CHAPTER 1
INTRODUCTION

1.1 Introduction

Statistical models are useful in predicting the yield of agricultural crops. Several models have been developed. Across use, specific role of a given model would be valid if it fits to the given niche and environmental context. This study aims to investigate linear regression model for rice production forecasting, to be used and applicable in Nepal. According to Ristanoski *et al*. (2013) amongst the wealth of available machine learning algorithms for forecasting, time series linear regression is one of the most important and widely used methods. This method is simple to construct easy to use and interpret. Indeed, Ristanoski *et al*. (2013) have concluded that time series forecasting was a classic prediction problem until logically used model developed and for that reason linear regression model has been one of the best known and most widely used methods.

Several models are in practice to estimate or predict yield of the given crop. Zhang *et al.* (2004) have used univariate linear regression to fit and forecast global and regional trends of rice for seven time series (1961-2000) variables; rough rice area, rough rice production, and rough rice yield. In addition to this the remaining four variables are rice insecticide sales, rice fungicide sales), rice herbicide sales and rice pesticide sales. In the case of Nepal, Nayava (2012) has used the time series data (1971 - 2000) of rainfall and production of rice and has argued that, impact of rainfall on rice yield and production was quite evident. However the author had simply used time series plots to draw such inferences. Similarly, Dahal and Routray, (2011) have developed multiple regression model to evaluate apparent strength of the relationship and to explain the variations on crop yield against the soil variables. In both of the cases the equations are limited up to explaining the phenomenon, but they have not been investigated for making forecasts.

In this context this research was conducted aiming to move forward from simply limiting the model to explain the phenomenon but rather to investigate a linear regression model for rice production forecasting in Nepal. Specifically a multiple linear regression model was optimized and tested for its reliability and efficiency for forecasting the production. There are multiple benefits of such kind of research whereas this research also has opened the door to further investigate similar mathematical models and or more

advanced tools for rice production forecasting. Such models can then be compared with the model we have investigated for their performance and efficiency in forecasting.

After all, this research has developed a multiple linear regression model for rice production forecasting in Nepal.

1.2 Objectives

The objective of this study is to develop and validate a multiple regression model for rice production forecasting in Nepal, and to compare this model with naïve forecast model; one of the common benchmark for model comparison. The specific objectives of the study are as follows:

- To understand the factors that are more relevant to contribute for rice production forecasting in Nepal
- To add turning point in the forecasting education in the country through the application of the theory of regression model building in available data
- To put forward a mathematical model of crop production forecasting as an alternative to so far in use methods of forecasting production such as: Crop cutting experiments and eye estimation methods

1.3 Rationale

Forecasting is designed to help decision making and planning in the present (Walonick, 1993). Forecasts are important to the planners and the policymakers. Through forecasts, one can modify variables now to make the future in accordance to ones need or wish. In this context to better manage the food security situation in the country, Nepal needs correct forecasts for its crop production. Accordingly, this study has a set up to investigate and establish a multiple regression model for rice production forecasting in Nepal. So far for crop production forecasting, especially for rice, Nepal is relying on rather primitive type of forecasting methods like, crop cutting experiments and eye estimation methods. This study aims to have a greater shift in this regards, i.e. making a giant leap towards mathematical models for crop production forecasting instead ancient subjective methods.

Regression is one of the widely used methods for forecasting. According to Hall (2015) regression and forecasting techniques yield new insight for managers by uncovering patterns and relationships that they had not previously noticed or considered. Regression analysis studies the variables individually and determines their significance with greater accuracy. Any complex questions can be easily answered through these techniques. As compared to the other sophisticated methods of forecasting, such as NN (Neural Network), ANN (Artificial Neural Network) etc. regression is a simple model that has power to produce more accurate results. More importantly, "Unlike the other statistical tools, regression analysis takes into account the risks of making assumptions and easily addresses the most complicated of problems due to its flexibility (Richa, 2015)."

Once investigated and established, the planners will have a forecast model at their hand to use it in real life, which in the future should bring further advancement both in the establishing optimized regression models for crop production forecasting and in aiding to comprehend the forecast education learning environment in the country.

1.4 Research questions

The following are the research questions

- What are the factors that are more relevant to contribute for rice production forecasting in Nepal?
- With time series data that could have higher degree of applicability, what are the theories behind linear regression model building for forecasting?
- What is variable selection and the model selection approaches? What different criterions are applicable for these procedures?
- What are forecast errors and how are forecast accuracy tested?
- Can the investigated regression models, which are based on time series data in the given context, accurately forecast annual rice production? And,
- Can linear regression models be sufficiently used for rice production forecasting in Nepal instead of so far in use forecasting methods such as crop cutting experiments and the eye estimation methods?

1.5 Scope limitations and assumptions

The potential use of model building with regression and other techniques is almost limitless. For this reason the study is limited up to the boundaries of multiple linear regression. This allowed the researcher to dig into the depth of this part and do every bits and pieces possible to optimize a regression model by selecting some key variables as the predictors for rice production forecasting.

The single and the most daunting limitation in this study was the availability and the reliability of data. However, best model fit has been obtained with what data has been available. It would have been better if there was a chance to enlarge the sample size than what has been considered.

1.6 Methodology

In the book 'Forecasting Methods and Principles' Makridakis *et al.* (1998) detail description and the methodology to develop regression models to accurately forecast with time series data have been explained. Similarly Hyndman and Athanasopulos (2014), Bowerman, *et al.* (2005) have also described in a specific way detailing regression model building process for forecasting and testing the validity of the so developed regression model.

To cope up with the thesis heading" Optimizing regression model for rice production forecasting in Nepal", every principles and methods developed so far for regression model optimization is employed in the study. Starting with the procedure of data collection, data refinement and data management this study has moved forward through variable selection to model selection approaches and finally has developed the model. After words the model so developed was examined using various forecast accuracy methods. For instance, Mean Error (ME), Mean Absolute Error (MEA), Mean Percentage Absolute Error (MAPE) and, the Tracking Signal (TS) to check on the model if there could be any update and or some appropriate change in the developed model as the final.

The methodology used while conducting the research is described in the followings.

All possible and available literatures were extensively reviewed including books, periodicals, journals, and also the internet sites. Accordingly data were downloaded from the websites of International Rice Research Institute (IRRI), Ministry of Agriculture (MOA), Food and Agriculture Organization (FAO), and from the office of the Department of Hydrology and Metrology (DHM). Afterward the data were organized and processed for analysis and then various methods were employed to the data for the selection of the most prominent variables. For instance, past research experience and experts views were considered; automated procedure (s) for example family of stepwise methods and the best subset regression method to select the most prominent variables for the analysis were employed. Because the model was to be cross validated with a different sample than with what the model was built for testing its reliability, the whole set of the data were separated into two samples 1st: training sample (first 35 data points) and the

test sample (last 20 data points). Using the training sample and applying every theory behind regression model building a multiple regression equation was investigated which after words, was cross validated taking the help of the test sample. Software used for the whole lot of the study included, SPSS (Vs. 20), Minitab (VS. 16), Stata (Vs. 12) and Microsoft Office Excel 2007.

Conclusively, this study was based on the collective methods of multiple regression model building for forecasting, distinctly partitioned into four apparent stages. Namely, Variable selection, establishing and validating of the regression model, testing the model's performance through the eye of forecast accuracy constraints such as smaller errors and superiority over a bench mark model and finally discussing of the models goodness of fit and its applicability and implications when used in real life situation.

1.7 Organization of the study

This thesis is comprised of six chapters. The first chapter is about introduction. The second chapter covered all available literature review and the theoretical frameworks for the study. In this chapter historical development of regression methods is described. Afterward, the theory of regression model with time series data for forecasting is discussed in detail. The third chapter is for materials and methods. This chapter explains essential materials and the methods for the study. For instance, software used in the study , methods of data collection, variable selection methods, model selection approaches including all algorithms for regression model building for forecasting are put in here. The fourth chapter is for the result section. This covered pertinent results, tables and charts constructed as outputs of the study. Chapter five is about discussion and conclusion. In this chapter results, tables, charts etc. obtained in chapter four are discussed in detail and conclusions are drawn to its end. And at the end the thesis was closed with chapter six, which covered the brief summary of the study. Finally the thesis was closed with chapter six, which covered the brief summary of the study.

CHAPTER 2
LITERATURE REVIEW AND THEORETICAL FRAMEWORK

2.1 Literature review

This study is focused to investigate a linear regression model for crop production forecasting. Accordingly, pertinent literature have been sought in between the historical development of regression as statistical methods and the development of regression methods in its own up to the application of regression methods for forecasting. This section of the thesis has covered the review of the available literature within the mentioned regression environment.

2.1.1Statistics and regression

In simple sense statistical methods are for collecting, summarizing, analyzing and interpreting data generated for a variable to draw meaningful inferences. Definition of statistics may vary from person to person however the central theme for everyone remains the same. Delorme (2006) defines statistics as "the body of analytical and computational methods by which characteristics of a population are inferred through observations made in a representative sample from that population" (para.1). Statistical methods are used for diverse fields and purposes. For instance, statistical methods are used in economics, agriculture, health sciences, forecasting are some examples. So far it is believed that statistics was originated due to a breakthrough in game of chance in the early 18th century. According to Aldrich (2000) the origin of probability and statistics were found during 1650-1700. But Denis (2000) reported that Galton's discovery of Regression and Correlation technique is to be considered for the origin of the subject. Legendre (1805) and Guass (1809) (as cited in Wikipedia, 2013) invented the method of least squares, the earliest form of regression. Also Wikipedia (2013) reports that later in 19th century the term 'regression' was coined by Francis Galton while describing the biological phenomenon. A similar discussion on the subject by Galton (1886) is given below:

> **It appears that Sir Francis Galton (1822-1911), a well known British anthropologist and meteorologist, was responsible for the introduction of the word "regression." Originally he used the term "reversion" in an unpublished address "Typical laws of heredity in man" to the Royal institution on February 9, 1877. The later term "regression" appears in his Presidential address made before section H of the British Association at Aberdeen, 1885, printed in *Nature*, September 1885, pp.507-510 and also in a paper "Regression towards mediocrity in hereditary stature".**

To continue on the debate about the origin of statistics, Carter's (Rice University, 1995-test book on linear algebra) (as cited in Talk Stat, 2013) explanation that Gauss developed least square regression supports Legendre's idea of the invention of the method of least square, however he [Carter] reveals that Gauss did not publish the method until 1809.

While circling on the origin of statistics as such, comparably a robust study carried out by Fenberg (1992) has put forward more convincing fact for this. In the epilogue of his work Fenberg reveals that invention of probability should not be taken as the invention of statistics. It is the mathematical part but not a statistical method. His idea is Gauss Laplace-synthesis, which combined the normal error theory with the curve fitting method of least square. This was an inferential approach to the analysis of data using linear models, the first and the foremost event invented in the history of statistics. So it was the method of least square which seeded out statistics and Gauss is the one who should be credited for this. Interestingly then, it was method of least square which originated first, and later in the 19^{th} century came out to be the method of regression in the history gave birth to the popular subject statistics. About the origin of regression technique Armstrong (2012) claims:

> **Regression analysis entered the social sciences in the 1870s with the pioneering work by Francis Galton. But "least squares" goes back at least to the early 1800s and the German mathematician Karl Gauss, who used the technique to predict astronomical phenomena.**

And Stanton (2001) states that Galton's work on inherited characteristics of sweet peas led to the initial conceptualization of linear regression or it was the imagination of Sir Francis Galton that originally conceived modern notions of correlation and regression.

As comprehension of the above discussion, we can say that, despite the popular belief so far, that statistics was originated from the game of chance, what else has to be most likely is: method of least square popped out at the beginning in 1800 and then was the regression methods invented from it. And after all, it was the regression methods which should have given the birth to the subject statistics rather than it can be revealed to start around 1749 as reported in the history of statistics can be said to start around 1749 as reported in the history of statistics-Wikipedia, the free encyclopedia.

2.1.2 Regression methods

Regression method was originated from methods of least squares. Francis Galton initially described biological phenomenon using regression technique. Stanton (2001) claims that Galton's work on inherited characteristics of sweet peas led to the initial conceptualization of linear regression and further work of Galton and Pearson brought other sophisticated statistical methods like, multiple regression and product-moment correlation coefficient.

Regression models are the relationship between one or more response variables. The most elementary type of regression is the simple linear regression. A linear regression is one in which all the parameters appear linearly (Prajneshu, n.d.). Simple linear regression refers to a regression on two variables. Regression which refers to two or more variables is multiple regression. It is another form of linear regression. Makridakis *et al.* (1998) reported that simple regression is a special case of multiple regression. According to Hastie *et al.* (2009) linear models were largely developed in the pre-computer age of statistics, but even in today's computer era there are still good reasons to study and use them since they are the foundation of more advanced methods.

According to Data Science Central (2015) linear regression is the oldest type of regression, designed 250 years ago for the computations on small data. This type of regression can be used for interpolation, but not suitable for predictive analytics; has many drawbacks when applied to modern data e.g. sensitivity to both outliers and cross-correlations (both in the variable and observation domains), and subject to over-fitting. A better solution is piecewise-linear regression, in particular for time series.

The other major type of regression is nonlinear regression. In non linear regression at least one parameters in the relationship appears nonlinearly. Non linear models are sometimes called 'intrinsically linear' meaning that these models can be transformed to linear relation by means of some mathematical transformation.

Regression is not limited to this. It now has evolved itself as a gigantic subject. Many more advanced regression techniques have been invented and applied for different problem solving. To shed light on some other regression methods Flizmoser (2008); Darper and Smith (1998); and Data Science Central (2015) can be summarized as the followings:

Principle Component Regression (PCR)

This method looks for transformations of the original data into a new set of uncorrelated variables called principal components. This transformation ranks the new variables according to their importance, and eliminates those of least importance. Then a least squares regression on the reduced set of principal components is performed.

Partial Least Squares (PLS) regression

This technique also constructs a set of linear combinations of the inputs for regression. But unlike principal components regression it uses y in addition to x for this construction.

Shrinkage Methods

These methods keep all variables in the model and assign different weights to obtain a smoother procedure with a smaller variability.

Ridge Regression

This procedure is intended to overcome "ill conditioned" situations. Ridge regression is a way of preceding that adds specific additional information to the problem to remove the ill conditioning. Generalized Linear Models (GLM) analysis comes into play when the error distribution, is not normal. GLM analysis provides a larger framework for estimation, which includes the cases of normally distributed errors.

Linear methods for classification

It is assumed that the K different classes exist, and that the class membership is known for the training data. The task then is to establish a classification rule that allows a reliable assignment of the test data to the classes. Linear classification tries to find linear functions of the form (1) which separate the observations into the different classes.

Robust regression

For many sets of data the departures, if they exist at all, are not serious enough for corrective actions, and we proceed with the analysis in the usual way. A least square analysis weights each observation equally in getting parameter estimates. Robust methods enable the observation to be weighted unequally.

Logistic regression

Logistic regression deals with the problem of classifying observations that originate from two or more groups. It is heavily used in clinical trials especially when the response is binary. It suffers same drawbacks as linear regression and computing regression coefficients is rather complex. However, this can be well approximated by linear regression after transforming the response (logit transform).

Ridge regression

A more robust version of linear regression, putting constraints on regression coefficients to make them much more natural, less subject to over-fitting, and easier to interpret.

Lasso regression

This is similar to ridge regression, but automatically performs variable reduction (allowing regression coefficients to be zero).

Ecologic regression

This method performs one regression per strata, if data is segmented into several large strata or groups.

Logic regression

This is used when all variables are binary, typically in scoring algorithms. It is a specialized, more robust form of logistic regression, where all variables have been binned into binary variables.

Bayesian regression

Similar to ridge regression, this method is more flexible and stable than conventional linear regression. It assumes some prior knowledge about the regression coefficients and the error term. However, in practice, the prior knowledge is translated into artificial priors - a weakness of this technique.

Quantile regression

This regression is used in connection with extreme events.

Generalized Additive Models (GAM)

In these methods, the weighted sum of the regressor variables are replaced by a weighted sum of transformed regressor variables. In order to achieve more flexibility, the relations between y and x are modeled in a non-parametric way, for instance by cubic splines. This allows identifying and characterizing nonlinear effects in a better way.

Tree based methods

Tree based methods are non-parametric estimation methods. These methods partition the space of the x-variables into a set of rectangular regions which should be as homogeneous as possible. Afterwards a simple model is fitted in each region.

In continuation to the brief description above, to comprehend regression methods a diverse range is found.

> **Starting from simple regression, that deals with only one predictor variable to the response variable to sophisticated regression methods like, "penalized regression methods to handle the problem of correlated variables and to cope with the panel structure of the data autogregressive processes and random effects are used.** (Hofmarcher, 2012 : Dissertation)

2.1.3 Regression methods and forecasting

Many forecasting methods use past or historical data in the form of time series. Whenever the necessary data are available, a forecasting relationship can be hypothesized, either as a function of time or as a function of independent variables, and tested (Markidakis & Whelwight , 1978). Hyndman (2014) mentions forecasting situation varies widely in their time horizons, factors determining actual outcomes, types of data patterns, and many other aspects. He reveals that forecasting methods can be very simple such as using the most recent observation as a forecast (which is called the "naïve method"), or highly complex such as neural nets and econometric systems of simultaneous equations.

In spite of numerous sophisticated methods of forecasting which are possible as mentioned above, for simplicity regression methods are popular. Innumerable studies have been carried out for forecasting in different situation with different perspective using the regression model. Developing and fitting a model and using the fitted model for forecasting are the two distinct phases in using regression as a forecasting tool. About regression method based forecasting Makridakis *et al.* (1998) reveals that regression models can be very useful in the hands of a creative forecaster. The authors further qualify regression analysis as a powerful method to model the effect of explanatory variables on the forecast variable.

National Agricultural Statistics Service (NASS) (2012) reports that linear statistical models are extensively used in forecasting and estimating crop yields and production. In crop production forecasting regression techniques could be used to test the effectiveness of incorporated variables and to test the rank of their efficacy in it. Ristanoski *et al.* (2013) have embarked that amongst the wealth of available machine learning algorithms for forecasting, time series linear regression has remained one of the most important and widely used methods simply due to its simplicity and the interpretability. In addition to this, Ristanoski *et al.* (2013) have concluded that time series forecasting was a classic prediction problem, for which linear regression was one of the best known and most widely used methods.

Regression models for forecasting are not limited to agricultural sector but are also equally used for economic research and as well in other fields too. Many regression

models in economics are built for explanatory purposes, to understand the relationships among relevant economic factors. Ramirez and Fadiga (2003) have argued that producing reliable forecasts was often a key objective in agricultural economic research and for which were used the time-series regression models.

In the light of the above principle of forecasting with regression models, Gommes (2001) has developed multiple (linear) regression model for crop yield with certain agro metrological variables and has described that the ultimate purpose of such modeling was for crop production forecasting. Lobell *et al.* (2007) has analyzed relationship between crop yield and three climatic variables; minimum temperature, maximum temperature and precipitation for 12 major Californian crops. In the study the authors have argued that yield-climate relationship could provide a foundation for forecasting crop production within a year and for projecting the impact of future climate changes. Similarly Zhang *et al.* (2004) have used univariate linear regression to fit and forecast global and regional trends of rice for the 6 variables with forty years of time series data. The variables namely are, rough rice area, rough rice production, rough rice yield, rice insecticide sales, rice fungicide sales, rice herbicide sales and rice pesticide sales.

To continue on the application of regression models for forecasting, Muhammad and Abdulah (2013) have used past forty years of time series data of paddy production and have developed various forecasting models including linear regression model. Bozrath (2011) has used single regression (univariate regression) to obtain a forecast for demand using sixteen months time series data of demand history. In the study he has built a regression model to handle trend and seasonality. Guenthner (1992) has developed models that can be used to forecast vegetable crop planting. In this study multiple linear regression analysis was used to determine the factors that influence planting of potatoes and onions. Shabri *et. al.* (2009) have used past thirty eight years (1971-2008) of time series records for rice yield data in Malaysia to have a comparative study on the hybrid methodology that combines the individual forecasts based on Artificial Neural Network (ANN) approach for modeling rice yields.

In addition to the sole use of regression models for forecasting, studies have been carried out to compare the predictive ability of various other sophisticated models with multiple regression model as a tool for forecasting. Kutsurelis (1991: Master's thesis) has

computed multiple regression model for forecasting financial markets together with Neural Network (NN) method and has compared with one another. Likewise Ulbrich (2010) has used regression models for calibrating which one of the two recommended math models of the calibration data is expected to have better predictive capabilities.

In the context of Nepal with time series data (1971 to 2000) Nayava (2012) had studied the relationship between rainfall and production of rice. However, the author had used simply time series plots to draw inferences and no objective measures were computed. The author argues that, impact of rainfall on rice yield and production was quite evident. The other example is, even that Dahal and Routray (2011) have developed multiple regression model to evaluate apparent strength of the relationship and to explain the variations on dependent variable (crop yield) against the soil variables.

As , depending upon the situation, priority and obviously in the availability of the required resources for forecasting, different studies have different approaches either in considering the number of variables or the type or the nature of the variables for estimating yield or forecasting production. Single model will never suffice from different perspective in the diverse need scenario. For this, Ramasubramanian (n.d.) suggested that multiple regression model based on, plant characters, weather indices, based on climatic variables such as rainfall, temperature etc. would be more relevant . Further, if the forecast was to be based on qualitative variables like presence or absence of affluent rain, logistic regression model could be used.

2.2. Theoretical frame work

This study has used secondary data sets and has built regression model for rice production forecasting. Specifically the study has applied regression model building theory to optimize the forecasting models. To achieve the objectives of the study numerous text books, papers, journal articles and other sources which explain regression model building for forecasting have been reviewed in detail. However, the main focus given is in the theories and the methods explained by Makridakis *et al.* (1998); Bowerman *et al.* (2005); Darper and Smith (1998); and, Hyndman and Athanasopoulos (2013). In addition; Young (2013); Liu (2009); and numerous other literature are explored thoroughly and applied, whatever applicable.

Markidakis and Whelwright (1978), states that time-series and regression (casual) models are two major types of forecasting models. In the first type, prediction of the future is based on past values of a variable and the objective of such time-series forecasting methods is to discover the pattern in the historical data series and extrapolate that pattern into the future. On the other hand, casual models assume that the factor to be forecasted exhibits a cause-effect relationship with one or more independent variables. Borazth (2011) explains that in time series models for forecasting, 'time period' is the independent variable and for casual model the independent variable is some other variable which have casual effect on the dependent variable than the 'time period'. In this study we will focus on building linear regression model/s using time period as the sole independent variable. In other cases we will use time series data of different casual variables to the time series data of production for model building.

Theoretically therefore, it will be dealt with two types of models for this study. Single regression model is the model which consist only one independent variable as the predictor and multiple regression model, the model which consists of two or more than two variables as the predictors.

2.2.1 Linear regression theory

Regression analysis has different purposes. If Y, the dependent variable in regression analysis is a phenomenon, this is explained with the help of the independent variable X. On the other hand, if the dependent variable Y is a forecast variable, to be predicted, regression analysis helps to predict the value of Y for a given value of the independent variable X. According to Baker (2010) regression analysis lets one to use data to explain and predict. To reflect the essence of the analysis at hand, depending upon the situation, the dependent and independent variables are common to have many other names. For instance, if the analysis is to study cause and effect relationship between the independent and the dependent variables, the independent variable is called the cause variable and the dependent variable the effect variable. For forecasting, the dependent variable is given the name, forecast variable and the independent variable, the predictor variable, and the like.

Linear regressions are of two types. In simple regression there is only one dependent variable X, i.e. for forecasting purpose, there will be only one predictor variable. For this reason simple regression is also called single regression. However, in practice most often there are many predictor variables to deal with, this is called multiple regression. In linear regression, the forecast variable is a linear function of one or more predictors plus an error introduced to account for all other factors. In the case of multiple regression the forecast variable Y is related to a linear combination of more than one predictor variables. As such multiple linear regression can be thought of an extension of simple linear regression, where number of predictor variables are more than one.

Restating that regression equations either explain a phenomenon or predict a forecast variable, single regression line estimates the effect on $Y$of a change in X. That estimated effect is b, the slope of the line. A change in X of 1 changes Y by b, on average. This explanation holds good for whatever phenomenon it is that Y represents. Similar explanation holds good in the case of multiple regression. And, if the regression equation is for prediction the regression line lets us calculate a predicted Y value that corresponds to any particular X value. Any linear regression equation works through the popular method of "least squares". Least square method provides way of choosing

regression coefficients by minimizing the sum of the squared errors to yield a line of best fit for the linear regression.

About the relationship between the variables in regression analysis, Baker (2010) has a brief summary. He reveals that, errors are the vertical distances from the points to the true line and residuals are the vertical distances from the points to the regression line and if regression line comes out horizontal, there is no linear relationship between X and Y. But if other than linear, there is an uphill and downhill like structure, this indicates non-linear relationship.

2.2.2 Linear regression as statistical modeling

Developing statistical model through linear regression theory is outstandingly described in Makridakis *et al*. (1998). Also, the theory is described well in Bowerman *et.al.* (2005) and Darper and Smith (1998). A conceptual summary for this is briefly mentioned in the followings.

Two equations that a linear regression holds are: A theoretical equation in which the parameters are the unknown constants and an equation in practice that is used for estimating these unknown parameters. For instance,

$$Y_i = \alpha + \beta X_i + \epsilon_i,$$

is the theoretical equation where, Y_i and $X_i$represent the ith observations of the variables Y and X respectively, α and β are fixed but unknown parameters and can be only estimated and ϵ_i is a random variable that is normally distributed with mean zero and having a variance σ_e^2. And,

$$Y_i = \alpha + bX_i + e_{i'} \text{ for } i = 1,2, \dots, n,$$

is the equation used in practice, where, a and b are estimates of α and β. And e_i is the estimated error, estimated variance of which is S_e^2. In the case of forecasting, when the parameters α and β are estimated, a regression line in practice is fitted and ultimately the forecast variable Y on average is predicted. At this point, the slope coefficient b means: A change in X of 1 makes Y change by b and the intercept coefficient a means: if X is 0 then Y is a.

According to Makridakis *et al*. (1998) in the practical process of estimating the coefficients a and b, the standard errors of these estimates tell how much these estimates are likely to fluctuate from sample to sample and the best way how much each estimate fluctuates is written in the form of a confidence interval. This leads to another notion of computing the confidence interval of the parameters α and β, the values of which are never known in practice. The values of a and b represent the best estimates of these parameters and the confidence interval of the parameters α and $\beta$are respectively written as:

$$\alpha: \quad a + t * s.e.(a)$$

$$\beta: \quad b \pm t * s.e.(b)$$

For the validity of the models described above several assumptions made in advance about the model are tested and at the end, fitted regression line with the obtained values of a and b, is tested with F-test for overall significance of the model. A large value of $F-$statistics will lead the slope b to be significantly different from zero, and the regression will explain a substantial proportion of the variance in the forecast variable Y. And the significance of the slope coefficients and the intercept are tested through two $t-$tests, one for each. In the procedure, a large absolute value of $t-$statistics for slope coefficient indicates the slope to be significantly different from zero and similar principle applies for the test of the intercept.

Moreover, Makridakis *et al.* (1998) has described that, the notions of statistical modeling of single regression are similar to multiple regression however the complexity in interpreting the coefficients and some additional theoretical concepts to arise. Coefficient of multiple determination, linearity concept, serial correlation and multicollinearity are notably added parts to consider while modeling multiple regression.

In multiple regression there is one variable to be predicted, but there are two or more explanatory variables, i.e. in multiple regression the forecast variable is the function of one or more of the explanatory variables.

In contrast to one regression coefficient in single regression, multiple regression consists of two or more regression coefficients (betas) in the equation, one coefficient per predictor. Each of these coefficients is interpreted as the change in the predicted value of Y for each-one unit change in that specific predictor, keeping all other predictors constant. For instance when we talk about the change made in the first predictor variable, this means that if X_1 differs by one unit, and other predictors did not differ, Y will differ by β_1 units, on average. Intercept in multiple is the predicted value of Y when the values of all predictors are 0.

2.2.3 Developing linear regression model for forecasting

Variable selection

Developing a regression model for real data is never a simple process, but some guidelines can be given (Makridakis *et al*., 1998). More specifically Karim (n.d.) reveals that where there is no clear cut theory, the problem of selecting variables for a regression equation becomes quite important. First of all through various means of variable selection procedures and strategies, (thoroughly discussed in the forth coming parts of this section) out of too many, a 'long list' of the potential predictors to impact on the forecast variable y , 'short list' of the most appropriate predictors are drawn up. For this however some amount of creativity, and a lot of feeling for the subject matter are not missed out at the first hand and basically variable selection stands upon the principles 1) Hunches of experts and other knowledgeable people in the general area come into the first place to counsel and give suggestion about which variables should be incorporated and which should not be of greater importance to get included for the analysis and 2) Availability of data; described by Makridakis *et al.* (1998), are employed at the first place of variable selection.

Scatter plot matrix and statistical significance

For the variable selection procedure, scatter plot matrix and the statistical significance of individual predictor variable with the forecast variable will help, but are not sufficient. From scatter plot it is not always possible to see the relationship, especially when the effect of other predictors has not been accounted for. And in the case of multiple linear regression, on all predictors, statistical significance does not always indicate predictive value. The reason for this is, when two or more predictors are correlated with each other the p-value is misleading.

Stepwise methods

In stepwise methods, in each step a single predictor variable is added to or deleted from a regression model, and a new regression model is evaluated. Different statisticians and researchers have different perspectives about how these methods perform. Bowerman *et al.* (2005) have mentioned that different computer packages carry out regression with these methods with slight variation.

Despite some controversies and dilemmas, in the use these methods, stepwise regression methods are helpful to get a guess of what are possible predictors when there are unmanageably large numbers of predictors for a multiple regression. According to "Stepwise Multiple Regression" (n.d.), stepwise methods can produce a predictive model that is parsimonious and accurate by excluding variables that do not contribute to explaining differences in the dependent variable. Mundry and Nunn (2009) have mentioned, "in addition to fundamental shortcomings with regard to finding the 'best' model, stepwise procedures are known to suffer from a multiple testing problem, yet the method is still widely used." Hyndman and Athanasopolous (2013) claims "It is important to realize that a stepwise approach is not guaranteed to lead to the best possible model. But it almost always leads to a good model." Also, Makridakis *et al.* (1998) mention that "stepwise regression is a method which can be used to help sort out the relevant explanatory variables from a set of candidate explanatory variables when the number of explanatory variables is too large to allow all possible regression models to be computed."

Coming along with all these differences and the different views about using the stepwise methods, to add one more, for Darper and Smith (1998) stepwise regression is the best among others however for Hyndman and Athanasopoulos (2014), backward elimination method is the best among the family of the stepwise methods. Together with these ideas incorporated critically, in this study we would be using family of stepwise methods for weeding out unimportant predictor variables from a group of large no of possible predictors in the analysis, primarily with the idea that Darper and Smith (1998) whom in support of their own idea "The techniques discussed in this chapter can be useful tools. However, none of them can compensate for common sense and experience, have presented the methods to be, stepwise procedure first and after words if exploration

of the equations around the stepwise choice is then desired, we do the 'best subsets procedure' (to be discussed later). Finally we come along the discussion of stepwise methods (Darper & Smith, 1998:343-344) as:

> **While the procedures discussed do not necessarily select the absolute best model, they usually select an acceptable one. However alternative procedures have been suggested in attempts to improve the model selection. One proposal has been: Run the stepwise regression procedure with given levels for acceptance and rejection. When the selection procedure stops, determine the number of variables in the final selected model. Using this number of variables, say, q, do all possible sets of q variables from the r original variables and choose the best set.**

What really are such stepwise methods? Now we advance this section discussing these methods in brief. In practice there are three types of stepwise regression. Forward stepwise, backward stepwise and stepwise (forward-with a backward look).Both forward and the backward methods pick one predictor at a time. Forward starts picking up the predictors one by one from the strongest to predict the forecast variable up to the one which has at least some specified amount of significant influence in the forecast variable, whereas the backward does this in a little different way. At first it enters all the predictors in the model and starts sorting out with the weakest predictor eradicated one by one in every step running a regression. And so on until there would be any predictor which won't meet the specified criterion i.e. the alpha level of significance to get included in the model.

To sum up, in stepwise methods we start with a model containing all potential predictors and try to subtract one predictor at a time. Keep the model if it improves the measure of predictive accuracy Iterate until no further improvement.

Model selection approach

Regression model building is setting possibly a large set of predictor variables to fit a parsimonious model that explains maximum possible variation in the forecast variable Y with as small set of predictors as possible. A model with too many predictors can be relatively imprecise while one with too few can produce biased estimates. After appropriate predictor variables were fixed (i.e. the short list of the predictors was prepared) using the stepwise /and other methods, we are then subjected to find the best subset of the predictors viz. the predictors to yield a parsimonious model.

Similar to variable selection approaches, for model selection too, there are numerous methods almost with the same amount of controversies and the debates. However, there is no any such unique way which outsmarts all the others. Darper and Smith (1998) explained techniques discussed for this purpose could be useful tools, however, none of them could compensate for common sense and experience. As such sometimes the model which consists of all selected variables becomes the optimum model, where as in many other times this might not be the case. This issue of selecting the best combination of the predictors for optimum model building comes under the theory of model selection criteria. In this context several procedures the authors (Darper and Smith) have described for model selection are 1) all possible regression using three criteria: R^2 , s^2, and the Mallow's Cp; 2) best subset regression using R^2, R^2 (adjusted) and Mallow's Cp ; 3) stepwise regression, 4)backward elimination; and 5) some variations on previous methods. Some other methods we can find commonly in other literature are Akaikis Information Criteria (AIC), and Basian Information Criteria (BIC). These procedures are valuable for quickly producing regression equations worth further consideration, however, the factors like common sense and basic knowledge of the data being analyzed cannot be put aside. In this section these measures are discussed in brief.

Adjusted R^2

Minimizing the standard error s is equivalent to maximizing R^2, the multiple coefficient of determination and will always choose the model with the most variables, and so is not a valid way of selecting predictors (Hyndman & Athnaspoluou, 2013). In fact a model with the largest value of R^2 is the model which contains all predictors included in the study. Every additional predictor variable will result in an increase in R^2.But, to investigate a parsimonious model; clearly not all these predictors should be included and the basic short coming about R^2, when one makes this the criteria for the best model is well explained by Bowerman *et al.* (2005). The authors say that even if the independent variables in a regression model were unrelated to the dependent variable, they would make R^2 somewhat greater than 0. And hence, to avoid overestimating the importance of the independent variables, is recommended calculating an adjusted multiple coefficient of determination. The problem with R^2 is, it does not take into account degrees of freedom. Adding any variable tends to increase the value of R^2, even if that variable is irrelevant. Adjusted R^2written as R^2(adj.), is given by the equation:

$$R^2(adj.) = 1 - \frac{(1 - R^2)(N - 1)}{N - k - 1}$$

Where, N is the number of observations and k is the number of predictors. This is an improvement on R^2 as it will no longer increase with each added predictor. Using this measure, the best model will be the one with the largest value of $R^2(adj.)$. We can compare the $R^2(adj.)$values for all the possible regression models and select the model with the highest value for this. As such maximizing $R^2(adj.)$ works quite well as a method of selecting predictors, although it does tend to err on the side of selecting too many predictors. Specifically $R^2(adj.)$ takes into consideration of parameters in the model, and punishes the models with too many terms.

The best subset regression

While building regression model for forecasting, stepwise regression methods help to sort out the important predictors from a long list of potential predictors. However even after selecting appropriate predictors the problem is not completely solved. For instance let us say ten predictors were selected for the final model. This means $2^{10} = 2024$ possible models are to be computed (considering all combinations of the predictors!) and the best among these was to be selected. This sounds quite impossible.

At this point to deal with such situation scientists have developed "the best subset regression" or "all possible regression" which even after using stepwise regression helps to bring the number of appropriate variables in a manageable size and the models yielded by such predictors are put forward for further in-depth analysis to chose the best one at the end.

As such the best subset regression provides various tools for all possible models which are useful to choose the best model among them all. Kandane, and Lazar (n.d.) have described that over the years numerous tools for selecting the "best model" have been suggested in the literature with many criteria. Out of these many tools some which are almost common are: R^2, R^2 (adj.), MSE, Mallow's Cp-statistics, $PRESS$ statistics, AIC, and BIC. The models that perform well according to this chosen criterion mandatorily are not the final models but are to be considered for an in-depth investigation.

The general idea behind best subsets regression is, we select the subset of predictors that do the best at meeting on these well-defined objective criterions and we end up with a reasonable and useful regression model. Cautions while using best subset regression are 1) in case the list of candidate predictor variables does not include all of the variables that actually predict the forecast variable, the model would be underspecified and therefore misleading, 2) sometimes the results do not point to one best model and needs best judgment, 3) as the number of potential predictors increase, the number of models to compare grows rapidly and this method limits its calculation speed in the software and, 4) issues of collinearity are major concern for explanatory model building, which is an even better reason than analysis speed to look at a correlation matrix and reduce the number of variables to include.

However numerous as above, for this study for model selection we confine to R^2(adj.), best subset criterion: Mallow's Cp-statistics, AIC and BIC which are described in brief in the following.

Mallow's Cp-statistics

Mallow's Cp-statistics is another criterion for model selection. Since the Cp-statistics for a given model is a function of the model's standard error, se and since se is to be small, we want Cp-to be small. So, by theory one should find a model for which Cp is as small as possible. Cp-statistics roughly equals p the number of parameters in the model. If a model has a Cp-statistic substantially greater than p, a different model for which Cp is slightly larger and more nearly equal to the number of parameters in that (different) model could be preferred. If a particular model has a small value of Cp and Cp for this model is less than p, then the model is considered desirable. After finding one or more potential final regression models, we check the regression assumptions and then identify outlying and influential observations. Based on this analysis, necessary improvements are made and eventually one or more final regression models are used to describe, predict and or control the dependent variable.

Mallow Cp is given by the formula,

$$Mallow's\ Cp \frac{RSS_p}{\hat{\sigma}^2} + (2p - n)$$

Where, $RSSp$ the residual sum of squares for a model with p terms, and $\hat{\sigma}^2$ the estimate of the error variance based on the full model. Cp itself should approximately be equal to p for an adequate model.

At the mid of the debates about which criterion is best for selecting best model, Darper and Simth (1998) have considered this ($Cp -$ statistics) to the best among the other mentioned ones. To identify "best" models recalling that p denotes the number of parameters in the model, given below are the reasonable strategies for using Cp presented by Kandane and Lazar (n.d.) :

Identify subsets of predictors for which the Cp value is *near p* (if possible).

- The full model always yields $Cp = p$, so don't select the full model based on Cp.
- If all models, except the full model, yield a large Cp not near p, it suggests some important predictor(s) are missing from the analysis. In this case, we are well-advised to identify the predictors that are missing!

- If a number of models have Cp near p, choose the model with the smallest Cp value, thereby insuring that the combination of the bias and the variance is at a minimum.
- When more than one model has a Cp value near p, in general, choose the simpler model or the model that meets your research needs.

Akaike's Information Criterion (AIC)s

Akaike's information criterion, commonly known as AIC is another tool which trades off between model fit and model complexity. This is defined as,

$$AIC = N \log\left(\frac{SSE}{N}\right) + 2(k+2)$$

Where N is the number of observations used for estimation and k is the number of predictors in the model.

The model with the minimum values of the AIC is often the best model for forecasting.

Corrected Akiaike's Information Criterion ($AICc$)

For small values of N, the AIC

tends to select too many predictors and so bias-corrected version of the AIC has been developed.

$$AICc = AIC + \frac{2(k+2)(k+3)}{N-k-3}$$

$AICc$ similar to AIC, should be minimized.

Bayesian Information Criterion (BIC)

A related measure is Baysian Information Criterion, commonly known as BIC is defined as,

$$BIC = N \log\left(\frac{SSE}{N}\right) + (k+2)\log(N).$$

Similar to AIC, minimized the BIC should give the best model. The model chosen by BIC is either the same as that chosen by AIC, or one with fewer predictors.

Regression assumptions

Linear regression focuses on the assumptions of errors, "Testing the assumptions in linear regression", (n.d.) and Makridakis *et al.* (1998) have described: Linearity, independence, constant variance and normality to be the four assumptions in linear regression. If these assumptions are valid, then the model at hand possibly is good one and it could provide reliable forecast. However when any of these assumptions are violated, then the forecasts, confidence intervals, and economic insights yielded by a regression model may not be as effective as expected. That is to say, if any of these assumptions do not satisfy for the given set of data, then it should be understood, In the case that these assumption are not satisfied, the statistical tests t-test, F-test, confidence interval, R^2 , etc. applied in the course of model building do not strictly go right. This leads to conclude, the model is not incorporating all the information in the data set to yield good model and more appropriate models for the data are yet to exist and they are to be re-estimated.

However, according to Bowerman *et al.* (2005) regression model building assumptions in very seldom hold exactly in any practical situation. And, it is continued that regression results are not extremely sensitive to mild departures from these assumptions. In practice, only pronounced departures from these assumptions require alteration and mild departures from the regression assumptions do not seriously hinder in the ability to use a regression model to make statistical inferences.

Out of the varieties of tools to check the assumptions in linear regression, in this study regression diagnostic and statistical tests of hypothesis are used. For this residual plots given by the model against 1)values of each independent variable, 2) values of the predicted value of the dependent variable and 3) the time order in which the data have been observed (if the regression data are time series data*)* are plotted. Regression diagnostic checks different aspects of the fitted model including underlying assumptions. In common the regression assumptions hold, the residuals should look like they have been randomly and independently selected from normally distributed populations having mean 0 and variance σ^2.

The basic principle of regression diagnostic is : Forecast error, i.e., the residual e , the difference between actual quantity 'A' and forecast $e = A - F$ is the

unpredictable random component of each observation and would expect these residuals randomly scattered without showing any systematic patterns while plotted variously in the course of checking these assumptions.

Linearity

Assumption of linearity is tested with the plots of the residuals versus the predicted values or each of the predictor variables. If these scatter plots show a pattern, then the relationship may be nonlinear and the model will need to be modified accordingly. To check if any potential predictors are missing the residuals against any predictors which are not in the model are plotted. If these show a pattern, then the predictors may need to be added to the model.

Lack of fit test is one of the statistical tests of hypothesis to test the model form in linear regression. For this test, if p value is larger than the significance level $\alpha = 0.05$, this concludes lack of linear fit is not significant.

Test of constant variance

The word homoscedasticity is used for constant variance. The regression model assumes that the residuals have the same variance throughout. And when the assumption of constant variance is violated, the problem is called heteroscedasticity or changing variance. This assumption is valid if the residuals in the regression model show same variance throughout. However, if a pattern is observed, the variance of the residuals may not be constant.

For the validity of constant variance assumption, plots of the residuals against the fitted values of x, y are examined. A residual plot with a horizontal band appearance suggests that the spread of the error terms around 0 is not changing much as the horizontal plot value increases. Such a plot tells us that the constant variance assumption approximately holds.

Fluctuation around 0 indicates the validity of the constant variance assumption. However, if a fanning-out pattern and or the funneling-in patterns is observed, both of this cases indicates an increasing error variance and the assumption is violated i.e. heteroscedasticity prevails. To overcome this problem, mathematical transformations such as a logarithm or square toot may be required.

Bruesh-Pagon test of constant variance (named after Trevor Breusch and Andrian Pgan) is the test of hypothesis for homoscedasticity. The procedure in Stata (Hamirick, 2013) follows: The command "estat hottest" followed by all predictor variables. If for the computed $Breusch - Pagan$ statistics, the p value is greater than the level of α (0.05) this indicates the assumption of homogeneity of error variance is valid.

The independence assumption

The independence assumption is most likely to be violated in the case of time-series data. This is also called the situation of serial correlation and or the autocorrelation. Plot of residuals against time is used for testing this assumption. In addition to this, a Durbin-Watson statistics (here after written as DW) obtained from Durbin-Watson test is the popular test of hypothesis for testing independence assumption. That shall remain valid in this case as well.

As residual diagnostic, if the plot of the time-ordered residuals displays a random pattern, the error terms have little or no autocorrelation, in such a case, it is reasonable to conclude that the independence assumption holds. However if this is not the case and if the residual plot displays some cyclical and or an alternating pattern, the independence assumption does not exists. It is violated.

Durbin Watson Test

This test, tests the null hypothesis: there is no serial correlation among the consecutive error terms against the alternative: the error terms are dependent to each other. In symbol we write,

$$H_0: \rho = 0 \text{ and } H_1: \rho \neq 0 \text{ .}$$

Among the adequate variations about the decision rule of the test, here we pick the one mentioned in Darper and Smith (1998, p. 186). According to this for the two-sided test the null hypothesis $H_0: \rho = 0$ is rejected at level $2(\alpha)$, against the alternative $H_1: \rho \neq 0$, if $DW > DW_U$ or $4 - DW < DW_U$.

For DW the computed value of the DW- statistics, DW_U is the upper bound value in the DW-statistics table and DW_L=lower bound .A similar rule is given in (Makridakis, et al. 1998, p. 268). According to which, the theory behind this statistics is complicated but readily usable in a practical setting. For which, DW Statistics ranges in value from 0 through 4, with an intermediate value of 2. And, the decision rule of the test is as the followings.

Compare DW or ($4 - DW$, whichever is closer to zero) with DW_L and DW_U from Durbin Watson statistics table.

1) If $DW < DW_L$, conclude that positive serial correlation s a possibility

2)If $DW > DW_U$ conclude that no serial correlation is indicated

3) If $4 - DW < DW_L$ conclude that negative serial correlation is possibility

4) If DW or $(4 - DW < DW_U)$ value lies between DW_L and DW_U the test is in conclusive.

An indication of positive or negative serial correlation would be cause for the model to be re-examined and the moral of DW test is: The more observations, the more likely we shall be able to make a definite decision via the Durbin-Watson test.

The normality assumption

According to (Makridakis *et al.*, 1998) normality assumption is not a serious assumption in that residuals are the result of many unimportant factors acting together to influence the forecast variable, and the net effect of such influences is often reasonably well modeled by a normal distribution. However if the assumption is seriously violated, it is inappropriate to do the significance testing. A mathematical transformation can help in correcting the problem of non-normality. According to Box and Cox (1964), family of power transformation, defined only for positive data values, would be useful to solve the problem of non normality.

Normal probability plot of the standardized residuals is used to check normality assumption and the Shaprio-Wilk test as hypothesis testing for normality with the null hypothesis, H_0: the residuals are normally distributed against the alternative, H_1: the residuals are not normally distributed.

Also the normality assumption is tested through, a histogram, stem-and-leaf display, and normality plot of the residuals. The histogram and the stem and display should look bell-shaped and symmetric about zero. The normal plot should have a straight –line appearance. A normal plot that does not look like a straight line, indicates that the normality assumption is violated viz. a curved pattern of a residual plot indicates that the functional form of the regression model is incorrect.

Shapiro-Wilk test is used for normality assumption. For this, statistical test of hypothesis, for computed Shapiro-Wilk statistics, for the given sample if $p > 0.05$. This proves the assumption of normality is valid.

Failure in normality assumption is commonly dealt with transforming the data into a new set and this possibly could satisfy the assumption. However, it should be kept in mind that effect might be due to one or more of the other assumptions are broken. Or, they are the outliers which cause a model not to be normal. This is therefore we test the normality assumption, it is important to realize that violations of the constant variance and correct functional form assumption etc. and the causes mentioned above should be kept in mind and do residual plot to check for non-constant variance and incorrect functional form before making any final conclusions about the normality assumption.

Outliers and influential points

Outliers take extreme values compared to the majority of the observations in a data set. If there is an outlier in the data, rather omit it, its effect are removed. The influence of outliers is identified by computing regression coefficients with and without outliers. Observations that have a large influence on the estimation results of a regression model are called "influential observations". Possible ways that any data point can be outlier are: it could have, an extreme x value, an extreme y value, an extreme x and y value and it might be distant from the rest of the data, even without x and y values. According to Bowerman *et al.* (2005) an observation may be an outlier with respect to its y value and /or its x value, but an outlier may or may not be influential.

When data set includes influential point, things to consider are: the influential point may be bad data viz. the measurement error, check the validity of the data point. Andersen (2012) and as well, Jacoby (n.d.) have described outlying observation can cause to misinterpret patterns in plots. More importantly, according to the author, separated points can have strong influence on statistical models viz. unusual cases can substantially influence on the fit of the Ordinary Least Square (OLS) model. And therefore, deleting outliers from a regression model can sometimes give completely different results. Cases that are both outliers and high leverage exert influence on both the slopes and intercept of the model, outliers may also indicate that our model fails to capture important characteristics of the data. Bowerman *et al.* (2005) recommend first dealing with outliers with respect to their y values, explaining that they could affect the overall fit of the model. According to whom if this was done first, other problems become much less important or disappear.

To know if an observation is an outlier with respect to its y value, studentized residual for the observation are useful. The authors continued explanation is: As a very rough rule of thumb, if the studentized residual for an observation is 2, in absolute value we have some evidence that the observation is an outlier with respect to its y value. At this end, the way to determine if an observation is influential is to calculate, Cook's distance measure written as $Cook's\ D$.

> **If none of these appears to be the case, two analyses—one with the influential cases in and one with these cases deleted—could be reported to emphasize the impact of these few points on the analysis. This is a case where researchers must use their training, intuition, reasoned argument, and thoughtful consideration in making decisions.** (Overbay & Osborne, 2004)

In addition to the above mentioned theory, the objective criterions used for outliers and influential points are: 1) If the studentized residual for an observation is greater than 2 in absolute value, there is some evidence that the observation is an outlier with respect to y value 2) If the leverage value for an observation is greater than $2(k+1)/n$, where k = number of independent variables and n = number of observation (in our case, $k = 3$, and $n = 35$) the observation is outlying with respect to x and 3) if Cook's Distances for the outliers are > 1 , then these outliers are considered to be the influential points.

Once identified and if there is a reason to believe that these cases arise from a process different from that for the rest of the data, then the cases should be deleted. For example, the failure of a measuring instrument etc. otherwise, two analyses—one with the influential cases in and one with these cases deleted—could be reported to emphasize the impact of these few points on the analysis.

Transformation

Often the data collected for model building for forecasting are not as simple. For example there might be variation in the data with time order or a different pattern could be observed than what we presume it to follow. In such cases transformation of data possibly could do better jobs. It might lead to simpler forecasting models and simpler forecasting models usually lead to more accurate forecasts. Some types of transformations that could be useful in this study are discussed in brief in the following.

Log transformation

Log transformation is followed taking the log of each observation base-10 logs (or base-*e* logs), base-10 logs in this study. Letting y^* denote the value obtained when the transformation is applied to y, log transformation is written as:

$$y^* = log\ y$$

The back transformation for this is to raise 10 to the power of the number. The log transformation is good for 'size' data and is useful both for making patterns in the data more interpretable and for helping to meet assumption in inferential statistics. For instance log transformation works for data where the residual gets bigger for bigger values of dependent variable. Log transformation not only tends to equalize the residuals but also tends to "straighten out" certain types of nonlinear data plots (Bowerman *et al.* 2005). This increases the importance and the usefulness of log transformation. However log transformation is simple and relatively easy to interpret, many a times other transformations are also need to be used.

Square root transformation

Square-root transformation is useful for count data. This consists of taking the square root of each observation.

$$y^* = \sqrt{y}\ = y^{\left(\frac{1}{2}\right)}$$

The back transformation is to square the number. For negative numbers, square root transformation cannot be taken. Some constant must be added to each number to make them all positive.

Box-Cox transformations

A useful family of transformations that includes logarithms and power transformations is the family of "Box-Cox transformations" (Makridakis *et al.*, 1998).

> **The statisticians George Box and David Cox developed a procedure to identify an appropriate exponent (Lambda = l) to use to transform data into a "normal shape." The Lambda value indicates the power to which all data should be raised. In order to do this, the Box-Cox power transformation searches from Lambda = -5 to Lamba = +5 until the best value is found.** (Buthmann, 2010)

As normally distributed data is always a preferred need most often in a number of statistical analysis, particularly the Box-Cox power transformation, is one of the remedial actions that may help to make data normal and get practitioners better prepared to work with non-normal data. As square root transformation, for Box-Cox transformation also, all the data are required to be positive and greater than 0.

Multicollinearity

Use of multiple regression is generally made for identifying the relative effects of the predictors to the forecast variable. In this context multicollinearity exists among the predictors when these predictor variables are related to or dependent upon each other. Following are the situations listed by Makridakis *et al.* (1998) when multicollinearity may arise in multiple regressions are:

> **1) Two explanatory variables are perfectly correlated 2) two explanatory variables are highly correlated 3) a linear combination of the explanatory variables is highly correlated with another explanatory variable. 4) a linear combination of one subset of explanatory variables is highly correlated with a linear combination of another subset of explanatory variables**

Multicollinearity is a matter of degree, not a matter of presence or absence (Paul, 2004). Hence, when multicollinearity if significant, the ordinary least squares estimators are imprecisely estimated. However, Makridakis *et al.* (1998) describe it is not a problem for forecasting and ability of a model to forecast is not affected by it. But when individual regression coefficients are of interest and also attempts are to be made to isolate the contribution of one explanatory variable to the forecast variable y multicollinearity then is a problem.

Some common methods for identifying and curing multicollinearity are: examination of correlation matrix, tolerance, variance inflation factor etc. Statistician often regard multicollinearity in a data set to be severe if at least one of simple correlation coefficient between the independent variable is at least 0.9 (Bowerman *et al.,* 2005). Another very common way to measure multicollinearity is to use variance inflation factor (VIF), defined as:

$$VIF = \frac{1}{1 - R^2}$$

Where, R^2 is multiple coefficient of determination

Remaining all other things equal, lower levels of VIF is desired. Higher levels of VIF are known to affect the models adversely i.e. results associated with a multiple regression analysis. Regarding VIF, as a tool for detecting multicollinearity varying literatures are found. Darper and Smith (1998) describe that all the guidelines given for

how large should VIF be to get notified with the multicollinearity problem in a regression model are essentially arbitrary and each person must decide for him or herself. Multicollinearity (2015) describes commonly given rule of thumb is: $VIFs$ of 10 or higher (or equivalently, tolerances of .10 or less) may be reason for concern. Further discussion related to rule of thumb is as the followings:

> **Unfortunately, several rules of thumb – most commonly the rule of 10 – associated with VIF are regarded by many practitioners as a sign of severe or serious multicollinearity (this rule appears in both scholarly articles and advanced statistical textbooks). When VIF reaches these threshold values researchers often attempt to reduce the collinearity by eliminating one or more variables from their analysis; using Ridge Regression to analyze their data; or combining two or more independent variables into a single index. These techniques for curing problems associated with multi-collinearity can create problems more serious than those they solve. (**O'brien, 2007)

As such however widely used, this method of rule of thumb is therefore suggested to use with caution as it works only in which context the model was formed how the data were collected and what in fact was the models objective were etc.

Several methods which are used to reduce the degree of multicollinearity are ridge regression; principal components regression, etc.

Forecast errors and the accuracy measure

Forecast accuracy measure investigate the suitability of a particular forecasting method/model for a given data set. At the end these measures help update and or even recommend changing the models. We define forecast error and forecast accuracy as the followings which in fact are the extracts from

Forecast error is the deviation of the actual from the forecasted quantity error (Demand Planning.Net, 2014). i.e.,

$$Forecast\ Error(E) = (A - F)$$

Where, A = Actual values and F = forecasted quantity, Forecast error equivalently is also called forecast bias. i.e.,

$$Forecast\ Bias\ (B) = (A - F)$$

However for these, while taking the deviation, absolute values are considered more useful because magnitude of the error is more important than the direction of the error. In such case,

$$Forecast\ Error(E) = (|A - F|)$$

In relative term,

$$Percentage\ Error(PE) = \frac{A-F}{A} \times 100$$

Error 100% implies a zero forecast accuracy or a very inaccurate forecast and error close to 0% tends to increasing forecast accuracy. Forecast accuracy is a measure of how close the actual are to the forecasted quantity. If actual quantity is exactly the same as the forecast there would be 100% accuracy. However if error exceeds 100% this tends to 0% accuracy.

Some commonly used forecast accuracy measure are:

Mean Error:	$ME = \frac{1}{n}\Sigma(\mathrm{A} - \mathrm{F})$
Mean Squared Error:	$MSE = \frac{1}{n}\Sigma(\mathrm{A} - \mathrm{F})^2$
Mean Absolute Error:	$MAE = \frac{1}{n}\Sigma\lvert(\mathrm{A} - \mathrm{F})\rvert$
Percentage Error:	$PE = \frac{A-F}{A} \times 100$
Mean Percentage Error:	$MPE = \frac{1}{n}\Sigma(\frac{A-F}{A} \times 100)$
Mean Percentage Absolute Error:	$MAPE = \frac{1}{n}\Sigma\frac{\lvert A-F\rvert}{A} \times 100$
Root Mean Squared Error:	$RMSE = \sqrt{1/\mathrm{n}\Sigma(\mathrm{A} - \mathrm{F})^2}$
Tracking Signal:	$TS = \frac{\Sigma(A-F)}{MAE}$

In the above, ME is used to measure forecast bias where as MAE indicates the absolute size of the errors and has a great use in calculating $Tracking\ Signal$ (TS). According to Borazth (2011). The ideal value of ME is$'0'$. If $ME > 0$, the model tends to under-forecast and if $ME < 0$, the model tends to over-forecast.

Whatever and whatsoever have been discussed above, these methods are neither sufficient nor universal for forecast accuracy measures. There are plenty of debates and controversies for these forecast accuracy measures, from defining them to applying them in real practice.

"Hyndman and Koehler (2006) recommend that the "symmetric" Mean Absolute Percentage Error ($sMAPE$) not be used; this is included here only because it is widely used, although we will not use it in this book (Hyndman & Athanasopoulos, 2014)." If so then, what actually is forecast error and how should forecast accuracy be considered? Chokalingam (2012) reveals forecast error is the deviation of the actual from the forecasted quantity whereas for Clements (2010) a good forecast is an accurate forecast. Preston (2014) describes forecast accuracy is to be about quantity accuracy and time accuracy. For Clements, because it was easy to calculate and the results were easily understood, Mean Percent Error (MPE) was the best forecast accuracy metric. A similar idea to MPE is given by Hyndman and Athanasopoulos (2014). The authors reveal that, percentage errors have the advantage of being scale-independent and so are frequently

used to compare forecast performance between different data sets. But for Borazth (2011) the basic measures of forecast accuracy were Mean Forecast Error (MFE), equivalent to Mean Error (ME), Mean Absolute Deviation (MAD) equivalent to Mean Absolute Error (MAE) and the Tracking Signal (TS) as a check for model improvement. Further discourses are:

> **MAPE stands for Mean Absolute Percent Error - Bias refers to persistent forecast error - Bias is a component of total calculated forecast error - Bias refers to consistent under-forecasting or over-forecasting - MAPE can be misinterpreted and miscalculated, so use caution in the interpretation.**
> (Demand Planning.Net, 2014)

In the above context, for this study those accuracy measures which are the most common in practice from both, calculation point of view and the meaning they reveal while interpreted are used.

Comparing forecast methods

Most often for a specified forecast variable, we could end up with numerous forecasting methods. In such situation to judge which of the available forecast methods was good, and why, none- of the forecast accuracy measures are designed such that, they could work independently to fix this problem.

Nonetheless, some simple methods work as the benchmark for such purpose. Use of naïve method is recommended for model comparison by Makridakis *et al.* (1998) and Hyndman and Athanasopoulos (2014). Naïve method is one of the simple forecasting methods, and many a times these simple forecasting methods are found to be incredibly effective. However, despite their simplicity and occasional affectivity, most often such simple forecasting methods are used for comparing different forecasting methods for a specified subject. They serve as the benchmarks rather than the method of choice itself. So whatever forecasting methods are there, they will be compared with such simple forecasting methods to ensure that the new method is better than these alternatives. If not, the new method is not worth considering.

In this section we have discussed naïve and the average methods of forecasting as the simple methods. These are some commonly used methods of forecasting which will later be used to make comparison between different forecasting methods that will evolve at the completion of this study.

Average method

According to average method mean of the historical data is the forecast of all future values. Not only for historical data but this method can also be used for cross-sectional data (when we are predicting a value not included in the data set). Then the prediction for values not observed is the average of those values that have been observed.

Naïve method

Naïve method of forecasting works only for time series data. According to this method forecasts are nothing but simply the value of last observation. That is, the forecasts of all future values are set to be Yt, where Yt is the last observed value. According to (Hyndman & Athanasopoluos, 2014), this method works remarkably well for many economic and financial time series related data. For comparison purpose, the difference between the MAE or $MAPE$ obtained from a more sophisticated methods of forecasting and that obtained using NF (Naïve Forecast) provides a measure of the improvement attainable through use of that more sophisticated forecasting method. This type of comparison is much useful than simply computing the $MAPE$ or MAE of the first method, since it provides a basis for evaluating the relative accuracy of those results.

Goodness of regression model

Goodness of fit a regression model is tested with R^2, the multiple coefficient of determination, given that every efforts were done to optimize the model before testing its fit. Nonetheless, there are no set rules of what a good R^2 value is for a model's good fit, closer the value of R^2 to 1 the better the fit is. i.e. if the prediction is close to the actual values we would expect R^2 to be close to 1. On the other hand if prediction is unrelated to the actual values, $R^2 = 0$. As such in all cases, R^2 lies between 0 and 1.

One caution to be taken is, the R^2 value is commonly used but often incorrectly, in forecasting. And despite its vital importance when used correctly, validating the model's out-of-sample forecasting performance is much better than measuring the in-sample R^2, value.

R^2(pred.) : Prediction Sum of Squares (or $PRESS$) is used for model validation. This helps in assessing model's predictive ability. In general smaller the $PRESS$ value, the better the model's predictive ability. However, in practice $PRESS$ is rather customarily used to calculate predicted R^2 denoted by R^2 (pred.) which is more meaningful to interpret than PRESS itself. It is defined as

$$R^2 \text{ (pred.)} = 1 - \frac{PRESS}{TSS}$$

This helps to validate the model without splitting the data into training sample and the test sample as in the other traditional way of model validation.

Together, $PRESS$ and R^2 can help prevent overfitting because both are calculated using observations not included in the model estimation. Overfitting refers to models that appear to provide a good fit for the data set at hand, but fail to provide valid prediction for new observations. (Young, 2013)

Standard error of the regression

Another measure of how well the model has fitted the data is the standard deviation of the residuals, which is often known as the "standard error of the regression" and is calculated by,

$$s_e = \sqrt{\frac{1}{N-2}\sum_i^n e_i^2}$$

The standard error is related to the size of the average error that the model produces. This error is compared to the sample mean of y or with the standard deviation of y to gain some perspectives on the accuracy of the model. However in mind should be kept that, evaluation of the standard error is highly subjective as it is required when generating forecast intervals.

CHAPTER 3
MATERIALS AND METHODS

This chapter discusses the materials and the methods used for the study. This study has developed a parsimonious multiple regression model that explains as much variation as possible in the forecast variable with as small number of predictors as possible. The issue of availability and the quality data and, selecting of the most appropriate variables for the model fitted out of many promising ones remained the biggest challenges of the study. However, as a breakthrough to this, every small theoretical and as well the methodological issues were addressed minutely.

Issue of variable selection to investigate as less biased model as possible was addressed considering all aspects of the variables. For instance, key variables, promising variables and the possible variables; these all were incorporated in the study taking an utmost care non of such highly probable predictor was missed out and also no any predictor which led the model to over fitting was incorporated. The key variables were those variables which ought to be included in the model in any way. Prior studies or the expert view or the research experience and or the researcher's creativity apparently located the key variable/s to be included in the model. Also, the variable selection procedure which we employed always captured the key variable to be considered in the model.

Accordingly the promising variables were identified through the techniques of automated procedures: forward selection, backward selection and by the use of stepwise regression methods. And, at the end the main focus in variable selection was given to the approach of best subset regression governed by Mallow's Cp statistic which located the best model, out of many other potential candidate regression models. This was the approach of selecting the most appropriate predictor variables out of the candidate/potential predictors.

As such the variables to be included in the model were determined and the raw multiple regression model was developed using regression option in the Minitab. Afterwards the regression model was set for assumption testing for which residual diagnostics and various statistical tests were conducted. And at the end, the model

developed as such was testified with forecast accuracy measure coupled with Tracking Signal (TS) for the possible final update and or the change in the model before recommending its use in forecasting.

3.1 Materials

The following software were used while computing analyzing and interpreting the data.

1. SPSS vs. 20
2. Minitab vs. 16
3. Stata vs. 12 and,
4. Excel 2007

Sometimes when specific script inside the software were needed for special cases to obtain the results and or the test statistics or so, such scripts are either built or burrowed from elsewhere available. Such special cases have been mentioned specifically in the proper places.

3.2 Methods

The methods section has covered the step by step description of the methods while building the model. There were four visible steps in this section. The first part was variable selection. This started from gathering all possible variables at the first hand reaching up to of identifying of the most appropriate predictors for the model. For this, personal judgments, automated procedures such as family of stepwise methods and best subset regression were used. When a list of the potential predictors were investigated, to end up with the most appropriate predictors, through best subset regression, Mallow's Cp-statistics was computed and applied.

Second part was building of the regression model. This meant regression model was obtained with the finally selected variables. This considered, assumption testing, pertinent mathematical transformation, multicollinearity testing and sorting out of the effect of outliers and the influential variables. The third part was out of sample cross validation of the model. At this stage, the regression model was tested for its performance with the observations that were not used while building the model. The out of sample is called the test sample.

Forth coming steps were the comparison of the investigated model with a benchmark, called the naïve forecast method and, testing of goodness of fit of the model . Detailed description of the mentioned steps follows.

Variable selection

At the first hand a bunch of variables from different categories were sorted out to have some impact on 'rice production' the forecast variable. For instance; from supply and utilization category: harvested area, yield-paddy, production-paddy, rice consumption-per capita, total consumption-milled rice stock exchange-milled rice, seed consumption, and number of released and registered varieties were collected. Trade category consisted of: export quantity and import quantity. Price consisted of: export price and farm harvest price. For the land use category it was: Irrigated rice area. Input category: Fertilizer consumption, tractors harvesters and threshers. Human resource: rural-population, male labor force in agriculture and female labor force in agriculture ed. Finally, from climatic variables: annual mean rain fall and annual mean temperature were considered in the list.

Afterwards the variables were given the names to use them in the software and data for each of the listed variables were obtained. Among above, the predictor variables including the forecast variable rice production (*rice_pordn*) in (000t) ; harvested area (*harv_area*) in (ha), price at harvest (*frmhv_price*) in (NRS/t), fertilizer consumption (*fert_consump*) (000t), number of tractors (*n_tractors*) in (000 tractors) were collected from International Rice Research Institute ([IRRI], 2014). Likewise, data for seed consumption (s*eed_consump*) in (000t) was collected from (FAOSTAT, 2014). The variables, rural population (*rural_popln*) in (000 people), male labor force in agriculture (*mlag_lbfrc*) in (000 people) and female labor force in agriculture (*fmlag_lbfrc*) in (000 people) were obtained from (FAOSTAT, 2013). Data for the variable number of released and registered varieties (*rr_varieties*) was collected from Ministry of Agriculture ([MOA], 2010). And the data for the climatic variables *annual_rain* (annual mean rain fall) in (ml), and *annual_temp* (annual mean temperature) (^{0}C) were collected from Department of Hydrology and Metrology ([DHM], 2014).

Accordingly, the data collected were integrated in single spreadsheet and are presented as the total sample of the study [Appendix 1: (A)].

For the reason that, the accuracy of the model could only be determined by considering how well the model performed on new data, i.e. the data that were not used when fitting the model, a portion of the data called the test sample was segregated out

from the total sample leaving the rest as the training sample. Training sample was used for building the model. According to Hyndman and Athanasopoulos (2014) however the size of the test sample was dependent to the length of the sample was, and in how far ahead one wants to forecast, the size of the test set should minimally be 20 % of the total sample. And since in the case of time series data the test sample compulsoraly be the last part or the observarion, for this case we set up the first 35 observarions (1961-1995) to be the trainig sample [Appendix 1: (B)] and the last 15 observarions (1996-2010) to be the test sample [Appendix 1: (C)]. The test sample was taken to be 30% (i.e. 15 observation out of 50) keeping in view that, Hyndman & Athanasopoulos (2014) have argued, the size of the test sample should ideally be at least 20% of the total sample size. This study has aimed a short term forecast horizon as there is always a trade off between the accuracy of the forecast and the length of the time horizon.

However the above mentioned variables were allotted as the first plan data collection, the variables after words were put through the window of researcher's feelings about the subject matter and the insight, available literature, and with the experts' idea and of course of the availability of the data. This consisted a list of the possible predictors. Afterwards matrix plot [Appendix 2: (A)] and the correlation matrix [Appendix 2: (B)] of these eleven possible predictors with the response (rice production) were drawn to have the general view of the predictors whether or not a variable was to be included in the model.

Forward selection was carried out in SPSS and for the reason that, the automated methods of variable selection vary on their own from one software to the other software, to have a cross check on it we again did the forward selection once, but in Minitab this time. The screened out predictors from the forward selection were then subjected to the backward selection of both of the software SPSS and Minitab. Variables not eradicated by either or the other software up to this point were then subjected to SPSS/Minitab stepwise (forward and backward) selection. As such from the list of eleven possible predictors a list of five potential predictors were yielded at the end of the automated procedure of variable selection.

Next to the investigation of list of appropriate predictors, model selection was vital. It was to choose the best combination of the predictors satisfying the criterion that the model was best among all other possible combination of selected predictors. For this best subset regression was conducted in Minitab. This computed three statistics needed for model selection: Mallows Cp, adjusted R^2 and the standard error of regression *s.*

Accordingly the appropriate predictors which were used for the model building were therefore *harvested area, rural population and price at harvest*. However the predictors were chosen to be the best, the model so comprised was yet the crude one. This was then taken through the whole lot of model optimizing.

Regression model optimizing

This step consisted analyzing of the scatter plot of the predictors drawn against the response variable and the correlation matrix (Table 10) interpretation. This was the preliminary part of model validation process. Afterwards we moved on to assumption testing and multicollinearity check; outliers and influential points including. . Assumption testing was carried out through residual diagnostic and one statistical test of hypothesis each for each of the four assumptions: linearity, homoscedasticity, independence and normality.

Out of sample cross validation of the mode

This was mainly the testing of forecast accuracy of the investigated model. Various forecast accuracy measures mentioned in the theoretical framework section were computed using Excel with the help of the investigated model and the test sample and they were interpreted. Accordingly the inferences have been drawn out.

Comparison of the model

Investigated model on its own could not represent to which reference then it came out to be superior or inferior. For this comparison with naïve forecast method (out of several other methods) was carried out. Naïve forecast model is a standard method to compare any other advanced models investigated using the same time series data. We therefore for this comparison purpose, computed the naïve forecasts using the naïve

forecast method and using the training sample observation data. Next to this some popular forecast accuracy measures for instance MAE and $MAPE$ for the naïve forecast model were computed and hence the results with their respective counter parts of the investigated multiple regression model were compared to investigate which model was comparatively good to the other model.

Testing the goodness of fit of the model

Accordingly when at the end, we were finished of doing any treatment on the investigated model, the final task remained was to test goodness of fit of the model. For this we accumulated the various statistics that had been calculated during the run of the above different model optimizing processes, and such statistics were then interpreted in concordance to the theory of goodness of fit of the model. Mainly the statistics used for testing the goodness of fit of the model R^2, R^2 (adj.), PRESS and R^2 (pred.) and the standard error of regression s of the model.

CHAPTER 4
RESULTS AND DISCUSSION

In this study a multiple regression model is optimized to use it for rice production forecasting in Nepal. Fifty years long time series data were used for investigating the model. All theories and principles to optimize a model were extensively searched and applied. This started from variable selection followed by various techniques of optimization of the model; for instance assumption testing, testing of multicollinearity and so on. The regression model so obtained was then set for, out of sample cross validation and was looked after for is goodness of fit. At the end the model got to be founded with three predictors: *harvested area*, *rural population* and *price at harvest*. The methods sequence how we reached to this end as such was discussed in detail in the methods section of chapter 3 (materials and methods).

Building of the multiple regression model, in this study , started with selecting of the appropriate predictors. Once the predictors were finalized, the issue of model selection i.e. finding an ultimate model from the right combination of the candidate/potential predictors was settled down. Not always the full model (i.e. the model which consisted all of the candidate predictors would be the best model, but many a times a model that consists the smaller subset of the selected predictors come out to be the best model. For this reason danger was always there that, we selected the unimportant predictors. If unimportant predictors were selected we would be overfitting our model and if important predictors were excluded, our model would have been underfitted.

Variable selection procedures that were adopted in the study yielded a parsimonious model leaving eight predictors behind out of eleven possible predictors at hand at the first stage. When it was kept on moving, at the end was obtained five potential predictors (Table 8). Accordingly, for these five potential predictors when we employed best subset regression, this handed us only the three predictors (*harvested area*, *rural population* and *price at harvest*) to be the most appropriate predictors to result in the best model.

Harvested area was selected possibly because at least in the context of the least developed country like Nepal, where other factors to influence production for instance,

advanced agricultural technology etc. are not in sufficient rate of increase, production is proportional to the amount of area harvested. Similar argument can be given for the predictor *rural population*. The selection of the predictor (*rural population*) should have caught the truth of co-integration of the this specifically with the possible strong hold predictors *male labor force in agriculture* and *female labor force in agriculture*. However, for the third predictor *price at harvest* perhaps in the past it was not much shown up as the prominent predictor for rice production forecasting in Nepal. But for now it should have been selected due to the right methodology we employed in selecting predictors.

In the case of many predictors linear regression model, the aim is to develop a good predictor but the number of predictors should be small. Accordingly for optimal result with the least bias as possible, in the selection of predictors we did cross check replicating the same procedure but with the different software or different approach of variable selection. For instance we did not only use the backward selection in SPSS but also we employed the backward selection with Minitab and the differences between these two approaches were noticed and treated keenly for further processing in variable selection.

This was done to confirm no any potential predictors were missed in the analysis. Variable selection at the starting phase was done through research experience, available literature and through experts' views. Despite all the efforts as above, because problem of variable selection has never a final answer in the so far statistical world, the findings of the study are along with this unsolved risk for what so ever we have concluded at this end.

Coming along to a logical end of this overview, in this chapter the results and the relevant information obtained while conducting the methods in sequence as mentioned previously (specifically in methods section of ‘chapter three: the materials and the methods’), are well organized and presented for discussion. To ease the discussion and for better understanding, the chapter is categorically presented into different subtitles the results are thoroughly discussed followed by their meaningful interpretation throughout.

4.1 Variable selection

Variable selection rightly started from the stage of gathering all possible predictors at the first hand without much plans and thought on them. Such was simply a raw collection of the predictors and were not organized. Afterwards the list was processed through numerous stages of variable selection techniques and the varieties of automated methods which at the end led to reach the most appropriate predictors for the model. Following was the first hand collection of the predictors which were thought to have some impact on rice production.

harvested area, rice yield, rice consumption per capita, total consumption milled rice, stock exchange-milled rice, seed consumption, number of released and registered varieties, export quantity, import quantity, export price, farm harvest price, irrigated rice area, fertilizer consumption, number of tractors, numbers of harvesters, number of threshers, rural population, male labor force in agriculture, female labor force in agriculture, annual mean rain fall and annual mean temperature

Clearly the above was a huge list of the predictors, a challenging task to select some out of these many. Karim (n.d.) reveals that where there is no clear cut theory, the problem of selecting predictors for a regression equation becomes quite important, especially where there were lot of predictors. Also, according to Makridakis *et al*. (1998) have mentioned that developing a regression model for real data was never a simple process.

In this context above listed predictors when scanned carefully through researcher's feelings about the subject matter, the insight, available literature, experts' idea and of course the availability of data of a particular predictor, yielded a list of possible predictors (Table 1).

Table1: List of possible predictors

1. harvested area
2. farm harvested price
3. fertilizer consumption
4. number of tractors
5. seed consumption
6. annual mean rainfall
7. annual mean temperature
8. number of registered and released varieties
9. rural population
10. male labor force in agriculture
11. female labor force in agriculture

Matrix plot [Appendix 2: (A)] and the correlation matrix [Appendix 2: (B)] of these eleven predictors, despite their weak power to judge whether or not a variable was to be included in the model, were conducted just to have some idea and to screen out the week predictors. This signified some of the variables in the list of possible predictors were possibly not significant. They were number of registered and released varieties, annual rain fall, and annual temperature. For these variables, neither scatter plot did show some specific trend to suspect their causality in the forecast variable, nor was the zero order correlation found strong with. This signaled the variables will not get included in the model. However, for the mentioned reason of uncertainty of scatter plot and the correlation matrix, none of the variables in the list of the possible predictors were eliminated but were taken up to the next criteria: the automated procedures of family of stepwise methods (Darper & Smith, 1998; Hyndman & Athanasopoulos (2014); and, Makridakis *et al.*, 1998 and Mundray & Nun, 1998) . And, for the first time it was the forward selection of the variables.

Forward selection was carried out in SPSS. This ended up with the selection of seven variables (Table 2) out of the eleven possible predictors in the list (Table 1).

Table 2: SPSS forward selection (alpha to enter = 0.25)

1. rural population
2. harvested area
3. female agricultural labor force
4. male agricultural labor force
5. annual mean temperature
6. farm harvest price
7. fertilizer consumption

Despite the above, just because the automated methods of variable selection vary on their own from one software to the other software (Bowerman, *et al.*, 2005) , to have a cross check on it we again did the forward selection once, but in Minitab this time. The results of the selected variables are given below (Table 3).

Table 3: Minitab forward selection (alpha to enter = 0.25)

1. harvested area
2. farm harvest price
3. rural population
4. seed consumption
5. fertilizer consumption

Accordingly for the above two steps three variables; number of tractors, annual mean rainfall and the registered and the released varieties were commonly rejected. The rest eight: rural population, harvested area, female agricultural labor force, male agricultural labor force, annual temperature, farm harvest price, fertilizer consumption, and seed consumption were then subjected to the backward selection of both of the software SPSS and Minitab. The outputs of these actions are presented in (Table 4) and (Table 5) respectively.

Table 4: SPSS's backward selection (alpha to remove=0.1)

1. female agricultural labor force
2. fertilizer consumption
3. farm harvest price
4. harvested area
5. male agricultural force
6. rural population

Table5: Minitab's backward selection (alpha to remove=0.1)

1. rural population
2. harvested area
3. farm harvest price and
4. seed consumption

Three variables: harvested area, rural population and farm harvest price were commonly selected. And male labor force, female labor force, fertilizer consumption and seed consumption were selected at least from one of the approaches above. So we kept them for further processing. The only variable commonly eradicated was annual mean temperature. The variables selected so far were therefore: rural population, harvested area, female agricultural labor force, male agricultural labor force, farm harvest price, fertilizer consumption, and seed consumption

Variables not eradicated by either or the other software were: rural population, harvested area, farm harvest price, seed consumption, female agricultural labor force, fertilizer consumption, and male agricultural labor force. These variables were then subjected to SPSS/Minitab stepwise (forward and backward) selection. Following (Table 6 and Table 7) are the respective outputs.

Table 6: SPSS stepwise selection (alpha to enter =0.05, alpha to remove = 0.1)

1. Rural population
2. Harvested area
3. Male agricultural labor force
4. Female agricultural labor force

Table 7: Minitab stepwise selection (alpha to enter =0.05, alpha to remove = 0.1)

1. Harvested area
2. Rural population
3. Farm harvest price

As such from the list of possible predictors when weeded out the unimportant variables, a list of the potential predictors (Table 8) was yielded:

Table 8: List of the potential predictors

1. Harvested area
2. Rural population
3. Farm harvest price
4. Male agricultural labor force and
5. Female agricultural labor force

Next to the investigation of list of appropriate predictors, model selection was vital. It was to choose the best combination of the predictors satisfying the criterion that the model was best among all other possible combination of selected predictors. However called the appropriate predictors, not always the model that comprised all predictors was the best model. Taylor (2004) has mentioned that the problem as such in statistics this is an "unsolved" issue and there are no magic procedures to get the "best model". However again, for getting a best model in optimum, several criterions have been purposed, among which, Mallows Cp obtainable from best subset regression option in the statistical software like Minitab is the main one. Other criterions we have used together with Mallows Cp statistics are: Adjusted R-square and *s* the standard error of the regression.

Statistical software, Minitab, has best subset regression option with the criterion: the model with smaller (most often the smallest) Mallows' Cp-statistic. However, according to (Pardon, 2015) the limitation is that, based on the C_p criterion, with respect to bias, the researcher might have different legitimate models from which to choose. That is, for the models for which C_p values to one another are similar there is little to separate

these models based on this criterion. The criterion for a model to be unbiased is : The model are all unbiased models if their C_p values are equal (or are below) the number of parameters p. And there might be more than one unbiased model at a time from which we will have to choose the best model. In such situation the compatibility of the other criterions for choosing a bet models are sought out in addition to the smaller value of Mallow's C_p statistics. Out of many some commonly employed such supplementary criterions are the model with the largest adjusted R^2, and the model with the smallest MSE (or s = square root of MSE).

However, no matter what effort was done to select a best single model, there sometimes is the danger that different criteria may lead to different "best" models. And in such situation the problem was dealt with sufficient amount of creativity and researchers experience as Darper and Smith (1998) have mentioned. And, last but not the least, while carrying the best subset regression, similar to stepwise regression procedure, a fundamental rule is that the list of potential predictor variables must include all of the variables that actually predict the response variable. Otherwise, we end up with a regression model that is underspecified and therefore misleading. This fact also was carefully addressed.

Best subset regression was conducted in Minitab. This computed three statistics needed for model selection: Mallows Cp, and also the adjusted R-square and the standard error of regression *s*.

The best subset regression output for the forecast variable rice production versus the five potential predictors (Table 9) is given below.

Table 9: Best subset regression

Best Subsets Regression: *production* versus *harvested area, farm harvest price, rural population and male labor force and female labor force in agriculture*

```
Response is production

                                                    f     f
                                                    r m   m
                                                  h m u l   l
                                                  a h r a   a
                                                  r v l g   g
                                                  v _ _ _   _
                                                  _ p p l   l
                                                  a r o b   b
                                                  r i p f   f
                          Mallows                 e c l r   r
Vars   R-Sq  R-Sq(adj)         Cp        S       a e n c   c
   1   81.1       80.5       55.3   227.14    X
   1   73.5       72.7       89.8   268.70      X
   2   84.9       83.9       40.0   206.30    X X
   2   82.2       81.1       52.3   223.84    X       X
   3   93.3       92.6        3.6   139.60    X X X
   3   90.2       89.2       17.9   169.13    X     X X
   4   93.4       92.5        5.1   140.77    X X   X X
   4   93.3       92.4        5.6   141.76    X X X X
   5   93.6       92.6        6.0   140.47    X X X X X
```

The fourth column in the output (Table 9) above reveals Mallows Cp statistics for the model with different combination of the candidate (appropriate) predictors. In this column the Cp value of 3.6 for the model consisting of the predictors *harvested area*, *price at harvest* and *rural population* is minimum among all. This complies the number of the parameters to this model to be: number of predictors + the intercept parameter (i.e. 3+1) = 4. This satisfies the criterion, Cp (3.6) < no. of parameters (4) for the model to be unbiased. This as such satisfies both the criteria that, the Cp value is the smallest and as well less than the number of parameters for the model; which reveals that he model is 'the best' model. Not only the criterion of Cp is satisfied for this model, but also it has minimum standard error $se(139.60)$ *and* the highest value for adjusted R-square (93%) as compared to the models with other possible alternative combination of the candidate predictors. This signified the model with the specified predictors (i.e. *harvested area*, *price at harvest* and *rural population*) was the best model among all others. Hence the ultimate predictors for the model to fit for rice production forecasting in Nepal are as such finalized and are accordingly mentioned in (table 10).

Table 10: List of the appropriate predictors

1. harvested area
2. rural population and
3. price at harvest

From the unmanageable number of potential predictors at the start now we had, was a parsimonious model. All options which were theoretically appropriate for weeding out the less important variables were used. And because the stepwise methods, forward, backward and or forward-backward, all have variation on their own, not only one software was used but these methods were tried on both SPSS and the Minitab.

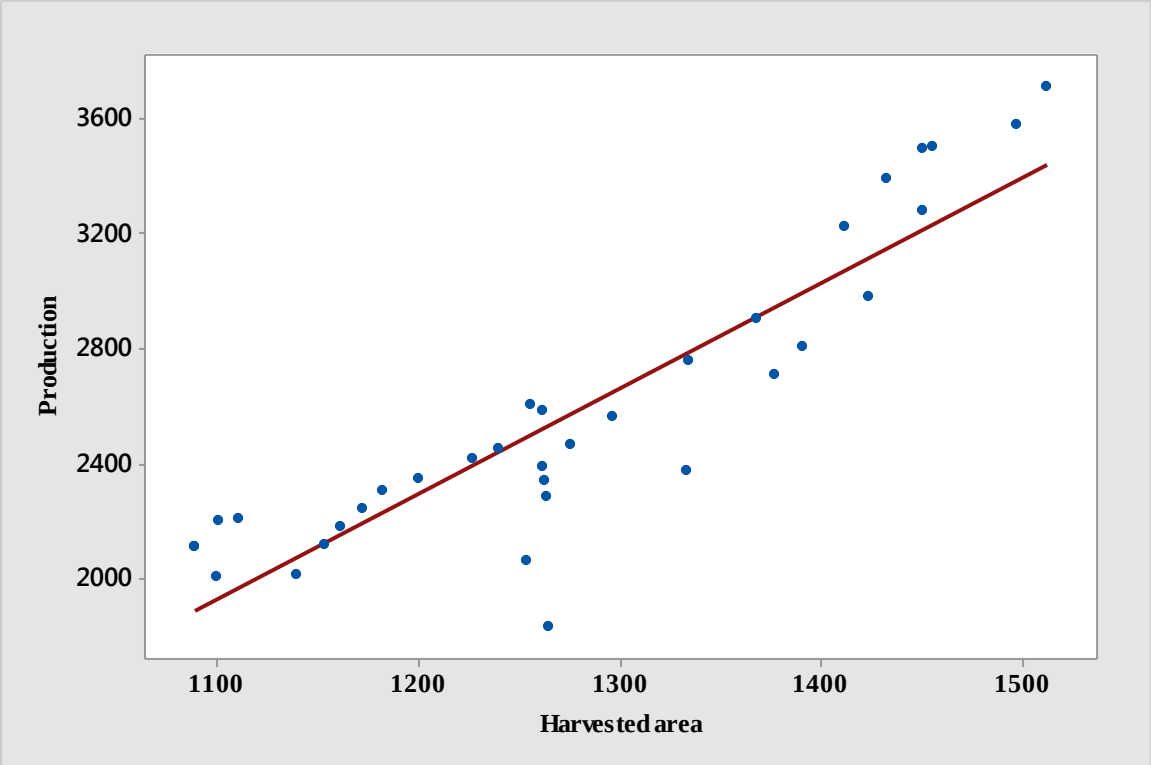
At the end as mentioned in above, however the predictors were chosen to be the best, the model so comprised yet was a crude one. This was therefore taken through the whole process of model optimizing (i.e. testing assumptions, checking for multicollinearity and dealing with the outliers and the influential variables etc.).

4.2 Optimizing of the model

Preliminary part of model validation

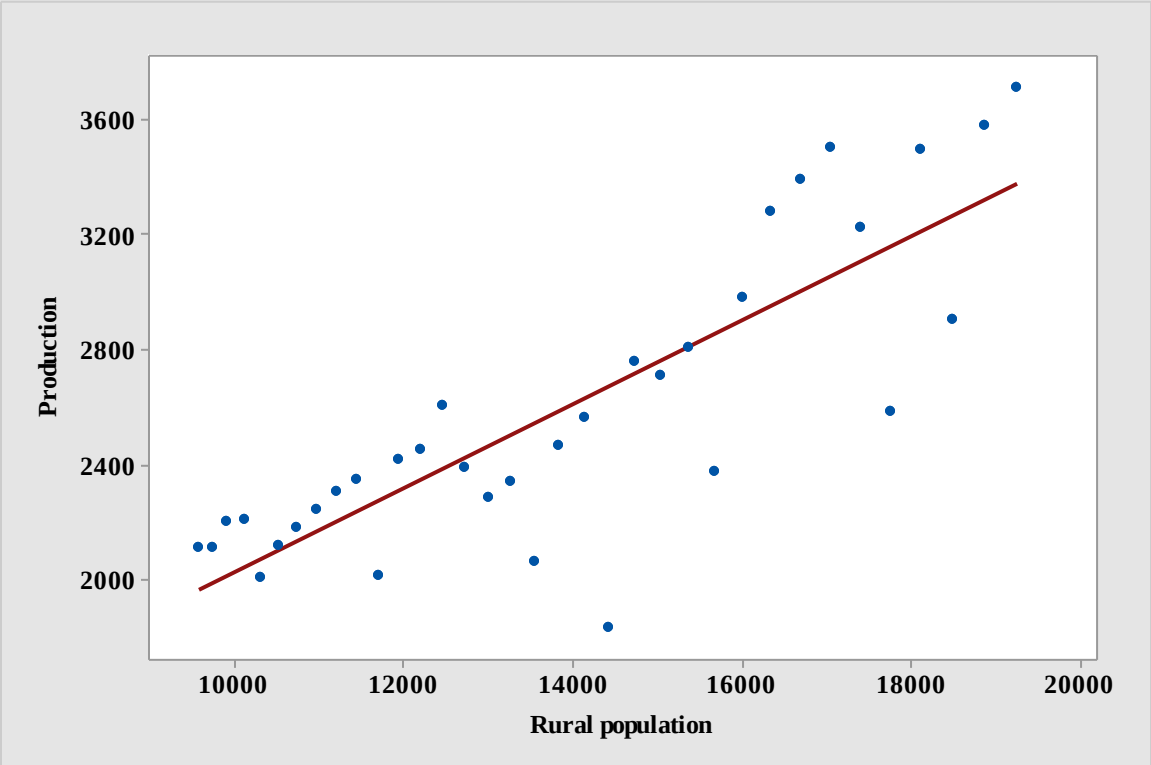
The scatter plots of the forecast variable *production* with the predictors: *harvested area* (Figure 1), *rural population*, (Figure 2) and with *price at harvest* (Figure 3) are the evidence of what was seen earlier in the regression model outputs, that they have positive and linear relationship with each other. All scatter plots show a clear upward straight line trend in the graph.

Figure 1: Scatter plot of production versus harvested area



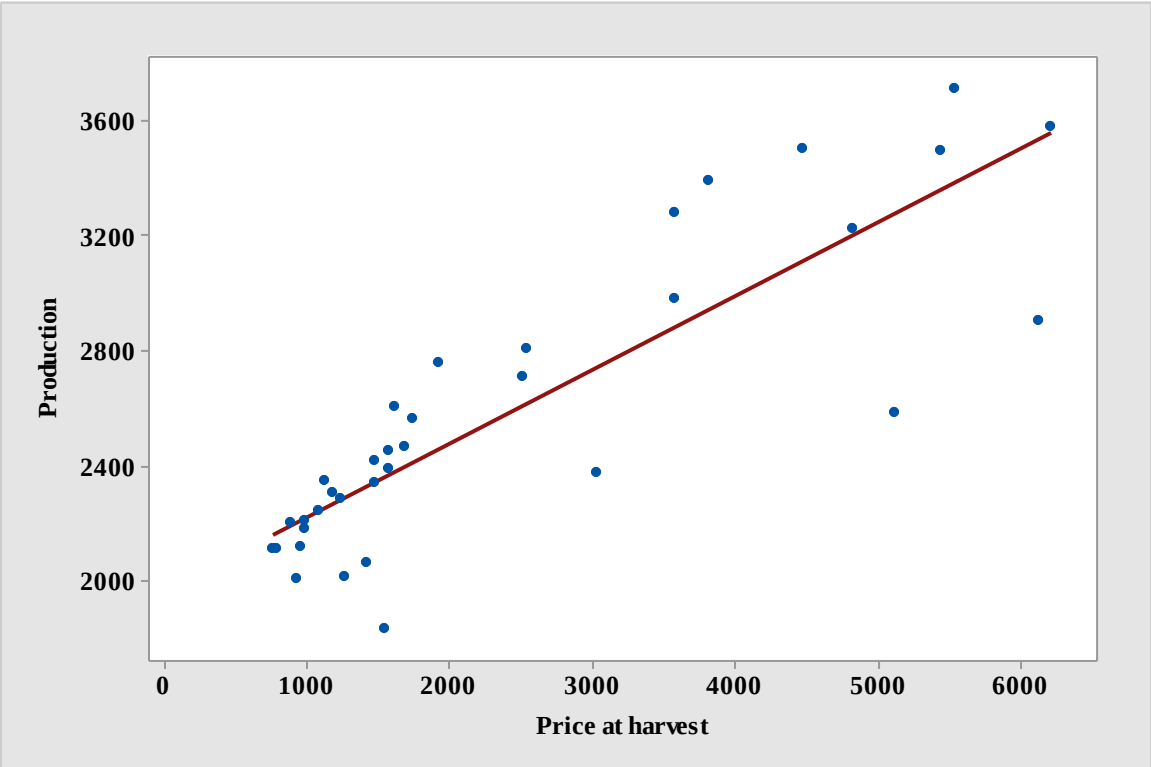
Scatter plot are subjective measures to locate the relationship between two variables. This scatter plot was drawn to indentify the relationship between the forecast variable *production* and the predictor *harvested area.* As the upward linear trend is clearly seen in the plot, this indicates an approximate linear relationship between these two variables.

Figure 2: Scatter plot of production versus rural population



This scatter plot was drawn to estimate the relationship between the forecast variable *production* and the predictor *rural population*. Again, as the plot shows an upward trend as a whole despite some discrepancy in the middle part a linear positive relation could easily be guessed between the variables.

Figure 3: Scatter plot of production versus price at harvest



This scatter plot was drawn to identify whether the forecast variable *production* was linearly related with the predictor *price at harvest* and the figure clearly showed an approximate linear relationship between the variables

Correlation matrix (Table 11) shows the significant correlation between the predictors: *harvested area* (r = 0.90, p < .001), *rural population* (r = 0.84, p < .001) and *price at harvest* (r = 0.86, p < .001) with the forecast variable *production.*

Table 11: Correlations matrix of the forecast variable and the predictors used in the study

	prodn_rice	harv_area	frmhv_price
harv_area	0.901		
	0.000		
frmhv_price	0.858	0.833	
	0.000	0.000	
rurl_popln	0.835	0.933	0.938
	0.000	0.000	0.000

Cell Contents: Pearson correlation p-value

The above table clearly shows that all predictors with forecast variable (rice production) and as well the predictors to each other have strong positive correlation with p < .001. None of the correlations are less than 0.80. This provides evidence about the relevancies and accuracy of indicators that all predictor variables were having strong relationship with the forecast variable and also between the predictor variables too. The later condition clearly indicates the sign of the presence of multicollinearity in the model.

Assumption testing

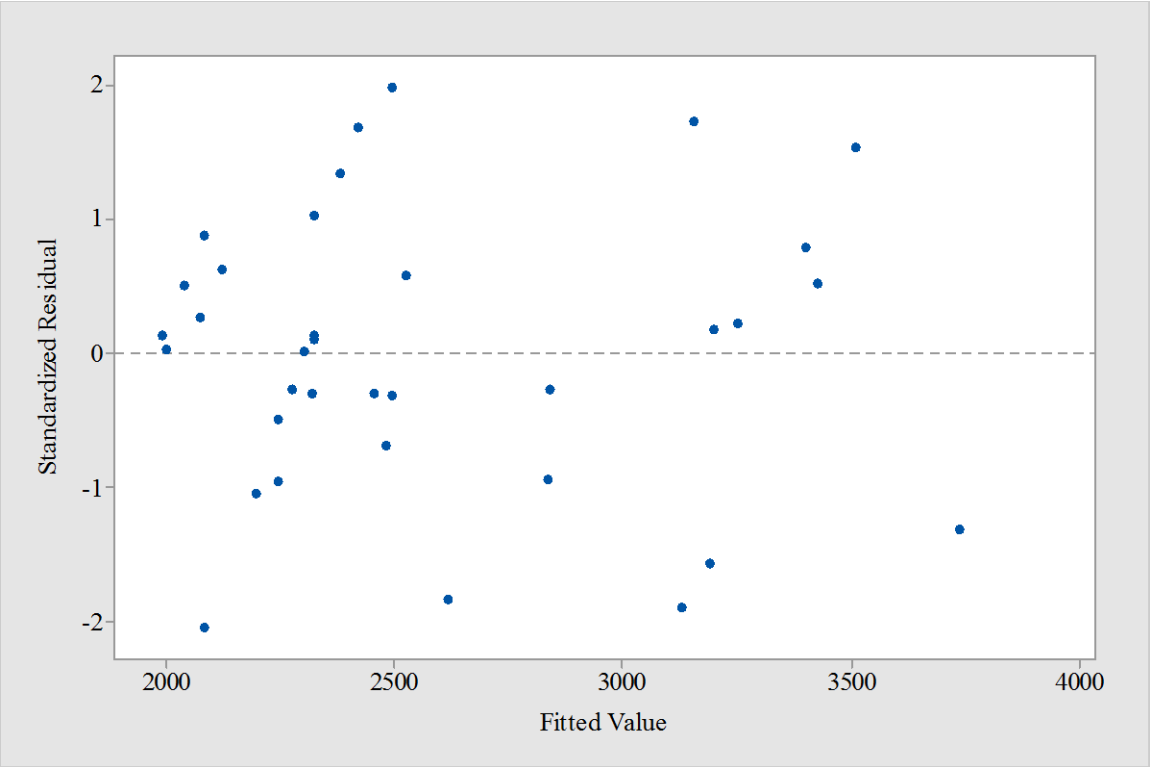
Assumption testing is essential and it should not be undermined for having reliable linear regression model. However, many a times ambiguity prevails on which methods to employ. One method we had used was regression diagnostic as the visual asses assumption testing tool. But for the reason that always this was not sufficient to rely on for a cross check , objective counterparts called the hypothesis testing methods for testing assumptions were also employed side by side.

Four basic assumptions to test were linearity, heteroscedasticity, independence and normality. According to Obsborne and Waters (2002) not all assumptions are of the same in degree to robustness in violation. Some are robust (for instance normal distribution of errors is more robust to violations) and some are easy to deal through the design of the study which researchers could easily check for substantial benefits. Specifically checking of these mentioned assumptions helps avoid Type I and II errors. A common belief is that when any of these mentioned assumptions is violated results and or the inferences yielded by a regression model would be either inefficient or badly misleading.

However, for further clarity in detail, Nau (2014) has been specific for what happens when an individual assumption is violated and has given the remedial in such cases. Violation of linearity causes problem when the fitted regression line needs extrapolation for prediction. In regression diagnostic we assess the scatter plots visually and draw inferences, accordingly whereas in the statistical tests of hypothesis about any assumption tested either it was to reject a null hypothesis set up or to fail to reject the null hypothesis and accordingly the inferences were drawn.

For linearity, plot of the residuals versus the fitted values (Figure 4) was examined. The residual in the fits plot and in the predictor variables plots did not show any particular trend. Rather the residuals were randomly startled around the horizontal line without showing any specific pattern. This indicated that the lack of fit did not hold and the assumption seemed valid.

Figure 4: Plot of residual versus fitted values



In the residual diagnostics, plots of standardized residuals against fitted values are used to check the assumption of linearity and as well the assumption of homoscedasticity in the fitted model. In the plot above the standardized residual are randomly scattered around the horizontal line with no any sign of showing some specific pattern of the residuals when the fitted values keeps on increasing. And this proves to its capacity that the assumptions; linearity and homoscedasticity are valid.

In addition to this, to test this assumption objectively, the lack of fit test was conducted in Minitab. The output came out to be "overall lack of fit test is significant at $p = .052$." (Table12). As this p-value is larger than the significance level $\alpha = 0.05$, we conclude that there was not sufficient evidence at this level of, ($p > .05$) to conclude there was lack of linear fit. Hence in overall the data in the population do not contradict in the model form.

Table: 12 Lack of fit test

Possible interaction in variable harvested area (p-value = .052)
Overall lack of fit test is significant at p = .052

Lack of fit test was conducted in Minitab. This test is the mathematical supplement to its residual plot against the fitted value and the other corresponding plots of regression diagnostic. The above result shows the test was not significant at five percent level of significance as $p > .05$

For homoscedasticity (constant variance) assumption the plot of residuals against the fitted values (Figure 4) was examined. The residual plot showed a horizontal band appearance scattered around the line zero. This suggested the spread of the error terms around 0 was not changing much as the horizontal plot value was increased and this proved the constant variance assumption hold good.

In supplement to above regression diagnostic, Bruesh-Pagon test of homoscedasticity (Table 13) named after Trevor Breusch and Andrian Pgan in Minitab with the command "estat hettest" followed by all independent variables was conducted. The result, Breusch-Pagan (3) = 9.64, (p = .02) > .01 indicated that the assumption of homogeneity of error variance was valid.

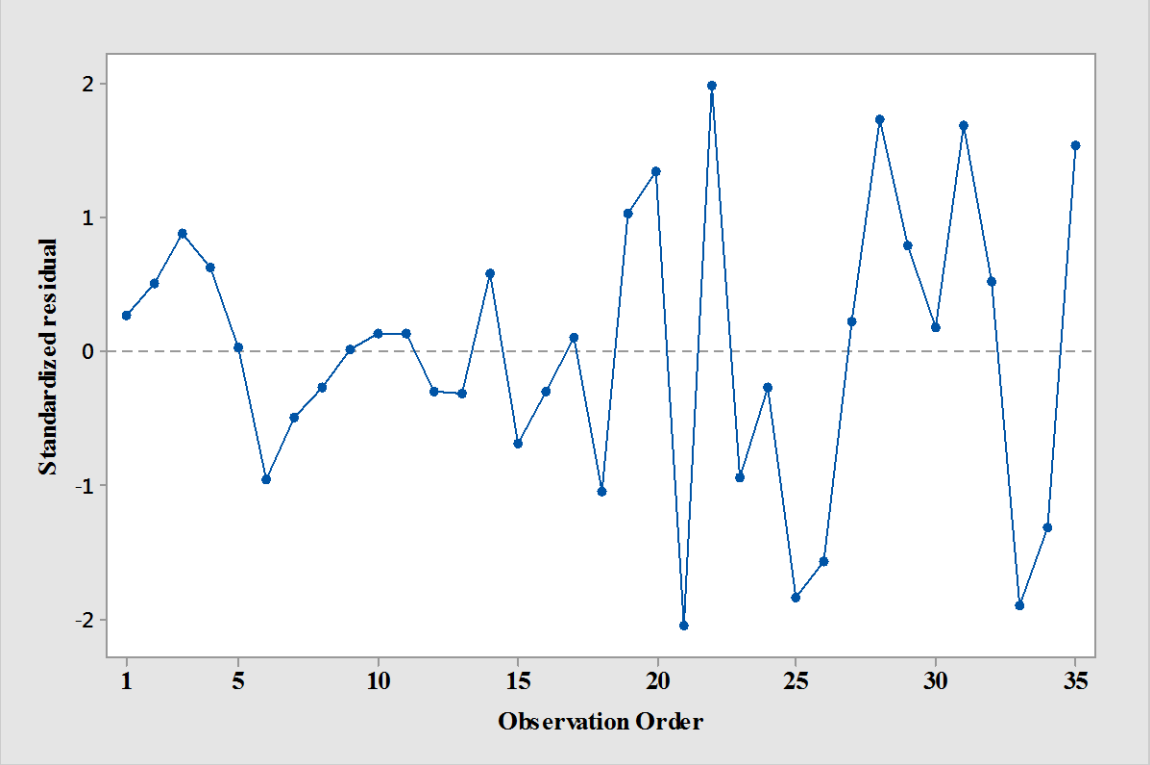
Table 13: Bruesh-Pagan Test

Breusch-Pagan / Cook-Weisberg test for heteroscedasticity Ho: Constant variance			
Variables:	harvested area	rural population	price at harvest
chi2(3)	= 9.46		
Prob > chi2	= .0238		

Breush-Pagon test of heteroscedasticity was conducted in Stata. This test is the mathematical supplement to its residual diagnostic of plot of residual against the fitted value. The above result shows the test was not significant at one percent level of significance as $p > .01$

The independence (serial correlation) assumption was tested with the plot of residuals against time (Figure 5). The plot of the time-ordered residuals displayed a random pattern the error terms have little or no autocorrelation. Reasonably it was therefore concluded that the independent assumption hold.

Figure 5: Plot of standardized residual versus observation order



The above is the order plot of residuals. This is the regression diagnostic for testing independence assumption (test of serial correlation). This plot shows a bit more alternating trend almost from the first half of the observations in order, signifying some evidence of autocorrelation in the residuals

As the mathematical supplement to this assumption testing, Durbin-Watson test (Table 14) was conducted. Durbin-Watson statistics (DW= 1.99) computed in Minitab was > the upper bound (DW_U = 1.58) of the table value of Durbin Watson statistics. This proved no correlation exists between the errors. Thant is the assumption of independence was valid.

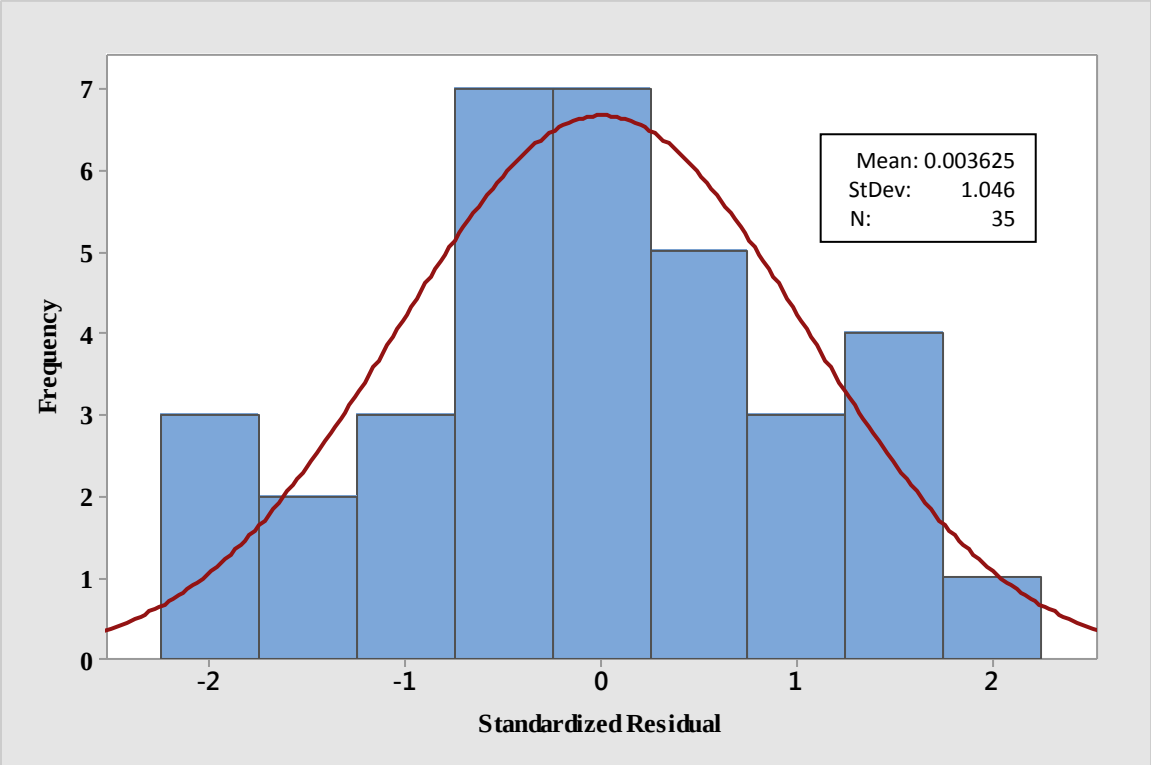
Table 14: Durbin-Watson test

Durbin-Watson statistic = 1.99511
For a 5% one-sided test of Durbin- Watson statistics : n=number of observations, k=number of parameters (so number of explanatory variables =k-1) Table value : Lower bound (DW_L :35, 3) =1.34 Upper bound (DW_U :35, 3) =1.58

For the test of independence assumption Durbin-Watson test was carried out in Minitab. The table values for the lower bound and the upper bound are taken from Makridakis *et al.* (1998, p.630). As the decision rule of the test is: If DW > upper bound, no correlation exists; if DW < lower bound, positive correlation exists; if DW is in between the two bounds, the test is inconclusive, the above result shows Durbin Watson statistics (DW=1.99) >1.58 (the upper bound) value of the Durban Watson statistics. A clear evidence of no significant autocorrelation between the residuals

At the end assumption of normality was tested. For this a histogram (figure 6) of the standardized residuals were examined. The histogram is uniform. It resembled a well bell shaped normal curve. This was the strong evidence of normality assumption was valid.

Figure 6: Histogram of standardized residuals



Above is the histogram of the residuals imposed with a normal curve. This is plotted as the test of normality assumption. The uniform histogram with a fine normal curve imposed on it is a clear evidence of the normality assumption of the residuals

Also, Shapiro-Wilk test (Table: 15) for normality was carried out in SPSS. This yielded Shapiro-Wilk (35) = 0.979, (p =.736) > .05. And this was the clear indication that the assumption of normality of error was valid too.

Table 15: Shapiro-Wilk test for normality

	Shapiro-Wilk		
	Statistic	Df	Sig.
Standardized Residual	.979	35	.736

Shapiro-Wilk test for normality was conducted in SPSS. This test is the mathematical supplement to its residual diagnostic of Histogram of the residual and or the normal probability plot of the residuals. The above result shows the test was not significant at five percent level of significance as $p > .05$

As such all four assumptions were found to be valid and no any transformation of the variables were needed.

Test of multicollinearity: Correlation matrix (Table 11) had showed the correlation amongst the predictors were quite high, a sign of multicollinearity. All three simple correlations between the predictors were not less than 0.08. They were all significant ($p < .001$) at $p = .001$.

VIFs (Table 16) however have showed a slightly different scenario. For the predictors, *harvested area* (8.69) and *price at harvest* (9.34) the VIFs were within the tolerable limit i.e. <10. However, for the predictor *rural population* the VIF (22.12) was found a bit high than it was desired.

Table 16: VIF for the predictors

Predictor	VIF
harvested area	8.658
rural population	22.112 > 10
farm harvest price	9.397

Variance influential factors for the predictors were computed from Minitab. A common rule of thumb: for any predictor $VIF > 10$ should be examined for possible multicollinearity problem

There are different schools of thoughts for deciding what to do basing on how much the VIFs to the predictors were found. One school of thought is VIFs more than 10 are taken to be significant and the predictors which resemble that are recommended for possible removal from the model. The other school of thought is

> **The rules of thumb associated with VIF (and tolerance) need to be interpreted in the context of other factors that influence the stability of the estimates of the *i*th regression coefficient. These effects can easily reduce the variance of the regression coefficients far more than VIF inflates these estimates even when VIF is 10, 20, 40, or more.** (O'brien, 2007)

In our case before we decided anything we removed the rural population from the model and saw its consequences (Table 17). The model left with the remaining two predictors i.e. *harvested area* and the *price at harvest* deteriorated. Lack of fit test was significant,

at p = .002. So no better model could be expected in the absence of rural population. This hindered in the aim of getting a multiple linear regression.

Table 17: Model after rural population was removed

	Description							
Model	S	R^2	$R^2_{(adj)}$	PRESS	$R^2_{(pred)}$	D-W statistic	Overall lack of fit test is significant at	VIF
1	139.60	93.3%	92.6%	897983	90.03%	1.96	P = 0.052	HA(8.66) RP(22.11) PH(9.40)
2	206.30	84.9%	83.9%	1674033	81.41%	1.65	P = 0.002	HA(3.26) PH(3.26)

1) Model with three predictors, harvested area, rural population and price at harvest 2) model with two predictors, harvested area and price at harvest

And for this reason we decided to sustain the model with three predictors. As multicollinearity was not much sensitive for forecasting we made this compromise and moved forward for checking the effect of outliers and the influential points in the model. During analysis (Table 18) the description of the suspicious observations, the possible outliers and or the influential points were observed.

Table 18: Description of unusual observations

Unusual observations	Studentized residual	Leverage	Cook's Distance
21	-2.04204	0.228139	0 .308127
31	1.68410	0.517597	0.760781

Above are the statistics related to locate outliers and the influential values for the suspicious observations 21, 31 in the training sample.

For this, following listed criterions were used to deal with the possible outliers and the influential points.

1) If the studentized residual for an observation is greater than 2 in absolute value, there is some evidence that the observation is an outlier with respect to *y* value

2) If the leverage value for an observation is greater than $2(k + 1)/n$, where k = number of independent variables and n = number of observation (in our case, k =3, and n = 35; so that $2(k + 1)/n = 0.23$ the observation is outlying with respect to x and

3) for the outliers which have Cook's Distances are > 1 , are the influential points

Table 19: Detecting influential observations

Observation	Studentized residual		Leverage value		Cook's Distance	
	Observed	Thresh hold	Observed	Thresh hold	Observed	Thresh hold
21	2.04	|2|	0.23	0.23	0.31	1
31	1.68	|2|	0.52	0.23	0.76	1

Essential statistics and their corresponding thresholds to detect whether or not the suspected observations were outliers and or the influential observations

During the analysis observations 21 and 31 were suspected to be the possible outlying observations. Observation 21was indicative of a large standardized residual i.e., and observation 31 was got to be one whose x value gave large leverage. As such these large standardized residual and large leverage values were checked against their respective threshold values as mentioned in the criterions and finally checked if they were influential observations or not, with Cook's distance thresh hold.

For this at first, these observations were checked back for their possible wrong recording, but were found correctly recorded. Then when looked into the table above, for observation 21: studentized residual (2.04) > 2 (however not significantly greater) and leverage value (0.23) which is not significant outlier either with respect to its y value or with that of the x value. Cook's Distance (0.31) < 1 further proved that observation 21 was not influential. Then, observation 31 studentized residual (1.68) < 2 was not an outlying due to its y value but clearly was a outlier due to its x value [leverage value (0.517) > 0.23]. But Cook's Distance (0.76) <1 showed that this too was no more the influential observation.

However, keeping in view that the criterions many a times (as they have been discussed in the theoretical section) could give malicious results, models with and without the suspected observations were computed (Table 20) and cross checked.

Table 20: Model with and without the outliers

	Description							
Model	S	R^2	$R^2_{(adj)}$	PRESS	$R^2_{(pred)}$	D-W statistic	Overall lack of fit test is significant at	VIF
Original	139.60	93.3%	92.6%	897983	90.03%	1.96 (1.58)	P = .052	HA(8.66) RP(22.11) PH(9.40)
Obs. 21 deleted	132.02	93.8%	93.2%	741806	91.19%	1.38 (1.58)	(P > = .1)	HA(9.527) RP(27.314) PH(11.460)
Obs.31 deleted	135.26	93.9%	93.3%	756545	91.60%	1.96 (1.58)	(P > = .1)	HA(16.04) RP(30.24) PH(8.78)
Obs. 21 and 31 deleted	130.69	94.1%	93.5%	702891	91.65%	1.37 (1.58)	P = .060	HA(18.95) RP(40.02) PH(11.07)

Figures in the parenthesis for the DW statistics column are the thresh hold value. If DW > upper bound (1.58) at 5% level of significance, no correlation exists between the predictor variables (Makridakis, *et al*. (1998).

And again it was confirmed that no any option of the removal, either one (this or that) or both of the observations did yield better model in the major aspects of model fitting. For instance, autocorrelation, lack of fit condition and the multicollinear situation were rather degraded in the newer model as compared negligible better measures for R-square statistics, standard error of regression and or in the PRESS statistics. And hence the model sustained as it was because the effect of the outlying observations came out to be conclusively insignificant.

4.3 Out of sample cross validation of the model

Optimization of the model through the careful check of assumption testing was followed by out of sample cross validation of the model. This meant testing of the forecast accuracy of the model with the data points which were not used for building the model. As such, the model computed at this end was subjected to forecast accuracy test using the test sample [Appendix 1(C)].

Accordingly, in the midst of the debates and the discussions about the definition of forecast accuracy measures, the study followed the most cited literature Makridakis *et al.* (1998) and Hyndman and Athanasopoulos (2014). According to which, forecast error

e is the difference between actual quantity A and the forecast F . For this study therefore, forecast error was computed using the expression:

Forecast Error: $e = A - F$

Various measures of forecast accuracy were then computed (Table 21) through Excel.

Table 21: Forecast accuracy measures

ME	MAE	MSE	MPE	MAPE	RMSE	TS
-92.7459	199.5848	50529.03	-2.16192	4.76993	224.7866	-6.97041

The forecast accuracy measures were computed using Excel

Smaller the errors are better the model is; i.e. if actual quantity is exactly the same as the forecast there would be 100% accuracy. However if error exceeds 100% the forecast accuracy tends to 0%. According to Makridakis *et al.* (1998) forecast accuracy is the accuracy of the future forecast. There are two types of accuracy measures: scale-dependent and percentage or the relative measures. Scale dependent measures cannot be used for comparison of the forecast models. Hence the later (relative measures) were invented.

According to Hyndman and Athanapolous (2014) two commonly used scale dependent forecast accuracy measures are Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). However, Makridakis *et al.* (1998) have mentioned Mean Error (ME), Mean Absolute Error (MAE), and Mean Squared Error (MSE) as the scale dependent error metrics. RMSE is the square root of MSE to comparatively ease in the interpretation.

By nature as the positive and the negative errors cancel one another, ME tends to be minimal and the ideal value of ME is 0. Hence the size of ME has not much significance except its direction leads to tell if there is systematic under – or over-forecasting. For the positive values of ME, model tends to under-forecast whereas for the negative value of ME, model tends to over-forecast. Based on this in our case as ME (-92.75) (Table 21) is negative, the model we have developed tends to over forecast.

Size of the error (smaller the better) is better measured with MAE. Other measures almost for the same are MSE (or $RMSE$). Over ME these error metrics are either more interpretable or are easier to handle mathematically. For instance recalling the negative value of ME (-92.75), with MAE (199.58) (Table 21) to a layman we could interpret the model to over-forecast, with an average absolute size of the error 199.58 units. Makridakis *et al.* (1998) reveals that MSE (or $RMSE$) are rather for statistical optimization than for the purpose of interpreting the size of the forecast error.

Relative forecast accuracy measures facilitate comparison between the models of time series that have different scales or different time intervals. According to Hyndman and Athanapoluos (2014) Mean Absolute Percentage Error ($MAPE$) is the most commonly used relative accuracy measure. However in addition to $MAPE$, Makridakis *et al.* (1998) have mentioned that Percentage Error (PE) and Mean Percentage Error (MPE) are the other commonly used relative accuracy measures. Compared to $MAPE$, MPE is less efficient metrics for comparison of the forecast models. Similar to ME because positive and negative percentage errors (PEs) tend to cancel out one another, MPE (-2.16192) is also likely to be small. Thus the values of $MPEs$ are also not much indicative to compare the different forecast models. And, to the date, $MAPE$ is considered as the most commonly used relative measure of forecast accuracy. In the case above $MAPE$ (4.76993) (Table 21) lets us to interpret that average absolute size of the error in the model investigated is approximately 5% which is quite sensible to understand even to a non-specialist than simply knowing that MSE of the model is 50529.03 and or the $RMSE$ is 224.79 . As such interpreting these relative measures makes much larger sense than absolute measures.

Bozarth (2011) has revealed that tracking signal are used to pinpoint forecasting models that need adjustment. In addition Bozarth (2011) claims that as long as -4 <TS< 4, the model is assumed working correctly. However, if otherwise some amendment is required in the model. In our case TS (-6.97) is minimally out of the required range. This indicates that the model investigated still has some room for its improvement to make it more reliable.

4.4 comparison of the model with a benchmark

Next to the completion of the test of forecast accuracy measure of the model investigated, it (the model) was compared with Naïve Forecast, a popular bench mark for model comparison. For this, observations in the test sample were the actual values and the forecast quantities were obtained through this test sample as well with the Naïve Forecast 1 approach. The forecasts were simply the most recent observations, in the past year, i.e. the forecast for the year 1996 was the actual observation of the year 1995, the forecast of the year 1997 was the actual observation of the year 1996 and so on. With these actual observations and the forecast quantities forecast error, $E = A - F$ was computed. After words, MAE and MAPE of the naïve forecast (Table 22) were computed and compared with the initially computed MAE and the MAPE of the multiple regression model investigated through above mentioned rigorous process of model building.

Table 22: Comparison of forecast models

Model	Forecast Error	
	MAE	MAPE
Naïve Forecast	240.44	5.80%
Multiple regression	199.5848	4.77%

The forecast errors MAE and MAPE useful for model comparison were computed using Excel

The model as such was cross validated and the errors were found to be minimal. However with what reference were the errors minimal? And with which other model was the model we investigated was good or bad? Popular measures for comparing a model we are Mean Absolute Error (MAE) and Mean Absolute Percentage Error ($MAPE$).

Accordingly, the task was to compare the two different models (multiple regression model and the naïve forecast model) generated from the same data set. For the multiple regression model, MAE is 199.58 (Table 22) which interprets as: the average size of the forecast error in the model regardless of its sign was nearly 200. However, when we consider the same (MAE) for the naïve forecast method to be 240.44 (Table 2), clearly we could say that the multiple regression model was better than the naïve forecast method. The mean of the absolute size of the error was greater in the bench mark model than in the multiple regression model. For comparison of the models generated through

the same data set, the absolute size of the error is more meaningful. It well decides whether or not a model was good or bad over the other model.

For the relative comparison between these models, Mean Absolute Percentage Error ($MAPE$) was put forward. For multiple regression model it was approximately 5% (Table 21) and the same (i.e. $MAPE$) for the naïve forecast method was nearly 6% (Table 22). Through these result we could see that in both cases the multiple regression model showed its superiority. And hence this was the better model than the benchmark naïve forecast method. However with a minimal difference though. This leaves us in an ambiguity if we had done a good job doing maximum work for investigating a multiple regression model which leads with the simple benchmark method just by 1% better error measure. We better think about it in the future.

4.5 Test of goodness of fit of the model

First of all our model was established with a rigorous exercise starting from the steps of variable selection up to out of sample cross validation, which were mentioned several times in the previous chapters. And because in the processes like model selection R-squared adjusted was used, in assumption testing Durbin Watson statistics was used and Cross validation had used Mean Absolute Percentage Error ($MAPE$) etc. and many more other criterions for model optimization were exploited up to here, the model at hand at this state we would say, was already plentifully a good fit. And at the end we got the model which we aspired even when the model was compared with standard benchmark, we got it superior to this. According to (Nau, 2014) the important criteria for a good regression model are (a) to make the smallest possible errors when predicting the future, and (b) to derive useful inferences from the structure of the model and the estimated values of its parameters.

However, as only these procedures are not sufficient, further we take the reference of some other objective statistical measures (Table 23) which particularly are designed for testing the model's goodness of fit. The question remained therefore was, if the model we had at hand was a good fit based on these standard measures that test model’s goodness of fit.

In addition, there are some standard measures which indicate whether or not a model is a good fit. For instance, R^2 indicates how well the model fits the data. Greater value of R^2 with as small value of standard error of regression s as possible is referred to be a good fit, whereas $PRES$ and $R^2_{(pred.)}$ help to check if the model was over fitting.

Accordingly metrics for goodness of fit statistics (Table 23) were calculated and extracted from the various regression outputs.

Table 23: Goodness of fit statistics

S	R^2	R^2 (adj.)	PRESS	R^2(pred.)
139.61	93.3%	92.6%	897983	90%

The first measure we used was the standard error of regression (se). In regression modeling, this is one of the best error statistics to look at. Smaller values of se are better because it indicates that the observations are closer to the fitted line (Frost, 2014). As such this leads us to check how wrong the regression model was on average using the units of the response variable. In this case, standard error of regression (se) was140 (Table 24) with an average 2586 ($SD = 514$) for 35 observations for the response variable. This gave a concept about how close the forecasts were to the observed values.

Some 140 of error compared to the average of 2586 of production in the response variable could be considered as minimal error, a good point of the model's goodness of fit. The reference to it according to Nau (2014) is: standard error of the regression is a lower bound on the standard error of any forecast generated from the model, however in common the forecast standard error will be a little larger as it also took into account the errors in estimating the coefficients and the relative extremeness of the values of the independent variables for which the forecast was being computed.

For R^2 value as the criterion for goodness of fit, Nau (2014) claims that good value of R^2 depends on the variable with respect to which we have measured it, and also on the decision-making context we have had. A very common concept is if we increased the number of fitted predictors in the model, R-squared will increase although the fit may

not improve in a practical sense. As a precaution to this, we have the degrees of freedom adjusted R^2 statistic called the R^2(adj.)

The interpretation of R^2 meant, 93% variance in the forecast variable (production) was explained by the three predictors included in the model (the model was parsimonious). Despite the fact that adjusted R^2 is a unitless statistic, there is no absolute standard for what is a "good" value. Together with this therefore we have included the standard error of the regression for better understanding of our models good fit. Up to now the figures we have obtained are in agreement with each other with R^2 (adj.) (nearly 93%) which is not at all deviated from the value of R^2.

Last but not the least, the R^2 (pred.) and the Predicted Sums of Square ($PRESS$) statistics are also mentioned here to support the evidence that the model we have at hand was a good fit. However, for $PRESS$ (897983) (Table 24) we do not have to do much as this is the absolute value, smaller the better. But, R^2 (pred.) is a very useful tool. This checked for the over fitting of the model. If R^2 (pred.) was very unusual (relatively very small) than the value we had for R^2, this would have questioned that the model was over fitted. The other way according to Frost (2014) is that R^2 (pred.) starts to fall as we add predictors, even if they were significant. And in our analysis we had stopped at the point when it happened during model selection. And after because we had R^2 (pred.) (90%) (Table 25) our model was a good fit in the mentioned circumstances.

Gain in predictive ability of the model

To have a check on how much gain had been there in the predictive ability due to the combination of the predictors in the model, the other check for how good the fit was, the results of part and partial correlations are presented in (Table 25).

Table 24: Zero-order partial and part correlations of the predictors

Predictors	Correlations			Partial coefficient of determination	Part coefficient of determination
	Zero-order	Partial	Part		
harvested area (000 Ha)	.901	.860	.436	74%	19%
farm harvest price (NRS/t)	.858	.803	.349	64%	12%
rural population (000people)	.835	-.746	-.290	56%	8%

The results of partial and the part (semi partial) correlations (Table 24) is the extract of one of the regression outputs (coefficients) of SPSS (Appendix 5). Last two columns in this table (Table 24) were extended afterwards as the SPSS output does not give the squared part and the partial correlations. The partial correlation presumes other predictors are kept constant where as part correlations are the correlations which presume the effect of the other predictors have been excluded out. So the partial coefficients of determinations are simply the unique contributions of the predictors. These part coefficients of determinations are helpful to identify whether or not the multiple regression used was beneficial. These coefficients of determinations (the last column in table 24) when added up (19+12+8) this means 39% of the variance in the response variable is accounted by these three predictors. And this percentage of variance in the response variable is different from the R-squared value (93%) in the model. Meaning that (93-39) 54% overlapping predictive work was done by the predictors. This proved that the combination of the variables had been quite good.

Contribution of the predictors

Finally we are also interested to have a check on the contribution of the predictors in the model. For this we computed the standardized regression equation as the followings. The standardized coefficients are brought from the regression output (Appendix) this simply compares the strengths of the predictors.

production = (1.28) * (**harvested area**) + (-1.36) (**rural population**) + (1.07) * (**farm price at harvest**)

The equation shows that the strongest predictor is the *rural population*, second strongest is the *harvested area* and the third one is *farm price at harvest*.

4.6 Interpretation of the model investigated

And at this end the model with selected appropriated predictors i.e. *production* regressed with *harvested area*; *rural population* and *price at harvest* came out to be:

production = - 1619 + (5.26)*(**harvested area**) – (0.239)*(**rural population**)
+ (0.321)*(**price at harvest**)

This is the model with unstandardized coefficients. The results (Table 25) for the model are extracted from the regression out puts, Anova and the Coefficients are straightforward. We have high p-values ($p < .001$) for the entire model and for all predictors as well.

Table 25: Regression outputs (Analysis of variance and the coefficients)

Anova		Coefficients			
F	*P*	Predictors	Coefficients	*t*	*P*
147.70	.000	Constant	-1619.3	-3.94	.000
		harvested area	5.2595	9.36	.000
		rural population	-0.23884	-6.24	.000
		price at harvest	0.32082	7.50	.000

The table suffices, overall model was significant, F (3, 34) = 147.70, p < .001. The strong evidence this revealed was: at least one of the predictors in the population was not equal to zero. Also this showed that, all three predictors (*harvested area, rural population* and *farm price at harvest*) in the model including the constant were highly significant, p < .001.

When interpret the unstandardized coefficients, for the predictor *harvested area*, it revealed: For each 000' hectare change in harvested area, there was 5.26 thousand tonnes increment in rice production. For this, however we were not estimating national productivity of rice, the figure 5.26 because in another way is simply the production of rice in tonne (t) per hectare (ha) of harvested area, it should serve nothing else but the national 'productivity' of rice. And because in the country the figure for productivity of rice for different situations and sometimes even for different time periods has ranged from 1.7 t/ha to 12 t/ha, [1.7 t/ha (Poudel, 2014); (2.6 t/ha (Uprety, 2008); 2.75 t/ha (Basnet, 2008); 2.98 t/ha (MOAC, 2011); and 12 t/ha (Mahato, 2011)] this revealed that the coefficient for harvested area which we have obtained is not a different one but a compatible figure to the surrounding scenario. A minimal difference from that of the national average (approximately 3 t/ha) might be just because this predictor (*harvested area*) here is co-integrated with the other two predictors, *rural population* and the *price at harvest* or there now is an upward shift in the figure: an indication of accumulated advancement in the country's agricultural sector.

Along with this, for every additional (000) *rural population* area in the production process, 0.239 thousand tonnes rice was diminished. This result is in agreement with the principle of diminishing marginal productivity. According to Encyclopedia Britanica (2014) this principle states that in a production process, as one input variable is increased there will be a point at which the marginal per unit output will start to decrease, holding all other factors constant.

In the above context at least for the country like Nepal where the other inputs including advancement in technology is somehow a fixed factor as compared to the increment in rural population, simply adding additional labor force to agriculture will not accommodate good in the production process and the return starts to diminish from a certain point than the previous return despite the possibility that efficient advancement in the technology was an asset for increased production. According to Pia *et al.* (2012) based on rice farms of two sample districts agricultural production can be increased by 26-33% through improving efficiency in a given technological condition.

The coefficient of rural population therefore, reveals that if the country gets more industrialized, to accommodate the surplus of the human resource in agriculture and the agricultural sector got more advanced in technology such accommodation seems to fit in the nation's development system. This will improve the productivity of the country in both ways i.e. from industrial sector and from agriculture with advanced technology. And one day the country would be one of the developed countries in the planet.

Accordingly, in the case of *price at harvest* (the remaining predictor in the model) for every additional per unit (NRS/t) *price at harvest*, some additional 0.321thousand tonnes of rice was produced. And the possible implication could be that *price at harvest* was not noticed much and future research should consider this as one of the subject to study for increased agricultural production in the country.

The intercept (i.e. the constant term) in the model has no any physical interpretation as none of the predictors can take their value near 0.

Strong points of the study

Through a rigorous and very scientific process we obtained the model and he results associated with it. These results are important mainly in three perspectives. Firstly, the results can then be applied in the real life. For instance the forecasting model can be practiced in the real life. This will replace the existing practice of crop cutting experiments and eye estimation methods for forecasting rice production ultimately a significant advancement in the process of agricultural production forecasting could be experienced. And when applied in real life the model could gain with further amendments and list of precautions investigated when to apply this.

Secondly the results will have some specific implication to the policy makers and or to the planners. They could now have the idea that the variables investigated were apparent to have strong influence in rice production forecasting in the country together with some new school of thought have been opened. So far *price at harvest* was in shed, and could not be seen much in the literatures as one of the influencing predictor in the process of rice production. Further investigations are now to be carried out in this perspective and when proved national policy could be set accordingly. Another important knowledge the results have added is the indication of negative coefficient of rural *population* in the model. May the planners now take care of the law of diminishing return and make policy in agreement with this. This should lead to go ahead for the right way of utilizing human resources in the country for its better overall development.

Thirdly, the results have their importance in the contribution of future research. In the subject area, as some of the results have new smell than before, in the future one can obtain the forecasting method using the same data set but with different approach and compare the model's efficiency in relative to this. The results have therefore created an opportunity to further investigate the related facts, to replicate and or to seek new or better methods for forecasting rice production in the country. Ultimately, the results investigated here has broken the traditional way of thinking in the agricultural production forecasting and added a new step to this.

The study was painstaking. However the positive side of this was, a parsimonious model (three predictors for 35 observations) with sufficient (93%) variance explained in the response variable (*production*) was established. The methodology of the research was clear and very well defined. Throughout the study, very authentic and the popular literatures for regression model building for forecasting [(Wheelwright *et al.* 1998); (Hyndman & Athanapoluous, 2014) (Bowerman *et al.* 2005), (Darper & Smith, 2001) etc.)] were employed.

We fitted a multiple linear regression model with time series data. The predictors in the model were very carefully chosen. The model computed is simple, easy to understand and to apply. No transformation of the variables was required hence simplicity was further enhanced while interpreting the essential elements of the model including the whole model itself. Nau (2014) mentions that for a regression model the outputs: standard error of regression, R^2 (adj.), significance of the estimated coefficients, values of the estimated coefficients, plots of fitted values and residuals, and out of sample validation is important. And, all of these outputs in our study are analyzed carefully in the due course of optimizing the model. Every bits and pieces have been done, assumptions are crosschecked with statistical hypothesis testing, out of sample cross validation of the model performed and the model had small error measures in both the estimation and validation periods compared to the bench mark method.

The model was parsimonious (three predictors) but explained relatively high (93%) percentage of the variation in the forecast variable comparably with a compatible R^2 (pred.) to avoid the suspicion of over fitting of the model. And when applied to a different sample the model did not show any loss of predictive power as we obtained the R2(adj.) (nearly 93%). And this at the same time had erased the possibility of omitted variable(s) because the predictors have already explained abundant variance. This indicates any other economic indicator variables added in the model in the name of improving the model's performance is less probable.

Lurking variables

According to Frost (2014) good rule of thumb for a model to be parsimonious was maximum of one predictor for every 10 data points. In our study we have three predictors for 35 data points which quite good fits in the rule of thumb is the other a strong point of the study. And hence there was very small chance of the presence of any lurking variable. Lurking variables are the variables not among the explanatory or response variables in a study that may influence the interpretation of relationships among the measured variables. Two variables may be correlated because both are affected by some other (measured or unmeasured) variable) in the study to hinder the model's performance in any respect.

4.7 Limitation of the study

The firs and the foremost limitation of the study was that we had to rely on secondary data. They are used as they were provided. And because of the unavailability of the data for the older years than what has been recorded here, and as there was no possibility to extend the sample size for the future years (time series data) we were limited relatively to small sample size (however by definition the sample was large (n=35)). Likewise, for the reason that we have fitted the linear regression model, results of the study have all the limitations that this procedure inherits. For example, problem in extrapolating, setting up of the forecast horizon for long term etc.

The next was, despite the study had rigorous chasing for an optimized model, yet there were some of the areas which could have deeper examination for the further advancing of the model. They include: the issue of multicollinearity, assumption of modal form, and the dealing with the unusual observations.

Similarly a variety of excuses might be found to suspect in the performance of the model. During assumption testing, however the mathematical test of hypothesis showed the marginal validity of the assumptions, the residual diagnostics in heteroscedasticity and autocorrelation assumption testing did not seem convincing and we had to have move forward along with this. The second case was, while we were dealing with the multicollinear situation in the model. Out of the three predictors the VIF of rural population was found to be out of the range of the threshold value but again we could not

remove this predictor as when tested the model got more deteriorated and the available literature suggested not to be much worried about the multicollinearity as the model was mainly developed for prediction purpose. Fitted model does not effect on the value of R^2 even in the presence of multicollinearity until and unless the estimated coefficients were on the high emphasis of the research.

And at some point when an attempt was done to check the unusual observations a similar situation rose. Minitab had signaled two observations (21 and 31) as unusual, but our analysis could not prove the observations were influential and we had to go alone with these observations while the study had to advance further. Besides this we did have to notice any feeble points in the study.

Chapter 5

SUMMARY AND CONCLUSIONS

5.1 Summary

The principal purpose of the study was to develop and optimize a multiple regression model for rice production forecasting in Nepal. For this it became necessary to locate the right trail for the research which remained as; locating of the all possible predictors; data collection and preparation for analysis; investigating of the potential predictors and; selecting of the most appropriate predictors for the model. Further steps were, testing the validity of the model, for instance assumption testing, test of multicollinearity etc.; discussing of the model's forecast accuracy situation and; comparison of the model with a benchmark. Once these fundamental steps were firmly formalized this research moved forward. This chapter reports the conclusions and the recommendations resulted from the study.

Past literature, experts' hunches, and the availability of the data, led to locate eleven possible predictors to influence in rice production. Fifty years (1961- 2010) time series data were used. Data set were divided into two samples: training sample and the test sample. This was done first to train and then to test the reliability of the model.

Automated procedures: forward selection, backward selection and stepwise selection methods were employed to the data set of the training sample to locate the list of five potential predictors out of the list of eleven possible predictors.

Table: 26 List of possible and the potential predictors

List of possible predictors	List of the potential predictors
Harvested area Farm harvest price Fertilizer consumption Number of tractors Seed consumption Annual mean rainfall Annual mean temperature Number of registered and released varieties Rural population Male labor force in agriculture Female labor force in agriculture	Harvested area Rural population Farm harvest price Male agricultural labor force Female agricultural labor force

At the end, when the five potential predictors were set for the best subset regression, this analysis successfully screened out three appropriate predictors, namely; *harvested area*, *farm price at harvest* and *rural population* to be placed in the model. Meaning that, the model that was aimed, has been purposed with these mentioned three predictors. And, this completed the variable selection stage of the study.

Next was testing of the assumptions of this purposed model. For each of these assumptions, regression diagnostic and the mathematical approach, the hypothesis test of significance were employed. The assumptions were found to be valid. Afterwards check for multicollinearity; and the effect of the outliers and the influential points in the model were conducted. The result was, none of these did signify any influence in the model at least for the forecasting purpose.

Accordingly, forecast accuracy was assessed and the model's superiority over the benchmark method, i.e. naïve forecast model, was tested. The model was over forecasting, mean error (-92.75). And, when comparison of the models was performed, with the benchmark method (naïve forecast model), the multiple regression model (i.e. the investigated model) showed its dominance in both MAE, and $MAPE$, the popular metrics for model comparison. Remarkably, the value for the Tracking Signal (-6.97) was slightly out the range $-4<TS<4$, meaning that the model was yet to upgrade.

To get continued on, a good fit, R^2 (93%) of the model was reached. Where, overall model and as well all of its predictors, *area harvested, rural population* and *price at harvest* were highly significant ($p < .001$). Model's predictive ability was (54%), R^2 (adj.) (93%) and the R^2 (pred.) (90%). Model's standardized equation showed, the first influential predictor was rural population (-1.36), the second was area harvested (1.28) and price at harvest (1.07) got to be the third strongest predictor to influence the rice production in Nepal

5.2 Conclusions

A forecast model was optimized. Out of the several potential predictors, apparently the predictors *area harvested, rural population,* and *farm price at harvest* were the more relevant to contribute in rice production forecasting in Nepal; with the predictor, *rural population,* having its diminishing coefficient. The model was cross validated with the test sample which initially was put aside. This ended up showing its superiority over the popular benchmark, the naïve forecast model. All assumptions were tested and were satisfied. The model was a good fit, with high degree of predictive ability. As per the strength of the predictors, rural population was on top influential position, the second was area harvested and price at harvest was in third position to influence rice production forecasting. The percentage variations of R^2, R^2 (adj.) and R^2 (pred.) were found close enough from each other. However, the value of the tracking signal was noticed to be moderately out of the wished range.

Given the mentioned results, this study finally has investigated a valid and optimized multiple regression model for rice production forecasting in Nepal. This model is a good fit and is concluded to be comparably more efficient than its counterpart, the naïve forecast model. With the evidence of the close values for the test of goodness of fit statistics; R^2, R^2 (adj.) and R^2 (pred.), a conclusion is drawn that, the model was not much susceptible to the sampling fluctuation, neither it was over fitting or, nor did it have any significant illusions to further investigate. However, the value of the tracking signal at the end indicated that there was some room to further improve the model.

Accordingly, the study has reached its destiny to understand the factors that are more relevant to contribute for rice production forecasting in Nepal, which came out to be *harvested area*, *rural population*, and *price at harvest*. As such a mathematical forecast model is put at hand, which now has added a turning point in the current practice of forecasting of rice production in the country with ancient approaches, the crop cutting experiments and the eye estimation method.

Consequently, a look back into the results showed some significant implications regarding the coefficients of the predictors. The coefficient of *area harvested* slightly shifted up, over the national average has indicated a possible collective advancement in the agricultural farming system. Moreover, it was interesting to locate that the predictor *price at harvest* was as one of the relevant contributor. It was important to know that as compared to other seemingly prominent predictors such as *fertilizer consumption, annual rainfall,* etc. *price at harvest* has the lead over these equivalents. This leads to have a significant implication of the model when used in real life; farmers encouraged at their farm have a positive effect in producing more rice in the country.

Similarly, one another interesting finding was the negative coefficient for the rural population. This finding also appears to contradict a common belief that more labor is employed more will be the production. Perhaps, it was because, rural population employed as the labor force to produce rice was in excess and it got to be unmanageably used. This got be in agreement with the theory the law of marginal diminishing return. Labor force results in decreased production when it continues to go further from its peak, keeping other factors constant. This suggests that country would have to accommodate it to better plan the labor force in agriculture and the possible shifting of it in the industrial sector to rise up country's revenue in multi sector basis rather than simply getting focused into agriculture and or the foreign employment sector.

Also, given that this dissertation asked to establish an optimized model for rice production forecasting, the overall approach adopted seemed to work well. The model was a best fit with its predictive ability ousting the possible presence of any lurking variables in the study. And, the value of the tracking signal which indicated further refinement of the model is considered to be the greatest asset of the study, as no any perfect situation is practicably attainable. And, obviously there were some limitations in

the study. The issue of having quality data was always there, as it was all form the secondary sources. Hence, taking up with this quality data issue, some possible alteration of the sample size scheme, and a total replication of the research in the future, should provide further confidence to use this forecast model in real life to get highly benefitted from.

5.3 Recommendation for future work

In the future, a forecast model using the same data set but with different approach could be developed and the model's efficiency could be compared. Further the various results obtained throughout the study, has created an opportunity to replicate the study as a whole or in part and to move forward and ultimately to have at hand a more improved model to better the confidence in its use.

Also, despite the fact that the model here is primarily designed for forecasting only with main effect model, there is every opportunity to future research to have other applications from the study besides forecasting. For instance, the explanation of the phenomenon and testing of the contribution of the predictors to the response variable could be validated in some more depth with some extra effort to widen the scope of the study in the same plot.

References

Aldrich, J. (2000). Figures from the history of probability and statistics. University of Southampton, UK. Retrieved from http://www.economics.soton.ac.uk/staff/aldrich/Figures.htm Accessed on 11.10.2013

Andersen, R. (2012). SOC6078 Advanced statistics: Outliers and influential cases. Department of Sociology, University of Toronto. Retrieved from http://individual.utoronto.ca/andersen/soc6708/6.DiagnosticsII.pdf Accessed on 11.01.2014

Armstrong, J.S. (2011). Illusions in regression analysis. Retrieved from http://repository.upenn.edu/cgi/viewcontent.cgi%3Farticle%3D1190%26context%3Dmarketing_pape Accessed on 11.10.2013

Baker, S.L. (2010). Simple regression theory I. Retrieved from http://hspm.sph.sc.edu/Courses/J716/pdf/716-1%2520Simple%2520Regression%2520Theory%2520I.pd...Accessed on 09.27.2014

Basnet, B.M. (2008). Environment friendly technologies for increasing rice productivity. *The Journal of Agriculture and Environment* (9). Retrieved from http://www.nepjol.info/index.php/AEJ/article/view/2114 Accessed on 10.25.2014

Best subset regression. (2014). *PENNSTATE, STAT 501-Regression Methods*. Retrieved from https://onlinecourses.science.psu.edu/stat501/node/89 Accessed on 03.24.2014

Borazth, C. (2011). Single regression approaches to forecasting : A tutorial. Retrieved from http://scm.ncsu.edu/scm-articles/article/single-regression-approaches-to-forecasting-a-tutorial Accessed on 09.13.2012

Bowerman, B. L., O'Connell, R. T., & Koehler, A.B. (2005). *Forecasting, time series, and regression: An applied approach.* (4th ed.). Thomson Brooks/Cole, 10 Davis Drive Belmont, CA 94002, USA.

Box, G.E., & Cox, D.R. (1964). What to do when data are non-normal. *Engineering statistics handbook.* Retrieved from http://itl.nist.gov/div898/handbook/pmc/section5/pmc52.htm Accessed on 09.30.2014

Buthmann, A. (2010). Making data normal using Box-Cox power transformation. *i Six Sigma*. Retrieved from http://www.isixsigma.com/tools-templates/normality/making-data-normal-using-box-cox-power-transformation/ Accessed on 05.19.2014

Chokalingam, M. (2010). Forecast accuracy and safety stock strategies. *Demand Planning LLC*. Retrieved from http://demandplanning.net/documents/dmdaccuracywebVersions.pdf Accessed on 01.15.2014

Clements, B. (2010). The absolute best way to measure forecast accuracy. *AXSIUM.* Retrieved from http://www.axsiumgroup.com/en/Axsium/Research/Blogs/2010/October/19/The%2520Absolute%2520Best%25 Accessed on 01.13.2014

Dahal, H., & Routray, J.K. (2011). Identifying association between soil and production variables using linear multiple regression models. *The journal of agriculture and environment* (12). Retrieved from www.moad.gov.np/geed/art4.pdf Accessed on 10.5.2012

Darper, N. R., & Smith, H. (1998*). Applied regression analysis*. (3rd ed.). John Wiley & Sons, Inc., USA.

David, L. (2008). Online statistics education: A multimedia course of study. Rice University, University of Houston & Tufts University. Retrieved from http://onlinestatbook.com/ Accessed on 10.06.2013

Delorme, A. (2006). Statistical methods. Swartz Center for Computational Neuroscience, INC, University of San Diego California. Retrieved from sccn.ucsd.edu/~arno/mypapers/statistics.pdf Accessed on 11.18.2013

Denis, D.J. (2000). The origins of correlation and regression: Francis Galton or Auguste Gravais and the error theorists? Paper presented at the *61st Annual Convention of the Canadian Psychological Association.* Retrieved from http://www.psychology.sunysb.edu/ewaters/301/11_correlation_coef/origin_corr.pdf Accessed on 11.10.2013

DHM. (2010). Government of Nepal, Department of Hydrology and Meteorology.

Encyclopedia Britanica. (2014). Diminishing returns. Retrieved from http://www.britannica.com/EBchecked/topic/163723/diminishing-returns Accessed on 10.27.2014

FAO. (2013). FAOSTAT. Retrieved from http://faostat.fao.org/site/567/DesktopDefault.aspx?PageID=567#ancor Accessed on 14.03.2013

Fienberg, S.E. (1992). A brief history of statistics in three and one-half chapters: A review essay. *Statistical Science 7*(2), 208-225. Retrieved from JSTOR http://www.jstor.org/discover/10.2307/2246307?uid=3738752&uid=2&uid=4&sid=21102913162817 Accessed on 11.13.2013

Flizmoser, P. (2008). Linear and nonlinear methods for regression and classification and applications in R. Department of Statistics and Probability Theory, Vienna University of Technology, Vienna, Austria. Retrieved from www.statistik.tuwien.ac.at/forschung/CS/CS-2008-3complete.pdf Accessed on 10.2.2013

Frost, J. (2013). Multiple regression analysis: Use of adjusted R-squared and predicted R-squared to include the correct number of variables. *The Minitab blog*. Retrieved from http://blog.minitab.com/blog/adventures-in-statistics/multiple-regession-analysis-use-adjusted-r-squared-and-predicted-r-squared-to-include-the-correct-number-of-variables Accessed on 10.28.2014

Frost, J. (2014). Regression analysis: How to interpret *s*, the standard error of the regression. The Minitab blog. Retrieved from http://blog.minitab.com/blog/adventures-in-statistics/regression-analysis-how-to-interpret-s-the-standard-error-of-the-regression Accessed on 10.29.2014

Galton, F. (1886). Regression towards mediocrity in hereditary stature .*The Journal of the Anthropological Institute of Great Britain and Ireland, 15* (1886), 246-263. Royal Anthropological Institute of Great Britain and Ireland. Stable URL: http://www.jstor.org/stable/2841583. Retrieved from http://www.jstor.org/stable/2841583 Accessed on 29.03.2012

Gommes, R. (2001). An introduction to the art of agrometeorological crop yield forecasting using multiple regression. *Crop Monitoring and Forecasting Group. Crop Yield Forecasting and Agrometeorology.* Sub-Project UTF/BGD/029, ASIRP/DAE Dhaka. Retrieved from http://www.fao.org/nr/climpag/pub/Crop%2520Yield%2520Forecasting%25202001%2520Gommes.pdf Accessed on 11.17.2013

Guenthner, J.F. (1992). Forecasting annual vegetable plantings. *Hort Technology 2 (1).* Department of Agricultural Economics; University of Indaho; Moscow, ID 83483. Retrieved from http://horttech.ashspublications.org/content/2/1/89.abstract Accessed on 03.27.2013

Hall, S. (2015). The advantages of regression analysis and forecasting. *Demand Media*. Retrieved from http://smallbusiness.chron.com/advantages-regression-analysis-forecasting-61800.html Accessed on 09.11.2015

Hastie, T., Tibshirani, R., & Friedman, J . (2009). Linear methods for regression. *The Elements of Statistical Learning. Springer Series in Statistics*, 43-99. Retrieved from http://link.springer.com/chapter/10.1007%2F978-0-387-84858-7_3 Accessed on 06.21.2014

Hofmarcher, P. (2012). *Advanced regression methods in finance and economics: Three essays.* Doctoral thesis. WU Vienna University of Economics and Business. Retrieved from http://epub.wu.ac.at/3489/ Accessed on 10.04.2014

Hyndman, R.J., & Athanasopoulos, G. (2014). *Forecasting: Principles and practice , a online text book*. Retrieved from http://otexts.com/fpp/ Accessed on 02.05.2014

IRRI. (2014).*World Rice Statistics Online Query Facility*. International Rice Research Institute. Retrieved from http://ricestat.irri.org:8080/wrs2/entrypoint.htm Accessed on 06.09.2014

Jacoby, W. G. (n.d.). Regression III: Advanced methods. Michigan State University. Retrieved from http://polisci.msu.edu/jacoby/icpsr/regress3/lectures/week3/11.Outliers.pdf Accessed on 03.20.2014

Kandane, J. B.; & Lazar, N. A. (n.d.). Methods and criteria for model selection. Department of statistics Carnegie Mellon University Pittsburgh. Retrieved from https://www.cs.cmu.edu/~tom/10-702/tr759.pdf Accessed on 03.24.2014

Karim, E. M. (n.d.). Selection of best regression equation by sorting out variables. Institute of statistical research and training: University of Dhaka Bagaladesh. Retrieved from http://www.angelfire.com/ab5/get5/selreg.pdf Accessed on 09.30.2014

Kutsurelis, J. E. (1998). Forecasting financial markets using neural networks: An analysis of methods and accuracy. United States Naval Academy, Department of System Management. Retrieved from http://www.wardsystems.com/financial%2520forecasting%2520thesis%2520kutsurelis.pdf Accessed on 10.15.2012

Lobel, D.B., Cahill, K.N.; & Fied , C.B. (2007). Historical effects of temperature and precipitation on Californian crop yields. *Climatic Change 81*(2), 87-203. Retrieved from link.springer.com/article/10.1007%2Fs10584-006-9141-3 Accessed on 10.02.2013

Mahato, R. (2011). Nepal's hunger solution. *Nepali TIMES* (559). Retrieved from http://nepalitimes.com/news.php?id=18312#.VEnaNvkozcg Accessed on 10.24. 2014

Makridakis, S., Wheelwritht, S.C., & Hyndman, R. J. (1998). Forecasting methods and application. (3rd ed.). John Wiley & Sons, Inc.

Markidakis, S.,& Whelwight S. C. (1978). *Forecasting methods and application.* Santa Barbara, New York: John Wiley and Sons, Inc.

McDonald, J.H.(2009). *Handbook of biological statistics.* (2nd ed.). Sparky House Publishing, Baltimore, Maryland. Retrieved from http://udel.edu/~mcdonald/stattransform.html Accessed on 05.19.2014

Minitab (Version 17). (2010). Statistical Software. State College, PA: Minitab, Inc.

MOAC. (2010). Statistical information on Nepalese agriculture. Ministry of Agriculture and Co-operatives, Agri-Business Promotion and Statistics Division [ABPSD]

Muhammad , & Abdulah. (2013). Modelling and forecasting paddy production in Kelantan under the implementation of system of rice intensification (SRI). *Academic Journal of Agricultural Research 1* (7), 106-113. ISSN: 2315-7739. Retrieved from http://www.academiapublishing.org/ajar Accessed on 13.07.2013

Multicollinearity. (2015). University of Notre Dame. Retrieved from https://www3.nd.edu/~rwilliam/stats2/l11.pdf Accessed on 02. 13. 2015

Mundry, R., & Nunn, C. L. (2009). Stepwise model fitting and statistical inference: Turning noise into signal pollution. *American Naturalist, 173*(1), 119-123. doi:10.1086/593303

NASS. (2012). The yield forecasting and estimating program of NASS [National Agriculture Statistics Service]. Statistical Methods Branch, Statistics Division, National Agricultural Statistics Service, U.S. Department of Agriculture, Washington, D.C. NASS staff report No. SMB 12-01. Retrieved from http://www.nass.usda.gov/Education_and_Outreach/Understanding_Statistics/Yield_Forcasting_Progr...Accessed on 10.05.2013

Nau, R. (2014). What is a good value of R-squared? Statistical forecasting: Notes on regression and time series analysis. Fuqua School of Business, Duke University. Retrieved from http://people.duke.edu/~rnau/rsquared.htm Accessed on 10.23.2014

Nayava, J.L. (2012). Variations of rice yield with rainfall in Nepal. In J.L. Nayava (Ed), *Climates of Nepal and their implication.*WWF report, 121-130. Retrieved from assets.panda.org/downloads/climates_of_nepal.pdf Accessed on 09.27.2014

O'brien, F.M. (2007). A caution regarding rules of thumb for variance inflation factors. *Quality and Quantity* (41), pp. 673-690. doi: 10. 10.1007/s11135-006-9018-6

Osborne, J.W., & Amy, O. (2004). The power of outliers (and why researchers should always check for them). Practical Assessment, Research & Evaluation. North Carolina State University. Retrieved from pareonline.net/getvn.asp?v=9&n=6 Accessed on 10.7.2013

Paul, R. K. (2004). Multicollinearity: Causes, effects and remedies. I.A.S.R.I. Library Avenue, New Delhi-110012. Retrieved from www.iasri.res.in/.../3.%20**multicollinearity**-%20**causes**,**effects**%20and%2... Accessed on 03.26.2014

Pia. S., Kiminami, A., & Yagi, H. (2012). Comparing the technical efficiency of rice farms in urban and rural areas: A case study from Nepal. *Science Direct*. doi: 10.3923/tae.2012.48.60

Poudel, M.N. (2014). Rice (*Oryza sativa* L) cultivation in the highest elevation of the world. *Nepal Journals Online.* Retrieved from http://www.nepjol.info/index.php/AJN/article/download/7519/6105 Accessed on 10.25.2014

Prajneshu. (n.d.). Non linear regression models and their applications. Indian Agricultural Statistics Research Institute. Library Avenue, New Delhi. Retrieved from http://www.iasri.res.in/ebook/EB_SMAR/e-book_pdf%2520files/Manual%2520IV/1Nonlinear%2520Re...Accessed on 09.27.2014

Preston, J. (2014). Measuring forecast accuracy. Retrieved from http://faculty.mu.edu.sa/download.php%3Ffid%3D78887 Accessed on 01.13.2014

Ramasubramanian, V. (n.d.). Forecasting techniques in agriculture. *Indian Agricultural Statistics Research Institute.* Library Avenue, New Delhi-110 012. Retrieved from http://www.iasri.res.in/ebook/eb_smar/e-book_pdf%2520files/manual%2520iv/2-forecasting%2520techn... Accessed on 10.5.2014

Ramirez, O.A., & Fadiga, M. (2003). Forecasting agricultural commodity prices with asymmetric-error GARCH models. *Journal of Agricultural and Resource Economics 28* (1), 71-85. Western Agricultural Economics Association. Retrieved from http://ageconsearch.umn.edu/bitstream/30714/1/28010071.pdf Accessed on 03.5.2014

Richa. (2015). Regression analysis: Find the best fit for your statistical model. Retrieved from https://blog.udemy.com/regression-analysis/ Accessed on 09.11.2015

Ristanoski , G., Liu, W., & Bailey, J. (2013). Time series forecasting using distribution enhanced linear regression. Retrieved from www.goceristanoski.com/uploads/2/0/8/9/.../pakdd2013ristanoski.pdf Accessed on 0 9.8.2013

Shabri, A., Samsudin, R., & Ismail, Z. (2009). Forecasting of the rice yields time series forecasting using artificial neural network and statistical model. *Journal of Applied Sciences*, (9), 4168-4173. doi: 10.3923/jas.2009.4168.4173

SPSS (Version 20). (2011). IBM SPSS Statistics for Windows. IBM Corp, Armonk, NY

Stanton, M. J. (2001). Galton, Pearson, and the Peas: A brief history of linear regression for statistics instructors. *Journal of Statistics Education, 9*(3). Retrieved from http://www.amstat.org/publications/jse/v9n3/stanton.html Accessed on 11.10.2013

Stata (Version 12). (2011). Stata Statistical Software. College Station TX: StataCorp LP

Stepwise multiple regression. (n.d.). Retrieved from www.utexas.edu/.../Solving%20Stepwise%20Regression%20Problems.pp Accessed on 17.03.2014

Talk Stats. (2013). Retrieved from http://www.talkstats.com/showthread.php/10617-What-is-the-history-of-regression analysis Accessed on 11.10.2015

Taylor, J. & Cobb, K. (2004). Model selection. Statistics 262: Intermediate biostatistics. retrieved from http://statweb.stanford.edu/~jtaylo/courses/stats262/spring.2004/notes/week9.pdf Accessed on 09.13.2014

The brief history of statistics. (2000). Retrieved from folk.uib.no/ngbnk/kurs/notes/node4.html Accessed on 10.13.2013

The MathWorks. (2013). Time series regression VII: Forecasting. Retrieved from www.mathworks.com/help/econ/examples/time-series -regression-vii-forexasting.html Accessed on 26.8.2013

Ulbrich, N., & Volden, T. (2010). Regression model term selection for the analysis of strain-gage balance calibration data. *Twenty seventh AIAA Aerodynamic Measurement Technology and Ground Testing Conference.* Chicago, Illinois. AIAA 2010-4545. Retrieved from http://enu.kz/repository/2010/AIAA-2010-4545.pdf Accessed on 04.05.2014

Uprety, R. (2008). Growing much more rice. *Nepali TIMES* (442). Retrieved from http://nepalitimes.com/news.php?id=18312#.VEnaNvkozcg Accessed on 10.24. 2014

Walonick, D. S. (1993). An overview of forecasting methodology. Stat Pack 1993.Retrieved from http://statpac.org/research-library/forecasting.htm Accessed on 09.11. 2015

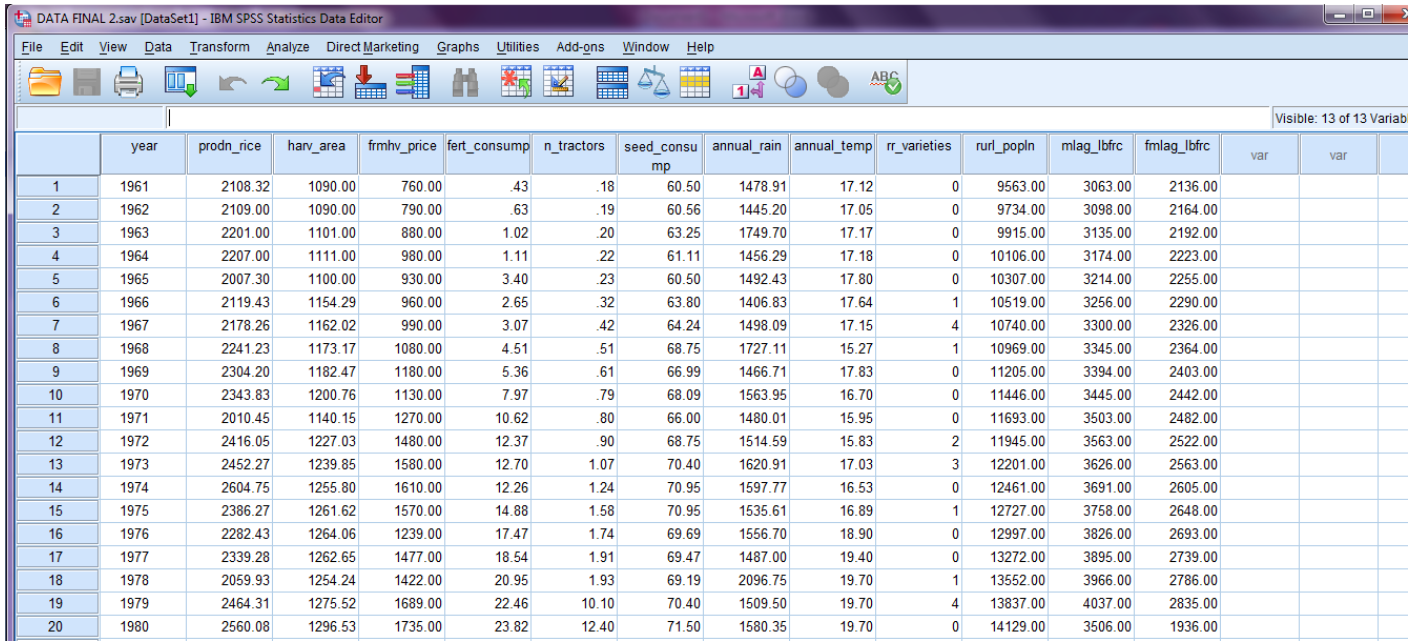
Wikipedia the free encyclopedia. (2013). Regression *analysis.* Retrieved from http://en.wikipedia.org/wiki/Regression_analysis Accessed on 11.10.2013

Young, D. (2013). STAT 501. Retrieved from https://onlinecourses.science.psu.edu/stat501/node/46 Accessed on 04.06.2014

Zhang,W., Qi. Y., & Lui, Y. (2004). Forecasting trend of rice production of the world and regions. Australian society of agronomy. Retrieved from http://www.regional.org.au/au/asa/2004/poster/0/1178_zhangwjok.htm Accessed on 10.06.2012

APPENDIX

Appendix 1(A): Total sample


DATA FINAL 2.sav [DataSet1] - IBM SPSS Statistics Data Editor

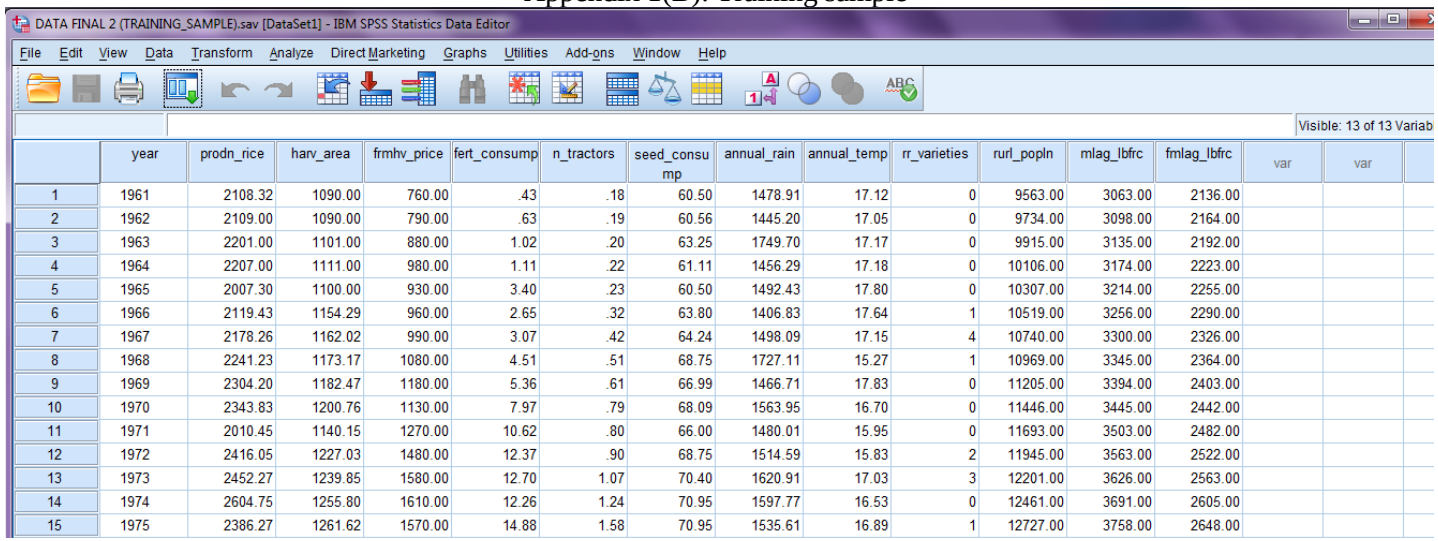
File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help

Visible: 13 of 13 Variab

	year	prodn_rice	harv_area	frmhv_price	fert_consump	n_tractors	seed_consump	annual_rain	annual_temp	rr_varieties	rurl_popln	mlag_lbfrc	fmlag_lbfrc	var	var
1	1961	2108.32	1090.00	760.00	.43	.18	60.50	1478.91	17.12	0	9563.00	3063.00	2136.00		
2	1962	2109.00	1090.00	790.00	.63	.19	60.56	1445.20	17.05	0	9734.00	3098.00	2164.00		
3	1963	2201.00	1101.00	880.00	1.02	.20	63.25	1749.70	17.17	0	9915.00	3135.00	2192.00		
4	1964	2207.00	1111.00	980.00	1.11	.22	61.11	1456.29	17.18	0	10106.00	3174.00	2223.00		
5	1965	2007.30	1100.00	930.00	3.40	.23	60.50	1492.43	17.80	0	10307.00	3214.00	2255.00		
6	1966	2119.43	1154.29	960.00	2.65	.32	63.80	1406.83	17.64	1	10519.00	3256.00	2290.00		
7	1967	2178.26	1162.02	990.00	3.07	.42	64.24	1498.09	17.15	4	10740.00	3300.00	2326.00		
8	1968	2241.23	1173.17	1080.00	4.51	.51	68.75	1727.11	15.27	1	10969.00	3345.00	2364.00		
9	1969	2304.20	1182.47	1180.00	5.36	.61	66.99	1466.71	17.83	0	11205.00	3394.00	2403.00		
10	1970	2343.83	1200.76	1130.00	7.97	.79	68.09	1563.95	16.70	0	11446.00	3445.00	2442.00		
11	1971	2010.45	1140.15	1270.00	10.62	.80	66.00	1480.01	15.95	0	11693.00	3503.00	2482.00		
12	1972	2416.05	1227.03	1480.00	12.37	.90	68.75	1514.59	15.83	2	11945.00	3563.00	2522.00		
13	1973	2452.27	1239.85	1580.00	12.70	1.07	70.40	1620.91	17.03	3	12201.00	3626.00	2563.00		
14	1974	2604.75	1255.80	1610.00	12.26	1.24	70.95	1597.77	16.53	0	12461.00	3691.00	2605.00		
15	1975	2386.27	1261.62	1570.00	14.88	1.58	70.95	1535.61	16.89	1	12727.00	3758.00	2648.00		
16	1976	2282.43	1264.06	1239.00	17.47	1.74	69.69	1556.70	18.90	0	12997.00	3826.00	2693.00		
17	1977	2339.28	1262.65	1477.00	18.54	1.91	69.47	1487.00	19.40	0	13272.00	3895.00	2739.00		
18	1978	2059.93	1254.24	1422.00	20.95	1.93	69.19	2096.75	19.70	1	13552.00	3966.00	2786.00		
19	1979	2464.31	1275.52	1689.00	22.46	10.10	70.40	1509.50	19.70	4	13837.00	4037.00	2835.00		
20	1980	2560.08	1296.53	1735.00	23.82	12.40	71.50	1580.35	19.70	0	14129.00	3506.00	1936.00		

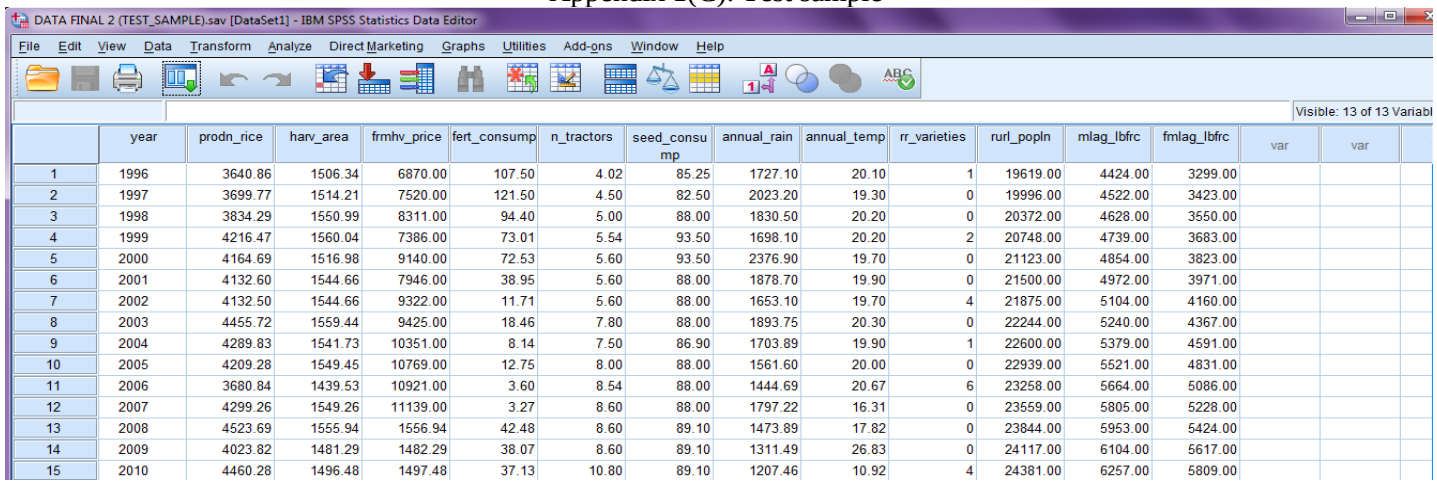
33	1993	2906.18	1368.42	6132.00	93.00	26.30	80.30	1806.55	19.80	0	18491.00	4190.00	2939.00		
34	1994	3578.83	1496.79	6208.00	93.70	3.20	85.25	1634.90	19.90	3	18865.00	4258.00	3057.00		
35	1995	3710.65	1511.23	5540.00	103.00	3.60	85.25	1625.85	20.00	0	19242.00	4336.00	3177.00		
36	1996	3640.86	1506.34	6870.00	107.50	4.02	85.25	1727.10	20.10	1	19619.00	4424.00	3299.00		
37	1997	3699.77	1514.21	7520.00	121.50	4.50	82.50	2023.20	19.30	0	19996.00	4522.00	3423.00		
38	1998	3834.29	1550.99	8311.00	94.40	5.00	88.00	1830.50	20.20	0	20372.00	4628.00	3550.00		
39	1999	4216.47	1560.04	7386.00	73.01	5.54	93.50	1698.10	20.20	2	20748.00	4739.00	3683.00		
40	2000	4164.69	1516.98	9140.00	72.53	5.60	93.50	2376.90	19.70	0	21123.00	4854.00	3823.00		
41	2001	4132.60	1544.66	7946.00	38.95	5.60	88.00	1878.70	19.90	0	21500.00	4972.00	3971.00		
42	2002	4132.50	1544.66	9322.00	11.71	5.60	88.00	1653.10	19.70	4	21875.00	5104.00	4160.00		
43	2003	4455.72	1559.44	9425.00	18.46	7.80	88.00	1893.75	20.30	0	22244.00	5240.00	4367.00		
44	2004	4289.83	1541.73	10351.00	8.14	7.50	86.90	1703.89	19.90	1	22600.00	5379.00	4591.00		
45	2005	4209.28	1549.45	10769.00	12.75	8.00	88.00	1561.60	20.00	0	22939.00	5521.00	4831.00		
46	2006	3680.84	1439.53	10921.00	3.60	8.54	88.00	1444.69	20.67	6	23258.00	5664.00	5086.00		
47	2007	4299.26	1549.26	11139.00	3.27	8.60	88.00	1797.22	16.31	0	23559.00	5805.00	5228.00		
48	2008	4523.69	1555.94	1556.94	42.48	8.60	89.10	1473.89	17.82	0	23844.00	5953.00	5424.00		
49	2009	4023.82	1481.29	1482.29	38.07	8.60	89.10	1311.49	26.83	0	24117.00	6104.00	5617.00		
50	2010	4460.28	1496.48	1497.48	37.13	10.80	89.10	1207.46	10.92	4	24381.00	6257.00	5809.00		

Appendix 1(B): Training sample



	year	prodn_rice	harv_area	frmhv_price	fert_consump	n_tractors	seed_consump	annual_rain	annual_temp	rr_varieties	rurl_popln	mlag_lbfrc	fmlag_lbfrc	var	var
1	1961	2108.32	1090.00	760.00	.43	.18	60.50	1478.91	17.12	0	9563.00	3063.00	2136.00		
2	1962	2109.00	1090.00	790.00	.63	.19	60.56	1445.20	17.05	0	9734.00	3098.00	2164.00		
3	1963	2201.00	1101.00	880.00	1.02	.20	63.25	1749.70	17.17	0	9915.00	3135.00	2192.00		
4	1964	2207.00	1111.00	980.00	1.11	.22	61.11	1456.29	17.18	0	10106.00	3174.00	2223.00		
5	1965	2007.30	1100.00	930.00	3.40	.23	60.50	1492.43	17.80	0	10307.00	3214.00	2255.00		
6	1966	2119.43	1154.29	960.00	2.65	.32	63.80	1406.83	17.64	1	10519.00	3256.00	2290.00		
7	1967	2178.26	1162.02	990.00	3.07	.42	64.24	1498.09	17.15	4	10740.00	3300.00	2326.00		
8	1968	2241.23	1173.17	1080.00	4.51	.51	68.75	1727.11	15.27	1	10969.00	3345.00	2364.00		
9	1969	2304.20	1182.47	1180.00	5.36	.61	66.99	1466.71	17.83	0	11205.00	3394.00	2403.00		
10	1970	2343.83	1200.76	1130.00	7.97	.79	68.09	1563.95	16.70	0	11446.00	3445.00	2442.00		
11	1971	2010.45	1140.15	1270.00	10.62	.80	66.00	1480.01	15.95	0	11693.00	3503.00	2482.00		
12	1972	2416.05	1227.03	1480.00	12.37	.90	68.75	1514.59	15.83	2	11945.00	3563.00	2522.00		
13	1973	2452.27	1239.85	1580.00	12.70	1.07	70.40	1620.91	17.03	3	12201.00	3626.00	2563.00		
14	1974	2604.75	1255.80	1610.00	12.26	1.24	70.95	1597.77	16.53	0	12461.00	3691.00	2605.00		
15	1975	2386.27	1261.62	1570.00	14.88	1.58	70.95	1535.61	16.89	1	12727.00	3758.00	2648.00		
25	1985	2372.02	1333.36	3030.00	45.05	2.42	79.75	1715.20	19.60	0	15685.00	3785.00	2234.00		
26	1986	2981.78	1423.29	3580.00	54.18	2.51	82.50	1547.75	19.60	0	16014.00	3836.00	2303.00		
27	1987	3283.21	1450.47	3580.00	56.29	2.59	82.50	1614.30	20.40	9	16349.00	3889.00	2375.00		
28	1988	3389.67	1432.85	3820.00	67.38	2.69	79.75	1824.35	19.90	0	16689.00	3941.00	2451.00		
29	1989	3502.16	1455.17	4470.00	72.51	2.77	82.50	1777.00	19.50	0	17037.00	3989.00	2532.00		
30	1990	3222.54	1411.81	4820.00	81.10	2.78	78.10	1810.50	19.60	3	17392.00	4032.00	2621.00		
31	1991	2584.90	1262.11	5121.00	82.00	21.70	77.00	1449.60	19.70	3	17753.00	4068.00	2717.00		
32	1992	3495.59	1450.45	5440.00	73.54	24.00	82.50	1351.05	19.40	0	18120.00	4129.00	2825.00		
33	1993	2906.18	1368.42	6132.00	93.00	26.30	80.30	1806.55	19.80	0	18491.00	4190.00	2939.00		
34	1994	3578.83	1496.79	6208.00	93.70	3.20	85.25	1634.90	19.90	3	18865.00	4258.00	3057.00		
35	1995	3710.65	1511.23	5540.00	103.00	3.60	85.25	1625.85	20.00	0	19242.00	4336.00	3177.00		

Appendix 1(C): Test sample

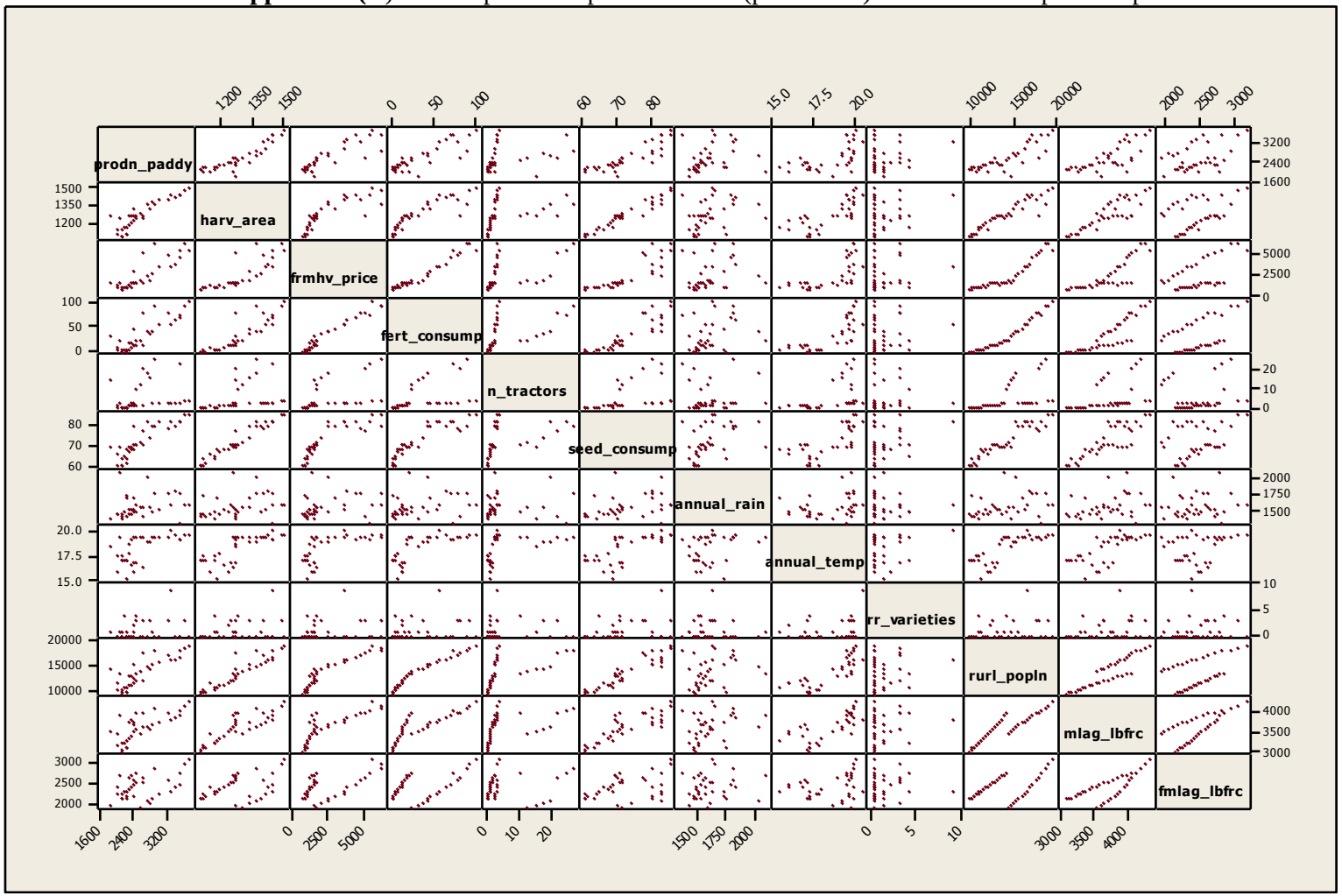

DATA FINAL 2 (TEST_SAMPLE).sav [DataSet1] - IBM SPSS Statistics Data Editor

File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help

Visible: 13 of 13 Variabl

	year	prodn_rice	harv_area	frmhv_price	fert_consump	n_tractors	seed_consump	annual_rain	annual_temp	rr_varieties	rurl_popln	mlag_lbfrc	fmlag_lbfrc	var	var
1	1996	3640.86	1506.34	6870.00	107.50	4.02	85.25	1727.10	20.10	1	19619.00	4424.00	3299.00		
2	1997	3699.77	1514.21	7520.00	121.50	4.50	82.50	2023.20	19.30	0	19996.00	4522.00	3423.00		
3	1998	3834.29	1550.99	8311.00	94.40	5.00	88.00	1830.50	20.20	0	20372.00	4628.00	3550.00		
4	1999	4216.47	1560.04	7386.00	73.01	5.54	93.50	1698.10	20.20	2	20748.00	4739.00	3683.00		
5	2000	4164.69	1516.98	9140.00	72.53	5.60	93.50	2376.90	19.70	0	21123.00	4854.00	3823.00		
6	2001	4132.60	1544.66	7946.00	38.95	5.60	88.00	1878.70	19.90	0	21500.00	4972.00	3971.00		
7	2002	4132.50	1544.66	9322.00	11.71	5.60	88.00	1653.10	19.70	4	21875.00	5104.00	4160.00		
8	2003	4455.72	1559.44	9425.00	18.46	7.80	88.00	1893.75	20.30	0	22244.00	5240.00	4367.00		
9	2004	4289.83	1541.73	10351.00	8.14	7.50	86.90	1703.89	19.90	1	22600.00	5379.00	4591.00		
10	2005	4209.28	1549.45	10769.00	12.75	8.00	88.00	1561.60	20.00	0	22939.00	5521.00	4831.00		
11	2006	3680.84	1439.53	10921.00	3.60	8.54	88.00	1444.69	20.67	6	23258.00	5664.00	5086.00		
12	2007	4299.26	1549.26	11139.00	3.27	8.60	88.00	1797.22	16.31	0	23559.00	5805.00	5228.00		
13	2008	4523.69	1555.94	1556.94	42.48	8.60	89.10	1473.89	17.82	0	23844.00	5953.00	5424.00		
14	2009	4023.82	1481.29	1482.29	38.07	8.60	89.10	1311.49	26.83	0	24117.00	6104.00	5617.00		
15	2010	4460.28	1496.48	1497.48	37.13	10.80	89.10	1207.46	10.92	4	24381.00	6257.00	5809.00		

Appendix 2(A): Matrix plot of response variable (production) with the eleven possible predictors



Appendix 2(B): Correlation matrix for the eleven possible predictors

		Prodn (tonnes)FAO	Farm harvest price of rough rice (Local currency/t)	Rural population-FAO (based on UN2009) (000 person)	Male labor force in agriculture-FAO (based on ILO	Female labor force in agriculture-FAO (based on ILO	Number of tractors (000 tractors)	Total fertilizer consumption from chemical sources (000 t)	Seed (tonnes)FAO
Prodn (tonnes)FAO	Pearson Correlation	1	.837	.817	.720	.455	.838	.837	.755
	Sig. (2-tailed)		.000	.000	.000	.006	.000	.000	.000
	N	35	35	35	35	35	35	35	35
Farm harvest price of rough rice (Local currency/t)	Pearson Correlation	.837	1	.929	.814	.568	.962	.978	.851
	Sig. (2-tailed)	.000		.000	.000	.000	.000	.000	.000
	N	35	35	35	35	35	35	35	35
Rural population-FAO (based on UN2009) (000 person)	Pearson Correlation	.817	.929	1	.915	.500	.984	.969	.939
	Sig. (2-tailed)	.000	.000		.000	.002	.000	.000	.000
	N	35	35	35	35	35	35	35	35
Male labor force in agriculture-FAO (based on ILO	Pearson Correlation	.720	.814	.915	1	.740	.888	.856	.845
	Sig. (2-tailed)	.000	.000	.000		.000	.000	.000	.000
	N	35	35	35	35	35	35	35	35
Female labor force in agriculture-FAO (based on ILO	Pearson Correlation	.455	.568	.500	.740	1	.529	.532	.399
	Sig. (2-tailed)	.006	.000	.002	.000		.001	.001	.018
	N	35	35	35	35	35	35	35	35
Number of tractors (000 tractors)	Pearson Correlation	.838	.962	.984	.888	.529	1	.984	.891
	Sig. (2-tailed)	.000	.000	.000	.000	.001		.000	.000
	N	35	35	35	35	35	35	35	35
Total fertilizer consumption from chemical sources (000 t)	Pearson Correlation	.837	.978	.969	.856	.532	.984	1	.896
	Sig. (2-tailed)	.000	.000	.000	.000	.001	.000		.000
	N	35	35	35	35	35	35	35	35
Seed (tonnes)FAO	Pearson Correlation	.755	.851	.939	.845	.399	.891	.896	1
	Sig. (2-tailed)	.000	.000	.000	.000	.018	.000	.000	

ANNEX

Annex 1: Publications/ Seminar Presentation

List of Published Papers

Dhakal C. P. Sthapit A. B, and Devkota N.R. (2014). Forecast accuracy measure: An overview. Samajiki Sandarsh, ISSN (2348-0076) (Oct.-Dec.) pp. 121-127. Future Fact Society, Vanarasi (U.P.) India.

Dhakal C. P., Sthapit A. B, and Devkota N.R. (2013). Single regression to forecast rice production: A case of time period and harvested area. Journal of institute of science and technology, Vol (18), pp. 132-137. Institute of Science and Technology, Tribhuwan University Kirtipur, Kathmandu, Nepal.

Dhakal C.P. (2013). Meaning of data given through a case: Multiple regression for rice production forecasting in Nepal. Oxford Aawaj, pp. (92-97). Oxford Higher Secondary School (Aadharsheela College), Nayabazar Town planning Kathmandu.

Seminar Presentation

Selecting Variables while Fitting Multiple Regression Model for Forecasting.

2016 March 29-31. 7th National Conference on Science and Technology. Nepal

Academy of Science and Technology [NAST], Khumltar, Lalitpur Nepal.

Annex 2: Published Papers