modified OLS is also low among all robust techniques.

5.2.3 Robust methods for Square root Production Function for Apple Data.

When dealing with outliers, analysts must use extreme caution. The parameters are frequently estimated using datasets. Outliers can affect estimates, resulting in potentially misleading results. When outliers are present, outlier accommodation promotes approaches that have little adverse influence but do not attempt to remove the outliers. The identification of outlier is the preliminary step in the process of data analysis. The mean and standard deviation of the study variable were found to be 2.05 and 2.18, respectively, implying that variation is greater than mean due to the presence of influential observations, according to the results of summary statistics Table (4.27) for independent and dependent variables. Acquiring a significant F statistic for a model in a linear regression problem is not always an indicative of best fit in many regression models, especially when determining the regression coefficient. Every observation must be thoroughly examined; even a single observation on a parameter can have a significant impact, and removing the set can totally alter the regression. The assumptions of multiple regression models are contradicted by both influential observations and outliers. Outliers were detected using Mahanablios Distance, Robust MCD Distance, and Standardized Robust Residual, cooks distance, studentized residual, WSSDI, Hat Diagonal, and Deleted Residual in Table (4.28). Table (4.29) shows the results of using ordinary least squares to estimate the square root production function of an apple (OLS). The presence of a significant F-value suggests that the model is adequate for the study variable. The estimated results for the square root production function of apple are shown in Table 4.30. With an R^2 of 0.99, S estimation outperforms the other robust methods proposed for this study. With an R^2 value of 0.98, the modified OLS after handling HLP also produces better results.

5.2.4 Robust methods for Square root Production Function for Maize Data.

In multiple regression models, achieving a significant F statistic for a model in a linear regression issue is not always an indicator of optimal fit, especially when determining the regression coefficient for each observation, which requires a thorough investigation. Outliers were found by Mahanablios Distance, Robust MCD Distance, and Standardized Robust Residual in table (4.31), as well as by cooks distance, studentized residual, WSSDI, Hat Diagonal, and Deleted Residual. The results of ordinary least squares estimation of the square root production function of maize data are shown in Table 4.32. The presence of a significant F-value indicates that the model is sufficient in terms of the study variable. After processing HLP employing the linear interpolation methodology for the Square root production function of maize data. Table (4.33) displays the outputs of M Estimation, MM, LTS, S, and modified OLS. Among Robust methods modified OLS estimation comes out to be best with R^2

value 0.86 followed by S estimation .AIC and BIC values are also low in modified OLS after handling HLP.

5.2.5 The Best Model of Linear Production Function for Simulated and Real Data for Square Root Production Function

The Best Model of Linear Production Function for Simulated Data P_1 for square root production function is as:

$$\hat{Y} = 5.13 + 5.03{x_1}^{1/2^*} + 3.06{x_2}^{1/2^*} - 0.12{x_1}^{NS} - 0.04{x_2}^{NS} + 0.04(x_1 * x_2)^{1/2^{NS}}$$

It is very similar to the quadratic form to ensure decreasing marginal productivity of each input factor, the parameters satisfy the same conditions as quadratic form and their interpretation is identical. In this method MM estimation performs better than other robust approaches. On the basis of the sign, size and significance of the variable in relation with OLS, robust regression has been selected.

The Best Model of Linear Production Function for Simulated Data $P_2$for square root production function is as:

$$\hat{Y} = 4.95 + 5.07{x_1}^{1/2^*} + 3.01{x_2}^{1/2^*} + 0.06{x_1}^{NS} + 0.02{x_2}^{NS} - 0.04(x_1 * x_2)^{1/2^{NS}}$$

In this function Modified OLS after handling HLP have shown good performance as compared to other robust methods .The factors satisfy the same constraints as quadratic form and their interpretation is equivalent to that of a quadratic form, which assures the decreasing marginal productivity of every input factor. Robust regression was based primarily on the sign, size and significance of the variable in relation to OLS.

The Best Model of Linear Production Function for Apple Data for square root production function is as:

$$\hat{Y} = 47.51 + 0.03{x_1}^{1/2^*} + 0.05{x_2}^{1/2^*} + 3.49{x_1}^{*} - 4.76{x_2}^{NS} - 0.09(x_1 * x_2)^{1/2^{NS}}$$

In this function S estimation outperforms other robust approaches. The parameters determined by classical OLS and robust regression estimation revealed that for apple production. The square root of both the variables i.e. $x_1$and x_2 are maximizing the yield. Overall the cost of labor variable is contributing to maximize the yield of apple. In case of Linear Production Function for Maize Data for square root production function the best model is as:

$\hat{Y} = 8.34 + 0.01{x_1}^{1/2^*} + 0.03{x_2}^{1/2^*} + 0.19{x_1}^{NS} - 0.05{x_2}^{NS} + 0.02(x_1 * x_2)^{1/2^*}$.

The variables satisfy the same conditions as a quadratic form, and their interpretation is the same as a quadratic form, ensuring that each input factor's marginal productivity decreases. In this function Modified OLS after handling HLP comes out to be best on the basis of higher R^2 value and low AIC and BIC values. Robust regression was essentially dependent on the variable's sign, size, and significance in relation to OLS. The square root of both the variables labor and capital are significantly contributing to maximize the output i.e. yield of the maize and their interaction is also coming significant.

It is concluded that the input variables individually contribute significantly in case of quadratic function whereas the interaction is significant in case of square root production function if data are contaminated with HLPS. Thus production functions should be used accordingly.

5.3 Robust Methods for Cobb Douglas Production Function in Case of Simulated/Real Data in Presence of HLP.

5.3.1 Robust methods for Cobb Douglas Production Function for Simulated Data (P_1).

From Table (4.34), it has been observed the mean and standard deviation of the study variable w.r.t simulated data were found to 0.50 and0.74 unit which clearly indicates variation is more in the data as the mean is smaller than the standard deviation due to the presence of influential observations. The Low standard deviation indicates that data points tend to be very close to the mean and high standard deviation indicates that the data points are spread out over a large range of value. The outliers have been presented in the table (4.35) It has been observed about 4 percent data were found as outliers and HLP. The different methods used for identify the outliers as per the techniques used were Mahanablios distance, Robust MCD Distance, Standardized Robust Residual, Cooks distance, studentized residual, WSSDI, Hat Diagonal and Deleted Residual. Robust methods need to be applied for studying the effect on the estimates of the parameter of quadratic function. Table (4.37) presents different robust methods M estimation, MM, LTS and S estimation. AIC and BIC values for S estimation is smaller as compared to other robust estimation methods Which makes this method best among other methods.

5.3.2 Robust methods for Cobb Douglas Production Function for Simulated Data(P_2).

Outlier detection methods formally determine if observations are outliers, enabling them to be removed. Adopt the standard convention that good observations are properly represented as a sample from a normal distribution as a basis for performing such tests at a defined significance level. The mean and standard deviation of the study variable w.r.t simulated data were found to be 0.48 and 0.74 unit respectively, indicating that variation is greater than mean due to the presence of influential observations, according to the summary statistics Table (4.38) for independent and dependent variables. Reaching a significant F statistic for a model in a linear regression problem is not always a sign of best fit in multiple regression models, especially when determining the regression coefficient for each observation, which requires a full inspection. Even a single observation on a parameter can have a big impact, and removing the set can modify the regression equation completely. The assumptions of multiple regression models are violated by both influential observations and outliers. Outliers were detected using Mahanablios Distance, Robust MCD Distance, and Standardized Robust Residual, cooks distance, studentized residual, WSSDI, Hat Diagonal, and Deleted Residual in Table (4.39). Estimates of the Cobb Douglas production function P_2using conventional least squares are presented in Table (4.40).The fact that the F-value was significant shows that the model is adequate in terms of the research variable. Table (4.41) shows the results of M Estimation, MM, LTS, S, and modified OLS after handling HLP using the linear interpolation methodology for the Cobb Douglas production function.AIC and BIC values in case of MM estimation are smaller as compared to other robust methods. Lower the value of AIC and BIC better the model is considered.

5.3.3 Robust methods for Cobb Douglas Production Function for Apple Data.

Researchers must exercise extra care when working with outliers. Datasets are widely used to estimate parameters. Outliers can distort estimates, sometimes leading to misleading conclusions. Outlier accommodation supports measures that have little adverse influence but do not strive to eradicate outliers when they are present. The detection of outliers is the first stage in the data analysis process. In many regression models, obtaining a significant F statistic for a model in a linear regression problem is not always indicative of optimal fit, particularly when determining the regression coefficient. Every observation must be reviewed closely; even a single observation on a parameter can have a big impact, and eliminating the set can completely change the regression. Influential observations and outliers both defy the assumptions of multiple regression models. Outliers were identified using Mahanablios Distance, Robust MCD Distance, and Standardized Robust Residual, as well as cooks distance, studentized residual, WSSDI, Hat Diagonal, and Deleted Residual, as shown in table (4.42). The outcomes of performing ordinary least squares to estimate the Cobb Douglas production function of apple data are shown in Table (4.43). A significant F-value indicates that the model is appropriate for the studied variable. Table 4.44 shows the estimated results for the Cobb Douglas apple production function. S estimate outperforms the other robust approaches proposed for this investigation on the basis R^2 value and AIC and BIC values.

5.3.4 Robust methods for Cobb Douglas Production Function for Maize Data.

A significant F statistic for a model in a linear regression problem is not always an indicator of optimal fit in multiple regression models, especially when determining the regression coefficient for each observation, which necessitates a comprehensive analysis. Mahanablios Distance, Robust MCD Distance, and Standardized Robust Residual, as well as cooks distance, studentized residual, WSSDI, Hat Diagonal, and Deleted Residual, were used to find outliers in table (4.45). Table shows the findings of the Cobb Douglas production function of maize data using ordinary least squares estimation (4.46). A significant F-value implies that the model is adequate in terms of the research variable. Following HLP processing, the Cobb Douglas production function of maize data was interpolated using a linear interpolation methodology. The outputs of M Estimation, MM, LTS, S, and modified OLS are shown in Table (4.47). Modified OLS after handling HLP has been found to be best among robust techniques as the values of AIC and BIC were very low in this method.

5.3.5 The Best Model of Non-Linear Production Function for Simulated and Real Data for Cobb Douglas Production Function

The Best Model of Non-Linear Production Function for Simulated Data P_1 for Cobb Douglas production function is as:

$$P(x_1, x_2) = 0.83{x_1}^{0.02^{NS}}{x_2}^{0.10^{*}}$$

x_1 is the labour input, x_2 is the capital input, b (0.83) is the total factor productivity and α (0.02) and β (0.10) are output elasticities of labour and capital respectively.The Cobb-Douglas produced a remarkably good fit. In this function S estimation outperforms the other robust approaches. These findings show that the Cobb-Douglas functional form is flexible enough that it can fit the data well.

The Best Model of Non-Linear Production Function for Simulated Data $P_2$for Cobb Douglas production function is as:

$$P(x_1, x_2) = 0.86{x_1}^{0.05^*}{x_2}^{0.06^*}$$

The labour input isx_1, the capital input isx_2, the total factor productivity is b (0.86), and the output elasticities of labour and capital are (0.05) and (0.06), respectively. The Cobb-Douglas developed an extremely accurate fit. These findings show that the Cobb-Douglas functional form is adaptive enough to accommodate the data. Among different robust estimation techniques MM estimation performed better on the basis of R^2value, AIC and BIC values.

The Best Model of Non-Linear Production Function for apple data for Cobb Douglas production function is as:

$$P(x_1, x_2) = -6.71{x_1}^{2.60^*}{x_2}^{-0.63^{NS}}$$

It is a homogeneous production function of degree one which takes into account two inputs, labour and capital for the entire output. Cobb-Douglas produced a remarkably good fit. These findings show that the Cobb-Douglas functional form is flexible enough that it can fit the data well the output elasticities of x_1(labour) and x_2(capital) are 2.60 and -0.63 respectively. In this S estimation method have shown good performance on the basis of maximum R^2 value and Minimum AIC and BIC values. The cost involved in labor maximizing the yield.

The Best Model of Non-Linear Production Function for Maize Data for Cobb Douglas production function is as:

$$P(x_1, x_2) = -1.28^{NS}{x_1}^{0.74^*}{x_2}^{-0.01^{NS}}$$

It is a homogeneous class one production function which takes two inputs, labour and capital into account for the entire yield. The efficiency of Cobb-Douglas had been excellent. These results suggest that the functional form of Cobb-Douglas is sufficiently flexible to match the data, with the output elasticities of x_1 (labour) and x_2(capital) respectively 0.74 and -0.01.In this function modified OLS after handling outperforms comes out to be best.

5.4 Robust Methods for Mitscherlich-Baule Production Function in Case of Simulated/Real Data in Presence of HLP.

5.4.1 Robust methods for Mitscherlich-Baule Production Function for Simulated Data$(\mathbf{P_1})$.

This function is characterized by continuously positive marginal productivities of the input factors. It does not exhibit negative marginal productivities .the function has been found to be preferable for use in an economic model because it allows for factor substitution following von Liebig's law of minimum. Mitscherlich Baule function regimes the most appropriate method the parameters are estimated by MLE procedure .it has sufficient flexibility to accommodate limited factor substitution if the data or production process suggest von Liebig's law of minimum : It states that growth is dictated not by total resources available, but by the scarcest resource (limiting factor).In least square method, Gauss Newton Method had been used and in Iteratively Reweighted Least Square method, Marquardt Method and Gradient Method had been used. Gauss newton method is the simple optimization procedure that we can use with any machine learning algorithm. It is used when training data models can be combined with every algorithm and is easy to understand and implement. Marquardt Method It is used to solve non linear least square problems. It is faster to converge than other methods. It can handle models with multiple free parameters which are not precisely known. It interpolates between the gauss Newton and gradient method. The primary application of Marquardt algorithm is in least square curve fitting problem .Gradient method an iterative algorithm to solve non linear least square problems. It requires Jacobean matrices and many partial derivatives to find a solution. In table 4.48 among three methods Gauss Newton Method outperforms other two methods. R^2 value (0.65) in case of Gauss Newton Method is remarkably higher than Marquardt method and Gradient Method. This indicates that 65 percent variation of study variable is explained through this method. The least were found to be in case of Gradient Method of 0.51.The AIC and BIC in case of Gauss newton method is also low. This method is more robust than other two methods.

5.4.2 Robust methods for Mitscherlich-Baule Production Function for Simulated Data(P_2).

The parameters of the Mitscherlich Baule function regimes are computed using the MLE process. If the data or manufacturing process reflects von Liebig's law of minimum, it has enough flexibility to accommodate limited factor substitution: It asserts that growth is determined not by the overall amount of resources available, but by the scarcest resource (limiting factor). The Gauss Newton Method was utilized in the least square approach, and the Marquardt Method and Gradient Method were employed in the Iteratively Reweighted Least Square method. The Gauss-Newton method is a simple optimization procedure that may be used to any machine learning system. The gauss Newton and gradient methods are interpolated. The Marquardt algorithm is most commonly used in the least square curve fitting problem. The gradient technique is an iterative strategy for solving problems with non-linear least squares. The Marquardt Method surpasses the other two methods in Table 4.49. This strategy is more reliable than the other two. The R^2 value of the Marquardt Method (0.99) is significantly greater than the R^2 values of the Gauss Newton and Gradient methods. In case of Marquardt Method approach, the AIC (3611.00) and BIC (3623.64) are also low.

5.4.3 Robust methods for Mitscherlich-Baule Production Function for Apple Data.

The marginal productivities of the input elements are continually positive in this function. It doesn't have any negative marginal productivity. The function has been determined to be preferred for usage in an economic model since it allows for factor replacement in accordance with von Liebig's rule of minimum. The parameters of the Mitscherlich Baule function regimes are computed using the MLE process. If the data or manufacturing process reflects von Liebig's law of minimum, it has enough flexibility to accommodate limited factor substitution: It asserts that growth is determined not by the overall amount of resources available, but by the scarcest resource (limiting factor). The Gauss Newton Method was utilized in the least square approach, and the Marquardt Method and Gradient Method were employed in the Iteratively Reweighted Least Square method. The Gauss-Newton method is a basic optimization procedure that may be used to any machine learning system. When training data models can be integrated with any algorithm and are simple to comprehend and implement, it is used. Method of Marquardt it is employed in the solution of nonlinear least squares problems. It converges faster than previous approaches. It can handle models with a large number of free parameters that are not exactly known. It is capable of interpolating between the Gauss Newton and gradient methods. In table 4.50, the Marquardt Method trumps the other two approaches. This approach is more resilient than the other two. The R^2 value (0.87) of the Marquardt Method is noticeably greater than that of the Gauss Newton Method and the Gradient Method. This technique accounts for 87 percent of the variation in the study variable. The Gradient Method yielded the lowest results of 0.53. In the instance of the Marquardt Method, the AIC (1227.96) and BIC (1237.91) are also low. The lower the AIC and BIC values, the better the technique.

5.4.4 Robust methods for Mitscherlich-Baule Production Function for Maize Data.

The MLE technique is used to calculate the parameters of the Mitscherlich Baule function regimes. If data or manufacturing processes follow von Liebig's law of minimum, they have enough flexibility to support restricted factor substitution: it states that growth is determined not by the total number of resources available, but by the scarcest resource (limiting factor). The least square approach used the Gauss Newton Method, whereas the Iteratively Reweighted Least Square method used the Marquardt Method and Gradient Method. Any machine learning system will benefit from the Gauss-Newton approach, which is a simple optimization procedure. There are interpolated between the gauss Newton and gradient methods. In the least quadruple curve fitting problem the Marquardt algorithm is most usually utilized. The technique of gradients is an iterative way for resolving non-linear minor square issues. The Marquardt approach in Table 4.51is better than other methods. The Marquardt method's R^2 value (0.94) is much higher than the Gauss Newton and Gradient's R^2 values. The AIC (362.96) and BIC (370.67) are also low in the case of the Marquardt Method.

5.4.5 The Best Model of Non-Linear Production Function for Simulated and Real Data for Mitscherlich-Baule Production Function

The Best Model of Non-Linear Production Function for Simulated Data $P_2$for Mitscherlich-Bauleproduction function

$$\hat{Y} = 442.23(1 - \exp(-0.07 * (7.58 + x_1{}^{*}))) * (1 - \exp(-14.60 * (72.11 + x_2{}^{NS})))$$

This production function follows von Liebig approach. Moreover, this functional form is characterized by continuously positive marginal productivities of the input factors. The coefficient describes the influence of the corresponding input factors on the study variable. Robust regression has been selected on the basis of sign, size and significance of the variable as compared to OLS. In this function Gauss newton method performs better than marquardt and gradient method on the basis of R^2value, AIC and BIC values.

The Best Model of Non-Linear Production Function for Simulated Data $P_1$for Mitscherlich-Baule production function

$$\hat{Y} = 504.30(1 - \exp(-0.07 * (4.86 + x_1{}^{NS}))) * (1 - \exp(-0.05 * (4.41 + x_2{}^{NS})))$$

This production function provides an overview of Liebig. In contrast, the marginal productivities of the variables of input are continually positive in this functional form. The coefficient describes the effect on the study variable of the respective input factors. In this function marquardt method performs better than Gauss newton and gradient method on the basis of R^2 value, AIC and BIC values. On the basis of sign, size and significance of the variable compared to OLS, a robust regression has been adopted.

The Best Model of Non-Linear Production Function for Apple Data for Mitscherlich-Baule Production Function

$$\hat{Y} = 609.12(1 - \exp(-6.75(-7.40 + x_1{}^{NS}))) * (1 - \exp(-7.64 * (15.81 + x_2{}^{*})))$$

The constantly positive marginal productivity of input elements distinguishes this functional form. The coefficient describes the influence on the variable in this study of the respective input factors. On the basis of sign, size and significance of the variable in relation to OLS the robust regression has been found. In this function marquardt method performs better than Gauss newton and gradient method on the basis of R^2 value, AIC and BIC values.

The Best Model of Non-Linear Production Function for Maize Data for Mitscherlich-Baule Production Function

$$\hat{Y} = 150.40^{*}\left(1 - \exp\left(-0.02 * (12.6 + x_1{}^{*})\right)\right) * \left(1 - \exp\left(-0.04 * (55.8 + x_2{}^{*})\right)\right)$$

In this function marquardt method performs better than Gauss newton and gradient method on the basis of R^2 value, AIC and BIC values. This function provides an instant overview of Liebig. In this functional form, however, the marginal productivities of the input variables always are positive. The coefficient expresses the impact of the various input elements on the variable in this study. A robust regression has been used to compare the sign, magnitude, and significance of the variable to OLS.

(5.5) Estimated Marginal Value Product (MVP) and Elasticity Of Linear and Non Linear Production Functions

The Marginal Value Product (MVP) and elasticity of simulated data and real data of linear and non linear production functions were found.MVP was calculated by using regression coefficients of each inputs of the functions and geometric mean value of inputs. The marginal value productivity were found to be positive in SF and MB thereby indicating that an increased used of these inputs could increase the output because these were sub optimally used. Marginal value productivity in QF and CD was found to be negative hence indicating excess use and should be avoided to check the fall of returns in production. The MVP more than one indicates that there is a chance to spend an additional cost on these factors to received additional income. Further, the elasticity was found to be greater than unity indicating that the output responds strongly to increases in the use of these inputs. Elasticity greater than unity indicating that the allocation and utilization of the variable was in the Stage I of production process. The elasticity for less than zero showed a negative decreasing function to the factors, indicating the over-utilization of the input implying that its allocation and utilization were in irrational stage of production (stage III) of the production process.

5.6 Performance Criteria of Linear and Non-Linear Production Functions for Simulated Data/Real using Cross Validation/ Technique.

Table 4.53 shows the performance for simulated data of linear and non-linear production functions using cross validation techniques. Cross-validation is to test the model's ability to predict new data that is not used in estimating it, in order to flag problems like over fitting or selection bias and to give an insight on how the model will generalize to an independent dataset. The model is best to study which has lowest values of RMSE. Lower values of RMSE indicate better fit. RMSE is a good measure of how accurately the model predicts the response, and it is the most important criterion for fit if the main purpose of the model is prediction. The RMSE is directly interpretable in terms of measurement units, and so is a better measure of goodness of fit than a correlation coefficient. One can compare the RMSE to observed variation in measurements of a typical point. Further in table 4.58 AIC, AICC, PRESS, SBC values have been found for simulated data of linear and non-linear production functions.AIC is a single number score that can be used to determine which of multiple models is most likely to be the best model. A lower AIC score is better.BIC is an estimate of a function of the posterior probability of a model being true, lower BIC means that a model is considered to be more likely to be the true model. The lowest values of PRESS indicating the best model. It has also been concluded that both the models are acting normally, because Training error is not higher than validation error. Further the performance of linear and non-linear production functions utilizing cross-validation approaches for simulated data is shown in Table4.54. Cross-validation is to test the capacity of the model to preview fresh data not used to estimate it, to indicate problems such as fitting or selection errors and to provide insights into how the model can generalize to an independent dataset. The easiest way to study the model is to have lowest RMSE values. Lower RMSE values show better efficiency. RMSE is a good indicator of how correctly the model predicts the result, and if the main objective of the model is to prediction it is the most essential fit criterion. In terms of units of measurement the RMSE is directly interpretable and therefore a better indicator of efficiency than a correlation coefficient. The RMSE can be compared to observable variations in normal point measurements. SBC values for simulated linear and non-linear production functions data were also found in table 4.58 AIC, AICC, PRESS. AIC is an individual numerical score which is most likely to be the best model of several models. Better is a lower AIC score. The BIC calculation is based on the subsequent likelihood that a model is correct; the lower BIC signifies that a model is more likely to be the real model.

5.7 Performance Criteria of Production Functions for Simulated Data/Real Data using K-Fold Cross Validation Technique.

The available learning set is partitioned into k subgroups of approximately equal size in k-fold cross-validation. The number of resulting subsets is referred to as fold in this case. This partitioning is done by taking a random sample of cases from the learning set and not replacing them. The model is trained using k -1 subsets that represent the training set as a whole. The model is then applied to the remaining subset, known as the validation set, and the performance is evaluated. This approach is repeated until all k subsets have been used as validation sets. The cross-validated performance is the average of the k performance measurements on the k validation sets. Using the k-fold cross validation technique, Table 4.55 displays the performance criteria of production functions for simulated data. K-Fold validation has been repeated up to 10 times depending on the complexity of the models under consideration. For the linear and non linear production functions, we have given the value of K for 5 fold and 10 fold. In each of these cases, 10 folded comes out on best as the R^2 value is higher in this case. In case of 5 fold cross validation, the coefficient of variation was found to be higher than in the case of 10 fold cross validation. As a result, 10 folded validation comes out to be best as compared to 5-Fold validation technique. In case of real data sets Table 4.56shows the performance criteria of production functions for simulated data using the k-fold cross validation technique. Depending on the complexity of the models under consideration, K-Fold validation has been done up to ten times. The K values for 5 fold and 10 fold linear and non linear production functions. In each of these circumstances, 10 folded is the best option because the R^2 value is higher. The coefficient of variation was found to be higher in the case of 5 fold cross validation than in the case of 10 fold cross validation. Lower the coefficient of variation better the method is. As a result, the 10-fold validation technique outperforms the 5-fold validation technique.

Chapter-6

Summary and Conclusions

The Present study was based on "Estimation and validation of linear and non linear production functions through Robust Regression". In order to estimate the production productions in case of HLP's. To summarize and concluded results for the production functions, the research utilizes robust regression methods for the estimation of the production functions of simulated data and real data sets contaminated with outliers and influential observations. It was observed the assumptions of the OLS were not achieved so robust regression model(s) have been proposed. Four functional forms of production functions commonly used linear viz the quadratic and square root, non linear viz Mitscherlich-Baule and Cobb Douglas functions. It has been seen that the major source of misleading results if the least squares criterion is used to estimate production function coefficients due to presence of HLP. It is observed that the use of robust estimation narrows the range of optimal input levels across the different functional forms. Thus, differences between functional forms are reduced by applying robust regression. Our study showed that, besides the functional form, the estimation method through OLS is decisive for production function comparisons. Therefore the properties of robust regression to ensure efficient and reliable coefficient estimation in presence of outliers might thus be particularly valuable for applications and economic assessments. Altogether, robust regression is a valuable tool for a wide range of modelling problems that require a proper representation of response functions to variable inputs. The different robust regression methods, such as M-estimation, MM-estimation, LTS estimation and S estimation were used for achieving the precise results. The techniques used namely mahalanobis distance, robust minimum covariance determinant (MCD) distance, standard robust residuals, cook's distance, studentized residual, WSSDI, hat Diagonals and deleted residuals are recommended and are effective to identify the bad leverages cum HLP. There are also HLP replaced with robust values estimated through interpolation techniques had shown the precise results. In case of linear production functions,QF outperforms SF on the basis of high R^2 value, minimum AIC and BIC while in case of non-linear production function CD and MB are usually used under different conditions so are not compared but the preference is to MB due the estimation of input variable and influenced in study variable.

In the last as per the validation techniques among linear production functions Quadratic function outperforms square root function on the basis of Minimum RMSE, maximum R^2value, and Minimum PRESS value. All models are acting normally, because training error is not higher than validation error. In K-fold validation the samples replicated more than 10 times. The value of K for 10 folded comes out to be best on the basis of minimum RMSE and maximum R^2.The marginal value productivity were found to be positive in SF and MB thereby indicating that an increased used of these inputs could increase the output because these were sub optimally used. Marginal value productivity in QF and CD was found to be negative hence indicating excess use and should be avoided to check the fall of returns in production. The MVP more than one indicates that there is a chance to spend an additional cost on these factors to received additional income vice versa. Elasticity greater than unity indicating that the allocation and utilization of the variable was in the Stage I of production process. The elasticity for less than zero showed a negative decreasing function to the factors, indicating the over-utilization of the input implying that its allocation and utilization were in irrational stage of production (stage III) of the production process.

Strength of the Thesis

I. Robust regressions are more appropriate in the presence of outliers, influential observations. It provides stable results in presence of outliers. Moreover it is observed that if the HLP are replaced by the robust values then the estimation of functions through OLS provides precise estimates at par with robust regression techniques.

II. The input variables individually contribute significantly in case of quadratic function whereas the interaction is significant in case of square root production function if data are contaminated with HLPs. Thus production functions should be used accordingly.

III. When input factors are characterized by positive marginal productivities Mitscherlich-Baule function should be considered while as in case of negative marginal productivities Cobb Douglas function is preferred.

Future Work

The role of robust methods like W estimation, L estimation R estimation and Least Median of Square (LMS) need to be studied in the presence of HLPs. The estimation of other linear and nonlinear production functions are also required when the data are contaminated by outliers.

References

Aigner, D., Lovell, C.K. and Schmidt, P. 1977. Formulation and estimation of stochastic frontier production function models. *Journal of econometrics*, ***6***(1):21-37.

Al Sayed, A.R., Isa, Z. and Kun, S.S. 2018. Outliers Detection Methods in Panel Data Regression: An Application to Environment Science. *International Journal of Ecological Economics & Statistics*, ***39***(1):73-86.

Alguraibawi, M., Midi, H. and Imon, A.H.M. 2015. A new robust diagnostic plot for classifying good and bad high leverage points in a multiple linear regression model. *Mathematical Problems in Engineering*, **29**(1):87-94.

Alma, Ö.G., 2011. Comparison of robust regression methods in linear regression. *Int. J. Contemp. Math. Sciences*, ***6***(9):409-421.

Barnston, A.G. 1992. Correspondence among the correlation, RMSE, and Heidke forecast verification measures; refinement of the Heidke score. *Weather and Forecasting*, *7*(4):699-709.

Battese, G.E. 1992. Frontier production functions and technical efficiency: a survey of empirical applications in agricultural economics. *Agricultural economics*, *7*(3):185-208.

Battese, G.E. and Coelli, T.J. 1995. A model for technical inefficiency effects in a stochastic frontier production function for panel data. *Empirical economics*, ***20***(2):325-332.

Battese, G.E., Rao, D.P. and O'donnell, C.J. 2004. A met frontier production functions for estimation of technical efficiencies and technology gaps for firms operating under different technologies. *Journal of productivity analysis*, ***21***(1):91-103.

Bayes, C. L., Bazan, J. L. and Garcia, C. 2012.A New Robust Regression Model for Proportions. *Bayesian Analysis*,**7**(4):841-866.

Belsley, D.A., Kuh, E. and Welsch, R.E. 1980. Regression diagnostics: Identifying influential data and sources of Collinearity.In. New York: John Willey & Son.

Bhatia, K., Jain, P., Kamalaruban, P. and Kar, P. 2017. Consistent Robust Regression. In *NIPS* **10**(2):2110-2119.

Bloom, D.E., Canning, D. and Sevilla, J. 2004. The effect of health on economic growth: a production function approach. *World development*, ***32***(1):1-13.

Blundell, R. and Bond, S. 2000. GMM estimation with persistent panel data: an application to production functions. *Econometric reviews*, ***19***(3):321-340.

Burnham, K.P. and Anderson, D.R. 2002. A practical information-theoretic approach. *Model selection and multimodel inference*, ***2***(3)24:28

Burton, A., Altman, D.G., Royston, P. and Holder, R.L. 2006. The design of simulation studies in medical statistics. *Statistics in medicine*, ***25***(24):4279-4292.

Cadila, J., Vltavska, K., Krejci, I., Hartman, D. and Brabec, M. 2017.Aggregate production function and income identity–empirical analysis. *International Journal of Economic Sciences*,**6**:1-17.

Caramanis, C., Mannor, S. and Xu, H. 2012.Robust optimization in machine learning. *Optimization for machine learning*, **25**:.369-375

Chaloner, K. and Brant, R. 1988. A Bayesian approach to outlier detection and residual analysis. *Biometrika,* **75**(4):651-659.

Chatterjee, S. and Mächler, M. 1997. Robust regression: A weighted least squares approach. *Communications in Statistics-Theory and Methods*, ***26***(6):1381-1394.

Chaudhuri, A. and Basu, A. 2017.Robust Regression by the Self Updating Process. Technical Report No.Isru/2017/1 ISI

Chen, C. 2002. Robust regression and outlier detection with the robustreg procedure. In *Proceedings of the Proceedings of the Twenty-Seventh Annual SAS Users Group International Conference*.**2**:265-27

Chen, C., NC, C. 1987.Robust regression and outlier detection with the ROBUSTREG procedure. *Statistics and Data Analysis,* **3***:*1-13

Chen, H., Meer, P. and Tyler, D.E. 2001. December. Robust regression for data with multiple structures. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001* (Vol. 1, pp. I-I). IEEE.

Chen, J.H., Chang, Y.C. and Hung, W.L. 2020. A self-organizing clustering algorithm for functional data. *Communications in Statistics-Simulation and Computation*, **49**(5):1237-1263.

Choy, B., Wan, W.Y. and Chan, C.M. 2009. Bayesian student-t stochastic volatility models via scale mixtures. *Wai Yin and Chan, Chun Man, Bayesian Student-T Stochastic Volatility Models Via Scale Mixtures*.

Christensen, B.J., Nielsen, M.Ø. and Zhu, J. 2015. The impact of financial crises on the risk–return tradeoff and the leverage effect. *Economic Modelling*, **49**:407-418.

Cobb, C.W. and Douglas, P.H. 1928. A theory of production. *The American Economic Review*, ***18***(1):139-165.

Cook, R.D. 1977.Detection of influential observation in linear regression. *Technometrics*, ***19***(1):15-18.

Cook, R.D. 1979. Influential observations in linear regression. *Journal of the American Statistical Association*, ***74***(365):169-174.

Corsi, F. and Reno, R. (2009).HAR volatility modelling with heterogeneous leverage and jumps. *Econometric Reviews,* **27**:1–31

Croux, C. and Haesbroeck, G. 2000. Principal component analysis based on robust estimators of the covariance or correlation matrix: influence functions and efficiencies. *Biometrika*, **87**(3): 603-618.

Cary, NC.2013.Current Directions in SAS/STAT Software Development. In *Proceedings of the SAS Global Forum 2013 Conference. Cary, NC: SAS Institute Inc. http://support. sas. Com/resources/papers/proceedings13/432-2013.*

Cuthbert, D., Wood, F.S. and Gorman, J.W., 1980. Fitting equations to data: Computer analysis of multifactor data. *New York: Wiley*.

dan Susanti, Y.Y., 2008. Estimasi-M dan Sifat-sifatnya pada Regresi Linear Robust. In *Jurnal Math-Info* (Vol. 50, No. 10). UNS.

Dempster, A.P., Laird, N.M. and Rubin, D.B. 1980.Iteratively Reweighted Least Squares for Linear Regression When Errors are Normal/Independent Distributed. *In Multivariate Analysis,* **5**:35-57.

Dempster, A.P., Laird, N.M. and Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, *39*(1), pp.1-22.

Deyoreo, M. and Kottas, A., 2018. Bayesian nonparametric modeling for multivariate ordinal regression. *Journal of Computational and Graphical Statistics*, *27*(1), pp.71-84.

Diskin, T., Draskovic, G., Pascal, F. and Wiesel, A. (2017).Deep Robust Regression.IEEE 7th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP).

Donoho, D.L. and Huber, P.J., 1983. The notion of breakdown point. A Festschrift for Erich L. Lehmann. *PJ Bickel, KA Doksum, and JL Hodges, Eds*, pp.157-184.

DuMouchel, W. H. 1983. Estimating the stable index $\alpha $ in order to measure tail thickness: A critique. *The Annals of Statistics*, *11*(4), 1019-1031.

El Karoui, N., Bean, D., Bickel, P.J., Lim, C. and Yu, B., 2013. On robust regression with high-dimensional predictors. *Proceedings of the National Academy of Sciences*, *110*(36), pp.14557-14562.

El Karoui, N., Bean, D., Bickel, P.J., Lim, C. and Yu, B., 2013. On robust regression with high-dimensional predictors. *Proceedings of the National Academy of Sciences*, **110**(36):14557-14562.

Ellisi, R. P. and Mookim, P. G. 2013.K-Fold Cross-Validation is Superior to Split Sample Validation for Risk Adjustment Models: 1-23.

Epple, D., Gordon, B. and Sieg, H. 2010.A new approach to estimating the production function for housing. *American economic review,* **100**: 905-924

Ferrari, S. and Neto, F. C. 2010.Beta Regression for Modelling Rates and Proportions. *Journal of Applied Statistics*.**31**(7): 799–815

Finger, R. and Hediger, W. 2008.An application of robust regression to a production function comparison. *The Open Agriculture Journal*, **2**:90-98

Fischler, M.A. and Bolles, R.C., 1981. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, *24*(6), pp.381-395.

Førsund, F.R., Lovell, C.K. and Schmidt, P. 1980. A survey of frontier production functions and of their relationship to efficiency measurement. *Journal of econometrics*, ***13***(1):.5-25.

Fox, J. and Weisberg, S. 2011. Multivariate linear models in R. *An R Companion to Applied Regression. Los Angeles: Thousand Oaks*.

Frank, E., Harrell, J. and Kerry L. Lee, K. L. 1984.Regression Modelling Strategies for Improved Prognostic Prediction. *Statistics In Medicine*, **3**(2): 143-152

Fung, W.K. 1993 unmasking outliers and leverage points: A confirmation. *Journal of the American statistical association,* **88**:515-519

Gauss, C. F. (1809). Theoria Motus Corporum Coelestium. Perthes, Hamburg. Translation reprinted as Theory of the Motions of the Heavenly Bodies Moving about the Sun in Conic Sections. Dover, New York, 1963.

Griffin, J. E. and. Brown, P. J. 2013.Some Priors for Sparse Regression Modelling. *Bayesian Analysis*, **8**(3):691-702.

Guerard, J.B., 2016. Investing in global markets: big data and applications of robust regression. *Frontiers in Applied Mathematics and Statistics*, **1**(14):1-16

Habshah, M., Norazan, M.R. and Rahmatullah Imon, A.H.M., 2009. The performance of diagnostic-robust generalized potentials for the identification of multiple high leverage points in linear regression. *Journal of Applied Statistics*, ***36***(5):507-520.

Hadi, A.S. and Simonoff, J.S. 1993. Procedures for the identification of multiple outliers in linear models. *Journal of the American statistical association*, ***88***(424):1264-1272.

Hadi, A.S., 1992. A new measure of overall potential influence in linear regression. *Computational Statistics & Data Analysis*, ***14***(1):1-27.

Hampel, F. R. 1985.The Breakdown Points of the Mean Combined With Some Rejection Rules. *Technometrics*, **27**(2):95-107

Harrell Jr, F.E., Lee, K.L., Califf, R.M., Pryor, D.B. and Rosati, R.A. 1984. Regression modelling strategies for improved prognostic prediction. *Statistics in medicine*, ***3***(2):143-152.

Hauck W. W and Anderson, S.1984.A survey regarding the reporting of simulation studies. *The American Statistician,* **38**(3):214–216.

Hoaglin, D. C. and Andrews, D. F. 1975.The reporting of computation-based results in statistics. *The American Statistician,* **29**(3):122–126.

Hoaglin, D.C. and Welsch, R.E. 1978.The Hat matrix In Regression and ANNOVA.*The American Statistician,* **32**:17-22.

Hogg, R.V. 1974. Adaptive Robust Procedure: A Partial Review and Some Suggestion for Future Applications and Theory. *Journal of the American Statistical Association*, **69**:909-923.

Holland, P. W. 1973.Weighted Ridge Regression: Combining Ridge and Robust Regression Methods. Working Paper No.11

Hoseemer, D.W. and Lemeshow, S. 1980.Applied logistic regression Wiley, New York 3rd ed.

Huang, G., Huang, G. B., Song, S. and You K. 2015. Trends in extreme learning machines: A review. *Neural Networks,* **61**: 32-48.

Huang,D., Cabral, R. S. and Torre, F. D. 2012.Robust Regression.*Springer*,**4**(5): 616–630

Huber, P.J. 1973.Robust Regression: Asymptotics, Conjectures and Monte Carlo.*In The Annals of Statistics*, **1**:799-821.

Hubert, M. and Rousseeuw, P. J. 1996.Robust regression with a categorical covariable. In: H. Rieder, Ed., *Robust statistics, Data analysis, and Computer Intensive methods*, Springer, New York.

Hubert, M.and Branden, K. V. 2003. Robust Methods for Partial Least Squares Regression. *Journal of Chemometrics,* **17**(10):537-549

Huggins, R. M. 1993.On The Robust Analysis of Variance Components Models for Pedigree Data.*Australian Journal of Statistic,* **35**(1):43-57

Imon, A.R. and Apu , M.R. 2016. Detection of High Leverage Points Using a Nonparametric Cut-off Point for the Robust Mahalanobis Distance. *Malaysian Journal of Mathematical Sciences*, ***10***(3):283-295.

Imon, A.R. and Hadi, A.S. 2013. Identification of multiple high leverage points in logistic regression. *Journal of Applied Statistics*, ***40***(12):2601-2616.

Kass, R.E. and Wasserman, L. 1995. A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *Journal of the American statistical association*, ***90***(431):928-934.

Kia, M., Banazadeh, M. and Bayat, M. 2018. Rapid seismic vulnerability assessment by new regression-based demand and collapse models for steel moment frames. *Earthq Struct*, ***14***(3):203-214.

Koehler, E.and Haneuse, B. E. 2009. On the assessment of Monte Carlo error in simulation- based statistical analyses. *The American Statistician,* **63**:155–162.

Koutsias, N. and Karteris, M. 1998.Logistic regression modelling of multitemporal Thematic Mapper data for burned area mapping. *International journal on remote sensing*, **19**(18): 3499- 3514

Lange, K. and Sinsheimer, J.S. 1993.Normal/Independent Distribution and Their Application in Robust Regression.*Journal of Computational and Graphic Statistics*.**2**:175-198.

Legendre, A.M. 1806. *Nouvelles méthodes pour la détermination des orbites des comètes; par AM Legendre...* chez Firmin Didot, libraire pour lew mathematiques, la marine, l'architecture, et les editions stereotypes, rue de Thionville.

Levin, D., 1998. The approximation power of moving least-squares. *Mathematics of computation*, **67**(224).1517-1531.

Levinsohn, J. and Petrin, A. 2003.Estimating Production Functions Using Inputs to Control for Unobservable .*Review Of Economic Studies*, **70**:317-342.

Lewis and Ward 2013. Improving epidemiologic data analyses through multivariate regression modelling. *Emerging Themes in Epidemiology*, **4** :1-10

Li, C. H. 2012. Validation of the Chinese version of the Life Orientation Test with a robust weighted least squares approach. *Psychological Assessment,* **24**:770–776

Marona., Martin, R. andYohai, V. J. 2006 .Robust Statistics Theory and Methods, JohnWiley & Sons Ltd., England.

Marubini, E. and Orenti, A. 2014.Detecting outliers and/or leverages points: a robust two stage procedure with bootstrap cut-off points. *Epidemiology biostatistics and public health,* **11**:1-17.

Midi and Mohammed 2015. The identification of good and bad high leverage points in multiple linear regression model. *Mathematical Methods and Systems in Science and Engineering,***13**:147-153

Mitra, K., Veeraraghavan, A. and Chellappa, R. 2010. March. Robust regression using sparse learning for high dimensional parameter estimation problems IEEE International Conference on Acoustics, Speech and Signal Processing **2**: 3846-3849. IEEE.

Montgomery, D.C., Peck, E.A. and Vining, G.G. 2006. Introduction to Linear Regression Analysis. John Wiley & Sons. *Inc. Hoboken, New Jersey*.

Morris, T. P., WhiteI. R. and Crowther, M. J. 2017.Using Simulation Studies to Evaluate Statistical Methods.*Tutorial in biostatistics,* **21**(2):1-28

Motulsky, H. J. and Brown, R. E. 2006.Detecting outliers when fitting data with nonlinear regression – a new method based on robust nonlinear regression and the false discovery rate.*BMC Bioinformatics,***7**:1-20

Nakaya, T., Fotheringham, A.S., Brunsdon, C. and Charlton, M. 2005. Geographically weighted Poisson regression for disease association mapping. *Statistics in medicine*, **24**(17):2695-2717.

Nguyen, T. and Welsch, R. E. 2010.Outlier detection and robust covariance estimation using mathematical programming.*Adv Data Anal Classif*, **4**:301–334

Nobre and Singer 2010.Leverage analysis for linear mixed models. *Journal of Applied Statistics,* **38***:* 1063-1072

Norazan.,Sanizah, A. and Habshah, M. 2012.Identifying bad leverage points in logistic regression model based on robust regression. *Mathematical Models and Methods in Modern Science,***14**:62-67.

Nurunnabi, A.A.M., Nasser, M. and Imon, A.H.M.R. 2016. Identification and classification of multiple outliers, high leverage points and influential observations in linear regression. *Journal of Applied Statistics*, ***43***(3): 509-525.

O'Leary, D.P. 1990. Robust regression computation using iteratively reweighted least squares. *SIAM Journal on Matrix Analysis and Applications*, ***11***(3), pp.466-480.

Olley, S. and Pakes, A. 1992. The dynamics of productivity in the telecommunications equipment industry.*Econometrica,* **64**:1263-1297.

Ostertagová, E. 2012. Modelling using polynomial regression. *Procedia Engineering*, ***48***:500-506.

Petrin, A., Poi, B.P. and Levinsohn, J. 2004. Production function estimation in Stata using inputs to control for unobservable. *The Stata Journal*, ***4***(2): 113-123.

Pollak, R.A. and Wachter, M.L. 1975. The relevance of the household production function and its implications for the allocation of time. *Journal of Political Economy*, ***83***(2): 255-278.

Preminger, A. and Franck, R. 2007 forecasting exchange rates: A robust regression approach. *International Journal of Forecasting,***23**:71–84

Ramalho, J.J. and Silva, J.V. 2011. Functional form issues in the regression analysis of financial leverage ratios. *Empirical Economics*, ***44***(2):799-831.

Rey, W.J. 2012. Introduction to robust and quasi-robust statistical methods. Springer Science & Business Media.

Ripley, B.D. 2009. Stochastic simulation (Vol. 316). John Wiley & Sons.

Ronchetti, E. 1979. "Robust heitseigenschatlen von Tests," Diploma thesis, ETH, Zurich

Ronchetti, E., Field, C. and Blanchard, W. 1997.Robust linear model selection by cross-validation. *Journal of the American Statistical Association,* **92**(439):1017-1023.

Rousseeuw, P. J. and Leroy, A. M. 1987.Robust Regression and Outlier Detection .Wiley Series in Probability and Statistics, Wiley &Sons, New York, NY, USA.

Rousseeuw, P.J. and Yohai, V. 1984.Robust Regression by Means of S Estimators in Robust and Nonlinear Time SeriesAnalysis. *Lecture Notes in Statistics*, **26**: 256-274

Rousseeuw, P.J., Debruyne, M., Engelen, S. and Hubert, M. 2006. Robustness and outlier detection in chemometrics. *Critical reviews in analytical chemistry*, ***36***(3-4):221-242.

Salibian-Barrera, M. and Zamar, R.H. 2002. Bootstrapping robust estimates of regression. *Annals of Statistics***5**:556-582.

Schaul, T., Zhang, S., & LeCun, Y. 2013. No more pesky learning rates. In *International Conference on Machine Learning* **22**: 343-351.

Schmidt, P. 1976.On the Statistical Estimation of Parametric Frontier Production Functions .*The Review of Economics and Statistics.***4**:238-239.

Schulz, E., Brink, M. S. and Krause, A. 2018. A tutorial on Gaussian process regression: Modelling, exploring, and exploiting functions. *Journal of Mathematical Psychology,* **85**:1-16

Simon, H.A., 2005. Darwinism, altruism and economics. *The Evolutionary Foundation of Economics* **2**(5*)*89-104.

Stewart, C. V. 1999. Robust parameter estimation in computer vision. *SIAM review*, ***41***(3), 513-537.

Sturm, J.E. and De Haan, J. (2001).How Robust is the Relationship between Economic Freedom and Economic Growth. *In Applied Economics,* **33**:834-44.

Susanti, Y. and Pratiwi, H. 2014. M estimation, S estimation, and MM estimation in robust regression. *International Journal of Pure and Applied Mathematics*, ***91***(3):349-360.

Susanti, Y., Pratiwi, H. and Liana, T. 2009. Application of M-estimation to Predict Paddy Production in Indonesia, presented at IndoMS InternationalConference on Mathematics and Its Applications (IICMA), Yogyakarta, Todd, P. E. and Wolpin, K. I. 2003.On the specification and estimation of the production function for cognitive achievement. *The Economic Journal*, **113**: 3-33

Tukey, J. 1960 .Contributions to Probability and Statistics, chapter A Survey of Sampling From Contaminated Distributions.Standford University Press: Stanford, California.

Vankeerberghen, P., Smeyers-Verbeke, J., Leardi, R., Karr, C.L. and Massart, D.L. 1995. Robust regression and outlier detection for non-linear models using genetic algorithms. *Chemometrics and Intelligent Laboratory Systems*, ***28***(1), pp.73-87.

Verardi, V. and Croux, C. 2008.Robust Regression in Stata.*Stata Journal*, **9**(3):439-453

Wager, T.D., Keller, M.C., Lacey, S.C. and Jonides, J. 2005. Increased sensitivity in neuroimaging analyses using robust regression. *Neuroimage*, ***26***(1):99-113.

Wang, J.J., Chan, J.S. and Choy, S.B. 2011. Stochastic volatility models with leverage and heavy-tailed distributions: A Bayesian approach using scale mixtures. *Computational Statistics & Data Analysis*, ***55***(1):852-862.

Weisberg, S. 2005. *Applied linear regression* (Vol. 528). John Wiley & Sons.

Wood, F.S., 1980. Linwood and nonlinwood-Linear and Nonlinear Least Squares Curve-Fitting Programs. *The American Statistician*, **34**(3):177-179.

Wooldridge, J. M. 1990.A Unified Approach to Robust, Regression-Based Specification Tests. *Econometric Theory,***6**(1): 17-43

Xu, H., Caramanis, C. and Mannor, S., 2008. Robust regression and lasso. *arXiv preprint arXiv:0811.1790*.

Yaffee, R. A. 2002.Robust Regression Analysis: Some Popular Statistical Package Options", Statistics, Social Science, and Mapping Group Academic Computing Services Information Technology Services.

Yang, T., Gallagher, C.M. and McMahan, C.S. 2019. A robust regression methodology via M-estimation. *Communications in Statistics-Theory and Methods*, ***48***(5):1092-1107.

Yang, Y. and He, X. 2012. Bayesian empirical likelihood for quantile regression. *The Annals of Statistics*, ***40***(2)1102-1131.

Yasar, H., Soyhan, H.S., Walmsley, H., Head, B. and Sorusbay, C. 2008. Double-Wiebe function: An approach for single-zone HCCI engine modeling. *Applied Thermal Engineering*, ***28***(11-12)1284-1290.

Yasar, M., Raciborski, R. and Poi, B. 2008. Production function estimation in Stata using the Olley and Pakes method. *The Stata Journal*, ***8***(2):221-231.

Yohai, V.J., 1987. High breakdown-point and high efficiency robust estimates for regression. *The Annals of statistics*,**2**(1):642-656.

Yu, C. and Yao, W. 2017. Robust linear regression: A review and comparison. *Communications in Statistics-Simulation and Computation*, ***46***(8):6261-6282.

Yuan, K.H. and Zhong, X. 2008.outliers, leverage observations, and influential cases in factor analysis: Using robust procedures to minimize their effect. *Sociological Methodology*, ***38***(1):329-368.

Zaman, A., Rousseeuw, P.J. and Orhan, M. 2001. Econometric applications of high-breakdown robust regression techniques. *Economics Letters*, ***71***(1):1-8.

Zhang, Y., Shi, D., Gao, J. and Cheng, D. 2017. Low-rank-sparse subspace representation for robust regression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* **2**(3):7445-7454.

VITA

Name of Student	:	Rizwan Yousuf
Father's Name	:	Mohd Yousuf War
Mother's Name	:	Raja Bano
Nationality	:	Indian
Date of birth	:	05 December,1994
Mobile	:	7006787919,9858236769
Email	:	rizwanwar50@gmail.com
Permanent home address	:	House No: 125, Peth-Kanihama Budgam Kashmir- 193401, (J&K)

EDUCATIONAL QUALIFICATION

Bachelor Degree	:	B.Sc.
University and year of award	:	University of Kashmir-(2015)
Percentage of marks (OGPA)	:	59.4%
Master's degree	:	M.Sc. Statistics
University and year of award	:	University of Kashmir-(2018
Percentage of marks (OGPA)	:	7.01/10.0
Ph. D Degree	:	Ph.D Statistics
University and year of award	:	SKUAST-Jammu (2021)
OGPA	:	8.33/10.0
Title of Ph.D. thesis	:	**Estimation and Validation of Linear and Non-Linear Production Function through Robust Regression**

CERTIFICATE – IV

Certified that all the necessary corrections as suggested by the external examiner and the advisory committee have been duly incorporated in the thesis entitled "**Estimation and Validation of Linear and Non-Linear Production Functions through Robust Regression**" submitted by **Mr. Rizwan Yousuf** Registration No. **J-17-D-24-BS**

24.11.21

Dr. Manish Kumar
Major Advisor and Chairman
Advisory Committee

Place: Jammu

Date..24-11-2021

Head of the Division 24.11.2021