ADDIS ABABA UNIVERSITY

GRADUATE STUDIES PROGRAM

FACULTY OF SCIENCE

DEPARTMENT OF STATISTICS

Predictors of Growth of Teledensity in Ethiopian Telecommunications Corporation

BY: ZEWDU MERESSA WAKENE

A THESIS SUBMITTED TO THE SCHOOL OF GRADUATE STUDIES PROGRAM OF ADDIS ABABA UNIVERSITY IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF SIENCE IN STATISTICS WITH A FOCUS ON APPLIED STATISTICS

June, 2009

ADDIS ABABA

ADDIS ABABA UNIVERSITY

GRADUATE STUDIES PROGRAM

FACULTY OF SCIENCE

DEPARTMENT OF STATISTICS

Predictors of Growth of Teledensity in Ethiopian Telecommunications Corporation

BY

Zewdu Meressa

Approved by the Board of Examiners:

_____ _____

Department Head **signature**

_____ _____

Internal Examiner **signature**

_____ _____

External Examiner **signature**

Tel.251-1-239465 **P.O.Box 1176 Addis Ababa, Ethiopia**

Acronyms

E.F.Y	Ethiopian Fiscal Year
ETC	Ethiopian Telecommunications Corporation
ICT	Information and Communications Technology
GDP	Gross domestic Product
GDPC	Gross Domestic Product per Capita
ML100	Main Lines per 100 inhabitants
LDC	Least Developed Country
LRM	Linear Regression Model
VIF	Variance Inflation Factor
sk	Skewness-kurtosis
OLS	Ordinary Least Square
BLUE	Best Linear Unbiased Estimator
MC	Multicollinearity
PACF	Partial Autocorrelation Function

AR (1)	Autoregressive of order 1
AC	Autocorrelation
GLS	Generalized Least Square
LM	Lagrange Multiplier

Acknowledgements

First and foremost, I would like to acknowledge the one and true God, my Lord Jesus Christ for blessing me with the ability and guidance to succeed this year.

Next, I sincerely thank my adviser, Dr Fentaw Abegaz, for providing me with his profound knowledge, direction and unrelenting encouragement. His thoughts were always valued and this study would not have been possible without his significant input.

Many thanks are due to my corporation the Ethiopian Telecommunications Corporation (ETC) which allows me to join MSc. Program and to utilize computer and other facilities with out limitation.

My appreciation and thanks are extended to other people who have contributed to this thesis directly or indirectly. I would like to acknowledge and thank my family for their support, both financial and otherwise.

An honorable mention must go out to Dr Emanuel, whose thoughts and comments both contributed to, and improved, this study.

Finally, I would like to thank the Department of Statistics for materials provided and allowing me to use the facilities. And lastly but not the least I would like to thank my family, friends for their special and continuous cooperation and encouragement throughout my study period.

Abstract

The Ethiopian Telecommunications Corporation is the sole provider of national and international telecommunications in Ethiopia since 1894. In Ethiopia, teledensity grew so slowly, as compared to developed countries or other developing countries. This study uses 31 years of yearly data starting from the beginning of 1970 in Ethiopian fiscal year.

The problems and strategic actions for growth in teledensity are discussed. Also, the opportunities for utilizing information and communication technologies to solve priority problems and to realize sustainable development in the country are examined. Parametric (Cochrane-Orcutt) and nonparametric (lowess) multiple regression models are employed.

Specifically, the findings of the parametric regression model based on Cochran-Orcutt transformation to handle serial correlation of residuals suggest that major determinant for growth of teledensity are higher GDPC and higher contribution of the service sector share to GDP in Ethiopia. And the average revenue generated by each telephone line (average used charge) is negatively related to teledensity in Ethiopia. In addition, the nonparametric regression based on lowess method fitted the teledensity data equally as good as the parametric method.

Table of Contents

Page

Chapter One **Introduction** 1

1.1. Telecommunication in Ethiopia: A brief History 2

1.2. Statement of the problem 4

1.3. Objectives 5

1.4. Application of the result 5

1.5. Organization of the thesis 5

1.6. Limitation of the study 6

Chapter Two **Literature Review** 7

Chapter Three **Data and Methodology** 14

3.1. The Data Issues 14

3.2. Variables in the Study 14

3.2.1 The Response Variable 14

3.2.2 Predictor Variables 14

3.3. Methodology 15

3.3.1. Linear regression model 16

3.3.1.1 The assumptions of linear regression model 18

3.3.1.2 Regression model diagnostic 21

3.3.2. Nonparametric Regression Analysis 39

3.3.2.1. The lowess method 39

Chapter Four Data Analysis 44

4.1. Model specification 44

4.2. Model diagnostic 45

4.2.1. Test of multicollinearity 45

4.2.2. Testing for normality 46

4.2.3.Testing the residuals for heteroskedasticity 47

4.2.4. Testing for outlier 48

4.2.5. Testing the residuals for autocorrelation 48

4.3. Fitting the final model 51

4.3.1. Cochrane-Orcutt and Prais-Winston method 51

4.3.2. Nonparametric regression using Lowess method 55

4.4. Graphical Comparisons of parametric and nonparametric analysis 56

Chapter Five Conclusion and Recommendation 58

5.1. Conclusion 58

5.2. Recommendation 59

References 61

Annex 64

List of Tables

Table **Page**

Table 1.1: Comparison of teledensity of Ethiopia with African countries 64

Table 1.2: Comparison of fixed line household access of Ethiopia with African countries 65

Table 4.1: Output of OLS estimates 66

Table 4.2: Bivariate Analysis of the dependent variable with independent variables 67

Table 4.4: Box-Ljung Statistic 68

Table 4.6: Output of Cochrane-Orcutt 69

Table 4.7: Output of Prais-Winstone 70

Table 4.8: Box-Ljung Statistic for transformed model 71

Table 4.9: Estimation using Lowess procedure 72

List of Figures

Figure **Page**

Fig. 1.1: Trend of teledensity 79

Fig. 4.1: Linearity checking graph 80

Fig. 4.2: Normal probability plot 81

Fig. 4.3: Plot of residuals versus time 82

Fig. 4.4: Plot of standardized residuals versus time 83

Fig 4.5: Partial autocorrelation function of unstandardized residuals 84

Fig. 4.6: Partial autocorrelation function of residuals of transformed model 85

CHAPTER ONE

1. INTRODUCTION

Telecommunications has a considerable impact in modern society. Enhancement in telecommunication technology coupled with the advanced computer and information systems makes the potential scale of its social, economic, and political impact unprecedented. Telecommunications has provided benefits to society. Saunders et .al. (1994) note that the benefits from better telecommunications are instant availability of market information, higher efficiency of transportation, better regional development, easier access in opening isolated areas, better emergency security facilities, and easier coordination of international activities. Also, telecommunications has been widely used to improve health care and the education sector. However, there remain vast differences in telecommunications facility throughout the world nowadays. People in developed countries enjoy having easy access and advanced services in telecommunications, while in some part of the world people have to wait for uncertainty of getting new telephone lines.

Teledensity refers to the number of main telephone lines for every one hundred inhabitants. One of the major prerequisites of economic integration in a modern, complex society is the development of a sound infrastructure in the telecommunications sector. The establishment of a modern, reliable, and rapidly expanding telecommunications infrastructure contributes considerably to the promotion of a variety of economic expansion activities. The International Telecommunications Union (ITU) (1994) shows the average level of teledensity among the least

developed countries is 0.29. This means just under one telephone main line for every 300 people. There exists a very wide gap between the teledensity of developing countries (like Ethiopia) and that of developed countries, (Mbarika, et al., 2003).

Mbarika, et al., (2003) and Callender, et al., (2003) investigated predictors of growth of teledensity in developing countries, they used linear regression models to examine the relationship between teledensity and various independent variables. Much study has not been done in the case of Ethiopia. Therefore, this study will investigate factors related to growth of teledensity using parametric and nonparametric regression methods.

1.1 Telecommunication in Ethiopia: A brief History

The introduction of telecommunication in Ethiopia dates back to 1894. In those years, the new technological scheme contributed to the integration of the Ethiopian society when the extensive open-wire line system was laid out linking the capital with important administrative cities of the country. Ethiopia became a member of International Telecommunication Union (ITU) in 1932. Most of the telecommunication network, however, was completely destroyed during the Italian Fascist aggression Later on Ethiopia developed its telecommunication facilities all over again. When the Imperial Telecommunication Board of Ethiopia was established by Proclamation 131/53 in 1953, it was granted full provision of administrative and financial autonomy. The major objectives of the Board were: to undertake the expansion of telecommunication services in the nation, to represent Ethiopia at all International fora

regarding telecom activities (except military communications), to allocate and construct communication frequencies, and to train the required personnel. In order to achieve its objectives, the organization had undergone through section development programs. Even though the institution had been granted full autonomy by the above proclamation, the right of the organizational had been violated during 1975, when it was reorganized and renamed as "Ethiopian Telecommunications Service" and in 1981, the organization was renamed again as "Ethiopian Telecommunications Authority". Finally, in 1996 it was established as a corporation "Ethiopian Telecommunications Corporation", * http://www.telecom.net.et/

Ethiopia Telecommunication Corporation (ETC) is the sole provider of telecommunications services in Ethiopia. Telecommunication is a key player in any development endeavor, has a great role in a different aspects, economical, social and technological development of one country. ETC has the mission to develop and maintain a modern information and communication network infrastructure capable of supporting voice, data and video services, equitably across the country and with high capacity digital connectivity to the rest of the world, provide world-class telecommunications services including basic telephony, mobile and Internet and multimedia services, provide training, education and research in the field of investments in information and communications technology (ICT), (Annual Statistical Bulletin, 2000 E.F.Y.).

At the end of the year 2000 E.F.Y (2007/08) the total installed exchange capacity of fixed telephone reached 1,146,555 out of which 99.78% are digital and the remaining 0.22% are

manual exchange lines. The country's telecom penetration for fixed telephone (teledensity) has reached 1.20% in the fiscal year 2000. The penetration figure rose to 3.78% when mobile subscription is included, see the trend in Figure 1.1, (Annual Statistical Bulletin, 2000 E.F.Y.).

1.2 Statement of the problem

Access to telecommunications is considered essential for development in rural and impoverished regions worldwide. Lack of telecommunications hinders progress in diverse areas including education, business development, health care, humanitarian issues, and quality of life. In the case of Ethiopia only 1.2 percent of the population has effective access to telecommunications. Over a 31 years period, from 1970 to 2000 E.F.Y, the Ethiopian teledensity of fixed telephone increased from 0.21 %(1970) to about 1.2 % in 2000 and the estimated population increased from 28.55 million to 75.84 million. As compared to North Africa, South Africa, Sub-Sahara and African average teledensity (2007 G.C.), as shown in Table 1.1, the average teledensity in African countries which is 3.77 in 2007 is three times greater than Ethiopian teledensity. In addition, the fixed line telephone use comparison with some African countries as presented in Table 1.2, fixed line household access is 7.6 percent in 2008 in Ethiopia.

1.3. Objectives

The objectives of the study are:

- To investigate the nature of teledensity in Ethiopia and to identify and examine the major socio-economic factors that could influence the growth of teledensity in Ethiopia.
- To fit parametric and nonparametric regression models based on data on growth of teledensity
- Predict teledensity based on major socio-economic factors.
- Make recommendations based on the findings that enable to boost teledensity in Ethiopia.

1.4. Application of the result

The result of the study will help

- To identify problems related to the growth of teledensity.
- To design policies and intervention strategies in the telecommunication sector in order to realize sustainable development.
- To help as basis for further studies in this area.

1. 5. Organization of the paper

The format of this thesis is as follows: The second Chapter deals with literature review. The third Chapter presents the data and methodology. In Chapter four analyses of the data with the

method discussed in Chapter three are presented and finally conclusion and recommendation are given in Chapter five.

1.6. Limitation of the study

This study employs parametric and nonparametric methods of analyses to examine the relationships between teledensity and the various independent variables. The independent variables test included proportion of telephones in residential use, subscriber of mobile telephone, technical staff and waiting list for telephone line that probably predict the growth of teledensity which are not included in this study due to unavailability of data, from all possible sources such as Ethiopian Telecommunications Corporations, Ethiopian Central Statistics Agency and Ministry of Finance and Economic Development. So the absence of data on these variables will have impacts on the result of this study.

CHAPTER TWO

LITERATURE REVIEW

In a study carried out by the international telecommunications union (ITU World Telecommunication Indicators, 1995), developing countries like Ethiopia were represented to be among the least developed in terms of the state of their telecommunication networks and limited range of services offered. The study shows evidence that least developed countries are falling farther behind other developing countries in the race to construct modern telecommunication networks. The same ITU study mentioned above shows further evidence that the falling of least developed countries behind other developing countries in the race to construct modern telecommunication networks arises not so much because they are not installing the latest equipment - in many cases the least developed countries have modern, state-of-the-art digital networks - but rather that they are not expanding fast enough to close the teledensity gap with other developing countries. Least developed countries are greatly behind other regions of the world in terms of their levels of teledensity, (International Telecomminications Union, 1995).

Mbarika, et al. (2003) employs a linear regression model to examine the relationships between teledensity and the various independent variables such as GDP per capita, telecommunications investment, telephone staff and waiting time for telephone installation. The data were taken from 119 developing countries. The findings suggest that there was a very weak negative

correlation between investment in telecommunications infrastructure and teledensity. The study, however, suggests a positive relationship between teledensity and other variables such as GDP per capita, telecommunications staff and length of waiting times to acquire and maintain telephones.

Many governments have made growth in teledensity. However, it has proven difficult to rapidly increase teledensity in rural areas of developing nations. Early strategies to increase teledensity focused on providing wire line telephone and data communications. In the 1990s these programs were largely supplanted by cellular telephone based approaches, due to the significant reduction in infrastructure and maintenance cost provided by wireless last-mile access. Despite the reduction in cost compared to wire line, cellular service has not yet become available in most impoverished rural areas.

Vanu (2005) explains the benefits of software radio devise for teledensity using India as a case study. Over a 10 year measurement period, March 1996 to September 2006, urban teledensity increased from 4% of the population to almost 33%. In the same period, rural teledensity increased from 0.3% to only 2%. In India the population considered as rural represents over 700 million people. Even in 2006, there were well over half a billion people in India without effective and affordable access to telecommunications. Providing communications services to such a huge number of people is a major business opportunity. The slow growth of teledensity as explained by Vanu (2005) is that it has been uneconomical to provide service due to factors like low revenues and high cost, government subsidies, limitation of traditional radios and lack of competition in the rural areas.

Several studies have been conducted to examine the link between teledensity and socio-economic factors. Clarke and Laufenberg (1983) showed that growth of teledensity brought a variety of social benefits in addition to economic benefits in rural Sub-Saharan Africa. In particular, the International telecommunications Union's CCITT (International Consultative Committee on Telephone and Telegraph) has sponsored several studies. The study by International Telecommunication Union (1998) reported a strong correlation between teledensity and variables such as Gross domestic Product (GDP) and a positive relationship between teledensity and economic development. Moreover, in the study the shortage of qualified personnel has been cited as a cause for inadequate maintenance of telecommunications infrastructure. This leads as indicated in another study by Mbarika, et al. (2003) that customers experienced extremely long waiting periods to obtain new telephone service and subsequent repairs. On the other hand Saunders, et al. (1983) found a negative correlation between teledensity and population size. Further they indicated that as GDP increases, telephone density increases more rapidly.

Sung and Lee (2002) examined the impact of rapid growth in mobile telephones on the demand for traditional fixed network access in South Korea. The analysis used South Korean regional panel data for the period 1991-98. The results showed that a 1 percent increase in the number of mobile telephones results in a reduction of between 0.1 and 0.18 percent in new fixed connections and an increase of between 0.14 and 0.22 percent in fixed disconnections.

Cadima and Barros (2000) studied the diffusion of fixed and mobile networks in Portugal, and concluded that mobile telephone growth slowed the growth of fixed telephone, but there was little impact of the latter on the former.

Wallsten (2001) carried out a study in Latin American and African countries from 1984 through 1997 and found that privatization, by itself, is not a driver of improved telecommunications performance but when it is combined with an independent regulator, the country experiences increased connection capacity and payphones per capita. Wallsten's proposal for an independent regulator comes from the ITU and characterizes a country as having an independent regulator if there is a separate telecommunications regulatory agency not directly under control of a ministry. The result suggests that ITU is correct to emphasize regulatory reforms along with privatization, since privatization without attention to regulation may be costly to consumers.

Gutierrez and Berg (2000) also investigated the impact that regulation has on sector performance but their focus is exclusively on Latin America from 1985 through 1995. They develop a dichotomous variable measuring the degree to which the regulatory framework in telecommunications affords: (1) enforcement power to regulators and (2) neutrality/independence. They stressed that the independent regulatory variable should not be interpreted as whether there is increased or decreased regulation but rather whether the regulator is independent. Their study showed that an independent regulator has significant positive impacts on telephone lines per capita.

Agustin (2003) examined the determinants of telecommunications sector performance in 20 Latin American countries from 1990 to 1998. The dependent variables used in this study are main lines per 100 inhabitants (ML100) – a measure of teledensity – growth in main lines per 100 inhabitants (ML100g) and main lines per employee (MLemp) – a measure of operating efficiency. The explanatory variables included are real GDPC, population density, real annual investment in telecommunications assets, percent of network with digital lines and percent of network that consists of residential lines. In addition regulatory variables such as privatization, competition, existence of an independent regulator and whether the country has a price cap regime in place were included. The data for this study consisted of cross-sectional and time series (panel) data.

A study by Mbarika (1999) examined 48 least developed countries worldwide with teledensity of less than one. The findings suggest that increased investment in telecommunication technologies is not a major determinant for growth of teledensity, but that higher GDPC and higher contribution of the service sector share to gross domestic product of least developed countries are major determinants for growth of teledensity.

The Maitland Commission (1984) described the teledensity gap as the unbalanced distribution of telephones across the world, with low teledensity, a shortage of exchange capacity, long waiting periods for acquiring a new telephone line, low quality of service, and imbalance of telecommunications infrastructure between urban and rural areas.

ITU world telecommunications report (1997) shows that over 80% of least developed countries telephone switching is made up of analog equipment which is unreliable and difficult to network with computers. One of the challenges of least developed countries telecommunications operators is how to successfully modernize their networks (Ogbe, 1990). Analog telecommunications equipments are becoming a pain for major telecommunications operators in least developed countries due to high maintenance costs and lack of spare parts for systems that need to be phased out. The lack of reliability and inconvenience of analog networks for data communications is a regular drain on network operators.

The monopolistic and parochial culture of least developed countries telecommunications operators is also reflected in the mounting tariff and awkward traffic situation in the region (Paltridge, 1994). The average telecommunications revenue per subscriber line in Europe is about 770 dollars while the average in least developed countries is roughly the double of that at 1,460 dollars. Given the lower per capita income in least developed countries, the ratio of revenue in least developed countries to that of Europe is extremely high. Restrictions further mount the high cost of telecommunications services. Statistics show that data transmission in Europe costs two to three times that of the United States of America due to more restrictions in Europe. In least developed countries, where restrictions are even higher, the costs are also exorbitantly high. Institutions and/or individuals are charged four to eight times that of Europe and four to twenty times that of the United States for the same loop of calls. An unanswered question here is whether complete privatization will bring down these costs and further

research along this line to use the experience of some developed countries, which have completely privatized their telecommunications industry, as a benchmark.

CHAPTER THREE

DATA AND METHODOLOGY

3.1 The Data Issues

There are generally three types of data that are available for empirical analysis: cross sectional, (2) time series, and (3) a combination of cross section and time series, also known as pooled/panel data. In this study the data used is yearly data from 1970 E.F.Y to 2000 E.F.Y, which was collected from Ethiopian Telecommunication Corporation, Ministry of Finance and Economic Development, and Central Statistics Agency.

3.2 Variables in the Study

3.2.1 The Response Variable

Teledensity is used to refer to the number of main telephone lines for every 100 inhabitants. During the year 2000 E.F.Y (2007/08), the number of telephone subscription under all categories i.e. residential, business, government and others reached 897,287 and the estimated population of Ethiopia was 75.84 million. As a result the teledensity was approximately 1.20%.

3.2.2 Predictor Variables

- Gross Domestic product per capita is measured as GDP divided by total population every year, measured in Ethiopian birr.

- Average revenue generated by each telephone line refers to the amount of total collection of fixed telephone lines divided by the total number of connected telephone lines every year, it is average yearly charge per line, measured in Ethiopian birr.
- Size of service sector contributed to GDP refers to annual telecommunication investment as a percent of GDP to be surrogate for telecommunication investment budget allocation in a country, to measure budget allocation.
- Percentage growth of GDP refers to percentage growth of GDP every year.
- Growth of investment in telecommunications infrastructure refers to growth of annual expenditure associated with acquiring ownership of property and plant used for telecommunication services, measured in million of Ethiopian birr.

3.3 Methodology

In this study we will employ both parametric regression model and nonparametric regression method to examine the relationship between teledensity and the various independent variables, and also to compare and choose the best model.

The study, first, will employ linear regression model to examine the relationship between teledensity and the various independent variables. If we are dealing with time series data, the data follow a natural ordering over time so that successive observations are likely to exhibit inter correlations, especially if the time interval between successive observations is short, such as a day, a week, or a month rather than a year.

3.3.1 Linear regression model

Linear regression is a form of regression analysis in which the relationship between one or more independent variables and a dependent variable. This function is a linear combination of one or more model parameters, called regression coefficients. Most real world phenomena are multi-factoral in nature, meaning there is more than one factor that impacts on, or causes changes in the dependent variable. Regression allows us to test how well we can predict a dependent variable on the basis of multiple independent variables.

The linear additive regression model relating dependent variable to (p-1) independent variables is:

$$Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + \ldots + \beta_{(p-1)} X_{(p-1)t} + e_t \qquad (3.1)$$

where t = 1, 2. . . , n.

The subscript t denote the observational unit (t=1,2,…,n) from which the observations on **y** and the (p-1) independent variables are taken. There are p parameters, β_j, j = 0, 1, 2. . . ,(p-1) , to be estimated when the linear model includes the intercept β_0.

This model can also be written as

$$\mathbf{Y = X \beta + e} \qquad (3.2)$$

where

Y is an (n x 1) vector of observations,

X is an (n x p) matrix of regressors,

β is (n x 1) vector of parameters,

e is an (n x1) vector of random error terms,

where E (**e**) = **0**, V (**e**) = **I** σ^2, so that the element of **e** are uncorrelated.

Since E (**e**) = **0**, an alternative way of writing the model is

$$E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta} \qquad (3.3)$$

The error sum of squares is then

$$\mathbf{e'e} = (\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})' (\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})$$

$$= \mathbf{Y'Y} - \boldsymbol{\beta}'\mathbf{X'Y} - \mathbf{Y'X}\boldsymbol{\beta} + \boldsymbol{\beta}'\mathbf{X'X}\boldsymbol{\beta} \qquad (3.4)$$

$$= \mathbf{Y'Y} - 2\,\boldsymbol{\beta}'\mathbf{X'Y} + \boldsymbol{\beta}'\mathbf{X'X}\boldsymbol{\beta} \quad \text{because} \quad \boldsymbol{\beta}'\mathbf{X'Y} = \mathbf{Y'X}\boldsymbol{\beta}$$

The least squares estimate of **β** is the value **b,** which, when substituted in equation (3.4), minimizes **e'e**. It can be determined by differentiating equation (3.4) with respect to the components of **β** and setting the resultant matrix equation equal to zero, and at the same time replacing **β** by **b**. This provides the normal equations.

$$\mathbf{X'X\,b} = \mathbf{X'Y}. \qquad (3.5)$$

Note that equation (3.5) consists of p independent equations in p unknowns or some equations may depend on others so that there are fewer than p independent equations in the p unknowns (the p unknowns are the elements of **b).** If some of the normal equations depend on others, **X'X** is singular so that $(\mathbf{X'X})^{-1}$ does not exist. If the p normal equations are independent, **X'X** is nonsingular, and its inverse exists. In this case the solution of the normal equations can be written as

$$\mathbf{b} = (\mathbf{X'X})^{-1}\mathbf{X'Y} \qquad (3.6)$$

3.3.1.1 The assumptions of Linear Regression modeling

There are several assumptions that have to be fulfilled for the classical linear regression model to be valid. When the regression model does not meet the fundamental assumptions, the prediction and estimation of the model may become biased. Residuals, differences between the values predicted by the model and the real data will be very large and seriously distort the prediction. When the residuals are extremely large or small, they are indicators of outlying observations, called outliers. The outliers will inflate the error variance. They inflate the standard errors. The confidence interval becomes stretched. The estimation cannot become asymptotically consistent. Outliers that bias the parameter estimates are those with leverage. They are called bad leverage points, whereas outliers that lie along the predicted line are those called good leverage points. When outliers inflate the error variance, they sap the model of power to detect the outliers.

The assumptions of linear regression modeling are:

1. Linearity

This is to mean that the relationship between the dependent variable **Y** and the independent variable(s) **X** is linear.

2. No multicollinearity

Suppose we wish to fit the model $\mathbf{Y = X\beta + e}$. The solution $\mathbf{b = (X'X)^{-}X'Y}$ would usually be sought. However, if **X'X** is singular, we cannot perform the inversion and the normal equations do not have a unique solution. This stems from the fact that there is at least one linear combination of the columns of **X** is linearly dependent on (i.e., is a linear combination of) the other columns. We would say that collinearity (or Multicollinearity) exists among the columns of **X**.

3. Homoskedasticity

One of the basic assumptions of regression analysis is equality of the error variance along the predicted line, a condition called homoskedasticity. If the residual distribution is normally distributed, the analyst can determine where the level of significance or rejection regions begin. Even if the sample size is large, the influence of the outlier can increase the local and possibly even the global error variance. This inflation of error variance decreases the efficiency of estimation.

4. No influential outliers

An outlier among residuals is one that is far greater than the rest in absolute value and perhaps lies three or four standard deviations or further from the mean of the residuals. The outlier is a peculiarity and indicates a data point that is not at all typical of the rest of the data. In linear regression, an outlier is an observation with a large residual. In other words, it is an observation whose dependent-variable value is unusual given its values on the predictor variables. An outlier may indicate a sample peculiarity or may indicate a data entry error or other problem.

Influence: An observation is said to be influential if removing the observation substantially changes the estimate of coefficients.

5. No autocorrelation of residuals

Another form of violation of assumptions is that when residuals lack independence, this lack of independence of the errors can manifest itself in terms of residual autocorrelation, which can further bias the estimation of significance tests as the error variance becomes artificially compressed by residual correlation. When this happens, the R^2, F and t values become inflated. Failures of these assumptions can predispose output toward false statistical significance.

6. Fixed independent variables-no measurement error

X is a deterministic data matrix (it is non-stochastic) that means the X_i's, i = 1, 2, . . . , (p-1) are a set of fixed values in the hypothetical process of repeated sampling, which underlies the linear regression model.

7. Normality of residuals

Another assumption is normality of the residuals. When there are violations of the assumption of normality of the residuals in ordinary least square regression analysis, the estimation of parameter becomes impaired.

3.3.1.2 Regression model diagnostic

We should not only know what assumptions have to be fulfilled for the model to be valid, we should also know how to test them for fulfillment. Frequently, fundamental assumptions- such as, independence of observations, linear functional form, no influential outliers, normality and homoskedasticity of the residual distribution are not adequately fulfilled.

Linearity

Preliminary testing prior to linear regression modeling is testing for linear functional form; we may graph the dependent variable with each of the independent variables. If the relationships are linear then we model them with linear models. If the relationships are nonlinear, then we model them with nonlinear or nonparametric models.

Normality

To test for **normality of the residuals**, we may examine a quantile-quantile plot or a p-p plot. The plots can be accompanied by tests based on the cumulative distribution of the residuals against that of the theoretical normal distribution with a chi-square test to determine whether there is a statistically significant difference. The null hypothesis is that there is no difference between the empirical and theoretical normal distribution. When the probability is less than .05, we need to reject the null hypothesis and infer that the residuals are not normally distributed. Or we may test for the normality of the residuals using formal test based on skewness and kurtosis.

Detection and assessment of influential outliers

Outlier detection involves the determination whether the residual (error = predicted – actual) is an extreme negative or positive value. One of the problems of outliers is that an outlier can really mess up the sample mean, but have relatively effect on the sample median and another problem of outliers may be that the data are not really normally distributed.

Outlier observations can be identified using standardized residuals, studentized residuals and Cook's d statistic. A standardized residuals are simply the residuals ($\hat{e}_t$) divided by the standard error of the regression ($\sqrt{\hat{\sigma}^2}$), that is, they are ($\frac{\hat{e}_t}{\hat{\sigma}}$). In large sample the standardized residuals is approximately normally distributed with zero mean and unit variance.

Rules have been proposed for rejecting outliers (i.e., for deciding to remove the corresponding observation(s) from the data, after which the data is reanalyzed without these observations). Automatic rejection of outlier is not always a very wise procedure. Limits of standardized residuals are, if the standardized residuals have values in excess of 3.0 and -3.0, they are outliers. If the absolute values are less than 3.0, then there are no outliers. Or using Cook's distance, the lowest value that Cook's D can assume is zero, and the higher the Cook's distance is, the more influential the point. The conventional cut-off point is 1.

Heteroskedasticity

Consider the properties of OLS estimator with the existence of heteroscedasticity, that is, the specified equation is

$$\mathbf{Y = X\beta + e} \quad \text{with } E(\mathbf{ee'}) = 0 \quad \text{and } E(\mathbf{ee'}) = \sigma^2 \Omega$$

For nonstochastic **X** the following results hold.

1. **OLS** estimator is unbiased and consistent.

$$\mathbf{b = \beta + (X'X)^{-1}X'e}$$

It follows directly that $\mathbf{E(b) = \beta}$, so the unbiased properties holds. Mean square consistency follows provided the variance matrix, var(**b**), has a zero prbability limit, as will be seen in point 3.

2. OLS estimator is inefficient.

Var (**b**) = E(**b**-E(**b**))^2 = E(**b-β**)^2 = E(((**X'X**)$^{-1}$**X'e**)((**X'X**)$^{-1}$**X'e**)')

= σ^2(**X'X**)$^{-1}$**X'ΩX**(**X'X**)$^{-1}$

= σ^2**X**$^+$**Ω(X**$^+$)' , where **X**$^+$ **= (X'X)**$^{-1}$**X'**

Now: (**X**$^+$**-(X'Ω**$^{-1}$**X)**$^{-1}$**X'Ω**$^{-1}$**) Ω (X**$^+$**-(X'Ω**$^{-1}$**X)**$^{-1}$**X'Ω**$^{-1}$**)'** ≥ 0

$\Rightarrow$ **(X**$^+$**)Ω(X**$^+$**)'-(X'Ω**$^{-1}$**X)**$^{-1}$ ≥ 0

$\Rightarrow$ σ^2 **(X**$^+$**)Ω(X**$^+$**)'-** σ^2 **(X'Ω**$^{-1}$**X)**$^{-1}$ ≥ 0

$\Rightarrow$ $\text{var}(b) - \text{var}(\tilde{b}) \geq 0$, where $\text{var}(\tilde{b})$ is the variance of the generalized least square (GLS) estimator, $\tilde{b}$.

Hence, the OLS estimaor is less efficient than the GLS estimator.

3. Standared errors of regression coefficient estimates based on OLS are incorrect, and the conventional test statistics based on them are invalied.

The correct variance matrix for the OLS coefficient vector is

Var (**b**) = E(**b**-E(**b**))^2 = E(**b-β**)^2 = E(((**X'X**)$^{-1}$**X'e**)((**X'X**)$^{-1}$**X'e**)')

= σ^2(**X'X**)$^{-1}$**X'**Ω**X**(**X'X**)$^{-1}$

The conventional formula calculates σ^2 **(X'X)**$^{-1}$, which is only part of the correct expression in the above equation.Thus the conventional test statistics are invalid.

The variance matrix may also be expressed as

$$\text{var}(b) = \frac{\sigma^2}{n}\left[\frac{(X'X)}{n}\right]^{-1}\left[\frac{(X'\Omega X)}{n}\right]\left[\frac{(X'X)}{n}\right]^{-1}$$

The probability of the first term in the above equation is zero. With stationary regressors the probability limit of the second term is a finit matrix. Consistency thus requires that the probability limit of **X'ΩX/**n also be a finite matrix, which in general will be true if the elements are finite.

Diagnostic method to assess heteroskedasticity

We may use a graph of residuals against time to obtain a graphical indication of heteroskedasticity. This displays any problematic patterns that might suggest heteroskedasticity. Or a formal statistical test like Breusch-Pagen test.

If after the diagnostic tests of heteroskedasticity, if we find that there is heteroskedasticity, then we can use estimation method which takes this in to account like the generalized least squares, method (GLS).

Multicollinearity

In the model, **Y = X β + e multicollinearity** means **X** contains two or more highly correlated columns. This problem occurs when the explanatory variables are very highly correlated with each other. One consequence of multicollinearity is non-identifiably of the regression coefficient vector **β**. In addition,

- the regression becomes very sensitive to small changes in the specification.
- R^2 will be high but the individual coefficients will have high standard errors.
- thus confidence intervals for the parameters will be very wide, and significance tests might therefore give inappropriate conclusions.
- the estimates may have unexpected or unrealistic signs (positive relationship might turn out to lead to negative signs and vice versa)

Diagnostics of multicollinearity

The easiest way to measure the extent of multicollinearity is simply to look at the matrix of correlations between the individual variables.

1. Condition number (**CN)**

The condition number of **X'X** is defined as

$$CN = \frac{\lambda_{\max}}{\lambda_{\min}}$$ where λ_{max} and λ_{min} are the largest and smallest eigenvalues of **X'X.**

A rule of thumb stipulates that if the condition number less than 100 this implies that there is no serious problem with multicollinearity and if it is greater than 1000 it becomes sever multicollinearity, but if it is in between 100 and 1000 it is called moderate to strong multicollinearity.

The condition indices of the matrix **X'X** are

$$CN_j = \frac{\lambda_{\max}}{\lambda_j} \qquad j = 1,2,\ldots,p-1$$

The number of condition indices that are large (say > 1000), are a useful measure of the number of near linear dependencies in **X'X**.

2. Variance inflation factor (VIF)

Variance inflation factors are the diagonal elements of cov ($\hat{\beta}$) with out σ^2, i.e

$$VIF(\hat{\beta}_i) = \sum_{j=1}^{p} \frac{p_{ij}^{\ 2}}{\lambda_j} \qquad i = 1, 2, \ldots, p.$$

If $VIF\ (\hat{\beta}_i) \in [0,10]$, then there is no problem in the estimated component of $\hat{\beta}$. On the other hand if any of VIF's exceeds 10, then this is an indication that the associated regression coefficients are poorly estimated due to multicollinearity.

Note that the VIF corresponding to the i^{th} regression coefficient can also be defined as:

$$VIR(\hat{\beta}_i) = \frac{1}{1 - R_i^2}$$

where R_i^2 is the coefficient of multiple determination obtained from regression X_i on the other regression variables X_j. j = 1, 2,…,p-1, i ≠ j.

The ways to "cure" the problems of multicollinearity are

- Drop one of the collinear variables.
- Include additional observations maintaining the original model such that a reduction in the correlation between variables is attained.
- Use information other than the regression coefficients from sources other than the sample at hand.
- Use biased (linear) estimators of β, such as ridge regression or principal components.

Ridge regression (algebraic method of overcoming MC)

One approach to the problem of MC involves trading a little bias for a large reduction in variance. The leading method of this kind is the so called ridge regression. The simplest and most common version of the method is the ordinary ridge regression (ORR).

The **(ORR) estimator of β** is

$$\hat{\beta}_R = \left[X'X + kI\right]^{-1} X'Y$$

where k > 0 is a constant. The ORR estimator is a linear transformation of the least square estimator since

$$\hat{\beta}_R = (X'X + kI)^{-1} X'X\hat{\beta} = Z_k\hat{\beta}$$

$$E(\hat{\beta}_R) = Z_K E(\hat{\beta}) = Z_K\beta$$

$\Rightarrow \hat{\beta}_R$ is a biased estimator of **β** and k is the biasing parameter.

$$E(\hat{\beta}_R) = E(X'X + kI)^{-1} X'(X\beta + e)$$

$$E(\hat{\beta}_R) = E\left[(X'X + kI)^{-} X'X\beta + (X'X + kI)^{-} X'e\right]$$

$$= \left[(X'X + kI)^{-} X'X\beta + (X'X + kI)^{-} X'E(e)\right]$$

$$= (X'X + kI)^{-} X'X\beta \qquad (3.7)$$

$$\hat{\beta}_R - E(\hat{\beta}_R) = (X'X + kI)^{-1} X'e$$

$$COV(\hat{\beta}_R) = E((\hat{\beta}_R - E(\hat{\beta}_R))(\hat{\beta}_R - E(\hat{\beta}_R))')$$

$$= (X'X + kI)^{-1} X'E(ee')X(X'X + kI)^{-1} \qquad (3.8)$$

From (3.7) and (3.8) we see that the trade off between bias and variance hinges on the value of k. The larger the value of k the larger the bias but the smaller the variance. Detail discussion can be found, in Draper and Smith (1998).

Autocorrelations

We apply Durbin-Watson d test or the Breusch-Godfrey (BG) test for autocorrelation or the autocorrelation function to test for autocorrelation in the random error terms.

Consequences of ignoring autocorrelation if it is present are:

(1) The OLS estimators are unbiased, but are not efficient.

(2) The conventionally estimated standard errors of OLS estimators are biased.

(3) As a result, the conventionally computed t and F tests are unreliable.

Sources of Autocorrelation

1. Lagged endogenous variables
2. Misspecification of the model
3. Simultaneity, feedback, or reciprocal relationships
4. Seasonality or trend in the model

Diagnostic method of autocorrelation

1. If there is a problem of autocorrelation, we must remove the problem before fitting the model using OLS. To apply method of correcting, we need to estimate the parameter of error autoregressive model. And, to determine the order of autocorrelation of the residual, we plot the partial autocorrelation function.

2. One can use the Durbin-Watson d statistic to test for first-order autocorrelation (AR (1)) after fitting the regression.

AR (1) scheme is given by:

$e_t = \rho e_{t-1} + u_t$ Where u_t is white noise

i.e $E(u_t) = 0$ and $E(u_t u_s) = \sigma_u^2$ for t=s and $E(u_t u_s) = \underline{0}$ for t ≠ s

- To test the null hypothesis that there is no autocorrelation

  H_0: $\rho = 0$ H_1: $\rho \neq 0$

  At α – level with use the Durbin – Watson test statistic is given by

$$d = \frac{\sum_{t=2}^{n} (\hat{e}_t - \hat{e}_{t-1})^2}{\sum \hat{e}_t^2} \qquad (3.9)$$

$$\hat{e}_t = Y_t - X_t b, \quad t = 1,2,\ldots,n,$$

e_t are the OLS residuals

$$d = \frac{\sum \hat{e}_t^{\,2} + \sum \hat{e}_{t-1}^{\,2} - 2\sum \hat{e}_t \hat{e}_{t-1}}{\sum \hat{e}_t^{\,2}} \qquad (3.10)$$

Since $\sum \hat{e}_t^{\,2}$ and $\sum \hat{e}_{t-1}^{\,2}$ differ in only one observation, they are approximately equal.

Therefore, setting $\sum \hat{e}_t^{\,2} \approx \sum \hat{e}_{t-1}^{\,2}$, the above equation may be written as

$$d \approx 2\left(1 - \frac{\sum \hat{e}_t \hat{e}_{t-1}}{\sum \hat{e}_t^{\,2}}\right) \qquad (3.11)$$

Now let us define

$$\hat{\rho} = \frac{\sum \hat{e}_t \hat{e}_{t-1}}{\sum \hat{e}_t^{\,2}} \qquad (3.12)$$

As the sample first-order coefficient of autocorrelation, an estimator of ρ.

Using equation (3.12), we can express equation (3.11) as

$$d \approx 2(1 - \hat{\rho}) \qquad (3.13)$$

But since $-1 \leq \rho \leq 1$, eqn. (5) implies that

$$0 \leq d \leq 4 \qquad (3.14)$$

These are the bound of d; any estimated d value must lie within these limits.

The assumptions underlying the d statistic

1. There are zones where the test is inconclusive.

2. The test assumes a constant in the regression.

3. The regressors must be non stochastic, or fixed in repeated sampling.

4. There cannot be any lagged endogenous variables.

5. Only tests for first order autocorrelation.

6. The error term e_t is assumed to be normally distributed.

Durbin and Watson derive upper limit ($d_u(\alpha)$) and lower limit ($d_l(\alpha)$) for different levels of significant, α.

Durbin-Watson d test decision rules

Null hypothesis	decision	if
No positive autocorrelation	Reject	$0 < d < d_l$
No positive autocorrelation	No decision	$d_l < d < d_u$

No negative autocorrelation	Reject	$4 - d_l < d < 4$
No negative autocorrelation	No decision	$4 - d_u < d < 4 - d_l$
No autocorrelation, positive or negative	Do not reject	$d_u < d < 4 - d_u$

3. **The Breusch-Godfrey (BG) or Lagrange Multiplier (LM) test for autocorrelation**

This test is a general test of autocorrelation and it allows for:

(1) Nonstochastic regressors, such as the lagged values of the regressand

(2) higher-order autoregressive schemes, such as AR (1), AR (2), etc and

(3) Simple or higher-order moving averages of white noise error terms

In the model, **Y = X β + e,**

Assume that the error term e_t follows the pth–order autoregressive, AR (p), scheme as follow:

$$e_t = \rho_1 e_{t-1} + \rho_2 e_{t-2} + ... + \rho_p e_{t-p} + \varepsilon_t \qquad (3.15)$$

where ε_t is a white noise error term as discussed above. This is simply the extension of AR (1) scheme.

The null hypothesis H_o to be tested is that

$$H_0 : \rho_1 = \rho_2 = ... = \rho_p = 0 \qquad (3.16)$$

That is, there is no serial correlation of any order.

The BG test involves the following steps:

1. Estimate a regression by OLS and obtain the residuals, $\hat{e}_t$

2. Regress residuals ($\hat{e}_t$) on the original all regressors and lagged residuals ($\hat{e}_{t-1}, \hat{e}_{t-2}, \ldots \hat{e}_{t-p}$). Note that to run this regression we will have (n-p) observations.

$$\hat{e}_t = \gamma_0 + \gamma_1 X_{1t} + \gamma_2 X_{2t} + \ldots + \gamma_P X_{Pt} + \hat{\rho}_1 \hat{e}_{t-1} + \ldots + \hat{\rho}_p \hat{e}_{t-p} + \varepsilon_t \qquad (3.17)$$

3. Obtain R^2 from this auxiliary regression

4. If the sample size is large (technically, infinite), Breusch and Godfrey have shown that

$$(n-p)R^2 \sim \chi_P^2 \qquad (3.18)$$

That is asymptotically, n-p times the R^2 value obtained from the auxiliary regression follows the chi-square distribution with p degree of freedom. If in application, (n-p) R^2 exceeds the critical chi-square value at the chosen level of significance, we reject the null hypothesis, in which case at least one rho in (3.15) is statically significantly different from zero.

Remedial measures with the existence of autocorrelation

If after applying one or more of the diagnostic tests of autocorrelation discussed above, we find that there is autocorrelation, then we have three remedial options:

1. If it is pure autocorrelation, one can use appropriate transformation of the original model so that in the transformed model we do not have the problem of (autocorrelation). As in the case of heteroscedasticity, we will have to use some type of generalized least square (GLS) method.
2. In large samples, we can use the Newey-West method to obtain standard errors of OLS estimators that are corrected for autocorrelation. This method is actually an extension of White's heteroscedasticity-consistent standard error method.
3. In some situations we can continue to use OLS method, as explained in Gujarati (2003).

Correcting for (pure) autocorrelation: The method of GLS

Knowing the consequences of autocorrelation, especially the lack of efficiency of OLS estimators, we may need to remedy the problem. The remedy depends on the knowledge one has about the nature of interdependence among the disturbances, that is, knowledge, about the structure of autocorrelation. If you know the form of autocorrelation (like the form of heteroscedasticity) you can use GLS. We can estimate ρ and use that value to transform **Y, X** and **e**. finally estimate the model parameters using OLS from the transformed model. This is the **Cochrane-Orcutt** iterative procedure.

The Cochrane-Orcutt estimation is an algorithm for estimating a time series linear regression in the presence of autocorrelated errors. Their procedure includes an improvement to include the first observation attributed to the Prais-Winston transformation. This version of the algorithm

can handle only first-order autocorrelation but the Cochrane-Orcutt method could handle more.

For illustration consider the two variables regression model:

$$Y_t = \beta_0 + \beta X_t + e_t \tag{3.19}$$

And assume that the error term follows the AR (1) scheme, namely,

$$e_t = \rho e_{t-1} + \varepsilon_t \qquad -1 \le \rho \le 1 \tag{3.20}$$

If (3.19) holds true at time t, it also holds true at time (t-1). Hence,

$$Y_{t-1} = \beta_0 + \beta X_{t-1} + e_{t-1} \tag{3.21}$$

Multiplying (3.21) by ρ on both sides, we obtain

$$\rho Y_{t-1} = \rho\beta_0 + \rho\beta X_{t-1} + \rho e_{t-1} \tag{3.22}$$

Subtract (3.22) from (3.19)

$$Y_t - \rho Y_{t-1} = \beta_0(1-\rho) + \beta(X_t - \rho X_{t-1}) + e_t - \rho e_{t-1}$$

$$Y_t^* = \beta_0^* + \beta^* X_t^* + u_t \tag{3.23}$$

Where

$$\beta_0^* = \beta_0(1-\rho), Y_t^* = Y_t - \rho Y_{t-1}, X_t^* = X_t - \rho X_{t-1}, \beta^* = \beta \text{ and } u_t = e_t - e_{t-1}$$

This model is called C**ochrane-Orcutt transformed model** where the error term u_t in equation (3.23) satisfy the usual OLS assumption. We can apply OLS to the transformed variables Y^* and X^* and obtain estimators with all the optimal properties, namely, BLUE (best linear unbiased estimators).

The regression in (3.23) is also known as the generalized difference equation. It involves regressing Y on **X**, not in the original form, but in the difference form, which is obtained by subtracting a proportion (ρ) of the values of variable in the previous time period from its value in the current time period. In this differencing procedure we lose one observation because the first observation has no antecedent. To avoid this lose of one observation, the first observation on Y and **X** is transformed as follows;

$$Y_1\sqrt{1-\rho^2} \text{ and } X_1\sqrt{1-\rho^2}$$

This transformation is known as the **Prias-Winsten transformation. This model** assumes only first order serial correlation (AR (1)). Detail discussion can be found, in Gujarati (2003).

3.3.2 Nonparametric Regression Analysis

Parametric regression model of the form $Y_t = f(\beta, X_t^{'}) + \varepsilon_t$ where $\beta = (\beta_0, \beta_2, \ldots, \beta_{p-1})^{'}$ is a vector of parameters to be estimated, and $X_t^{'} = (x_0, x_1, x_2, \ldots, x_{p-1})$ is a vector of; predictors for the t^{th} of n observations, the errors ε_t are assumed to be normally and independently distributed with mean 0 and constant variance σ^2. The function *f (·)*, relating the average value of the response Y to the predictors, is specified in advance, such specification can be linear or nonlinear type. The general nonparametric regression model is written in a similar manner, but the function *f* is left unspecified:

The object of nonparametric regression is to estimate the regression function *f* (·) directly, rather than to estimate parameters. Most methods of nonparametric regression implicitly assume that *f* (·) is a smooth, continuous function. As in nonlinear regression, it is standard to assume that ε_t ~ NID (0, σ^2). There are many nonparametric regression models, but this study uses the lowess method.

3.3.2.1.The Lowess Method

The Lowess method is a nonparametric method for estimating regression surfaces pioneered by Cleveland, Devlin, and Grosse (1988), Cleveland and Grosse (1991), and Cleveland, Grosse, and Shyu (1992). The Lowess method allows great flexibility because no assumptions about the parametric form of the regression surface are needed. You can use the Lowess method for

situations in which you do not know a suitable parametric form of the regression surface. Furthermore, the Lowess method is suitable when there are outliers in the data and a robust fitting method is necessary.

In the Lowess method, weighted least squares is used to fit linear or quadratic functions of the predictors at the centers of neighborhoods. The radius of each neighborhood is chosen so that the neighborhood contains a specified percentage of the data points. The fraction of the data, called the smoothing parameter, in each local neighborhood controls the smoothness of the estimated surface. Data points in a given local neighborhood are weighted by a smooth decreasing function of their distance from the center of the neighborhood.

In a direct implementation, such fitting is done at each point at which the regression surface is to be estimated. A much faster computational procedure is to perform such local fitting at a selected sample of points in predictor space and then to blend these local polynomials to obtain a regression surface.

You can use the Lowess method to perform statistical inference provided the error distribution satisfies some basic assumptions. In particular, such analysis is appropriate when the e_i are i.i.d. normal random variables with mean 0. By using the iterative reweighing, the Loess procedure can also provide statistical inference when the error distribution is symmetric but

not necessarily normal. Furthermore, by doing iterative reweighing, you can use the Loess procedure to perform robust fitting in the presence of outliers in the data.

The local regression smoothing process follows these steps for each data point:

1. Compute the regression weights for each data point in the span. The weights are given by the tricube function shown below.

$$w_i = \left(1 - \left|\frac{x - x_i}{d(x)}\right|^3\right)^3$$

x is the predictor value associated with the response value to be smoothed, x_i are the nearest neighbors of x as defined by the span, and d(x) is the distance along the abscissa from x to the most distant predictor value within the span. The weights have these characteristics:

- The data point to be smoothed has the largest weight and the most influence on the fit.
- Data points outside the span have zero weight and no influence on the fit.

2. A weighted linear least squares regression is performed where the regression uses a first degree polynomial.

3. The smoothed value is given by the weighted regression at the predictor values of interest.

In order to fit the Lowess method to the teledensity data, the study employs the Loess procedure in SAS that has the following main features:

- fits nonparametric models
- supports the use of multidimensional data
- supports multiple dependent variables
- performs statistical inference
- performs iterative reweighting to provide robust fitting when there are outliers in the data
- supports multiple score statements

Detail discussion can be found, in http://www.d.umn.edu/math/docs/saspdf/stat/chap38.pdf

Advantages of Lowess

As discussed above, the biggest advantage Lowess has over many other methods is the fact that it does not require the specification of a function to fit a model to all of the data in the sample. Instead the analyst only has to provide a smoothing parameter value and the degree of the local polynomial. In addition, Lowess is very flexible, making it ideal for modeling complex processes for which no theoretical models exist. These two advantages, combined with the simplicity of the method, make Lowess one of the most attractive of the modern regression methods for applications that fit the general framework of least squares regression but which have a complex deterministic structure.

Although it is less obvious than for some of the other methods related to linear least squares regression, Lowess also accrues most of the benefits typically shared by those procedures. The most important of those is the theory for computing uncertainties for prediction and

calibration. Many other tests and procedures used for validation of least squares models can also be extended to lowess models.

Disadvantages of Lowess

Although Lowess does share many of the best features of other least squares methods, efficient use of data is one advantage that Lowess doesn't share. Lowess requires fairly large, densely sampled data sets in order to produce good models. This is not really surprising: However, since Lowess needs good empirical information on the local structure of the process in order to perform the local fitting. In fact, given the results it provides, Lowess could arguably be more efficient overall than other methods like nonlinear least squares. It may simply frontload the costs of an experiment in data collection but then reduce analysis costs.

Another disadvantage of Lowess is the fact that it does not produce a regression function that is easily represented by a mathematical formula. This can make it difficult to transfer the results of an analysis to other people. In order to transfer the regression function to another person, we need the data set and software for Lowess calculations. In nonlinear regression, on the other hand, it is only necessary to write down a functional form in order to provide estimates of the unknown parameters and the estimated uncertainty. Depending on the application, this could be either a major or a minor drawback to using lowess.

CHAPTER FOUR

DATA ANALYSIS

4.1 Model specification

Before we fit the linear regression model, first we check for linear functional form based on graphical displays of the dependent variable with each of the independent variables. The plots displayed in Figure 4.1 (Appendix B) indicate that the relationship between the dependent and explanatory variables is near linear.

Next we fit a multiple linear regression model, for the dependent variable when all the explanatory variables are included, the functional form is:

$$TD_t = \beta_0 + \beta_1 GDPC_t + \beta_2 SISS_t + \beta_3 PGGDP_t + \beta_4 ARGET_t + \beta_5 GINVST_t + e_t \quad (4.1)$$

where t = year, t = 1970, 1971 . . ., 2000

TD_t = teledensity of fixed telephone in year t;

$GDPC_t$ = GDP per capita in year t;

$PGGDP_t$ = Percentage growth of GDP in year t;

$ARGET_t$ = Average revenue generated by each telephone in year t;

$GINVST_t$ = Growth of investment in year t;

$SISS_t$ = Size of service sector contributed to GDP in year t;

The result of OLS estimates for the multiple linear regression model given in equation (4.1) are shown in Table 4.1 (Appendix A).

4.2. Model Diagnostic

4.2.1. Test of Multicollinearity

The correlation matrix is presented in Table 4.2 (Appendix A). Some of the correlation coefficients between the explanatory variables are large and statistically significant. To check whether these correlations create the so called multicollinearity problem, a formal assessment is done using variance inflation factors and condition number criteria.

In this assessment, first VIF is applied to detect multicollinearity in the model. It has been noted that if any of the VIF is greater than 10, those variables are highly related to the other regressors. But in all cases as shown in Table 4.3, it is found that none of the variance inflation factor is greater than 10. Hence there is no problem of **colinearity** between independent variables**.**

Table 4.3 variance inflation factors

Variable	GDPC	SISS	ARGET	PGGDP	GINVST
VIF	4.64	2.97	2.84	1.86	1.68
1/VIF	0.215695	0.336265	0.351916	0.537347	0.594317

We also used the condition number criterion to check the problem of multicollinearity. The eigenvalues and the condition numbers are obtained as follows:

Eigenvalues of **(X'X)** are: λ_1 λ_2 λ_3 λ_4 λ_5

0.684 0.321 0.200 0.086 0.050

$$\text{Condition number} = \frac{\max \text{eigenvalue}}{\min \text{eigenvalue}}$$

$$= 13.68 < 100$$

Hence, there is no serious problem of multicollinearity.

4.2.2 Testing the residuals for normality

i. Using p-p plot

As displayed in Figure 4.2 (Appendix B) for the p-p plot, we observe that the cumulative probability from the data and expected cumulative probability from the theoretical normal distribution are along the 45^0 line. This indicates that there is no violation of the assumption of normality.

ii. Formal test of normality based on skewness-kurtosis test

The test results are given as

Skewness/Kurtosis tests for Normality

Variable	Pr(Skewness)	Pr(Kurtosis)	adj chi2(2)	Prob>chi2
resid	0.906	0.291	1.20	0.5479

The chi-square test is used to test whether the assumption of normality is accepted or not. In this assessment, the chi-square value is not significant (p = 0.5479) at the 0.05 level of significant. This means that we do not reject the normality assumption (Gujarati, 1995). In other words, it indicates that the data are not inconsistent with a sample from a normal distribution (Norusis, 1995).

4.2.3. Testing the Residuals for heteroskedasticity

i. A Graphical check of heteroskedasticity

Plots of residuals against time can be used to examine whether the variance of the error term is constant or not. As shown in Figure 4.3 (Appendix B), the spread of the residual does not follow an increasing or decreasing pattern, thus we can say the residuals are homoscedastic.

ii. Formal test of heteroskedasticity

The Breusch-Pagan test for heteroskedasticity states the null hypothesis as

Ho: Constant variance

The resulting chi-square test statistic value is 0.03 with p-value = 0.8610. This result indicates absence of heteroskedasticity at 0.05 level of significance. That is, such a result indicates the presence of equal variance of the residuals along the predicted line.

4.2.4. Checking for outliers

From the plot of standardized residuals shown in Figure 4.4 (Appendix B) there is no observation whose value of standardized residuals greater than 3.0 or less than -3.0, thus there is no problem of outlier.

Using Cook's d statistic one **observation** (observation 31) which is the last and most recent one is identified as possible influential outlier since Cook's d value greater than 1. However, removing this observation and fitting the regression model did not change the parameter estimates. Therefore the observation is retained in the data.

4.2.5. Testing the residuals for Autocorrelation

I. Graphical check of residuals for autocorrelation

From the scatter plot of residuals versus time in Figure 4.3 (Appendix B), neighboring residuals are clustering on either side of the line, $\hat{e} = 0$. This might be a sign that the errors are autocorrelated. However, the graphical interpretation needs to be supported by formal tests of autocorrelation.

II. Formal test of residuals for autocorrelation.

1. One can use the Durbin-Watson d test, to test for first-order autocorrelation, that is

H_0: ρ=0; H_1: ρ≠0 at $\alpha = 0.05$

Note that under the null the errors at t-1 and t are independent or the observations are not serially dependent. The Durbin-Watson d statistic is obtained as 1.153. The corresponding tabulated value for n = 31, k = 5 and $\alpha = 0.05$ is $d_l = 1.09$ and $d_u = 1.83$. Since the calculated DW d statistic is between the upper and lower value it is in the inconclusive region hence we need other statistical test to check serial correlation.

2. Breusch-Godfrey test

We also used Breusch-Godfrey LM test for autocorrelation the results are given as

lags(p)	Chi2	Df	Prob> chi2
1	6.735	1	0.0095

Similarly, since the p-value is very small (p = 0.0095) as compared to the 0.05 level of significance, we reject the null hypothesis and conclude that there is autocorrelation.

3. Ljung-Box Q statistic

The Durbin-Watson test of autocorrelation is based on autoregressive model of order 1, AR (1). Now we use autocorrelation, partial autocorrelation, and Ljung-Box Q-statistic to determine the order of autoregressive terms for the residual. The results are presented in Table 4.4 (Appendix A) and Figure 4.5 (Appendix B).

From Table 4.4, the significance of the autocorrelations for lags up to 16 using Ljung-Box Q-statistic is shown in the prob column. As can be seen from this column the autocorrelation only at lag 1 is significant and all other lags of these residuals have no any significant autocorrelation at 0.05 level of significant. Similarly, the partial autocorrelation function (PACF) shown in Figure 4.5, indicates only lag 1 is significant, since it is out of the confidence limit all other partial autocorrelation found not to be significant.

So, from both graphical and formal test of autocorrelation, it is appropriate to fit AR (1) model, such that

$$e_t = \rho e_{t-1} + \varepsilon_t$$ where ε_t is white noise (which satisfies all assumption of LRM)

To estimate ρ, we regress the estimated residuals ($\hat{e}_t$ on $\hat{e}_{t-1}$ without a constant). The estimated value of $\hat{\rho}$ is equal to 0.413.

Coefficients[a,b]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	Lres	.413	.174	.403	2.374	.024

a. Dependent Variable: Unstandardized Residual

b. Linear Regression through the Origin

4.3. Fitting the final model (autocorrelation remedy model)

4.3.1 Cochrane- Orcutt and Prais-Winston

In order to estimate the parameter efficiently, we apply the following (Cochrane- Orcutt) transformation:

$$TD_t = \beta_0 + \beta_1 GDPC_t + \beta_2 SISS_t + \beta_3 PGGDP_t + \beta_4 ARGET_t + \beta_5 GINVST_t + e_t \qquad (4.1)$$

Note e_t the disturbance term follows first-order autoregressive process and multiplying (4.1) by ρ we have

$$\rho TD_{t-1} = \rho\beta_0 + \rho\beta_1 GDPC_{t-1} + \rho\beta_2 SISS_{t-1} + \rho\beta_3 PGGDP_{t-1} + \rho\beta_4 ARGET_{t-1} + \rho\beta_5 GINVST_{t-1} + \rho e_{t-1} \qquad (4.2)$$

Further, subtract equation (4.2) from (4.1) we get:

$$TD_t - \rho TD_{t-1} = \beta_0(1-\rho) + \beta_1(GDPC_t - \rho GDPC_{t-1}) + \beta_2(SISS_t - \rho SISS_{t-1}) + \beta_3(PGGDP_t - \rho PGGDP_{t-1}) + \beta_4(ARGET_t - \rho ARGET_{t-1}) + \beta_5(GINVST_t - \rho GINVST_{t-1}) + (\varepsilon_t - \rho e_{t-1}) \quad (4.3)$$

$$where \; e_t - \rho e_{t-1} = \varepsilon_t.$$

Note that equation (4.3) fulfills all basic assumptions of the LRM and thus, we can estimate the parameters in this equation efficiently by OLS using two stages. Since ρ is not known, in the first stage we estimate ρ as discussed above. In the second stage we replace ρ by the estimated value ρ = 0.451 in the above transformation and we then regress

$TD_t - 0.413TD_{t-1}$ on the transformed explanatory variables

$$GDPC_t - 0.413\ln GDPC_{t-1},\;\; PGGDP_t - 0.413\ln PGGDP_{t-1},\;\; SISS_t - 0.413\ln SISS_{t-1},$$
$$ARGET_t - 0.413\ln ARGET_{t-1},\;\; GINVST_t - 0.413\ln GINVST_{t-1},$$

and the constant term $\beta_o(1\text{-}0.413)$ to obtain estimates of the regression coefficients. The results are shown in Table 4.5. In addition, using the Prais-Winston method the estimated regression coefficients are presented in Table4.5. Note that the results of Cochrane-Orcutt and Prais-Winston transformations provided results that are close to each other.

Table 4.5

Result of Cochrane-Orcutt and Prais-Winston model

Variables	Cochrane-Orcutt		Prais-Winston	
	$B_{(C.O)}$ coefficient estimates	**P-value**	**$B_{(P.W)}$ coefficient estimates**	**P-value**
constant	0.18392	0.000	0.17676	0.000
GDPC	.000398	0.000	.0003822	0.000
SISS	.0.04287	0.000	.04353	0.000
PGGDP	-.000734	0.403	-.000466	0.399
ARGET	-.000088	0.001	-.000081	0.000
GINVST	.0000687	0.238	.000061	0.252
F	143.87	0.000	88.34	0.000
$\mathbf{R^2}$ adjusted	0.9610		0.9357	
D.W	1.5240		1.6187	

From the results in Table 4.5 using both Cochrane-Orcutt and Prais-Winston method it can be seen that variables GDPC, SISS and ARGET are highly significant ($p < 0.05$), but PGGDP and GINVST are not significant. The adjusted $\mathbf{R^2}$ is very high and **F**-value indicates that the regression is significant in general.

In the same table **D .W** values of Cochrane-Orcutt and Prais-Winston methods are given as 1.524 and 1.619, respectively. Eventhough D.W value of Prais-Winston method are slightly higher than the value obtained by Cochrane-Orcutt method, these results fall in the inconclusive region we need further statistical test to check for serial autocorrelation.

Detail output of Cochrane-Orcutt method and Prais-Winsten method are presented in Table 4.6 and Table 4.7, respectively (Appendix A).

To check whether the autocorrelation is removed we used the partial autocorrelation function of the residuals of the transformed model displayed in Figure 4.6. There is no significant partial autocorrelation among the residuals at all lags. This, in turn, explains that there is no problem of autocorrelation in the transformed model. Similarly, as can be seen from Table 4.8, none of the autocorrelations from the transformed model are significant.

Moreover, the Breusch-Godfrey LM test for autocorrelation using the residuals of the transformed model is given below:

lags(p)	chi2	Df	Prob> chi2
1	1.183	1	0.1754

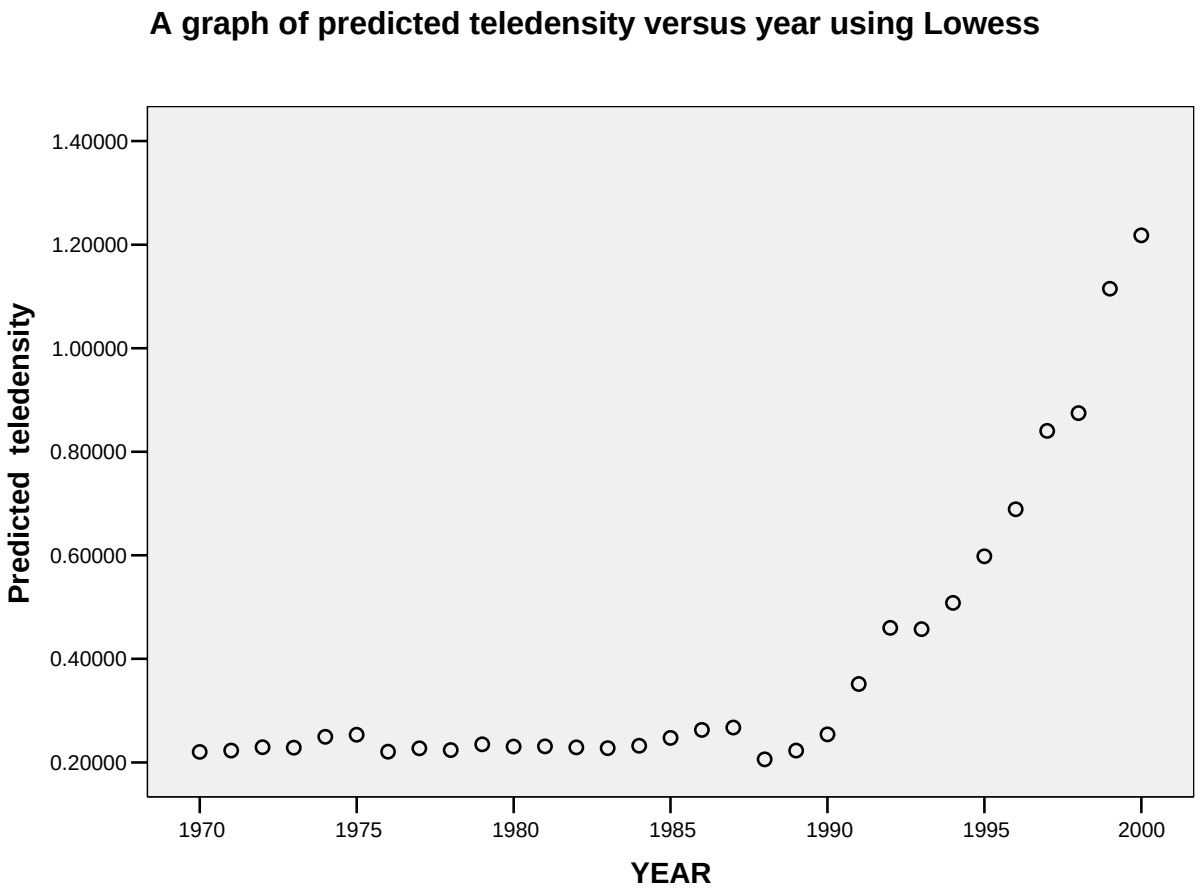
From the result the p value which is (p = 0.1754) is large as compared to the 0.05 level of significance, which means we cannot reject the null hypothesis of no serial correlation and conclude that the autocorrelation problem is no longer present. Therefore, the regression coefficient estimates and test results from the transformed model are taken as the final model estimates that helps to make conclusion and recommendations as discussed in the next chapter.

4.3.2. Nonparametric regression using Lowess method

As shown the output in Table 4.9, the optimal smoothing parameter and the residual sum of square are 0.91935 and 0.02833, respectively. When we compare this residual sum of square with the others residual sum of square having smoothing parameter different from 0.91935, the optimal smoothing parameter is the one having the minimum residual sum of square.

As shown in Figure 4.7 below, the graph of predicted teledensity using Lowess versus year is almost similar to the actual teledensity versus year in Figure 1.1.

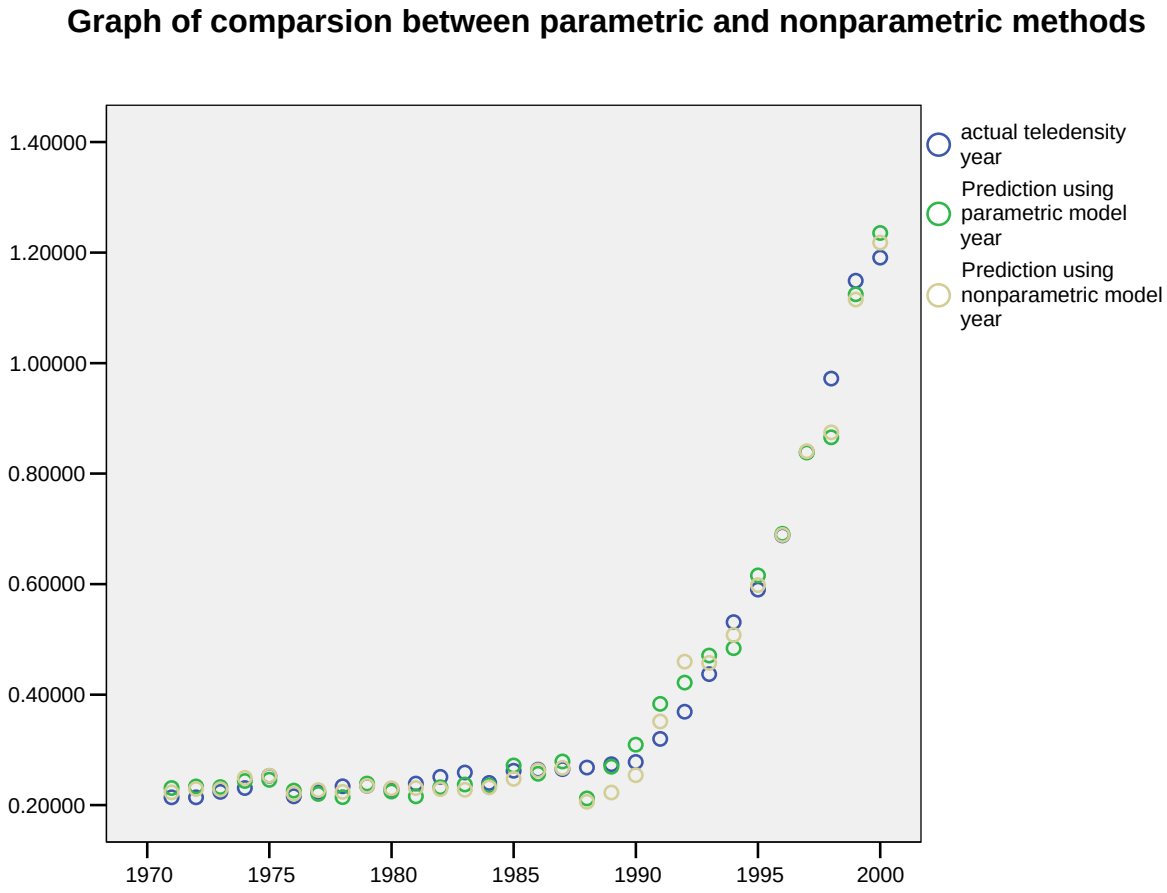
Figure4.7



4.5. Graphical Comparisons of parametric and nonparametric analysis

The actual teledensity and predicted teledensity using parametric models based on Cochrane-Orcutt transformations and nonparametric method using Lowess versus time are displayed in Figure 4.8.

Figure 4.8



From the plots in Figure 4.8, we observe that both the nonparametric and the parametric regression results are very close to the actual values which implies that both the nonparametric and the parametric regression model efficiently fit the teledensity data under consideration. This is further supported by approximately equal residual sum of square error values 0.028 and 0.032 for the nonparametric and the parametric regression model, respectively.

CHAPTER FIVE

Conclusion and Recommendation

5.1 Conclusions

This paper analyzes the effect of GDP per capita, the percentage contribution of the service sector to GDP, average revenue generated by each telephone line, growth of investment in telecommunications infrastructure and percentage growth of GDP on teledensity. The analysis was done based on parametric and nonparametric method.

In this study, we employed multiple linear regression analysis to these factors that are expected to facilitate the growth of teledensity. The results of both Cochrane-Orcutt AR (1) regression and Prais-Winston AR (1) regression estimates indicate that gross domestic product per capita (GDPC), size of service sector's share in the economy of Ethiopia (SISS) and average revenue generated by each telephone line (ARGET) have statistically significant effects on the growth of teledensity. Both Cochrane-Orcutt AR (1) regression and Prais-Winston AR (1) regression estimates explained more than 90 percent of the variation in the dependent variable, teledensity by the whole set of predictors.

As illustrated in the output of Cochrane-Orcutt AR (1) regression and Prais-Winston AR (1) regression, it is found that two variables namely, percentage growth of gross domestic product (PGGDP) and growth of investment in telecommunication infrastructure (GINVST) have poor performance in determining teledensity.

Specifically, the findings suggest that increased investment in telecommunications infrastructure in Ethiopia is not a major determinant for growth of teledensity, but that higher GDP per capita and higher contributions of the service sector share to Gross Domestic Product (GDP) of Ethiopia are major determinants for growth of teledensity and there is a significant negative relationship between average revenue generated by each telephone line (ARGET) and level of teledensity in Ethiopia.

In this study, as shown in Figure 4.8, prediction values based on nonparametric regression analysis (Lowess) and parametric regression analysis (Cochrane-Orcutt) are closer to the actual teledensity. Hence we can say that, nonparametric regression analysis almost equally fits the data like the parametric regression analysis. In addition, as shown in the output of Cochrane-Orcutt and Prais-Winston for parametric case and Lowess for nonparametric case, the residual sum of square for nonparametric regression analysis is approximately equal to that of parametric regression residual sum of square. Therefore, from the point of fitting the data, we also used the nonparametric regression analysis alternatively with the parametric regression analysis for teledensity data. But the parametric model has the advantage of explaining the relationship between each explanatory variable with the dependent variable.

5.2. **Recommendation**

Based on the result of the study we forward some recommendations. The study shows that average yearly usage charge per telephone line has a negative significant relationship with the

growth of teledensity. It seems that a lower usage charge is critical in attracting the consumers to the service; as a result it is critical for faster growth of teledensity.

Another point is that investment in telecommunications infrastructure is not significant. This does not mean that investment is not important, but the other factors are vital to the growth of teledensity. We cannot expect to improve teledensity by embarking on investment, unless we take into account qualified staff to take care of needed maintenance and providing quality service to reduce the waiting time for fixed line.

REFERENCE

- Agustin, R.S. (2003). The Impact of the Regulatory Process and Price Cap Regulation in Latin American Telecommunications Markets, Near Economic Consulting, Digital Graf Press.
- Cadima, N and Barros, P.P. (2000). Impact of Mobile Phone Diffusion on the Fixed-link Network, Center for Economic Policy Research, London.
- Callender, J., Golden, B., Lele, S. and Wasil, E. (2003). Identifying Investment Opportunities in International Telecommunications Markets Using Regression Models: ICS Conference.
- Clarke, D.G. and Laufenberg, A. (1983). The Role of Telecommunications in Economic Development: With special Reference to Rural Sub-Sahara Africa, Case study No.4 referred to in the synthesis Report on the ITU-OECD project "Telecommunications for Development."
- Cleveland, W.S., Devlin, S.J., and Grosse, E. (1988), "Regression By Local Fitting," Journal of Econometrics, 37, 87–114.
- Cleveland, W.S. and Grosse, E. (1991). "Computational Methods for Local Regression, "Statistics and Computing, 1, 47–62.
- Cleveland, W.S., Grosse, E. and Ming-Jen Shyu (1992), "A Package of C and Fortran Routines for Fitting Local Regression Models," unpublished paper.
- Draper, N.R. and Smith.H. (1998). Applied Regression Analysis: John Wiley and Sons. Inc.
- Ethiopia Telecommunication Corporation (2000). Annual statistical bulletin.

- Gujarati, D.N. (1995). Basic Econometrics New York: McGraw-Hill.
- Gujarati, D.N. (2003). Basic Econometrics: Gary Burko.
- Gutierrez, L. H. And Berg, S. (2000). Telecommunications Liberalization and Regulatory Governance: Lessons from Latin America, Telecommunications Policy, *www.ingentaconnect.com/content/els/03085961/**2000**/.../art00069*
- http://www.telecom.net.et/ (2001). About ETC History.
- International Telecommunications Union (ITU) (1995). Yearbook of Public Telecommunication Statistics, Geneva: ITU
- International telecommunications Union (ITU) (1994). World Telecommunications Development Report, Geneva: ITU.
- International Telecommunications Union (ITU) (1997). Yearbook of Public Telecommunication Statistics, Geneva: ITU.
- International telecommunications Union (ITU) (1998). World Telecommunications Development Report, Geneva: ITU.
- The Maitland Commission) (1984). 'Report of the Independent Commission for World Wide Telecommunication Development' The Missing Link Geneva: ITU.
- Mbarika. V (1999). Factors that Influence Growth of Teledensity in Least Developed Countries: Proceeding of the 22nd *Conference on Information Systems Research in the Scandinavia,* Jyvaskyla, Finland, 383-396.

- Mbarika, V., Kah, M.M., Musa, P.F., Meso, P. and Warren, J. (2003). Predictors of Growth of Teledensity in Developing Countries: A Focus on Middle and Low-Income Countries, *The electronic Journal on Information Systems in Developing Countries*, 12, 1-16.
- Norusis, M.J. (1995). SPSS 6.1 Guide to Data Analysis New Jersey: Prentice Hall
- Ogbe, O. (1990). Introductory Remarks, in proceedings of Africa Telecom '90, ITU, Harare. Geneva: ITU.
- Paltridge, S. (1994). A survey of tariff structures in Africa - comparison to the rest of the world. Africa Telecom '94, Cairo.
- Saunders, R., Warford, J.I. and Wellinius, R. (1994). Telecommunications and Economic Development (2nd ed) Baltimore.MD: John Hopkins University Press.
- Sung, N. and Lee, Y.H. (2002). Substitution between Mobile and Fixed Telephones in Korea, Review of Industrial Organization, 20, 367-374.
- Vanu, I. (2005) Announces Commercial Availability of Any wave Base Station; First FCC-Certified Software Radio Device Deployed Live in Mid Tex Cellular Network, *www.marketwire.com/press-release/**Vanu**-**Inc**-791779.html*
- Wallsten, S. J. (2001). "An Econometric Analysis of Telecom Competition, Privatization, and Regulation in Africa and Latin America," The Journal of Industrial Economics: 49, 1 - 19

APPENDIX TABBLES AND GRAPHS

Appendix A: Tables

Table 1.1

	Teledensity of	Ethiopian 2000 E.F.Y.	African Countries average (2007) forecast			
			North Africa	South Africa	Sub-Sahara	African
1	Fixed telephone	1.20	11.91	9.56	1.65	3.77
2	Mobile	2.56	53.39	87.08	18.28	27.48
	Total	3.78	65.30	96.64	19.93	31.25

Source ETC (2000), Annual bulletin

Table 1.2

Fixed Line Household Access	
Uganda	0.3%
Mozambique	1.7%
Cameroon	1.8%
Kenya	2.3%
Ghana	2.6%
Nigeria	2.7%
Benin	4.6%
Burkina Faso	4.7%
Cote dIvoire	4.8%
Ethiopia	7.6%
Botswana	11.0%
Namibia	17.4%
South Africa	18.2%

Source: LINK Centre, 2008

Research ICT Africa Household Survey

Table 4.1

Regress lnTD GDPC SISS PGGDP GINVST ARGET

```
    Source |       SS       df       MS              Number of obs =      31
-------------+------------------------------           F(  5,    25) =  265.31
       Model |  8.48587906     5  1.69717581            Prob > F      = 0.0000
    Residual |  .159922022    25  .006396881           R-squared     = 0.9815
-------------+------------------------------           Adj R-squared = 0.9778
       Total |  8.64580108    30  .288193369           Root MSE      =  .07998

------------------------------------------------------------------------------
        lnTD |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        GDPC |   .0006067   .0000491    12.34   0.000     .0005055    .0007079
        SISS |   .1194837   .0105834    11.29   0.000     .0976868    .1412807
       PGGDP |  -.0012205   .0021466    -0.57   0.575    -.0056415    .0032005
      GINVST |   .0000657   .0001421     0.46   0.648    -.0002269    .0003583
       ARGET |  -.0001214   .0000232    -5.24   0.000    -.0001691   -.0000737
       _cons |   -1.559006   .0351613   -44.34   0.000    -1.631422   -1.48659
------------------------------------------------------------------------------
```

Table 4.2

Correlations

		lnTD	ARGET	GINVSTinmill	GDPC	SISS	PGGDP
lnTD	Pearson Correlation	1	.676**	.626**	.940**	.882**	.590**
	Sig. (2-tailed)		.000	.000	.000	.000	.000
	N	31	31	31	31	31	31
ARGET	Pearson Correlation	.676**	1	.380*	.714**	.759**	.350
	Sig. (2-tailed)	.000		.035	.000	.000	.054
	N	31	31	31	31	31	31
GINVSTinmill	Pearson Correlation	.626**	.380*	1	.621**	.493**	.403*
	Sig. (2-tailed)	.000	.035		.000	.005	.024
	N	31	31	31	31	31	31
GDPC	Pearson Correlation	.940**	.714**	.621**	1	.740**	.657**
	Sig. (2-tailed)	.000	.000	.000		.000	.000
	N	31	31	31	31	31	31
SISS	Pearson Correlation	.882**	.759**	.493**	.740**	1	.403*
	Sig. (2-tailed)	.000	.000	.005	.000		.025
	N	31	31	31	31	31	31
PGGDP	Pearson Correlation	.590**	.350	.403*	.657**	.403*	1
	Sig. (2-tailed)	.000	.054	.024	.000	.025	
	N	31	31	31	31	31	31

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

Table 4.4

Autocorrelations

Series: Unstandardized Residual

Lag	Autocorrelation	Std.Error(a)	Box-Ljung Statistic		
			Value	df	Sig.(b)
1	.390	.171	5.178	1	.023
2	-.023	.168	5.196	2	.074
3	-.101	.165	5.569	3	.135
4	.019	.162	5.583	4	.233
5	-.093	.159	5.925	5	.314
6	-.201	.156	7.577	6	.271
7	-.202	.153	9.324	7	.230
8	-.121	.150	9.974	8	.267
9	-.083	.147	10.292	9	.327
10	-.010	.143	10.297	10	.415
11	-.147	.140	11.404	11	.410
12	-.191	.136	13.359	12	.343
13	-.073	.133	13.661	13	.398
14	-.052	.129	13.826	14	.463
15	.044	.125	13.952	15	.529
16	.071	.121	14.296	16	.577

a The underlying process assumed is independence (white noise).
b Based on the asymptotic chi-square approximation.

Table 4.6 Cochrane-Orcutt model

Cochrane-Orcutt AR(1) regression -- twostep estimates

Source	SS	df	MS	Number of obs =	30
Model	.96054119	5	.192108238	F(5, 24) =	143.87
Residual	.032046552	24	.001335273	Prob > F =	0.0000
Total	.992587742	29	.034227164	R-squared =	0.9677
				Adj R-squared =	0.9610
				Root MSE =	.03654

TD	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
GDPC	.0003976	.0000255	15.58	0.000	.0003449	.0004502
SISS	.0428738	.0063628	6.74	0.000	.0297416	.0560059
PGGDP	-.0007341	.0008632	-0.85	0.403	-.0025157	.0010474
GINVST	.0000687	.0000568	1.21	0.238	-.0000485	.0001859
ARGET	-.0000886	.0000157	-5.65	0.000	-.0001209	-.0000562
_cons	.1839228	.0249066	7.38	0.000	.1325182	.2353274
rho	.412775					

Durbin-Watson statistic (original) 1.153279

Durbin-Watson statistic (transformed) 1.524087

Table 4.7

Prais-Winsten AR(1) regression -- iterated estimates

Source	SS	df	MS	Number of obs =	31
Model	.557298795	5	.111459759	F(5, 25) =	88.34
Residual	.031542516	25	.001261701	Prob > F =	0.0000
Total	.588841311	30	.019628044	R-squared =	0.9464
				Adj R-squared =	0.9357
				Root MSE =	.03552

TD	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
GDPC	.0003822	.0000273	13.99	0.000	.000326	.0004385
SISS	.0435331	.0070381	6.19	0.000	.0290379	.0580284
PGGDP	-.0006637	.0007735	-0.86	0.399	-.0022569	.0009294
GINVST	.0000601	.0000512	1.17	0.252	-.0000453	.0001655
ARGET	-.0000805	.000017	-4.74	0.000	-.0001155	-.0000455
_cons	.1767672	.0272917	6.48	0.000	.1205589	.2329754

rho	.591009

Durbin-Watson statistic (original) 1.153279

Durbin-Watson statistic (transformed) 1.618216

Table 4.8

Autocorrelations

Series: Unstandardized Residual

Lag	Autocorrelation	Std.Error[a]	Box-Ljung Statistic		
			Value	df	Sig.[b]
1	.072	.174	.170	1	.680
2	-.133	.171	.780	2	.677
3	-.110	.168	1.209	3	.751
4	-.049	.165	1.299	4	.862
5	.010	.161	1.303	5	.935
6	-.019	.158	1.317	6	.971
7	-.104	.155	1.766	7	.972
8	-.146	.151	2.697	8	.952
9	-.227	.148	5.042	9	.831
10	.105	.144	5.570	10	.850
11	.026	.141	5.603	11	.898
12	-.047	.137	5.721	12	.930
13	.171	.133	7.381	13	.881
14	-.036	.129	7.459	14	.916
15	.013	.125	7.469	15	.943
16	.038	.121	7.569	16	.961

a. The underlying process assumed is independence (white noise).

b. Based on the asymptotic chi-square approximation.

Table 4.9

Teledensity data 20:33 Sunday, December 9, 2001

Independent Variable Scaling

Statistic	year	GDPC	ARGET	SISS	PGGDP	GINVST
Minimum Value	1970	227.17900	472.00000	0.09300	1.76200	-250.33000
Maximum value	2000	3016.82000	3659.33710	7.92000	33.86000	497.12600

Teledensity data 20:33 Sunday, December 9, 2001

Optimal Smoothing

Criterion

Smoothing

Parameter

-5.02173 0.91935

Output Statistics

Obs	GDPC	ARGET	SISS	PGGDP	GINVST	TD	Pred
1	227.17900	484.80000	0.23100	5.72700	4.83800	0.21300	0.22022
2	242.81100	472.00000	0.21100	9.60800	4.83800	0.21400	0.22299
3	249.02600	554.00000	0.36900	7.63800	10.35500	0.21400	0.22919
4	320.12400	606.00000	0.24700	31.68600	6.27400	0.22400	0.22856
5	324.31500	705.42000	0.14200	5.52400	0.92500	0.23100	0.24944
6	349.62700	713.80000	0.09600	10.71500	-18.15600	0.25300	0.25329
7	260.46300	756.90000	0.09300	1.80200	56.50900	0.21600	0.22060
8	300.49600	882.90000	0.12000	8.66600	12.85000	0.22400	0.22721
9	304.00300	1097.00000	0.23100	4.21200	2.66400	0.23400	0.22386
10	313.12900	1067.00000	0.48700	6.00900	11.10600	0.23500	0.23495
11	316.55800	1119.00000	0.33600	4.02700	13.99200	0.22900	0.23042
12	319.23100	1239.00000	0.38000	5.15400	-0.10400	0.23900	0.22093
13	331.58900	1097.00000	0.16500	6.88300	18.81900	0.25100	0.22899
14	367.62700	1213.00000	0.27400	14.08300	15.85000	0.25900	0.22764
15	386.98400	1414.00000	0.41800	8.31800	-45.58900	0.24000	0.23205
16	518.59600	1512.00000	0.58100	28.27700	53.94100	0.26200	0.24741
17	534.30600	1913.00000	0.77900	6.21500	24.91600	0.26400	0.26275
18	620.03700	2125.00000	1.10400	19.61300	15.24100	0.26500	0.26734
19	673.01000	3129.00000	1.18700	11.96000	5.51900	0.26800	0.20601

20	713.43900	3340.00000	1.91000	9.29800	22.32000	0.27400	0.22266
21	748.73600	3525.00000	3.06100	8.14000	58.56700	0.27800	0.25408
22	791.36200	3062.84100	3.98000	8.83800	4.77700	0.32000	0.35135
23	952.83500	2999.96700	3.87000	23.95800	85.48300	0.36900	0.43974
24	941.38400	2929.18000	4.27700	1.76200	347.61900	0.43700	0.45736
25	860.04700	2495.77200	5.40300	10.15700	-250.33000	0.53100	0.50792
26	1062.23000	2490.67200	5.29000	8.28500	19.02700	0.59000	0.59816
27	1218.86000	2496.85500	5.91700	18.01500	150.98500	0.68800	0.68887
28	1454.54000	2880.01300	7.92000	22.86100	181.84300	0.83900	0.84031
29	1750.95000	3337.55300	6.46600	23.66700	328.33100	0.97200	0.87482
30	2216.87000	2405.34600	5.05100	29.80900	497.12600	1.14900	1.11499
31	3016.82000	3659.33700	3.81700	33.86000	135.63200	1.19100	1.21805

Fit Summary

Fit method	**Interpolation**
Number of Observations	**31**
kd Tree Bucket Size	**5**
Degree of Local Polynomials	**1**
Smoothing Parameter	**0.91935**
Points in Local Neighborhood	**28**
Residual Sum of Squares	**0.02833**

Independent Variable Scaling

Statistic	year	GDPC	ARGET	SISS	PGGDP	GINVST

Minimum Value	1970	227.17900	472.00000	0.09300	1.76200	-250.33000
Maximum value	2000	3016.82000	3659.33710	7.92000	33.86000	497.12600

Fit Summary

Fit method	**Interpolation**
Number of Observations	**31**
kd Tree Bucket Size	**1**
Degree of Local Polynomials	**1**
Smoothing Parameter	**0.2000**
Points in Local Neighborhood	**6**
Residual Sum of Squares	**5.76727**

Fit Summary

Fit method	**Interpolation**
Number of Observations	**31**
kd Tree Bucket Size	**1**
Degree of Local Polynomials	**1**
Smoothing Parameter	**0.3000**
Points in Local Neighborhood	**9**
Residual Sum of Squares	**0.21217**

Fit Summary

Fit method	**Interpolation**
Number of Observations	**31**
kd Tree Bucket Size	**2**

Degree of Local Polynomials 1

Smoothing Parameter 0.4000

Points in Local Neighborhood 12

Residual Sum of Squares 0.04080

Fit Summary

Fit method Interpolation

Number of Observations 31

kd Tree Bucket Size 3

Degree of Local Polynomials 1

Smoothing Parameter 0.5000

Points in Local Neighborhood 15

Residual Sum of Squares 0.03117

Fit Summary

Fit method Interpolation

Number of Observations 31

kd Tree Bucket Size 3

Degree of Local Polynomials 1

Smoothing Parameter 0.6000

Points in Local Neighborhood 18

Residual Sum of Squares 0.03541

Fit Summary

Fit method Interpolation

Number of Observations 31

kd Tree Bucket Size	4
Degree of Local Polynomials	1
Smoothing Parameter	0.7000
Points in Local Neighborhood	21
Residual Sum of Squares	0.04492

Fit Summary

Fit method	Interpolation
Number of Observations	31
kd Tree Bucket Size	4
Degree of Local Polynomials	1
Smoothing Parameter	0.8000
Points in Local Neighborhood	24
Residual Sum of Squares	0.03123

Appendix B: Graphs

Figure 1.1

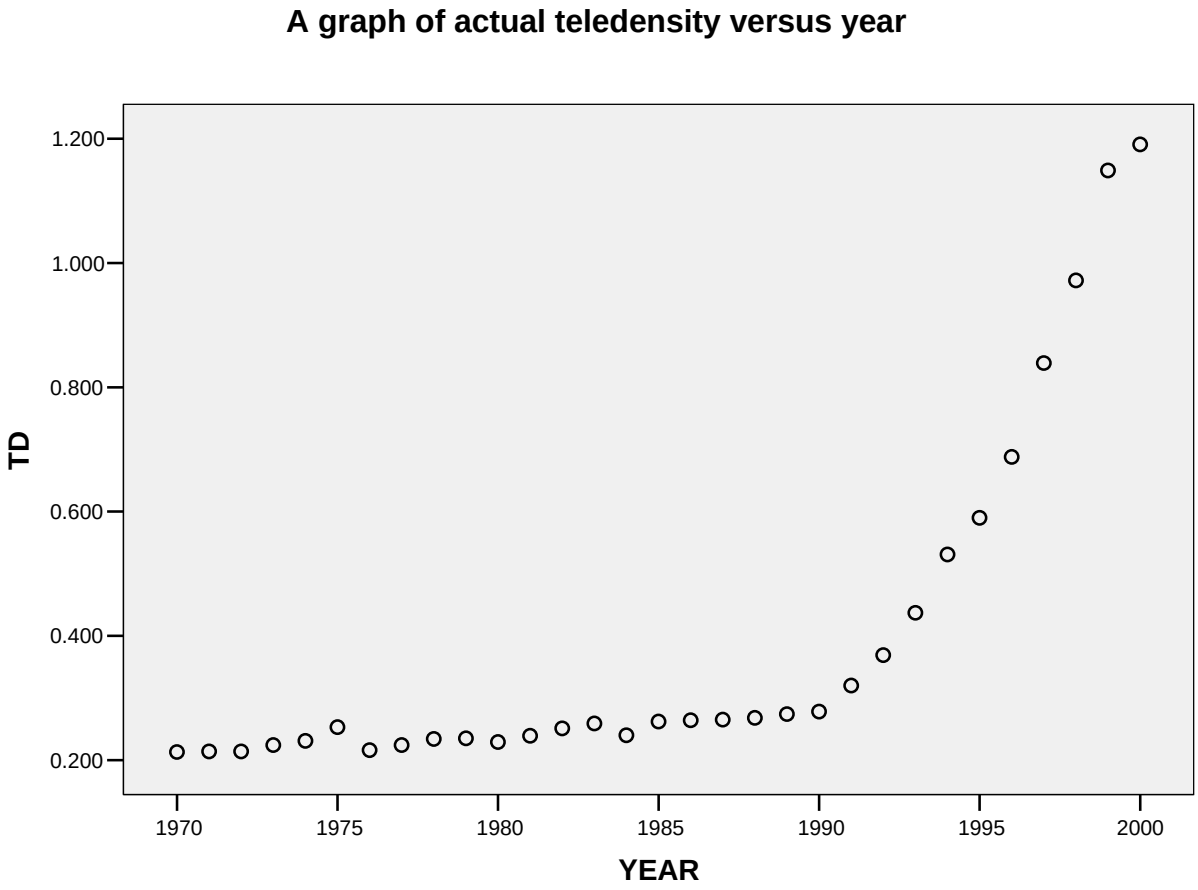

A graph of actual teledensity versus year
1.200
1.000
0.800
0.600
0.400
0.200
TD
1970
1975
1980
1985
1990
1995
2000
YEAR

Figure 4.1

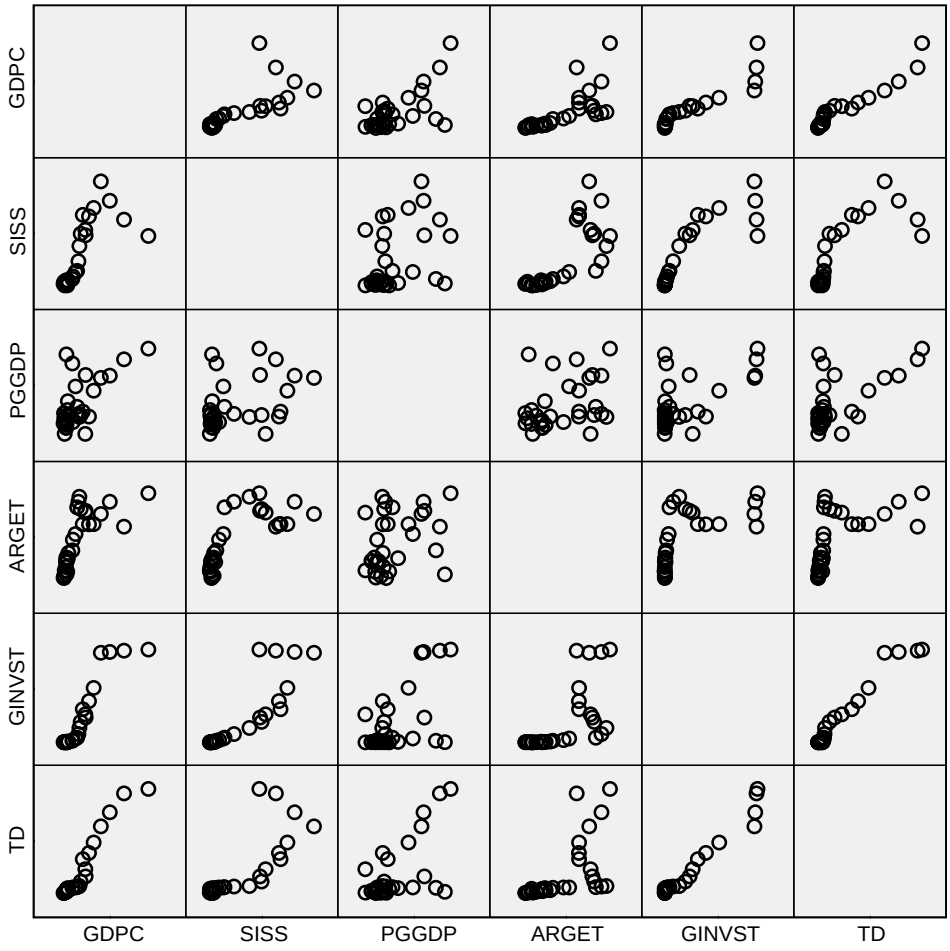

GDPC
SISS
PGGDP
ARGET
GINVST
TD
GDPC
SISS
PGGDP
ARGET
GINVST
TD

Figure 4.2

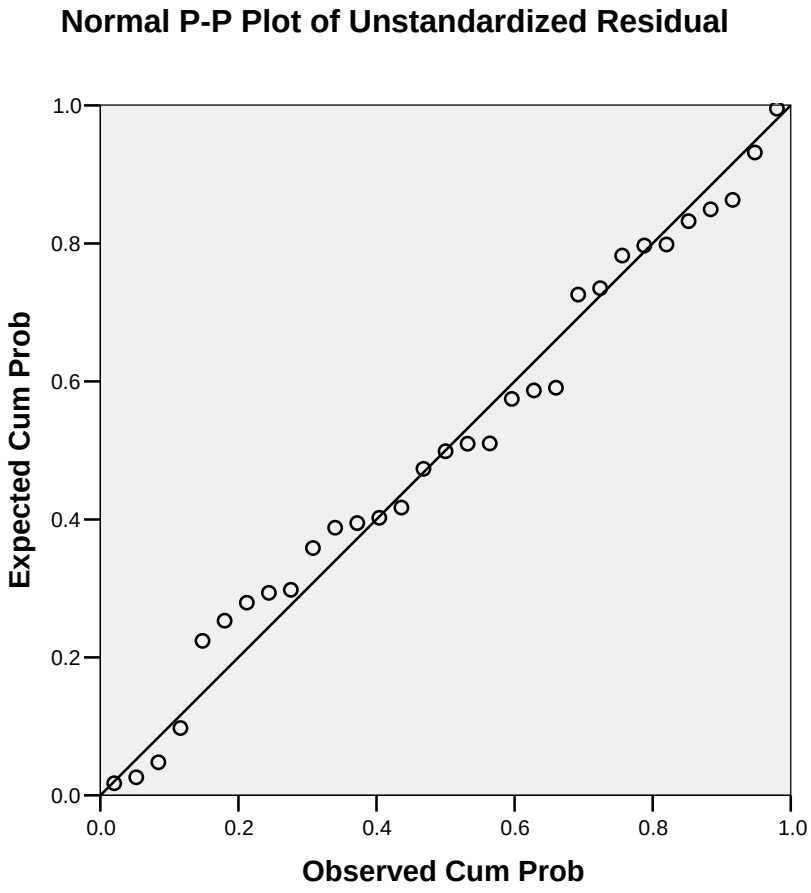


Figure 4.3

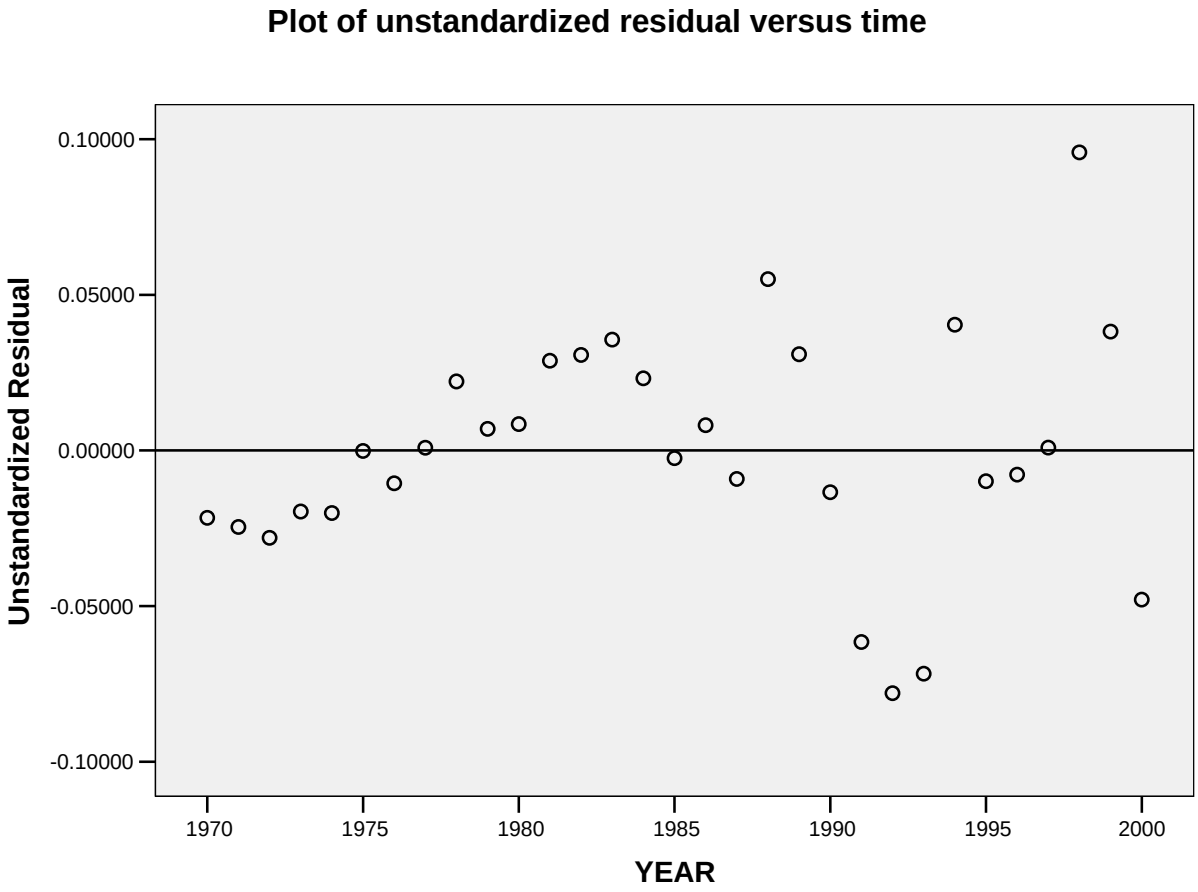

Plot of unstandardized residual versus time
Unstandardized Residual
0.10000
0.05000
0.00000
-0.05000
-0.10000
1970
1975
1980
1985
1990
1995
2000
YEAR

Figure 4.4

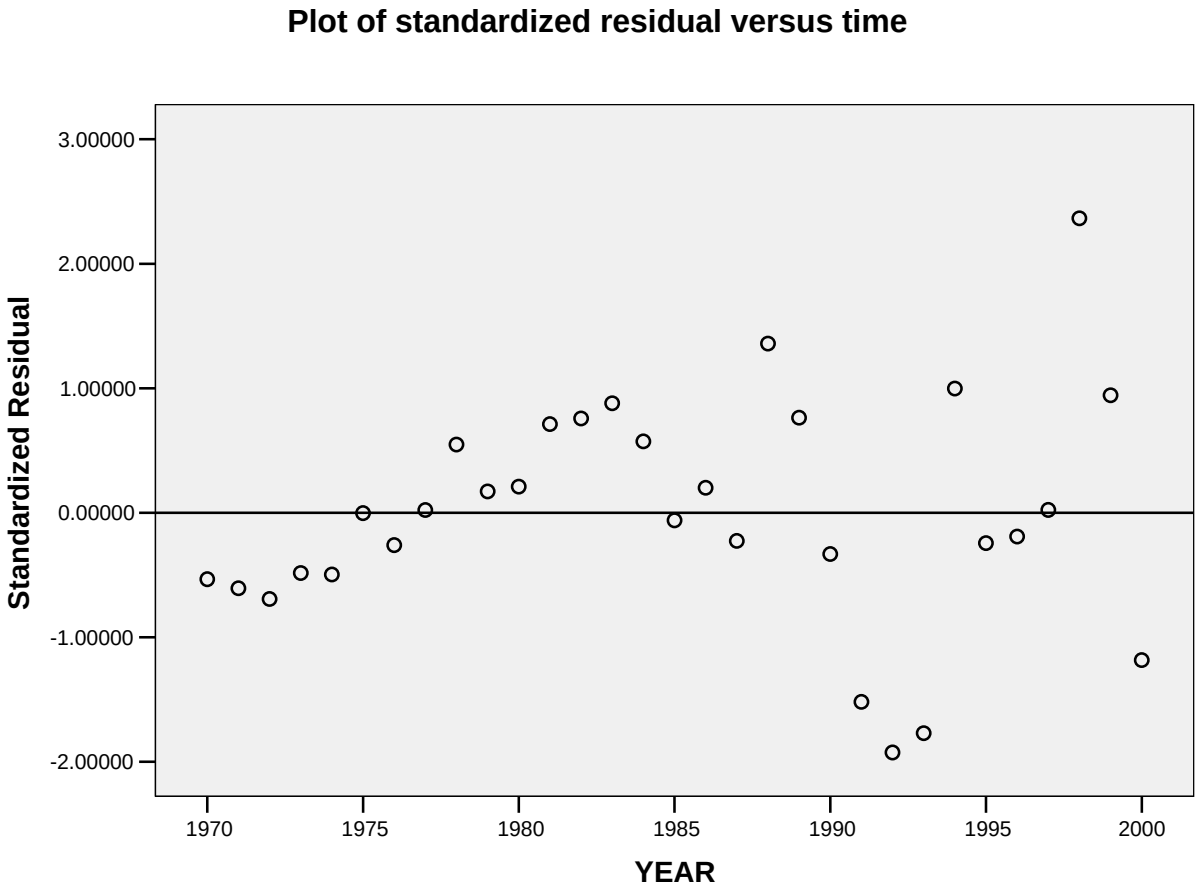

Plot of standardized residual versus time
Standardized Residual
3.00000
2.00000
1.00000
0.00000
-1.00000
-2.00000
1970
1975
1980
1985
1990
1995
2000
YEAR

Figure 4.5

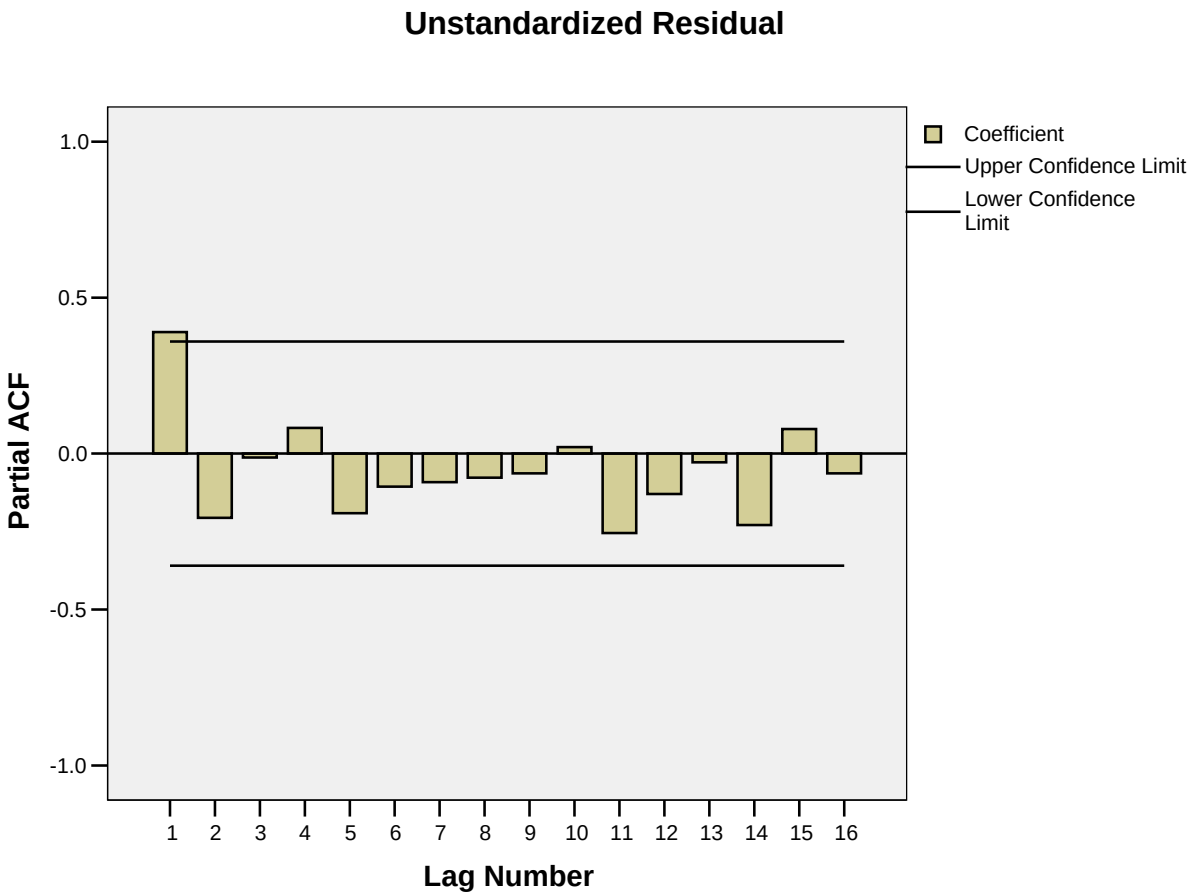
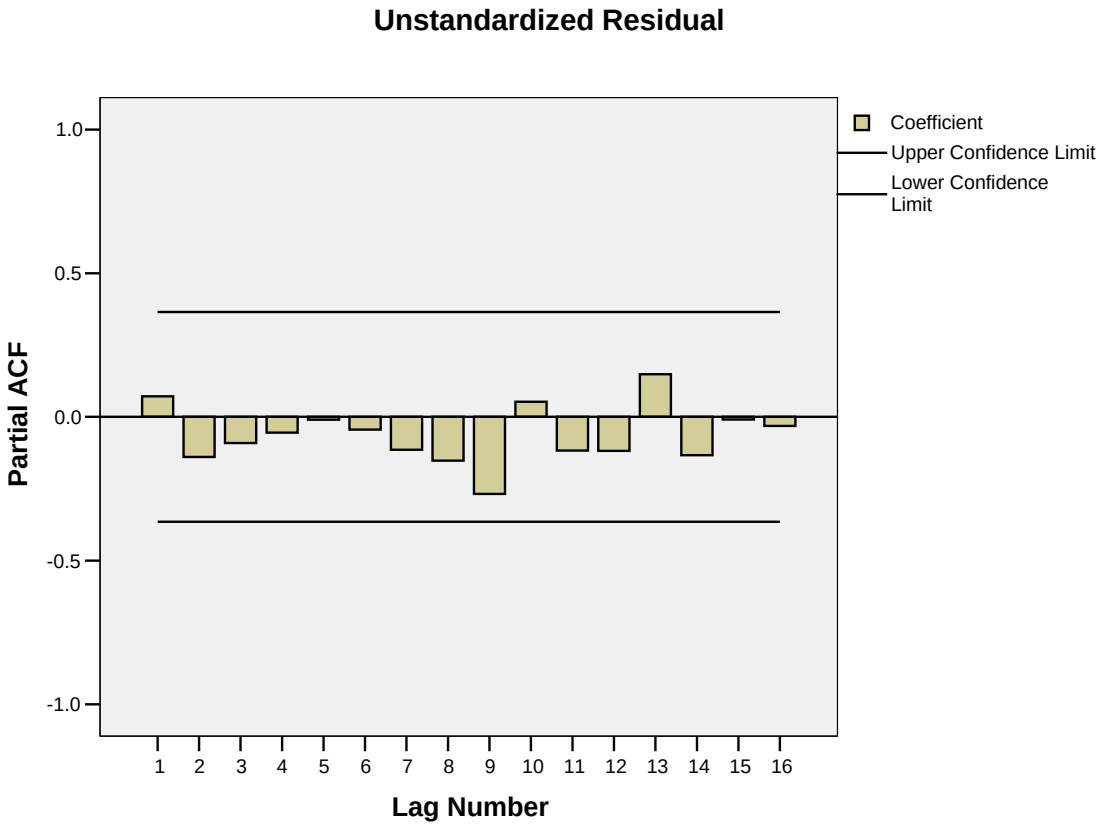

Unstandardized Residual
Partial ACF
1.0
0.5
0.0
-0.5
-1.0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
Lag Number
Coefficient
Upper Confidence Limit
Lower Confidence Limit

Figure 4.6



Declaration

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

Name: Zewdu Meressa

Signature: .....

Place: Faculty of Science, Addis Ababa University

Date: June 2009

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

.....

Dr. Fentaw Abegaz