UNLV | UNIVERSITY LIBRARIES

UNLV Retrospective Theses & Dissertations

1-1-2007

Determination of transformation function for predictor variables in multiple linear regression

Vimatha Ravi
University of Nevada, Las Vegas

Follow this and additional works at: https://digitalscholarship.unlv.edu/rtds

Repository Citation

Ravi, Vimatha, "Determination of transformation function for predictor variables in multiple linear regression" (2007). *UNLV Retrospective Theses & Dissertations*. 2136.
http://dx.doi.org/10.25669/2nuw-5jd6

This Thesis is protected by copyright and/or related rights. It has been brought to you by Digital Scholarship@UNLV with permission from the rights-holder(s). You are free to use this Thesis in any way that is permitted by the copyright and related rights legislation that applies to your use. For other uses you need to obtain permission from the rights-holder(s) directly, unless additional rights are indicated by a Creative Commons license in the record and/or on the work itself.

This Thesis has been accepted for inclusion in UNLV Retrospective Theses & Dissertations by an authorized administrator of Digital Scholarship@UNLV. For more information, please contact digitalscholarship@unlv.edu.

DETERMINATION OF TRANSFORMATION FUNCTION FOR PREDICTOR VARIABLES IN MULTIPLE LINEAR REGRESSION

by

Vimatha Ravi

Bachelor of Technology
National Institute of Technology, Warangal
2004

A thesis submitted in partial fulfillment
of the requirements for the

Master of Science Degree in Mathematical Sciences
Department of Mathematical Sciences
University of Nevada, Las Vegas

Graduate College
University of Nevada, Las Vegas
May 2007

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

UMI Number: 1443785

INFORMATION TO USERS

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleed-through, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

UMI®

UMI Microform 1443785

Copyright 2007 by ProQuest Information and Learning Company.

All rights reserved. This microform edition is protected against unauthorized copying under Title 17, United States Code.

ProQuest Information and Learning Company
300 North Zeeb Road
P.O. Box 1346
Ann Arbor, MI 48106-1346

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

UNLV
UNIVERSITY OF NEVADA LAS VEGAS

Thesis Approval

The Graduate College
University of Nevada, Las Vegas

April 10, 20 07

The Thesis prepared by

Vimatha Ravi

Entitled

Determination of Transformation Functions for Predictor Variables in Multiple Linear Regression

is approved in partial fulfillment of the requirements for the degree of

Master of Science in Mathematical Science

Examination Committee Chair

Dean of the Graduate College

Examination Committee Member

Examination Committee Member

Graduate College Faculty Representative

1017-53

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

ABSTRACT

Determination of Transformation Function for Predictor Variables in Multiple Linear Regression

by

Vimatha Ravi

Rohan Dalpatadu, Ph.D., P.E., Examination Committee Chair
Associate Professor, Department of Mathematical Sciences
University of Nevada, Las Vegas

In multiple linear regression involving several predictor variables, finding a suitable non-linear transformation of the predictors might be helpful to present the model in a simple functional form which is linear in the transformed variables. In this thesis, a computer code in C++ is developed to automate the process of finding a suitable transformation for the predictors. This is done by finding the transformation that yields the maximum correlation between the response and the transformed predictor. Several simulated examples are included to illustrate the method. A prime concern in calculating the correlation between two data sets is statistical accuracy. Correlation coefficients reveal the degree of correlation between two data sets. They are valued from -1 to 1. A positive value indicates correlation and negative values indicate anti-correlation.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

TABLE OF CONTENTS

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

LIST OF FIGURES

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

ACKNOWLEDGEMENTS

I wish to express my sincere gratitude to my advisor, Dr. Rohan Dalpatadu for his able support and expert guidance not only through the course of my thesis, but also during my graduate program at UNLV. I would like to thank Dr. Ashok Singh for his valuable advice during my M.S. program. I would also like to extend my profound gratitude to the members of my thesis committee, Dr. Murphy and Dr. Xin Li, for their guidance in my thesis related efforts. I would also like to thank Dr Laxmi Gewali for his guidance in my thesis. I thank all the students and staff at the UNLV Mathematics Department for their friendship and support. I also thank the UNLV Mathematics Department for providing me the financial support throughout my M.S. program. Last, but not the least I thank my husband, my parents and cousins, who have been a source of constant motivation.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

CHAPTER 1

INTRODUCTION

The Regression Model

Multiple linear regression analysis is widely used to describe statistical relationships between a response variable and two or more predictor variables. The purposes of regression analysis can be thought of as: (1) description, (2) control, and (3) prediction. The general first-order multiple linear regression model with $p-1$ predictor variables is shown below (Kutner, Nachtsheim, Neter; 2004):

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \cdots + \beta_{p-1} X_{i,p-1} + \varepsilon_i \qquad (1.1)$$

where:

Y_i = the value of the response variable in the i^{th} trial

$\beta_0, \beta_1, \cdots, \beta_{p-1}$ are the parameters

$X_1, X_2, \cdots, X_{p-1}$ are the predictor variables

ε_i is a random error term with mean 0 and variance σ^2

To define a model from Equation 1.1 and to define a statistical relation between the response variable Y, and the predictor variables X's, two basic steps need to be followed.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

The first step is to identify the relevant predictor variables from a large group of independent variables which can explain the behavior of the response variable and the next step is to obtain a statistical relationship between them.

Research Objective

Simple transformations of variables (response variable Y, predictor variable X, or both) are sufficient to make the linear regression model represented by Equation 1.1 more appropriate. The convenient and well known transformation is for a response variable. Transformation of a response variable is expected to stabilize the variance of error terms or to reduce the model to linearity. Frequently the assumption of normality of error terms is not in question and in such a case, transformation of one or more predictor variables can be attempted to reduce the model to a simple functional form. This method could be of more use when the true form of the model is unknown. The objective of this research is to obtain a simple functional form which is linear in the transformed scale by finding transformation functions of the predictor variables. The transformation function is obtained by looking at the degree of statistical dependency between the response variable and the predictor variable considered. The direction of dependency is ignored when taking the statistical dependency between the response and predictor variables into consideration. The variables in the model represented by Equation 1.1 are replaced with certain transformed functions of the predictor variables. The measure of dependency is given by the linear correlation coefficient calculated using the data of the response and the predictor variables. In this regression model, the higher the magnitude of the correlation is the higher is the degree of dependency. Therefore a transformation with a

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

high degree of dependency is selected with regard to a particular predictor variable. In other words, the transformations of the predictor variables were selected by maximal correlation theory.

Literature Review

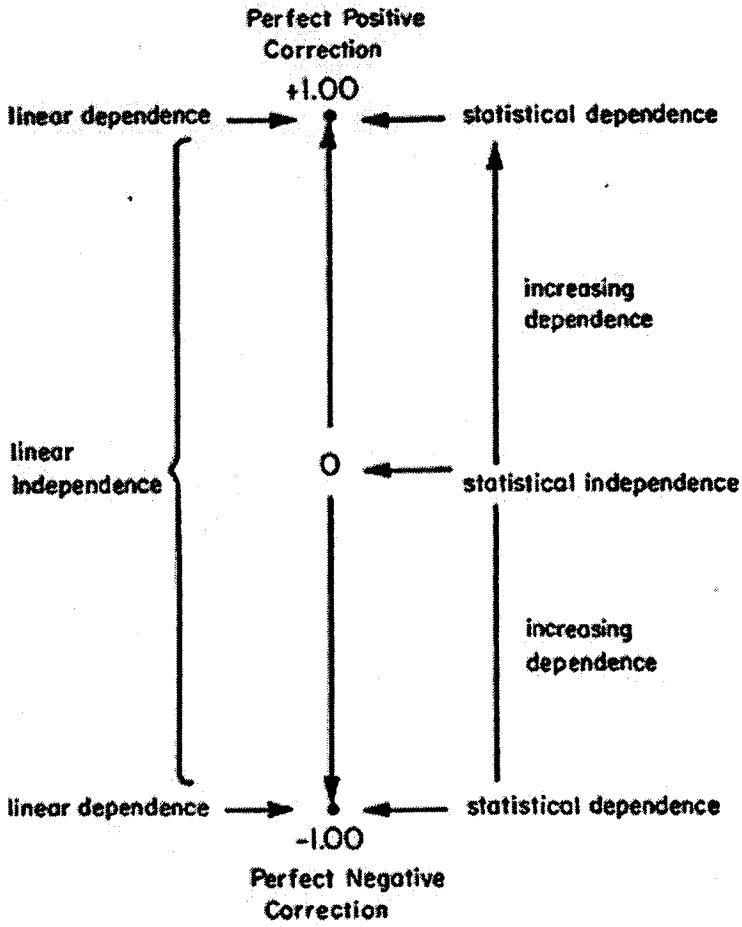
Correlation can be defined as the tendency towards concomitant variation and the correlation coefficient is simply a measure of such tendency. Correlation coefficient quantifies the direction and magnitude of linear association between two quantitative variables X & Y.

The range of correlation coefficient ρ between any two variables is $-1 \leq \rho \leq 1$. When $\rho = -1$, the two variables vary perfectly in the opposite direction. When $\rho = 1$, the two variables vary perfectly and positively in the same direction. Finally, when $\rho = 0$ (or a value near zero) it can be said that there is no correlation between the two variables. In other words, the two variables are independent and they vary separately. This is better illustrated in the Figure 1.

The squared correlation coefficient ρ^2 can be interpreted for the degree of dependency instead of correlation coefficient ρ. This is because the correlation coefficient misleads by suggesting a higher degree of co-variation than the existing co-variation. This problem gets worse as the correlation approaches zero. Moreover, the squared correlation gives the proportion of common variance between the two variables.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Figure 1: Range of correlation coefficient



Regression analysis is generally performed to fit a model that is adequate for the purpose intended. A transformation function of the predictor variable is obtained by looking at the correlation coefficient between the response variable and the transformed predictor variable. The transformation function of the predictor variable will have maximum squared correlation coefficient, which is used as a general measure of dependency (Breiman, Friedman 1985).The maximal correlation coefficient has the following properties:

1. $0 \le \rho_{\max}(X,Y) \le 1$.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

2. $\rho_{\max}(X,Y) = 0$ if and only if X and Y are independent.

3. If there exists a relation of the form $u(X) = v(Y)$, where u and v are Borel-measurable functions with $\text{var}[u(X)] > 0$, then $\rho_{\max}(X,Y) = 1$.

Organization of the Thesis

This thesis includes four chapters. Chapter 1 gives the introduction to the objective of the research study. The methodology of determining the transformation function for a predictor variable is presented in Chapter 2. Chapter 3 presents the results obtained from several simulated examples. Chapter 4 summarizes the conclusion and provides some recommendations of the thesis.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

CHAPTER 2

METHODOLOGY

Transformation of the Predictor Variable

Transformation of either a response variable or a predictor variable sometimes gives a more appropriate regression model instead of a model with the original variables. In this thesis, transformation of the predictor variables in regression is considered. The transformation function is obtained by comparing the squared correlations. An automated program in C++ is developed to find the transformation function with the maximum squared correlation among the transformation functions considered.

The method of determining the transformation function of the predictor variable involves:

1. transforming the predictor variable and obtaining the transformed data.
2. calculating the squared linear correlation coefficient, ρ^2, for all the transformations of the predictor variables considered.
3. selecting the transformation function which gives the maximum squared correlation, ρ^2.

To perform these steps, an automated computer program is developed using the programming language C++.

Transformation of a predictor variable sometimes helps us to better understand the statistical relationship between the response variable and the predictor variable. By

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

transforming the predictor variable the dependency between the two variables can be explained.

Transformation function for the predictor variable is obtained by maximal correlation theory. In this thesis, to demonstrate the method following transformation functions are considered:

1. $g(X) = \log X$
2. $g(X) = X^{\beta}$, $\beta = 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 2, 3, 4$

The data might have several predictor variables. A matrix plot of all the variables is studied to identify the predictors that require a transformation. The predictors which need a transformation are considered one at a time. Transformed data for all the transformation functions considered is obtained. Now, the squared correlation coefficient is calculated and compared.

Once the data for X is read and transformed, the next step is calculating the squared correlation coefficient, ρ^2. So now there are two sets of data, one is the Y data, which is stored in an array, and the other one is $g(X) = X'$, the transformed data. The squared correlation coefficient is calculated to measure the degree of statistical dependency between these two variables.

The estimate of squared correlation coefficient is given by:

$$R^2 = \frac{(\sum (x_i - \bar{x})(y_i - \bar{y}))^2}{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}$$

where:

x_i is the value of X in the ith trial.

y_i is the value of Y in the ith trial.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

The mean of X values, $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$

The mean of Y values, $\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$

While calculating R^2 for Y and $g(X)$, X is replaced by $g(X)$ in the above formula.

A function in C++ is defined to calculate the squared correlation. So when we pass the X (or $g(X)$) data, the Y data, and the size of the data as arguments to this correlation function, the function calculates and returns R^2 value. Another function is defined to calculate the means of X *(or* $g(X)$*)* and Y data. This mean function takes the data and size of the data as arguments, calculates the mean and returns the mean value.

The squared correlation is calculated for each $g(X)$ and Y. Each squared correlation is compared with the squared correlation between X and Y and also with the other squared correlations. The maximal squared correlation is obtained by these comparisons and the transformation which yields this maximal correlation is considered to be fit into the regression model instead of X.

The measure R^2 is called the coefficient of multiple determination. R^2 may be interpreted as the proportionate reduction of the total variation of the response variable, Y associated with the use of the set of predictor variables. The limiting values are 0 and 1. The closer it is to 1, the greater is the degree of linear association between the response and the predictor variables.

The measure of SSE can be interpreted as the variation in the response variable, Y that is present when the predictor variable is taken into account. It denotes the error sum of squares. The greater the variation of the Y observations around the fitted regression

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

line, the larger is the SSE. Lower values of SSE are expected for a better regression model.

One can always increase R^2 by adding more predictors as with more X variables SSE can never be larger. For this reason, it is often suggested to use the adjusted coefficient of multiple determination, R^2_{adj}. The formula to calculate R^2_{adj} is given below:

$$R^2_{adj} = 1 - \left(\frac{n-1}{n-p}\right)\frac{SSE}{SSTO}$$

where,

n = No. of data points

p = No. of β's

Regression analysis with the original and the transformed predictors is compared. The R^2_{adj} values and the residual error (SSE) terms are compared. If there is an increase in the R^2_{adj} values and a decrease in the residual error terms then, the model with the transformed predictors is considered to be a better model. Later, the residual plots are also compared to support the previous argument.

Residual Plots

The residual plots are used to examine the following:

1. Linearity of the regression model.
2. Constant variance of the error terms.
3. Outliers.
4. Normality of error terms.
5. Independence of error terms.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Figure 2: Prototype of Residual Plots

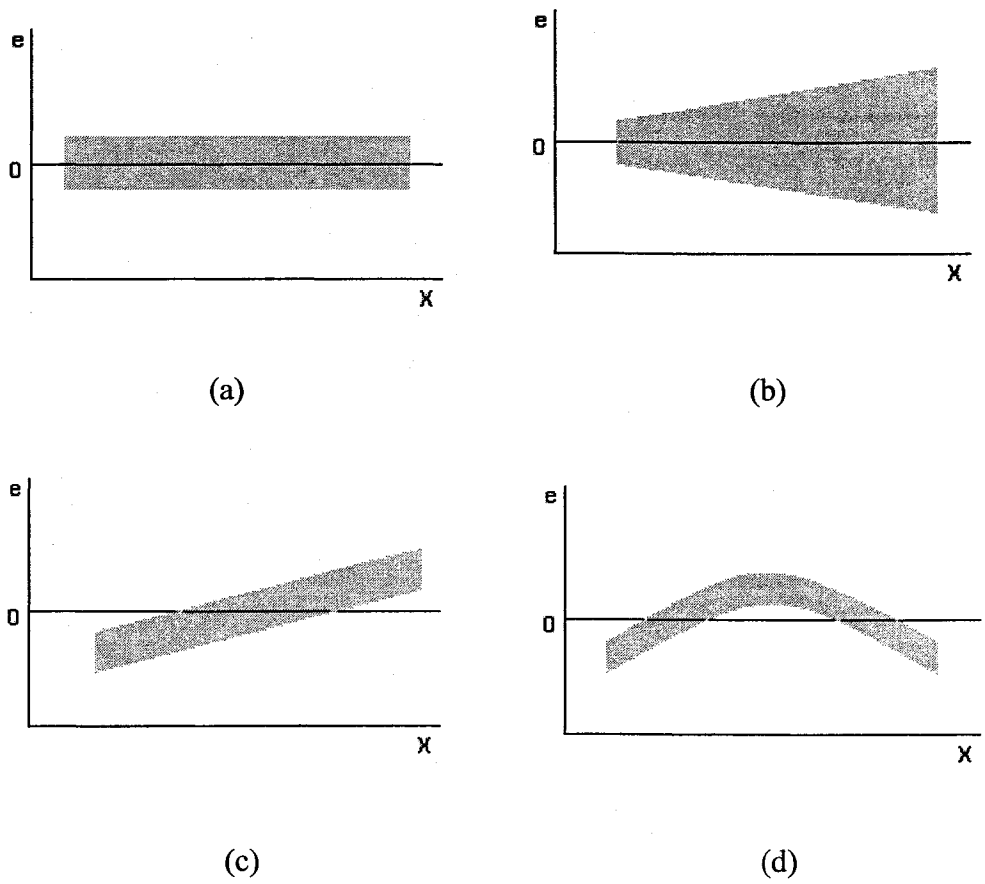


Figure 2(a) shows a prototype of the residual plot against the predictor variable, *X*, when a linear regression model is appropriate satisfying all the basic assumptions. The residuals are with in a horizontal band centered about 0, and do not display any systematic tendencies to be positive or negative. Figure 2(b) shows a prototype where there is a departure from the linearity of the model. It indicates the need for a curvilinear model. Figure 2(c) shows a prototype where the error variance is not constant. Figure 2(d) shows a prototype where there is a correlation between the error terms. Negative residuals are associated with the early trials and the positive residuals are associated with the later trials.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Data

This section provides the documentation for the data used in this thesis. The methodology is demonstrated on three different data sets. The highway data and the dwaste data are obtained from the book Applied Linear Regression, 3rd edition by Sanford Weisberg. Crop Yield data is taken from Applied Linear Regression Models, 4th edition by Michael H. Kutner, Christopher J. Nachtshiem & John Neter.

The highway data relates the automobile accident rate, in accidents per million vehicle miles to several potential terms. The response variable is the rate. The predictors are length of the highway segment in miles (X1), average daily traffic count in thousands (X2), truck volume as a percent of the total volume (X3), speed limit in 1973 (X4), width in feet of outer shoulder on the roadway (X5), and number of signalized interchanges per mile (X6).

The crop yield data is concerned with the effects of moisture and temperature on the yield of a hybrid tomato. The response variable is the yield of tomato. The predictor variables are moisture (X1) and temperature (X2).

The dwaste data is from an experiment conducted to study difference in the measurement of the oxygen uptake in milligrams of oxygen per minute, using five different chemical measurements. The response variable is the oxygen uptake measurement. The predictor variables are biological oxygen demand (X1), total kjeldahl nitrogen (X2), total solids (X3), Total volatile solids (X4), and chemical oxygen demand (X5).

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

CHAPTER 3

RESULTS

This chapter presents the results. Sections one to three present the results for various data sets. Section four summarizes the results for all the data sets. Summary tables are provided in section four.

Results for Highway data

A matrix plot is obtained for the data to identify the predictors that need a transformation. The plot is presented in Figure 3.

After observing the plot, one might want to find a transformation function for X1, X2, X3, and X6. By transforming these variables, we can obtain a simple linear model in the transformed variables. To obtain the transformation function for these variables, C++ code is used. The C++ output is shown in Figure 4.

The program suggests using a Log transformation on X1, power transformations on the rest of the variables selected. The transformed variables are obtained and regression analysis is performed. The regression models before and after the transformations are obtained and compared.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Figure 3: Matrix Plot for Highway Data

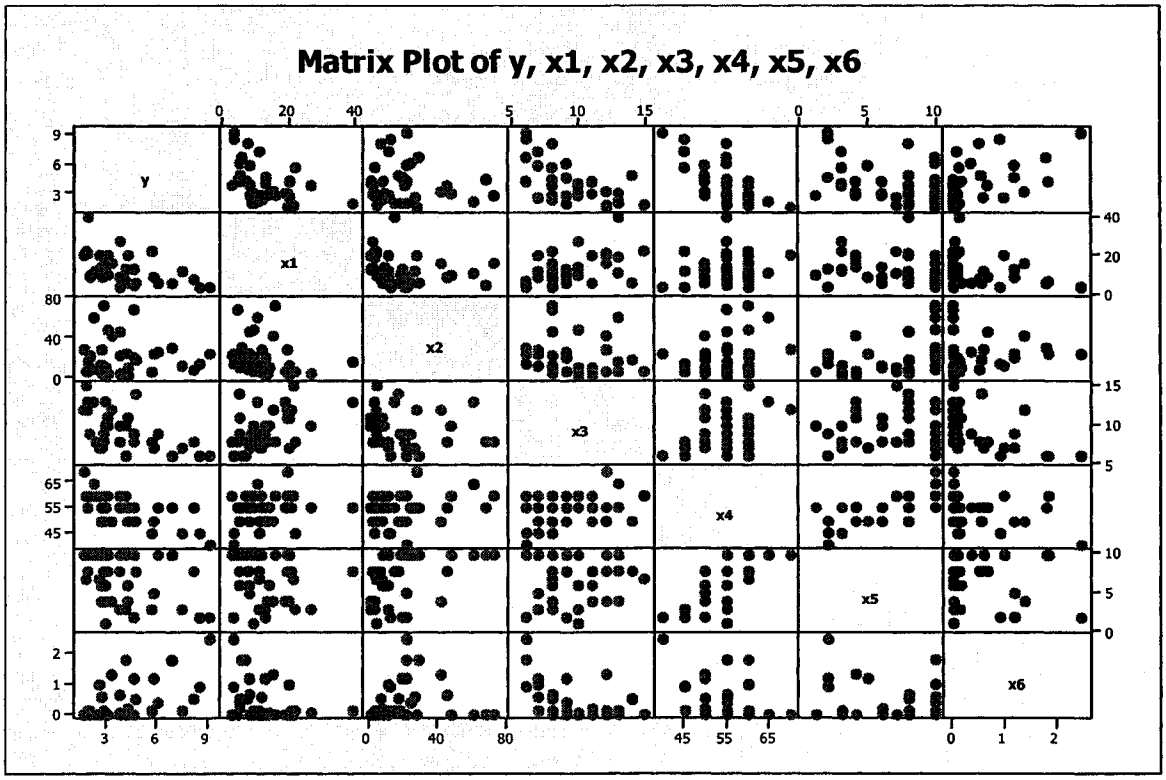
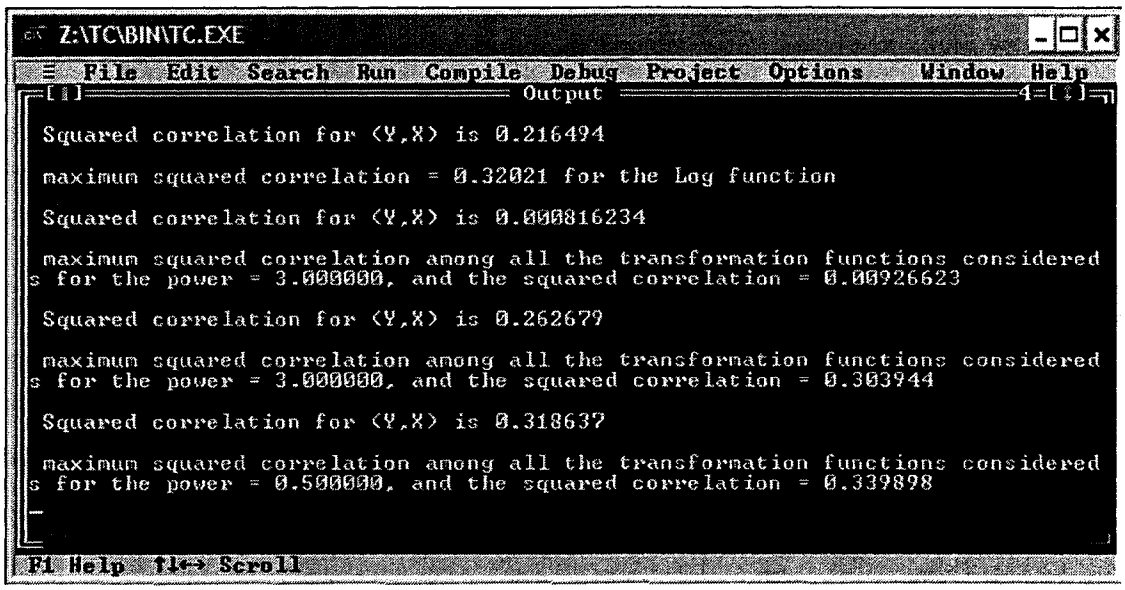


Figure 4: C++ Ouput for Highway Data



Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Regression Analysis: y versus x1, x2, x3, x4, x5, x6

The regression equation is
y = 14.4 - 0.0620 x1 + 0.0002 x2 - 0.136 x3 - 0.152 x4 - 0.048 x5 + 0.694 x6

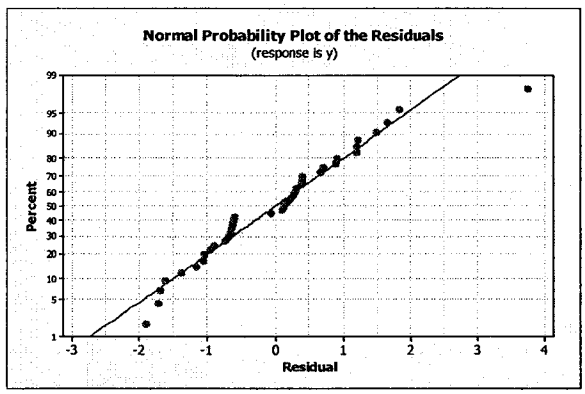
Predictor	Coef	SE Coef	T	P
Constant	14.418	2.699	5.34	0.000
x1	-0.06201	0.03309	-1.87	0.070
x2	0.00022	0.01316	0.02	0.987
x3	-0.1364	0.1103	-1.24	0.225
x4	-0.15208	0.05711	-2.66	0.012
x5	-0.0481	0.1074	-0.45	0.657
x6	0.6939	0.3989	1.74	0.092

S = 1.27715 R-Sq = 65.2% R-Sq(adj) = 58.6%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	6	97.690	16.282	9.98	0.000
Residual Error	32	52.196	1.631		
Total	38	149.886			

Figure 5: Normal Probability Plot



Regression Analysis: y versus Log x1, x2^3, x3^3, x4, x5, x6^0.5

The regression equation is
y = 13.4 - 1.17 Log x1 + 0.000001 x2^3 - 0.000301 x3^3 - 0.109 x4 - 0.121 x5
+ 1.06 x6^0.5

Predictor	Coef	SE Coef	T	P

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

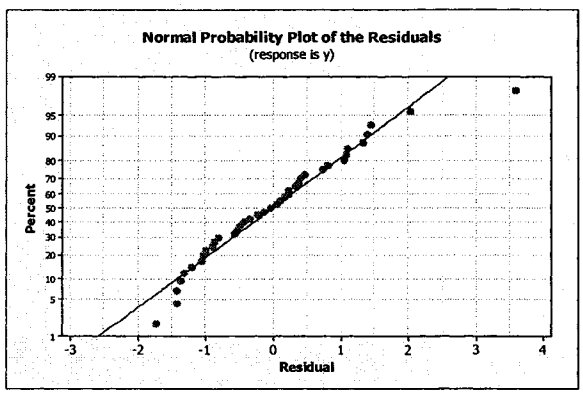
Constant	13.367	2.761	4.84	0.000
Log x1	-1.1700	0.3920	-2.98	0.005
x2^3	0.00000123	0.00000244	0.51	0.617
x3^3	-0.0003006	0.0002938	-1.02	0.314
x4	-0.10910	0.05900	-1.85	0.074
x5	-0.1209	0.1040	-1.16	0.254
x6^0.5	1.0603	0.5074	2.09	0.045

S = 1.21239 R-Sq = 68.6% R-Sq(adj) = 62.7%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	6	102.850	17.142	11.66	0.000
Residual Error	32	47.036	1.470		
Total	38	149.886			

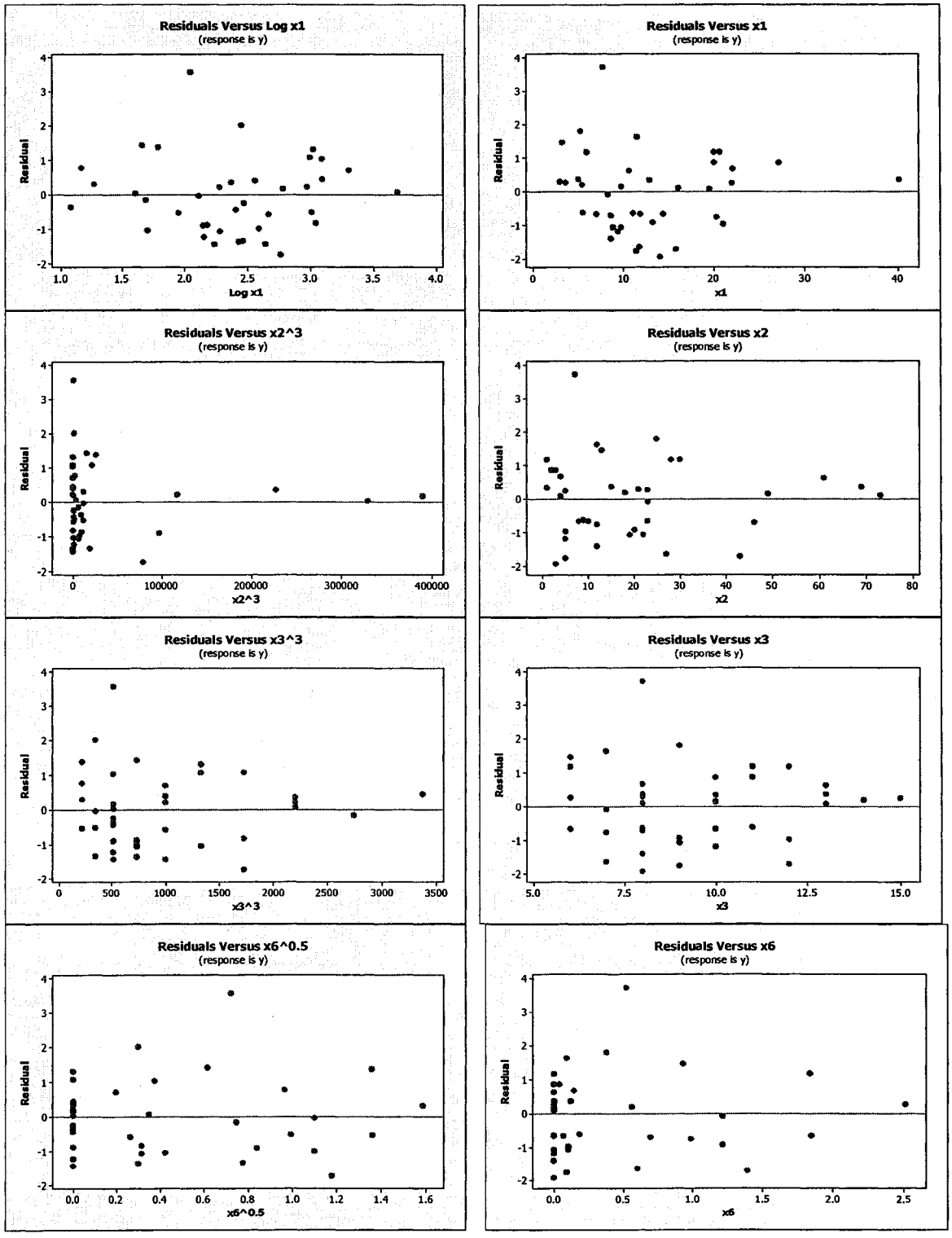
Figure 6: Normal Probability Plot



Adjusted coefficient of multiple determination, R^2_{adj}, is compared from both the models. Clearly, there is an increase in the R^2_{adj} value which suggests an improvement in the fit. Residual error, SSE values are also compared. SSE also decreased suggesting an improvement in the fit. For a better understanding, the residual plots are also compared. Residual plots also show an improvement, which suggests a better fit to the data. They are presented below.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Figure 7: Residual Plots for Highway Data

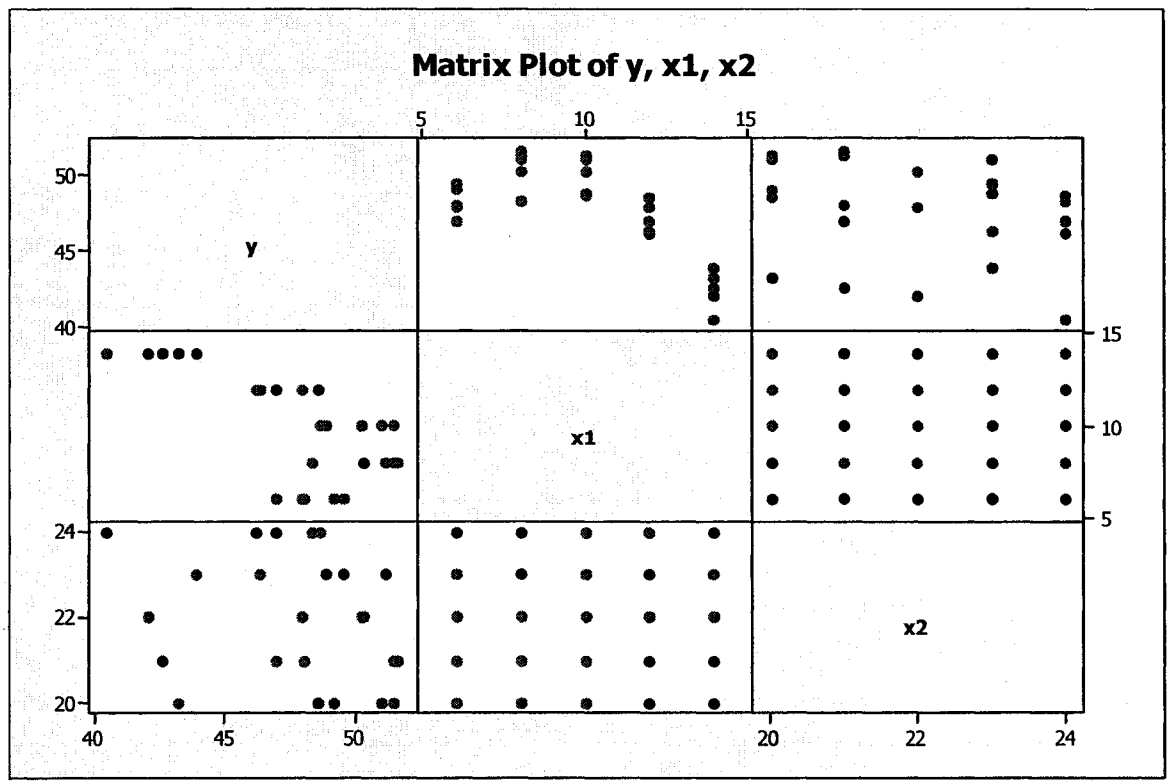


Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Results of Crop Yield Data

A matrix plot is obtained for the data to identify the predictors that need a transformation. The plot is presented below.

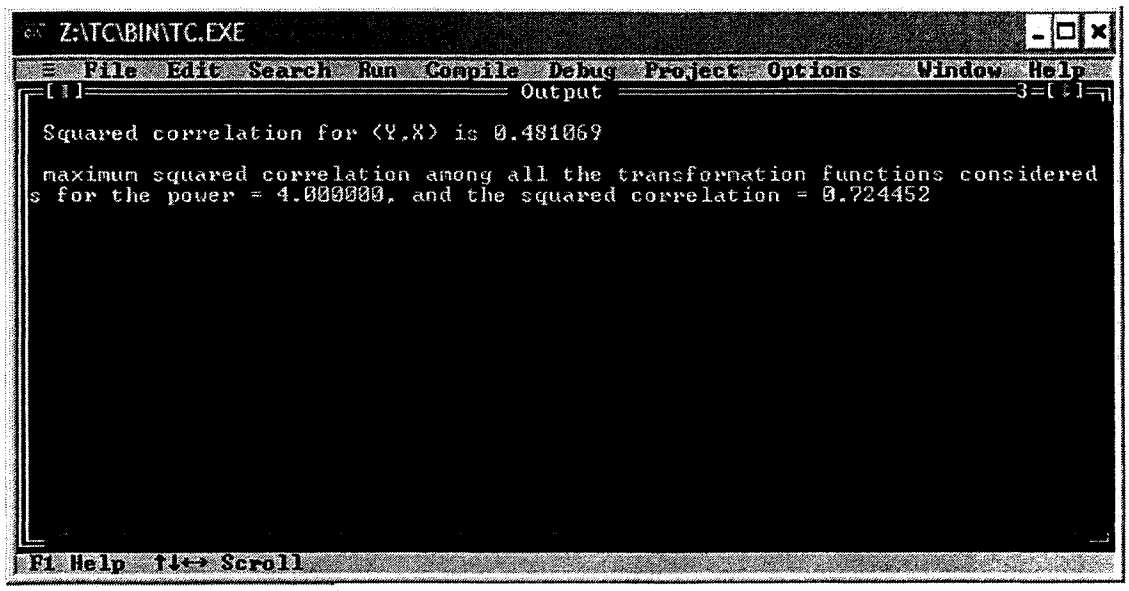
Figure 8: Matrix Plot for Crop Yield Data



The non linearity of the variable X1 is evident from the matrix plot. By transforming X1, we can achieve a linear function in the transformed X1. The transformation function for the variable X1 is obtained from the C++ code. The C++ output is shown below.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Figure 9: C++ Output for Crop Yield Data



The program suggests using a power transformation on X1. The transformed variable is obtained and regression analysis is performed. The regression models, before and after the transformation, are obtained and compared.

Regression Analysis: y versus x1, x2

The regression equation is
y = 67.0 - 0.762 x1 - 0.530 x2

Predictor	Coef	SE Coef	T	P
Constant	67.044	7.188	9.33	0.000
x1	-0.7620	0.1590	-4.79	0.000
x2	-0.5300	0.3180	-1.67	0.110

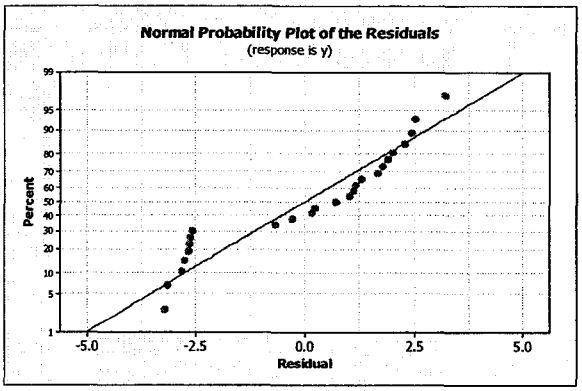
S = 2.24847 R-Sq = 53.9% R-Sq(adj) = 49.7%

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	130.174	65.087	12.87	0.000
Residual Error	22	111.224	5.056		
Total	24	241.398			

Figure 10: Normal Probability Plot



Regression Analysis: y versus x1^4, x2

The regression equation is
y = 62.3 - 0.000196 x1^4 - 0.530 x2

Predictor	Coef	SE Coef	T	P
Constant	62.342	4.827	12.92	0.000
x1^4	-0.00019574	0.00002286	-8.56	0.000
x2	-0.5300	0.2184	-2.43	0.024

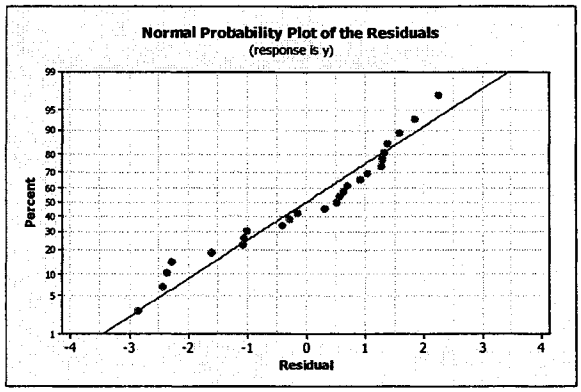
S = 1.54437 R-Sq = 78.3% R-Sq(adj) = 76.3%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	188.926	94.463	39.61	0.000
Residual Error	22	52.472	2.385		
Total	24	241.398			

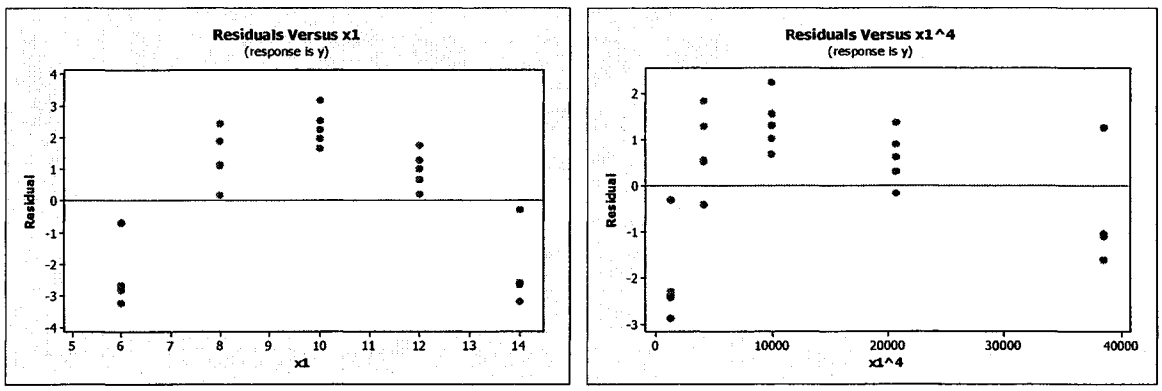
Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Figure 11: Normal Probability Plot



Adjusted coefficient of multiple determination, R^2_{adj}, is compared from both the models. Clearly, there is an increase in the R^2_{adj} value which suggests an improvement in the fit. Residual error, SSE values are also compared. SSE also decreased suggesting an improvement in the fit. For a better understanding, the residual plots are also compared. They are presented below.

Figure 12: Residual Plots for Crop Yield Data

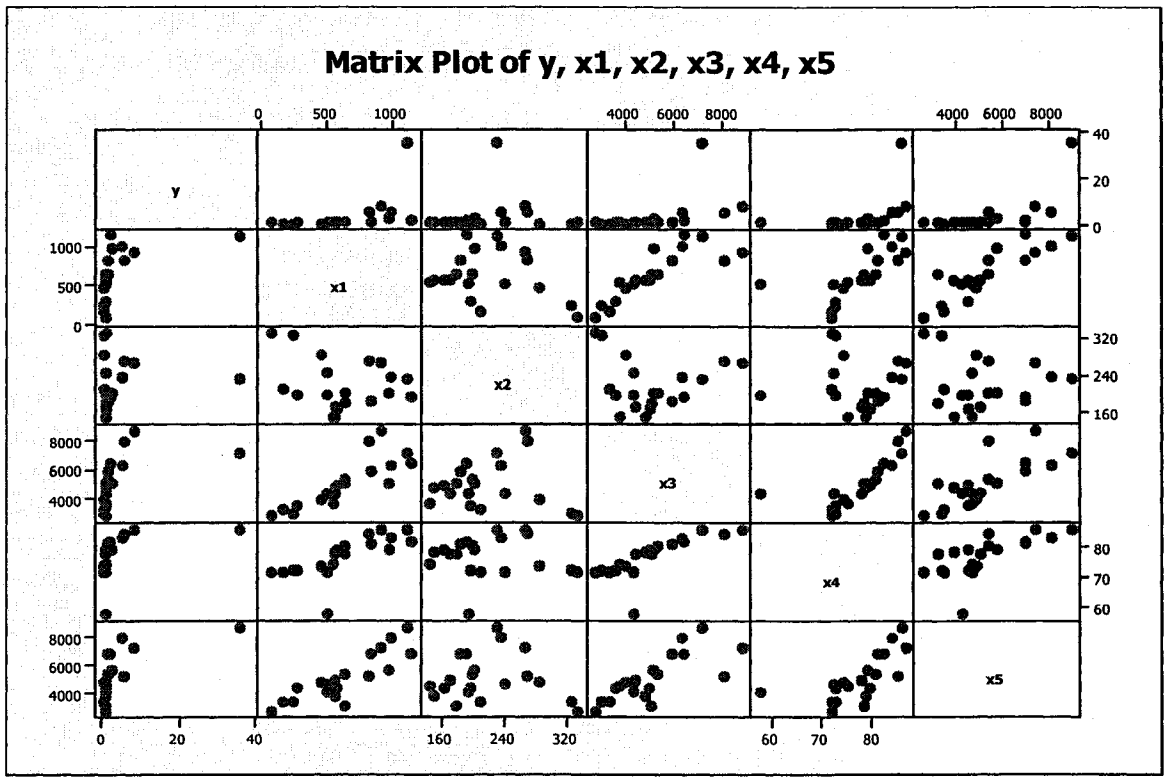


Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Results for Dwaste Data

A matrix plot is obtained for the data to identify the predictors that need a transformation. The plot is presented below.

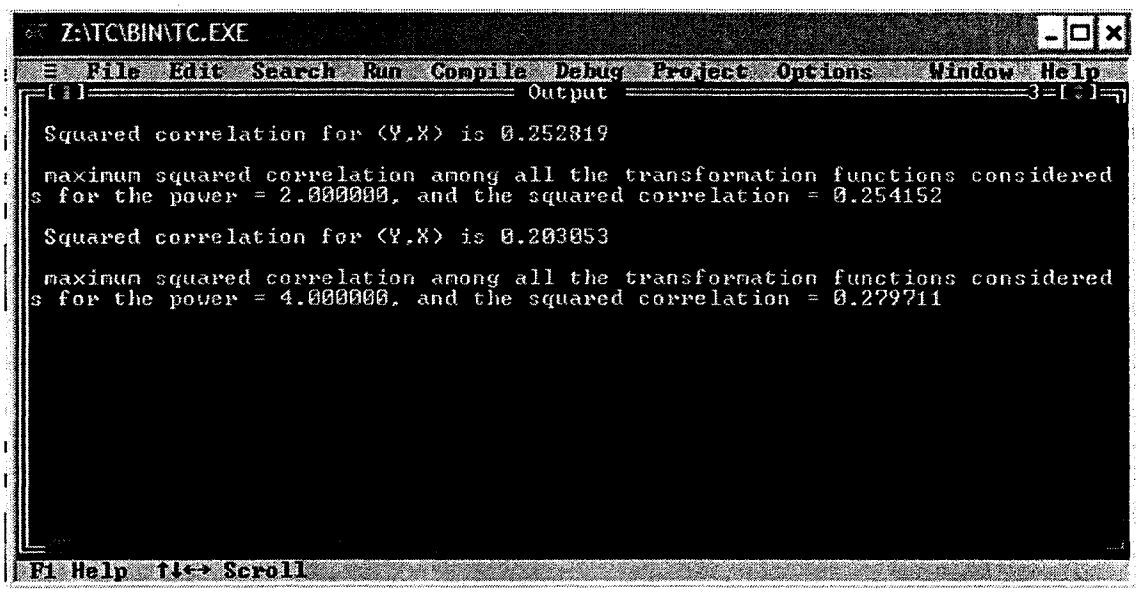
Figure 13: Matrix Plot for Dwaste Data



After observing the plot, one might want to find a transformation function for X3 and X4. To obtain the transformation function for this variable, C++ code is used. The C++ output is shown below. The program suggests using a power transformation for both the variables considered. The transformed variables are obtained and regression analysis is performed. The regression models, before and after the transformation, are obtained and compared.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Figure 14: C++ Output for Dwaste Data



Regression Analysis: y versus x1, x2, x3, x4, x5

The regression equation is
y = - 21.5 - 0.0043 x1 + 0.0194 x2 + 0.00019 x3 + 0.060 x4 + 0.00345 x5

Predictor	Coef	SE Coef	T	P
Constant	-21.50	23.77	-0.90	0.381
x1	-0.00431	0.01349	-0.32	0.754
x2	0.01937	0.03288	0.59	0.565
x3	0.000189	0.002001	0.09	0.926
x4	0.0599	0.3643	0.16	0.872
x5	0.003455	0.001919	1.80	0.093

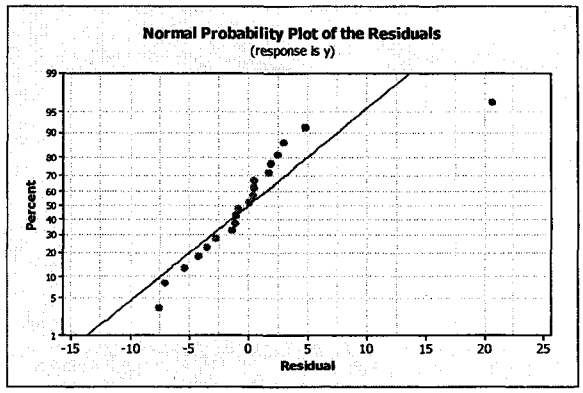
S = 6.81277 R-Sq = 45.5% R-Sq(adj) = 26.1%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	5	543.40	108.68	2.34	0.096
Residual Error	14	649.79	46.41		
Total	19	1193.19			

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Figure 15: Normal Probability Plot



Regression Analysis: y versus x1, x2, x3^2, x4^4, x4, x5

The regression equation is
y = 246 - 0.0073 x1 + 0.0210 x2 - 0.000001 x3^2 + 0.000004 x4^4 - 4.94 x4
+ 0.00200 x5

Predictor	Coef	SE Coef	T	P
Constant	245.6	128.1	1.92	0.078
x1	-0.00728	0.01144	-0.64	0.536
x2	0.02100	0.03054	0.69	0.504
x3^2	-0.00000050	0.00000029	-1.76	0.103
x4^4	0.00000387	0.00000183	2.11	0.055
x4	-4.943	2.390	-2.07	0.059
x5	0.001996	0.001852	1.08	0.301

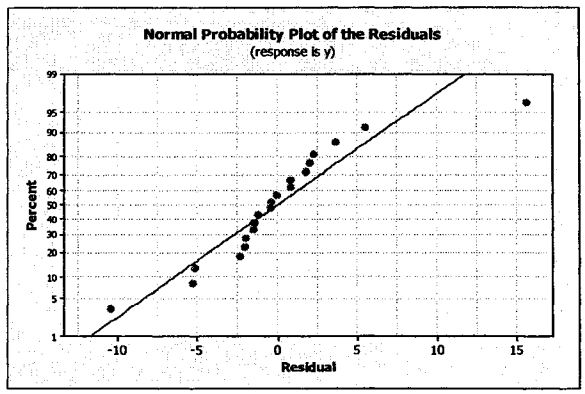
S = 6.09888 R-Sq = 59.5% R-Sq(adj) = 40.8%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	6	709.64	118.27	3.18	0.038
Residual Error	13	483.55	37.20		
Total	19	1193.19			

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

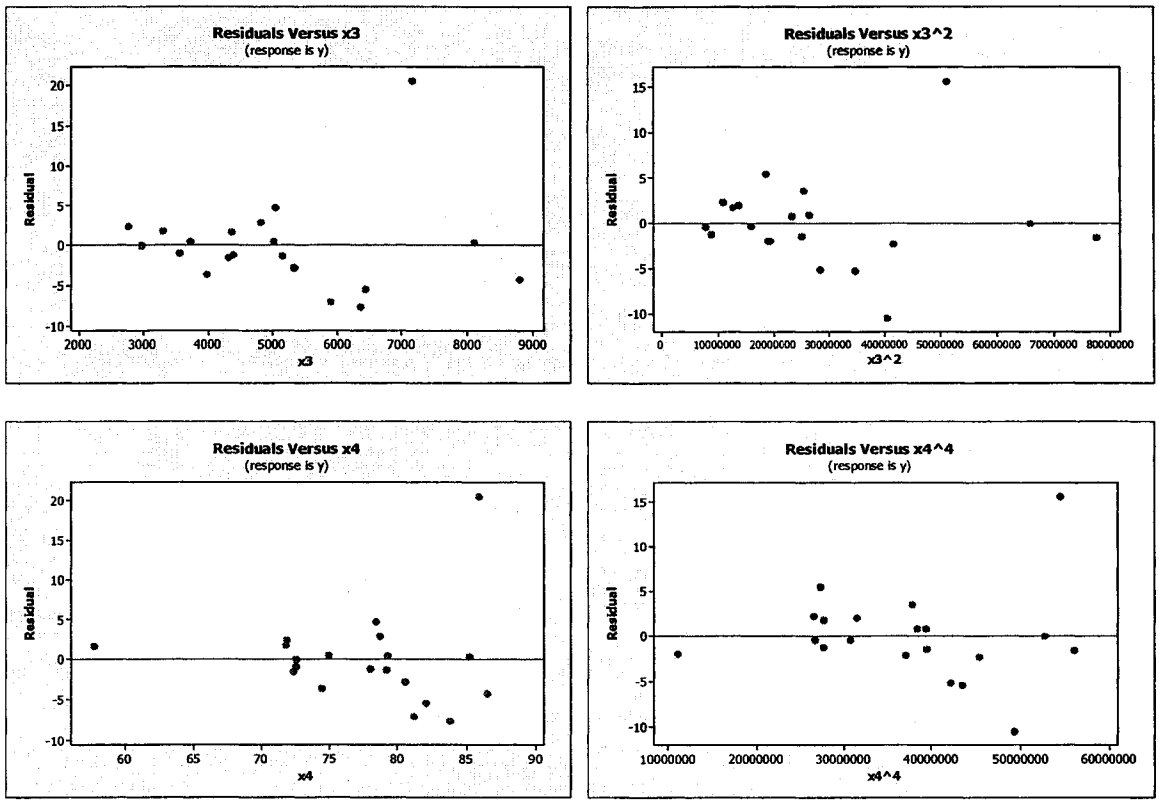
Figure 16: Normal Probability Plot



Adjusted coefficient of multiple determination, R^2_{adj}, is compared from both the models. Clearly, there is an increase in the R^2_{adj} value which suggests an improvement in the fit. Residual error, SSE values are also compared. SSE also decreased suggesting an improvement in the fit. For a better understanding, the residual plots are also compared. They are presented below.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Figure 17: Residual Plots for Dwaste Data



Summary

Summary tables for all the data sets are presented below.

Table 1: Summary Table for Highway Data

	Before Transformation	After Transformation
Sq.Corr(Y,X1)	0.216494	0.32021
Sq.Corr(Y,X2)	0.000816234	0.00926623
Sq.Corr(Y,X3)	0.262679	0.303944
Sq.Corr(Y,X6)	0.318637	0.339898
Adjusted Coefficient of multiple determination, R^2_{adj}	58.6%	62.7%
Residual Error, SSE	52.196	47.036

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Table 2: Summary Table for Crop Yield Data

	Before Transformation	After Transformation
Sq.Corr(Y,X1)	0.481069	0.724452
Adjusted Coefficient of multiple determination, R^2_{adj}	49.7%	76.3%
Residual Error, SSE	111.224	52.472

Table 3: Summary Table for Dwaste Data

	Before Transformation	After Transformation
Sq.Corr(Y,X3)	0.252819	0.254152
Sq.Corr(Y,X4)	0.203053	0.279711
Adjusted Coefficient of multiple determination, R^2_{adj}	26.1%	40.8%
Residual Error, SSE	649.79	483.55

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

CHAPTER 4

CONCLUSIONS AND RECOMMENDATIONS

The objective of this thesis is to obtain a transformation function for a predictor variable based on maximal correlation theory. A transformation function of the predictor will be able to provide a better fit and also might be able to reduce the model to linearity in the transformed variables. As we know, an appropriate regression model is obtained by understanding the relationship between the response variable and the predictor variables. This relationship between the variables is given by correlation coefficient. Hence, obtaining the transformation function of a predictor variable based on maximal correlation theory will be appropriate and helpful in obtaining a better regression model. In this thesis, a code in C++ is developed to automate the process of obtaining the transformation function that provides a better fit. The C++ code developed will transform the predictor variables and compares the squared correlation coefficient between the response variable and the transformation functions of a predictor variable. To demonstrate the method a set of transformation functions are considered here. This set does not include negative powers, but, while modeling, including the negative powers might be more appropriate. The transformation function with the highest squared correlation is then selected. Later, regression analysis is performed with the transformation functions of the predictors. The results showed a significant improvement in the regression model. This concludes that, correlation can be helpful in identifying a

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

transformation function for a predictor variable and there by obtaining a better fitted model.

Obtaining a confidence interval for the range of transformation functions used will be more helpful in identifying the transformation function for each variable. Also, apart from the R^2 criteria, some other criteria like AIC should be used to be able to justify the transformation function and the model.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

REFERENCES

1. Michael H. Kutner, Christopher J. Nachtsheim & John Neter, Applied Regression Models (Fourth Edition), McGraw Hill 2003.

2. Leo Breiman, Jerome H. Friedman, Estimating Optimal Transformations for Multiple regression and Correlation, Journal of the American Statistical Association, Vol. 80, No. 391. (Sep., 1985), pp. 580 – 598.

3. R. J. Rummel, Understanding Correlation, Honolulu; Department of Political Sciences, University of Hawaii, 1976.

4. G. E. P. Box and Paul W. Tidwell, Transformation of the Independent Variables, Technometrics 1962, Vol. 4, No. 4, Pages 531 – 550.

5. Sanford Weisberg, Applied Linear Regression (Third Edition), Wiley Series 2005.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

APPENDIX

C++ code for the transformation function:

```
#include<stdio.h>
#include<stdlib.h>
#include<math.h>
/*FUNCTION FOR CALC MEAN:*/
long double mean(long double data[],int s)
{
   int i;
   long double datamean;
   long double sum=0;

   for(i=0;i<s;i++)
   {
      sum=sum+data[i];
   }
   datamean=sum/s;
   return (datamean);
}

/*FUNCTION FOR CALC CORRELATION:*/
long double correlation(long double dataX[],long double dataY[],int s)
{
   int i;
   long double diffsumXY=0;
   long double sqdiffsumY=0;
   long double sqdiffsumX=0;
   long double a,b,r,r2;
   a=mean(dataX,s);
   b=mean(dataY,s);
   for(i=0;i<s;i++)
   {
      diffsumXY=diffsumXY+((dataX[i]-a)*(dataY[i]-b));
      sqdiffsumY=sqdiffsumY+(pow((dataY[i]-b),2));
      sqdiffsumX=sqdiffsumX+(pow((dataX[i]-a),2));
   }
```


Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

```
    r=(diffsumXY/(sqrt(sqdiffsumX*sqdiffsumY)));
    r2=pow(r,2);
    return (r2);
}

typedef struct
{
    long double val1,val2,val3,val4,val5,val6,val7,val8,val9,val10;
}value;

int main(void)
{
    float p[]={0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9,1,2,3,4};
    long double sqcorr=0,c=0,x=0;
    float power=0;
    int i=0,j=0,k=0,n=0,count;

    FILE *fp;
    value v;
    char headStr[200];
    char headVal[100][200];
    long double values[200][10], Y[200], X[200];
    long double g[200];
    fp=fopen("Prob#70.csv","r");
    if(fscanf(fp,"%s",headStr)!=EOF)
    {
        while(headStr[i]!='\0')
        {
            if(headStr[i]==',')
            {
                i++;
                headVal[j][k]='\0';
                j++;
                k=0;
                continue;
            }
            headVal[j][k++]=headStr[i++];
        }
        headVal[j][k]='\0';
    }
    else
    {
        printf("\n There is no data in the input file");
        printf("\n Hence exiting the program..");
        exit(0);
    }
```


Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

```
i=0;
while(fscanf(fp,"%Lg,%Lg,%Lg",&v.val1,&v.val2,&v.val3)!=EOF)
{
   values[i][0]=v.val1;
      values[i][1]=v.val2;
      values[i][2]=v.val3;
      i++;
}
k=i;
i=0;
for(count=1;count<3;count++)
{
for(i=0;i<k;i++)
{
   Y[i]=values[i][0];
   X[i]=values[i][count];
   n=k;
      fclose(fp);
}
for(i=0;i<n;i++)
{
 if(X[i]!=0)
 {
      g[i]=log(X[i]);
 }
 else
 {
      g[i]=X[i];
 }
}
sqcorr=correlation(g,Y,n);
for(i=0;i<13;i++)
{
   for(j=0;j<n;j++)
   {
         g[j]=pow(X[j],p[i]);
      }
      c=correlation(g,Y,n);
      if(p[i]==1)
      {
               x=c;
      }
      if(sqcorr<c)
      {
         sqcorr=c;
```


Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

```
            power=p[i];
        }
    }
    printf("\n Squared correlation for (Y,X) is %Lg\n",x);
    if(power==0)
    {
        printf("\n maximum squared correlation = %Lg for the Log function\n",sqcorr);
    }
    else
    {
        printf("\n maximum squared correlation among all the transformation functions
considered is for the power = %f, and the squared correlation = %Lg\n",power,sqcorr);
    }
    }
    return(0);
}
```


Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

VITA

Graduate College
University of Nevada, Las Vegas

Vimatha Ravi

Local Address:
7000 Paradise Road, Apt # 2129
Las Vegas, NV 89119

Degree:
National Institute of Technology, Warangal, India
Bachelor of Technology (Mechanical Engineering), June 2004

Thesis Title:
Determination of Transformation Functions for the Predictor Variables in Multiple Linear Regression

Thesis Examination Committee:
Chairperson, Dr Rohan Dalpatadu, Ph. D.
Committee Member, Dr Dennis Murphy, Ph. D.
Committee Member, Dr Xin Li, Ph. D.
Graduate Faculty Representative, Dr Laxmi Gewali, Ph. D.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.