

T-4577

**APPLIED SCORE STATISTIC TRANSFORMATION  
IN MULTIPLE REGRESSION MODELING**

**by**

**Michael A. Ringler**

ProQuest Number: 10784007

All rights reserved

INFORMATION TO ALL USERS

The quality of this reproduction is dependent upon the quality of the copy submitted.

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



ProQuest 10784007

Published by ProQuest LLC (2018). Copyright of the Dissertation is held by the Author.

All rights reserved.

This work is protected against unauthorized copying under Title 17, United States Code  
Microform Edition © ProQuest LLC.

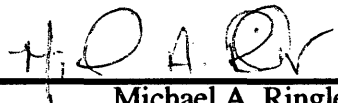
ProQuest LLC.  
789 East Eisenhower Parkway  
P.O. Box 1346  
Ann Arbor, MI 48106 – 1346


T-4577

A thesis submitted to the Faculty and Board of Trustees of the Colorado School of Mines in partial fulfillment of the requirements for the degree of Master of Science (Mathematical and Computer Sciences).

Golden, Colorado


Date 8/8/94

Signed:   
Michael A. Ringler

Approved:   
William R. Astle  
Thesis Advisor

Golden, Colorado

Date 8 August 1994

  
Dr. Ardel J. Boes  
Professor and Head,  
Department of Mathematical  
and Computer Sciences

## ABSTRACT

This thesis represents the culmination of the statistical analysis performed on data required by vendors to quote prices for printed circuit boards used in the manufacturing process at Advanced Energy, Inc., Fort Collins, Colorado. The purpose of this analysis is to create a model for use primarily by the Engineering Design Group during the initial design stages to effectively estimate the potential cost of purchasing printed circuit boards. A secondary function that this study provides is a mechanism for keeping track of each vendor from whom A.E. currently purchases its circuit boards.

The data used for this analysis is historic in nature, and was compiled from price quote requests sent to A.E.'s two major circuit board vendors R.J.R. and Lundahl. The method employed to derive the pricing model is a multiple regression ordinary least squares approach. Techniques like best subsets regression, data transformations on both the predictors and the response variables, and analysis of residuals and scatter plots all contributed to the model's development.

The model that has been developed to approximate cost is based on eight variables including the vendor, order quantity, size of the smallest hole on the board, size of the largest hole on the board, area in square inches, number of conducting layers in the board, and the copper composition of the board.

It must be noted that this model provides only an estimate of potential cost, and in no way can account for the volatile price fluctuations associated with high technology markets. Any external factors, such as inflation or individual problems related to each vendor at any given time cannot be included in the model.

## TABLE OF CONTENTS

	<b>Page</b>
ABSTRACT . . . . .	iii
LIST OF FIGURES . . . . .	vi
LIST OF TABLES . . . . .	vii
ACKNOWLEDGMENTS . . . . .	viii
DEDICATION . . . . .	ix
Chapter 1. INTRODUCTION . . . . .	1
1.1 General . . . . .	1
1.2 Simple Regression Model . . . . .	2
1.3 Data Transformations . . . . .	5
1.4 Purpose . . . . .	7
Chapter 2. METHODOLOGY IN PROBLEM SOLVING . . . . .	9
2.1 Zero-One Indicator Variables . . . . .	10
2.2 Multicollinearity . . . . .	11
2.3 Best Subsets Regression: $C_p$ Criterion . . . . .	14
2.4 Residual Plot Analysis . . . . .	22
2.5 Score Statistic Transformation . . . . .	23
2.6 Resulting Model . . . . .	37
Chapter 3. CONCLUSION . . . . .	39
REFERENCES CITED . . . . .	42
SELECTED BIBLIOGRAPHY . . . . .	44

## LIST OF FIGURES

<b>Figure</b>		<b>Page</b>
1	Scatter Plot of Cost vs. Width . . . . .	13
2	Scatter Plot of Cost vs. Length . . . . .	13
3	Residual Plot of the 6 Variable Model (Residuals vs. Fitted Costs) . . . . .	16
4	Residual Plot of the 6 Variable Model (Residuals vs. Costs) . . . . .	16
5	Regression Summary of the 10 Variable Model . . . . .	19
6	Residual Plot #1 for the 10 Variable Model (Residuals vs. Fitted Values) . . . . .	20
7	Residual Plot #2 for the 10 Variable Model (Residual vs. Response) . . . . .	20
8	Residual Plot #3 of the 10 Variable Model (Response vs. Fitted Response) . . . . .	21
9	Normal Probability Plot of the 10 Variable Model . . . . .	21
10	Regression Summary for Transformation of Response . . . . .	27
11	Regression Summary of the 9 Variable Model . . . . .	30
12	Residual Plot of 8 Variable Model (Transformed) . . . . .	34
13	Residual Plot of 9 Variable Model (Transformed) . . . . .	34
14	Regression Summary of the Final 8 Variable Model . . . . .	36

**LIST OF TABLES**

<b>Table</b>		<b>Page</b>
1	Correlation Matrix for the Data Set . . . . .	11
2	Best Subsets Regression Results . . . . .	14
3	Best Subsets Regression #2 . . . . .	18
4	Best Subsets Regression Using the Transformed Response . . . . .	29
5	Correlation Matrix with All Variables Considered . . . . .	31
6	Best Subsets Regression Omitting C18, and C19 . . . . .	32
7	Best Subsets Regression Omitting C3, C4, C5, and C6 . . . . .	33

## ACKNOWLEDGMENTS

I would like to express my gratitude to Advanced Energy, in Fort Collins, Colorado and to Doug Mader for the opportunity to conduct this study. I would also like to thank my Advisor, Professor William Astle and my committee members, Dr. Ruth Maurer, and Dr. R.E.D. Woolsey for their patience and guidance during my work towards this degree.

I would especially like to thank Erdem Ince and the rest of the Guild members for their support. We are all examples of how communication, cooperation, and goodwill can easily overcome any obstacle.

Finally, I cannot thank Shawn Bennett enough for her ultimate patience and understanding. I wonder how she put up with such a nut like me.

## DEDICATION

I would like to dedicate this work to my Dad, who has taught me that timing is everything. To my Mother, who has taught me that sincerity, dedication, and discipline create good cause and effect. To my niece, Keiko, who exemplifies that life is truly a constant struggle from the moment we set foot on this planet. Finally to the rest of my family who makes it all worth while.

## Chapter 1

### INTRODUCTION

#### **1.1 General**

The research and development of emerging technology in the business world is, without a doubt, a vital aspect that determines the ultimate success of growing organizations. As markets respond to the global economy, the research and design of new technology follows suit. Since the pursuit of research and development is speculative in nature, it is essential to remain focused on how much cost can be absorbed against the potential revenues that can be generated if the technology were to be implemented at the mass production level. Often, there is some question about the costs that are used in determining whether to continue with a given project. As new products are engineered, there are numerous designs that accomplish the same function. Consider the hundreds of different automobile models which all perform the same fundamental function of providing transportation from point A to point B. It is the task of the design team to evaluate the most effective solution for the given function while considering any budget constraints that may be imposed upon the project.

Statistics is a tool that engineers and scientists can use to quantify their questions concerning cost, and the probabilistic nature of the research and design process. Statistics

is used in a variety of functions within the business world from quality control measures, design of experiments, and the forecasting of demand. The branch of statistics that is commonly used for these types of applications is the concept of linear or nonlinear regression analysis.

### **1.2 Simple Linear Regression Model**

Regression analysis is useful in deriving the relationships between measurable variables and for the prediction of future values (Weisberg, 1980). Specifically, linear regression describes the class of relationships in which the variables can be related by linear functions. The British anthropologist and meteorologist Sir Francis Galton was the first to coin the term regression in his work that was published in 1885 referring to the relationship in parent and offspring sizes. The method of least squares estimation, however, dates back to Gauss and Legendre who independently claimed to have derived its utility around 1805. The variables considered in regression analysis are distinguished as either predictor or response variables.

The predictor variables are also called independent variables. These variables are the measurable data set from which their inter-relationship predicts the value of the response. The response variable is also known as the dependent variable, since its value is a function of the predictor variables.

The data used in this thesis is composed of almost entirely measurement data, i.e. length, width, thickness, etc. There is no doubt that measurement errors will always be present, because it is next to impossible to measure quantitative variables with perfect accuracy. It is also reasonable to believe that random errors due to natural variability exist within the data set.

Regression analysis relies on certain fundamental assumptions to insure model validation. These assumptions are essential when making any inferences about the model such as hypothesis testing and confidence interval estimates. Weisberg (1980) and Carroll and Rupert (1988) both give an elegant discussion of the classical assumptions. As mentioned before, data will inherently have some kind of random error associated with the observations. For our case we will call  $\epsilon_i$  the statistical error associated with the  $i^{\text{th}}$  case where  $i = 1, 2, \dots, n$ . The first assumption about regression is that the expected values of the error components are all equal to zero.

$$E(\epsilon_i) = 0 \text{ for } i = 1, 2, 3, \dots, n \quad (1.1)$$

The second assumption is that the errors are independent of each other. We can write this in terms of the covariance operator as follows:

$$\text{Cov}(\epsilon_i, \epsilon_j) = 0 \text{ for } i \neq j. \quad (1.2)$$

The third assumption is that the errors have a common constant variance as follows:

$$\text{Var}(\epsilon_i) = \sigma^2 \quad (1.3)$$

The final assumption is that the errors are normally distributed.



$$\varepsilon_i \sim \text{Normally distributed population} \quad (1.4)$$

This final assumption is particularly important when constructing confidence intervals and conducting hypothesis tests when testing the model's integrity. At this point, we can combine the four assumptions into one statement that describes the limitations that must be imposed on the random error components of the observations as follows.

$$\varepsilon_i \sim \text{N I D } (0, \sigma^2) \quad (1.5)$$

Given the previous assumptions concerning regression models, linear regression by least squares estimation attempts to fit the predictor variables to the response variable with the following linear equation.

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad (1.6)$$

This simple regression model is then easily expanded for more than one independent variable to give the multiple regression model in the following equation.

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \dots + \beta_{p-1} X_{i,p-1} + \varepsilon_i \quad (1.7)$$

In each case,  $Y_i$  is the response variable,  $X_i$  are the predictor variables, 'p-1' indicates the number of predictor variables, the  $\beta_i$ 's are the coefficients related to each predictor variable and  $\beta_0$  is the constant. There are often instances where the data may not lend itself to a simple linear model. In these instances it is beneficial to turn to more advanced methods of data manipulation in order to create a sound model. Data transformations are common methods of data manipulation that may aid in model development.

### **1.3 Data Transformation**

During regression analysis and the model building process, certain instances may arise where the model may violate the classical assumptions of least squares estimation as introduced in equations 1.1 through 1.5. It becomes necessary to turn to data transformations in an attempt to re-establish the least squares assumptions. Data transformations are commonly used tools that aid in reducing complex models down to simple linear models. There are two general types of data transformation, both of which are employed within this work.

The first type of transformation is accomplished by manipulating the predictor variables. It may be that certain predictor variables do not have a strong linear relationship to the response variable. Although this may seem to be a problem, there may be a very simple mathematical relationship between the predictor and response. The most common relationships and resulting transformations are the multiplicative, exponential or logarithmic, reciprocal, and higher ordered relationships (i.e. squares, cubes, square roots, etc.). It is often necessary to plot the predictor against the response (also known as scatter plots) to determine what kind of relationship exists between the variables. This is done by visual inspection, and a certain amount of analytical reasoning based on prior experience with the subject matter. Trial and error approaches to determine what effect a given transformation will generate between the predictor and response are generally the most effective way of determining an effective transformation.

The second general type of transformation is used when there is strong evidence to suggest that the model shows signs of nonlinearity, non-normality, and heteroscedasticity (non-constant error variance). Recall that these are direct violations of assumptions 1.3 and 1.4. A clear indication of the existence of non-normality would be a response that is non-negative. Atkinson (1985) suggested that a transformation of the response might be helpful when the response is in all cases non-negative.

In this particular case, the response which will always be non-negative and consequently cannot strictly follow a normal distribution is cost. Atkinson (1985) goes on to suggest that with a wide range of non-negative response values, a natural choice for a transformation would be in terms of a logarithm of some kind. A general method for choosing the desired response transformation would be to consider the parametric family of power transformations first introduced by Tukey (1957) and later Box and Cox (1964).

$$h(y, \lambda) = \begin{cases} (y^\lambda - 1) / \lambda & \text{if } \lambda \neq 0 \\ \ln y & \text{if } \lambda = 0 \end{cases} \quad (1.8)$$

When employing this method of transformation, the question arises about what value of  $\lambda$  should be used to complete the transformation. Further discussion of this specific transformation follows in Chapter 2.

Although this method of transformation can be very effective, it must be noted that this technique may not restore the classical least squares assumptions in all cases. There are numerous transformation techniques available in the mathematical community that are

not explored in this specific work. There are also specific circumstances where the possibility of employing generalized linear models (GLM) may be appropriate in a given model building process. Please be aware that the methods employed in this body of work do not imply that they are appropriate for all cases.

#### **1.4 Purpose**

Advanced Energy Inc. is a small but growing producer of industrial power supplies located in Fort Collins, Colorado. A.E. is the industry leader for providing power supplies to applications like surface mount technology within the computer chip industry, as well as applications for the medical industry. They have continually been recognized by Fortune Magazine as one of the nations leading “up and coming” companies.

A.E. currently relies on two vendors (R.J.R. and Lundahl) to manufacture specific printed circuit boards that are designed and engineered in-house, and then implemented in the manufacturing process. At the present time, the only method of determining the cost of purchasing printed circuit boards is by compiling certain vital information about the board, and then sending it via fax to each of Advanced Energy's circuit board suppliers. The suppliers would then return a fax with the price quote. Typically, this process has a twenty-four hour turnaround.

As Advanced Energy continues to grow forward, an aspect that has driven the company's previous success is its innovative designs, and the leading edge technology

employed in its products. Because future product development is such an important part of this company, the need for a model that estimates potential costs of purchasing printed circuit boards would aid in the early stages of product development and the decision making processes. A simple model will allow instant cost estimation and alleviate any disputes related to a specific design, by giving the engineers the ability to maintain a perspective on the costs associated with a specific design during the earliest stages of development. Another potential benefit of such a model would be the continual updating of a database that would enable the company to directly evaluate costs between its vendors.

Initial work towards the development of a model had been conducted by Advanced Energy employees Doug Mader and Dan Sutherland. The project was presented to me by both Doug and Dan, and this report represents the culmination of my summer internship with Advanced Energy. The purpose of this project was to compile several sets of printed circuit board design specifications, and to develop a single model to predict the cost of purchasing boards in the future. The purpose of this document is to provide the reader with a perspective on what tools and methodology were used in this specific project, how to apply statistics, and more specifically how to apply linear regression to a genuine problem in the business sector.

## **Chapter 2**

### **METHODOLOGY IN PROBLEM SOLVING**

The method employed to develop the resulting model was multiple regression, least squares estimation. This method was chosen because regression is an effective way to describe the relationship between two or more variables. One variable, known as the dependent variable, can be predicted from the other independent variables. The initial analysis conducted by Doug Mader and Dan Sutherland was done in this fashion, so it was logical to continue with the same methodology.

The first task at hand was to consolidate all of the data into a database for easy manipulation. This included the removal of incomplete records, because the statistical software utilized in the model building process cannot handle missing data. The data are historic, and were obtained from price quotes based on the desired circuit board specifications. The data were originally divided into three separate design criteria based on the whether or not the boards were single sided, double sided, or multi-layered. By utilizing indicator variables to ascertain with which design criteria each board was originally associated, the three data sets were compiled into a single file.

## **2.1 Zero-One Indicator Variables**

Zero-one indicator variables are very useful in analysis because they can quantitatively identify what is really a qualitative variable. For example a quantitative variable takes on a value on a defined scale such as temperature, age, or cost. Similarly qualitative variables generally refer to a state of existence such as gender (male or female), or switch state (on or off). Neter, Wasserman, Kutner (1990) suggest that there are many ways to quantitatively identify qualitative variables.

For this analysis, the variables will take on values of either zero or one. This convention implies that if the state applies to the particular record then the variable takes on a value of one. Otherwise, the variable takes on the value of zero. This is analogous to saying if the state is true then one, otherwise zero. Note that within the data set used in this analysis there are several columns of data that are quantified by zero-one indicator variables. Specifically note the vendor column, which was used to identify which company was solicited in the given quote. Zero-one indicators were also utilized in the Test, Tab Route, Two Layers, Three Layers, Copper 1, and Copper 2 columns. Refer to the data diskette to inspect the entire data set using a Microsoft Excel ver. 5.0 or higher format.

Once the data were in a suitable format, a correlation matrix and scatter plots were generated to look for any obvious signs of how to proceed in the model building process. The correlation matrix will identify any presence of pair-wise multicollinearity within any given pair of predictor variables.

## 2.2 Multicollinearity:

Multicollinearity is defined as the situation when independent variables have a strong relationship or correlation between one another. Ideally, it is desirable to have independent variables completely uncorrelated. However, in reality this situation is more than unlikely. For example, the variables Square Inches (C12) and Volume (C15) will naturally have a higher correlation. Similarly observe the high correlation between Width (C3), and Square Inches (C12) and Length (C4) and Square Inches (C12). Refer to Table 1 to observe the given examples within the correlation matrix.

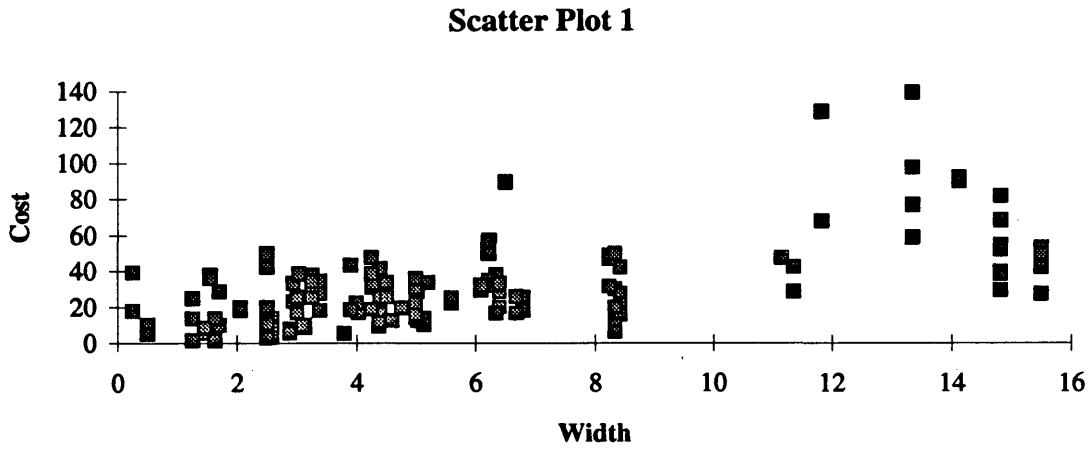
**Table 1 Correlation Matrix for the Data Set**

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	C16	C17
C2	-0.08																
C3	0.119	0.36															
C4	-0.04	0.4	0.185														
C5	0.01	0.027	0.468	0.426													
C6	-0.05	-0.12	-0.13	-0.22	-0.07												
C7	0.043	0.337	0.446	0.252	0.166	-0.02											
C8	-0.07	0.233	-0.26	-0.04	-0.33	0.119	-0.34										
C9	-0.03	0.262	0.211	0.213	-0.04	0.137	0.659	-0.01									
C10	-0.1	-0.07	-0.14	0.338	0.433	0.107	0.173	-0.2	0.109								
C11	0.042	-0.01	0.052	0.096	0.037	-0.03	0.05	-0.11	-0.07	0.049							
C12	0.048	0.504	0.799	0.629	0.56	-0.14	0.378	-0.17	0.214	0.041	-0.01						
C13	0.078	0.109	0.219	-0.27	-0.24	-0.06	-0.16	0.183	-0.04	-0.8	-0.13	0.073					
C14	-0.03	-0.17	-0.11	0.21	0.314	0.003	0.079	-0.19	-0.02	0.561	0.279	0.002	-0.7				
C15	0.008	0.659	0.765	0.605	0.45	-0.12	0.41	-0.11	0.253	0.006	-0.02	0.969	0.113	-0.05			
C16	0.082	-0.19	0.283	-0.16	0.191	-0.04	-0.11	-0.29	-0.19	-0.07	0.083	0.089	0.152	-0.06	0.02		
C17	-0.05	0.21	-0.19	0.076	-0.11	0.044	0.027	0.262	0.075	0.039	-0.11	-0.03	-0.08	-0.01	0.035	-0.85	
C19	0.134	0.266	0.592	0.513	0.514	-0.31	0.208	-0.23	0.038	0.071	0.048	0.755	-0.09	0.149	0.675	0.161	-0.15

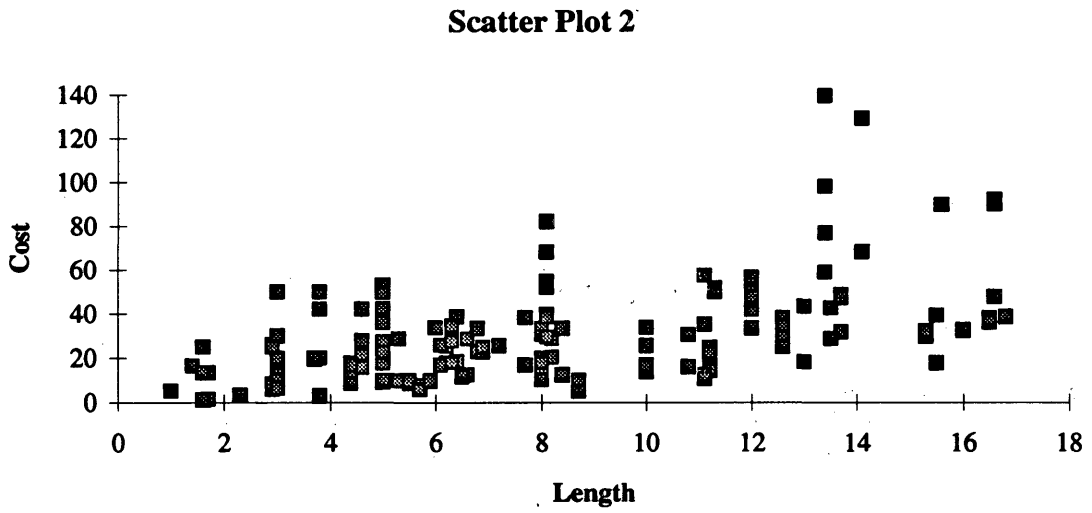
**Note:** C1 = RJR vs. Lundahl      C2 = Thickness      C3 = Width  
 C4 = Length      C5 = Holes      C6 = Order Qty  
 C7 = Drills      C8 = Smallest Hole      C9 = Largest Hole  
 C10 = Test 0or1      C11 = Tab Route 0or1      C12 = Sq Inch  
 C13 = Two Layers      C14 = Three Layers      C15 = Volume  
 C16 = Copper 1      C17 = Copper 2      C19 = Cost

Although multicollinearity is undesirable, Neter, Wasserman, and Kutner (1990) suggest its presence generally does not inhibit the ability to obtain a good fit between the response and independent variables or even make inferences about mean responses and predict new observations as long as the inferences are made within the region of the observations. Its presence does inhibit the ability to obtain an accurate assessment of the true individual regression coefficients. Multicollinearity also limits the ability to conduct any sensitivity analysis in terms of measuring the effects on the mean response when a specific independent variable is increased by one unit while holding the rest of the independent variables constant.

The scatter plots give a graphical representation of the relationship between the dependent variable and each of the independent variables, and often times the mathematical function that relates the variables is easily ascertained from the plot. Refer to Figures 1 and 2 and the data diskette for examples of scatter plots.



**Figure 1: Scatter Plot of Cost vs. Width**



**Figure 2: Scatter Plot of Cost vs. Length**

**2.3 Best Subsets Regression:  $C_p$  Criterion**

At this point, the first best subsets regression was performed on the data set. Best subsets regression is a very powerful tool that is employed in some statistical software packages such as Minitab. It iteratively evaluates all of the possible combinations of regression models based on the number of potential predictor variables in the data set. It then displays the best models based on the highest  $R^2$  in a given “n” variable model. The analyst is then able to assess what models might warrant further inspection. Table 2 shows the results of this initial best subsets regression command as output by Minitab.

**Table 2 Best Subsets Regression Results**

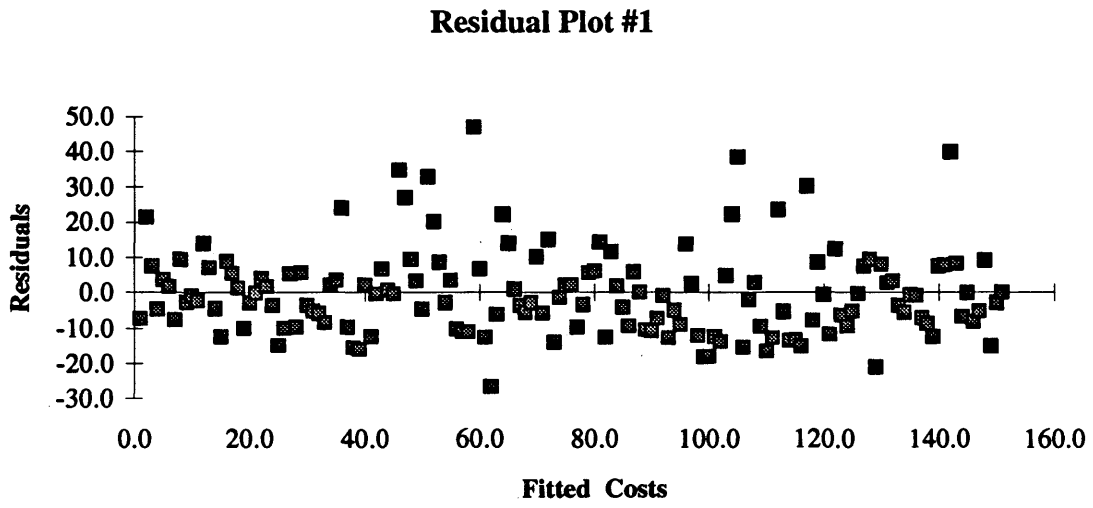
Vars	R-sq	Adj. R-sq	C-p	s	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1	57.0	56.7	52.2	14.855												X					
2	62.1	61.6	30.8	14												X			X		
3	65.9	65.2	15.1	13.318						X						X			X		
4	67.4	66.5	10.1	13.064						X						X	X		X		
5	68.6	67.5	6.7	12.867						X				X		X	X		X		
6	<b>69.3</b>	<b>68.0</b>	<b>5.3</b>	<b>12.763</b>						X				X		X	X		X		X
7	69.8	68.3	5.2	12.71						X	X			X		X	X		X		X
8	70.3	68.6	5.0	12.656	X					X	X			X		X	X		X		X
9	70.4	68.5	6.4	12.674	X	X				X	X			X		X	X		X		X
10	70.6	68.5	7.5	12.674	X	X				X	X	X		X		X	X		X		X
11	70.8	68.4	8.7	12.682	X	X	X	X		X	X			X		X	X		X		X
12	71.0	68.5	9.6	12.68	X	X	X	X		X	X	X		X		X	X		X		X
13	71.2	68.4	10.8	12.686	X	X	X	X		X	X	X		X		X	X		X	X	X
14	71.3	68.3	12.3	12.708	X	X	X	X	X	X	X	X		X		X	X		X	X	X
15	71.3	68.1	14.1	12.746	X	X	X	X	X	X	X	X		X	X	X	X		X	X	X
16	71.4	67.9	16.0	12.791	X	X	X	X	X	X	X	X		X	X	X	X	X	X	X	X
17	71.4	67.7	18.0	12.838	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X

Note: The Variable numbering convention follows the column numbering defined in Table 1.

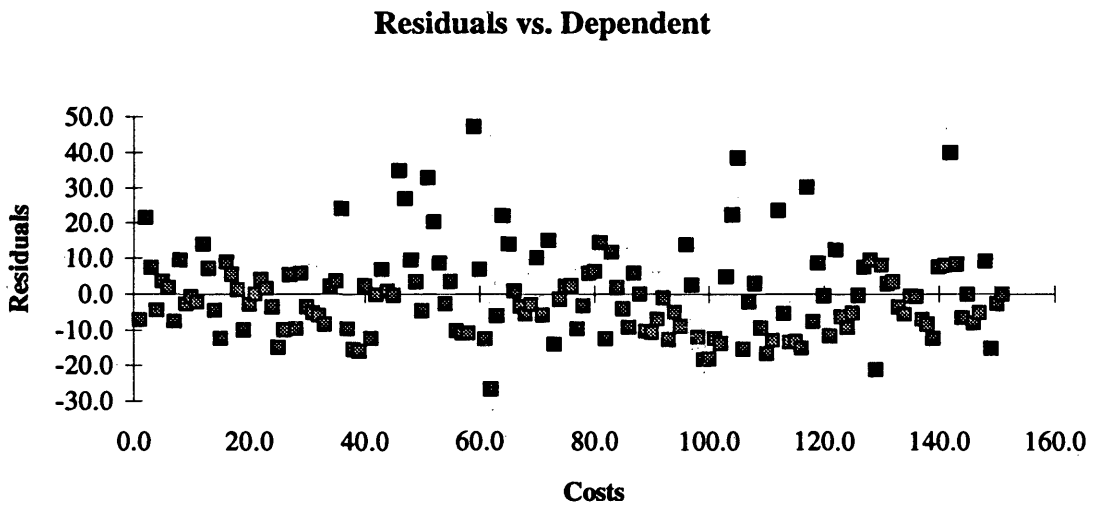
In addition to using the standard criteria of looking for a strong  $R^2$  and low regression estimated variance 's', another criterion used to determine the most desirable model as displayed by Table 2 is called the  $C_p$  Criterion. The criterion was first introduced by C.L. Mallow and is sometimes referred to as the Mallow's  $C_p$ . The criterion is derived from the total mean squared error of the "n" fitted values for each subset regression model. For a complete derivation of the criterion, refer to Miller (1990) or Neter, Wasserman, and Kutner (1990).

The  $C_p$  Criterion is used as a preliminary tool for making a choice about what models deserve a closer look. When used in conjunction with what are typically the most common criteria for evaluating a model, the  $C_p$  can be a very useful tool to make a more informed decision. The criterion suggests that the models that warrant further consideration are those whose  $C_p$  value approaches the number of predictor variables in the model including the constant ( $C_p = P$ ). For the case shown in Table 2, the model that warrants further consideration based on the  $C_p$ ,  $R^2$  and estimated variance would be the six variable model that is highlighted.

At this point, the complete regression was performed with Minitab and associated residual plots were generated. However, the relatively low  $R^2$  and the residual plots suggest that this model may not be very robust. More specifically the residual plots do not appear to be entirely random. Refer to Figures 3 and 4 as well as the data diskette.



**Figure 3: Residual Plot of the 6 Variable Model (Residuals vs. Fitted Costs)**



**Figure 4: Residual Plot of the 6 Variable Model (Residuals vs. Costs)**

Due to the lack of confidence in the six variable model at this point, a careful re-examination of the scatter plots indicated that individual transforms could be made to some of the independent variables. A trial and error process of looking at simple regressions using a variety of predictor variables and perturbations of the original predictors resulted in the construction of seven potential independent variables to consider in the model selection. Due to the power of the statistical software available on the Apple MacIntosh, the trial and error process was used to generate immediate graphics to see how a given transform fit the response. Simple regression was performed on a given predictor and the response. Then, numerous transformations were applied to the predictor variable and simple regressions of the constructed variables were generated against the response. By comparing the  $R^2$  for each of the simple regressions, as well as generating graphics of the response versus the fitted response from each simple regression, seven constructed variables were chosen to add to the data set. It was not clear at this point whether the new variables would even have an impact on the final model; however, subsequent best subset regressions would be able to include the new constructed variables in an effort to generate a robust final model. The seven variables that were included within the data set were,  $(\text{Ln}(\text{Width}))^2$ ,  $\text{Ln}(\text{Length})$ ,  $(\text{Ln}(\text{Holes}))^2$ ,  $(\text{Order Qty})^2$ ,  $\text{Ln}(\text{drills})$ ,  $(\text{smallest})^2$ , and  $\text{Ln}(\text{sq inches})$ .

Once the new variables were added to the database, another best subsets regression was performed on the data to get another idea of potential models to pursue. The results of this best regression command from Minitab are shown below in Table 3.

**Table 3 Best Subsets Regression #2**

Vars	R-sq	Adj. R-sq	C-p	s	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	
3	69.1	68.5	78.7	12.68									X	X								X							
4	74.3	73.5	43.9	11.613									X	X								X				X			
5	77.0	76.2	26.2	11.012									X	X								X		X	X	X			
6	78.2	77.2	19.9	10.77									X	X								X		X	X	X	X		X
7	79.2	78.2	14.3	10.54									X	X		X						X	X	X	X	X			
8	79.8	78.7	12.3	10.432		X							X	X		X						X	X	X	X	X			
9	80.4	79.1	10.0	10.312		X							X	X			X					X	X	X	X	X			X
10	81.0	79.6	7.8	10.193	X	X							X	X			X					X	X	X	X	X			X
11	81.3	79.8	7.4	10.139	X	X							X	X				X	X			X	X	X	X	X			X
12	81.6	80.0	7.4	10.099	X	X							X	X				X	X	X		X	X	X	X	X			X

The highlighted row in Table 3 indicates that the ten variable model would warrant further consideration based on the previously defined decision criteria where the number of predictors ‘P’ in the model approaches the ‘C<sub>P</sub>’ value as well as R<sup>2</sup> and estimated variance. At this point, a complete regression was run on the ten variables denoted by the model highlighted above. The results are summarized in Figure 5. By looking at the R<sup>2</sup> alone, there is some indication that the transformations that were made have had a significant impact on the model at this point.

The Regression equation is:

$$\text{Cost} = 32.3 + 3.53 C1 + 0.318 C2 - 0.397 C9 + 0.00108 C10 - 0.00147 C14 \\ + 0.918 C18 - 5.98 C19 - 11.4 C20 - 0.00623 C22 - 4.59 C24$$

<i>Predictor</i>	<i>Coef</i>	<i>Stdev</i>	<i>t-ratio</i>	<i>p</i>	<i>VIF</i>
Constant	32.28	10.34	3.12	0.002	
RJR (1) vs.Lundahl (0)	3.532	1.708	2.07	0.04	1
Thickness	0.3179	0.149	2.13	0.035	5.2
Order Qty	-0.39716	0.04091	-9.71	0	7.6
(order qty)^2	0.00108	0.0001388	7.79	0	8.1
(smallest)^2	-0.00147	0.0004462	-3.3	0.001	1.6
Sq Inch	0.9183	0.1261	7.28	0	54.7
Ln (sq inches)	-5.975	1.836	-3.26	0.001	5.5
Two Layers	-11.379	2.349	-4.84	0	1.1
Volume	-0.00623	0.00152	-4.1	0	59.6
Copper 2	-4.592	1.989	-2.31	0.022	1.1

$s = 10.19$

R-sq = 81.0%

R-sq(adj) = 79.6%

Analysis of Variance Tables:

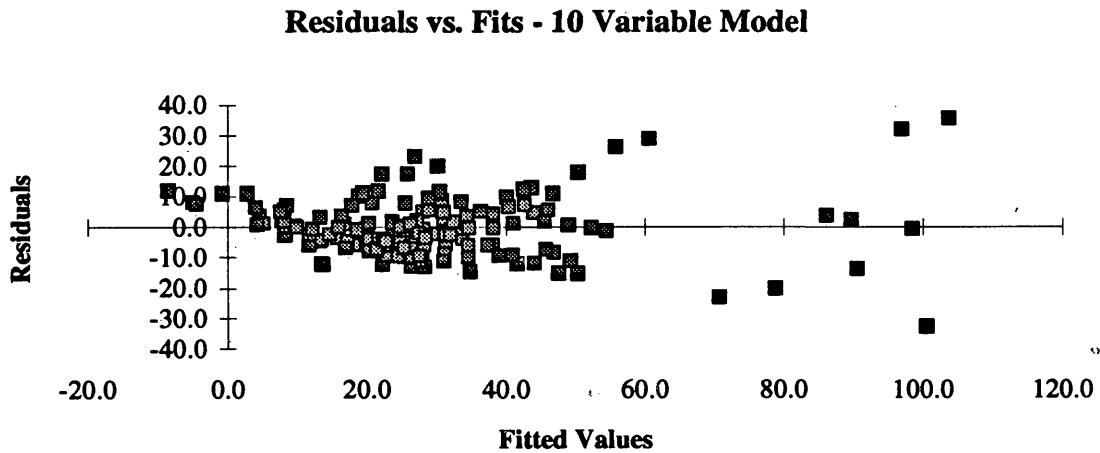
<i>SOURCE</i>	<i>DF</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>p</i>
Regression	10	61517.4	6151.7	59.2	0
Error	139	14443	103.9		
Total	149	75960.5			

<i>SOURCE</i>	<i>DF</i>	<i>SEQ SS</i>
RJR (1) vs.Lundahl (0)	1	1358.8
Thickness	1	5870.9
Order Qty	1	5675.5
(order qty)^2	1	5047.7
(smallest)^2	1	5969.7
Sq Inch	1	31557.6
Ln (sq inches)	1	196.9
Two Layers	1	3160.8
Volume	1	2125.9
Copper 2	1	553.8

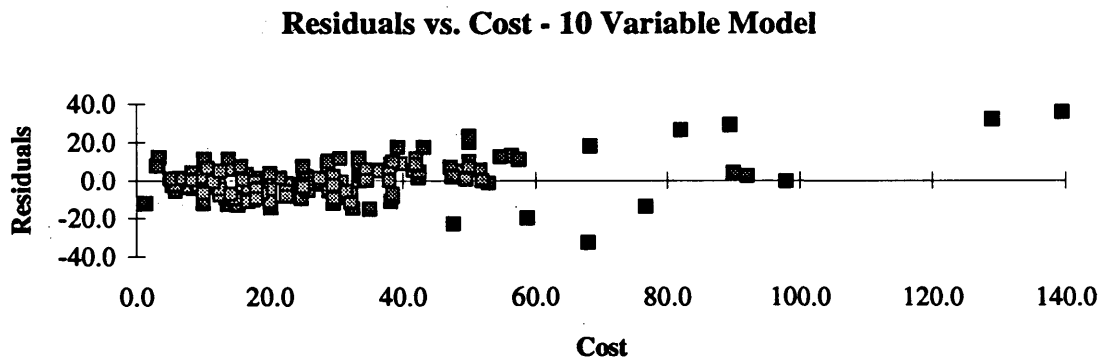
Note: The Column indexes correspond to the variable names of the Predictor column in the first table.

Figure 5: Regression Summary of the 10 Variable Model

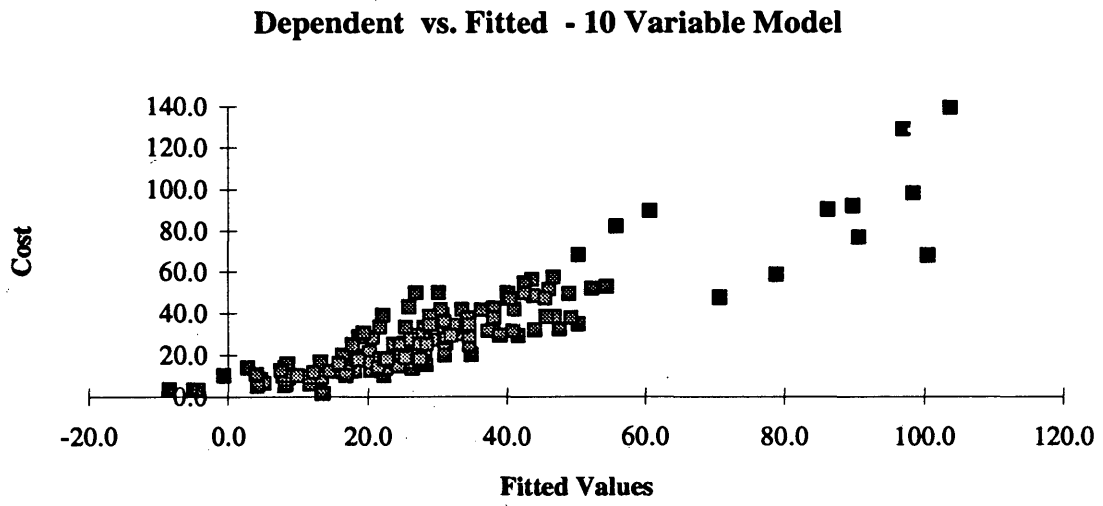
Along with the complete regression that was performed above, residual plots were generated to be evaluated as part of this model. Refer to Figures 6-9.



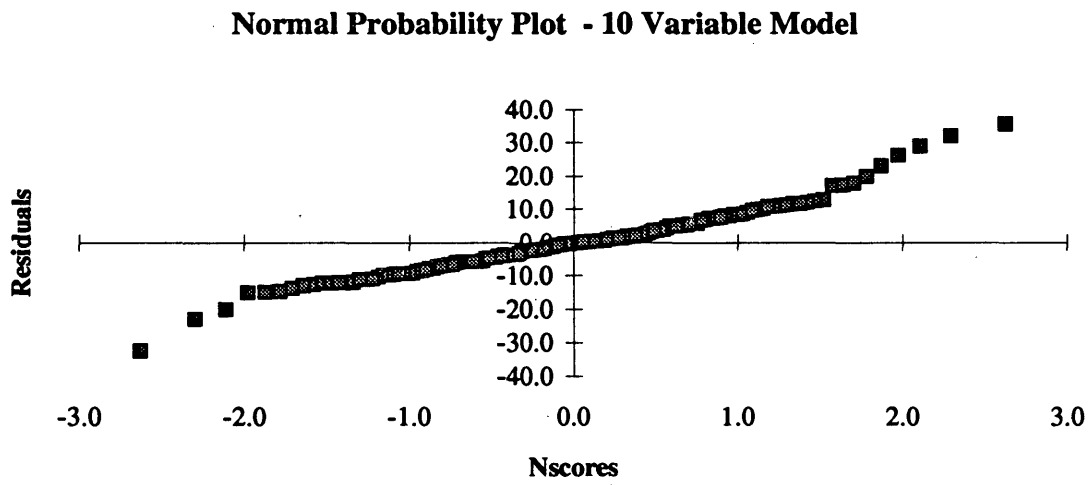
**Figure 6: Residual Plot #1 for the 10 Variable Model (Residuals vs. Fitted Values)**



**Figure 7: Residual Plot #2 for the 10 Variable Model (Residual vs. Response)**



**Figure 8: Residual Plot #3 of the 10 Variable Model (Response vs. Fitted Response)**



**Figure 9: Normal Probability Plot of the 10 Variable Model**

## **2.4 Residual Plot Analysis**

Since statistics and linear regression are inexact at best, analysts must draw from their own experience and intuition in the model building process. The value of residual plot analysis from data generated by a least squares estimation model is immense. Weisberg (1980) suggests that residual plots are the most important diagnostic tool available to the analyst. Graphical means of analysis provide the analyst the ability to see the data as a whole, and to be able to visualize what is often a very complex relationship.

When looking at residual plots, the general rule of thumb is that the graph should not exhibit any obvious trends. In fact, they should look highly random no matter if you are plotting the fitted response ( $\hat{y}_i$ ) or any of the predictors ( $x_i$ ) against the residuals generated by the regression. The residual is the difference between the observed and the predicted response for a given observation.

As mentioned before, it is desirable to have random looking residual plots. Anything that exhibits obvious trends would suggest that the given model is violating one or more of the original least squares assumptions that were defined in Chapter 1, equations 1.1-1.4.

The residual plots displayed in Figures 6 and 7 above exhibit potential problems with the ten variable model as defined. Close inspection indicates that each plot exhibits signs of a funnel shape that expands outward and to the right from the origin. Weisberg (1980) and Cook and Weisberg (1982) identify this funnel shape in residual plots as the classical signs of heteroscedasticity (non-constant error variance).

The direction of the funnel indicates that the variance is increasing as a function of the quantity displayed along the horizontal axis. It must be noted that funnel shapes are not the only trend that suggest problems with a given model. Weisberg(1980) is an excellent source for a more detailed discussion of other trends that occur within data sets and the remedial measures that can be taken for each case.

Figure 8 exhibits problems, because one would expect that the plot of fitted response against the actual response would follow a stronger linear trend. The normal probability plot displayed by Figure 9 is good way to check the normality assumption equation 1.4. If the plot does not show a strong linear trend then, there would be strong evidence to suggest that the errors may not be normally distributed. Figure 9 looks as though the normality assumption is not a problem.

These results should not come as a great surprise based on the discussion in the previous chapter about the nature of the data itself. Weisberg (1980) and Atkinson (1985) indicate the most commonly used solution to this problem would be a variance stabilizing transformation on the response variable.

## **2.5 Score Statistic Transformation**

Since all of the analysis up to this point has not produced a robust regression model that adheres to the assumptions set forth in equations 1.1-1.4, more complex data transformation is explored. Specifically a transformation of the response variable was

performed based on the parametric family of power transformations that was briefly discussed in the previous chapter.

As indicated in section 1.3 of Chapter 1, Box and Cox (1964) are principally responsible for the work done with this family of transformation. Recall the form of the transformation from equations 1.7 as follows.

$$h(y, \lambda) = \begin{cases} (y^\lambda - 1) / \lambda & \text{if } \lambda \neq 0 \\ \ln y & \text{if } \lambda = 0 \end{cases} \quad (1.8)$$

As mentioned previously, the question is now what value of  $\lambda$  should be used to complete the transformation. Box and Cox suggest using methods of maximum likelihood function or a Bayesian approach to complete the transformation. These methods are derived and summarized in all of the selected references of this paper with the exception of Aitchison (1986), Miller (1990), Neter, Wasserman, and Kutner (1990), and Pankratz (1991).

Atkinson (1985) suggests an alternate approach to complete the transform based on the Atkinson score statistic and a constructed variable,  $w_p$ , generated from the statistic. This approach is much easier conceptually and computationally versus the iterative maximum likelihood approach, however it is only an approximation. Any confidence interval estimation would require the use of the likelihood approach to insure the proper completion of the transform.

Atkinson's approach is only an approximation because he uses a truncated Taylor series expansion to arrive at the final score statistic. The complete derivation is conducted within Atkinson (1985), Atkinson and Lawrance(1989), and Lawrance (1987).

The transformation is computed by first calculating the constructed variable and adding it to the data set. The next step is to conduct a complete regression of the response against all of the available predictor variables that are in the data set including the Atkinson's score variable. Once the regression is performed a simple hypothesis test is conducted to verify whether or not the score variable is significant in the regression. The hypothesis test simply verifies if the coefficient associated with the Atkinson's score variable is zero or not zero. This is written symbolically:

$$\begin{aligned} H_o: \phi &= 0 \\ H_a: \phi &\neq 0 \end{aligned} \tag{2.1}$$

If the result of the hypothesis test implies that the Atkinson's score variable is indeed significant, then the coefficient  $\phi$  is used to determine the value of  $\lambda$  to complete the transformation. This process is summarized symbolically below.

Recall the basic multiple regression model from equation 1.7.

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \dots + \beta_{p-1} X_{i,p-1} + \epsilon_i \tag{1.7}$$

Atkinson's approach fits the following multiple regression model.

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \dots + \beta_{p-1} X_{i,p-1} + \epsilon_i + \phi w_p \tag{2.2}$$

Note that this model contains the constructed variable and its associated coefficient. We define the constructed variable,  $w_p$ , in the following equation.

$$w_p = y \{ \ln(y / \dot{y}) - 1 \} \quad (2.3)$$

Where  $\dot{y}$  is the geometric mean of the response variable defined as follows.

$$\dot{y} = \left( \prod_{i=1}^n y_i \right)^{1/n} \quad (2.4)$$

Once the Atkinson's score variable is added to the data set the full regression is performed with the results summarized in Figure 10.

The Regression Equation is:

$$\begin{aligned}
 C25 = & 27.0 + 3.75 C1 + 0.225 C2 + 1.21 C3 - 0.88 C4 - 0.299 C5 + 6.43 C6 \\
 & - 0.00406 C7 + 0.339 C8 - 0.345 C9 + 0.000995 C10 + 1.11 C11 \\
 & - 6.57 C12 + 0.148 C13 - 0.00218 C14 + 0.0058 C15 - 2.02 C16 \\
 & - 1.42 C17 + 0.242 C18 - 2.17 C19 - 9.79 C20 + 1.61 C21 \\
 & - 0.00172 C22 - 3.46 C23 - 7.51 C24 + 0.653 C26
 \end{aligned}$$

<i>Predictor</i>	<i>Coef</i>	<i>Stdev</i>	<i>t-ratio</i>	<i>p</i>
<b>Constant</b>	27	15	1.8	0.074
<b>RJR (1) vs.Lundahl (0)</b>	3.753	1.198	3.13	0.002
<b>Thickness</b>	0.2254	0.1179	1.91	0.058
<b>Width</b>	1.213	1.551	0.78	0.436
<b>(Ln (Width))^2</b>	-0.878	2.789	-0.31	0.753
<b>Length</b>	-0.2993	0.7372	-0.41	0.685
<b>Ln (length)</b>	6.429	5.663	1.14	0.258
<b>Holes</b>	-0.004056	0.003412	-1.19	0.237
<b>(Ln (Holes))^2</b>	0.3389	0.1943	1.74	0.084
<b>Order Qty</b>	-0.34485	0.03231	-10.67	0.000
<b>(order qty)^2</b>	0.0009947	0.0001075	9.25	0.000
<b>Drills</b>	1.1085	0.7608	1.46	0.148
<b>Ln (drills)</b>	-6.568	5.546	-1.18	0.239
<b>Smallest</b>	0.1481	0.2397	0.62	0.538
<b>(smallest)^2</b>	-0.002179	0.001279	-1.7	0.091
<b>Largest</b>	0.00585	0.01098	0.53	0.595
<b>Test</b>	-2.019	2.881	-0.7	0.485
<b>Tab Route</b>	-1.416	2.151	-0.66	0.512
<b>Sq Inch</b>	0.2421	0.1378	1.76	0.081
<b>Ln (sq inches)</b>	-2.169	2.187	-0.99	0.323
<b>Two Layers</b>	-9.788	3.598	-2.72	0.007
<b>Three Layers</b>	1.606	3.209	0.5	0.618
<b>Volume</b>	-0.001716	0.00141	-1.22	0.226
<b>Copper 1</b>	-3.463	3.538	-0.98	0.330
<b>Copper 2</b>	-7.506	2.983	-2.52	0.013
<b>Atkinson's Scores</b>	<b>0.65254</b>	<b>0.05389</b>	<b>12.11</b>	<b>0.000</b>

Figure 10: Regression Summary for Transformation of Response

With the results of the regression summarized above in Figure 10, the hypothesis test can be performed.

$$\left. \begin{array}{l} H_o: \phi = 0 \\ H_a: \phi \neq 0 \end{array} \right\} = \left\{ \begin{array}{l} H_o: C26 = 0 \\ H_a: C26 \neq 0 \end{array} \right. \quad (2.5)$$

On the basis of the p-value for the Atkinson Scores, C26, that was generated by the regression, it becomes obvious that the Atkinson Scores are significant. In effect, the p-value tells us that the probability that the regression will generate the given coefficient of 0.65254 when in fact it should actually be zero is zero. The coefficient generated for the Atkinson Scores variable is then used to complete the transform as follows:

$$\hat{\lambda} = (1 - \phi) \quad (2.6)$$

Using the coefficient that was generated by the regression,  $\phi = .65254$ , then  $\hat{\lambda} = .34746$ .

The resulting model then takes on the form displayed in equation 2.7 below.

$$\left( \frac{y_i^{\hat{\lambda}} - 1}{\hat{\lambda}} \right) = \beta_0 + \beta_1 X_{i,1} + \dots + \beta_{p-1} X_{i,p-1} + \epsilon_i \quad (2.7)$$

With the transformed cost added to the data set, further best subset analysis was conducted to try to identify other potential models. The result of this best subsets regression is summarized below in Table 4.

**Table 4 Best Subsets Regression Using the Transformed Response**

Vars	R-sq	Adj. R-sq	C-p	s	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
4	73.8	73.1	63.3	1.1848									X	X								X	X					
5	77.1	76.3	40.3	1.1133									X	X			X					X	X					
6	78.7	77.8	29.5	1.0763									X	X			X					X	X					X
7	80.3	79.3	19.2	1.0393	X								X	X			X					X	X					X
8	81.1	80.0	15.0	1.0217	X								X	X			X		X			X	X					X
9	<b>82.1</b>	<b>80.9</b>	<b>9.4</b>	<b>0.99848</b>	X			X	X				X	X			X					X	X					X
10	82.3	81.0	9.6	0.99559	X			X	X				X	X			X				X	X		X				X
11	82.6	81.2	9.5	0.99153	X	X		X					X	X			X				X	X		X			X	X
12	82.7	81.2	10.3	0.99078	X	X		X	X				X	X			X				X	X	X		X			X
13	83.0	81.3	10.5	0.98739	X	X	X		X				X	X			X				X	X	X	X		X		X
14	83.1	81.4	11.0	0.98552	X	X	X		X				X	X			X				X	X	X	X		X	X	X
15	83.4	81.5	11.4	0.98316	X	X					X	X	X	X			X	X	X	X	X	X		X		X	X	X

When considering all of the decision criteria that was defined in previous best subsets regressions, the highlighted model from Table 4 indicates the nine variable model warrants further consideration. A complete regression was run based on the variables in the given highlighted model. The result is summarized in Figure 11 below.

The Regression Equation is:

$$C28 = 6.29 + 0.578 C1 + 0.243 C4 + 0.666 C6 - 0.0498 C9 + 0.000145 C10 - 0.000139 C14 + 0.0184 C18 - 1.49 C20 - 0.663 C24$$

Predictor	Coef	Stdev	t-ratio	p	VIF
Constant	6.293	0.5464	11.52	0	
RJR (1) vs.Lundahl (0)	0.5779	0.1658	3.49	0.001	1
(Ln (Width))^2	0.2428	0.08	3.04	0.003	4.1
Ln (length)	0.6664	0.2118	3.15	0.002	2.5
Order Qty	-0.04979	0.004	-12.45	0	7.6
(order qty)^2	0.00014	0.00001356	10.66	0	8
(smallest)^2	-0.00014	0.00003677	-3.78	0	1.2
Sq Inch	0.01841	0.003914	4.7	0	5.5
Two Layers	-1.4903	0.2478	-6.01	0	1.3
Copper 2	-0.6634	0.1941	-3.42	0.001	1.1

s = 0.9985

R-sq = 82.1%

R-sq(adj) = 80.9%

Analysis of Variance Tables:

SOURCE	DF	SS	MS	F	p
Regression	9	638.505	70.945	71.16	0
Error	140	139.575	0.997		
Total	149	778.08			

SOURCE	DF	SEQ SS
RJR (1) vs.Lundahl (0)	1	18.562
(Ln (Width))^2	1	260.902
Ln (length)	1	136.102
Order Qty	1	23.094
(order qty)^2	1	106.086
(smallest)^2	1	30.619
Sq Inch	1	15.445
Two Layers	1	36.048
Copper 2	1	11.648

Note: The Column indexes correspond to the variable names of the Predictor column in the first table.

Figure 11: Regression Summary of the 9 Variable Model

Although the variance inflation factors given on the previous table do not look entirely bad, the given model may still contain variables that are correlated. For example, refer to Length/Width and Square Inches. With this possibility still a concern, another correlation matrix was generated that includes all of the variables in consideration at this point to try to identify other pair-wise instances of collinearity. This matrix is given below in Table 5.

**Table 5 Correlation Matrix with All Variables Considered**

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	C16	C17	C18	C19	C20	C21	C22	C23	C24	
C2	-.08																								
C3	.12	.36																							
C4	.12	.36	.99																						
C5	-.04	.40	.19	.23																					
C6	-.04	.39	.22	.26	.95																				
C7	.01	.03	.47	.49	.43	.42																			
C8	.04	.05	.54	.57	.48	.52	.86																		
C9	-.05	-.12	-.13	-.16	-.22	-.31	-.07	-.20																	
C10	-.06	-.09	-.16	-.19	-.21	-.33	-.09	-.26	.93																
C11	.04	.34	.45	.47	.25	.33	.17	.31	-.02	-.12															
C12	.04	.25	.44	.47	.27	.37	.22	.43	-.12	-.25	.94														
C13	-.07	.23	-.26	-.26	-.04	-.05	-.33	-.58	.12	.21	-.34	-.56													
C14	-.04	.22	-.23	-.21	-.04	-.03	-.23	-.46	.05	.13	-.31	-.52	.95												
C15	-.03	.26	.21	.24	.21	.29	-.04	.03	.14	.02	.66	.59	-.01	-.07											
C16	-.10	-.07	-.14	-.15	.34	.35	.43	.40	.11	.03	.17	.20	-.20	-.14	.11										
C17	.04	-.01	.05	.03	.10	.05	.04	.10	-.03	.02	.05	.10	-.11	-.09	-.07	.05									
C18	.05	.50	.80	.80	.63	.59	.56	.56	-.14	-.14	.38	.38	-.17	-.15	.21	.04	-.01								
C19	.05	.40	.76	.77	.64	.72	.56	.70	-.30	-.34	.50	.56	-.34	-.31	.29	.16	.05	.84							
C20	.08	.11	.22	.25	-.27	-.27	-.24	-.23	-.06	.00	-.16	-.19	.18	.12	-.04	-.80	-.13	.07	-.04						
C21	-.03	-.17	-.11	-.12	.21	.19	.31	.26	.00	-.02	.08	.11	-.19	-.11	-.02	.56	.28	.00	.05	-.70					
C22	.01	.66	.77	.76	.61	.56	.45	.48	-.12	-.11	.41	.39	-.11	-.12	.25	.01	-.02	.97	.79	.11	-.05				
C23	.08	-.19	.28	.27	-.16	-.19	.19	.27	-.04	-.05	-.11	-.03	-.29	-.16	-.19	-.07	.08	.09	.07	.15	-.06	.02			
C24	-.05	.21	-.19	-.21	.08	.10	-.11	-.20	.04	.07	.03	-.04	.26	.16	.08	.04	-.11	-.03	-.02	-.08	-.01	.04	-.85		
C25	.13	.27	.59	.59	.51	.48	.51	.50	-.31	-.19	.21	.25	-.23	-.18	.04	.07	.05	.76	.62	-.09	.15	.68	.16	-.15	

Note: C1 = RJR vs. Lundahl  
 C2 = Thickness  
 C3 = Width  
 C4 = (Ln(Width))^2  
 C5 = Length  
 C6 = Ln(length)  
 C7 = Holes  
 C8 = (Ln(Holes))^2  
 C9 = Order Qty  
 C10 = (order qty)^2  
 C11 = Drills  
 C12 = Ln(drills)  
 C13 = Smallest Hole  
 C14 = (smallest)^2  
 C15 = Largest Hole  
 C16 = Test 0or1  
 C17 = Tab Route 0or1  
 C18 = Sq Inch  
 C19 = Ln(sq inches)  
 C20 = Two Layers  
 C21 = Three Layers  
 C22 = Volume  
 C23 = Copper 1  
 C24 = Copper 2  
 C25 = Cost

The problem of multicollinearity and its effect on the model development was discussed earlier in this chapter. As stated earlier, its presence does inhibit the ability to obtain an accurate assessment of the true individual regression coefficients.

Multicollinearity also limits the ability to evaluate the change in the expected value of the dependent variable when a corresponding independent variable is increased or decreased by one unit while all of the other independent variables are held constant.

In order to avoid this correlation problem, two more best subsets regression models were evaluated. The first best subsets regression suppresses the availability of the variables Length & Width and Ln(sq inches) to see what models will be generated. The variables 'Width', '(Ln (Width))^2', 'Length', 'Ln (length)' are suppressed in the second best subsets regression to see what resulting model will be generated. Both of the results follow in Tables 6 and 7.

**Table 6 Best Subsets Regression Omitting C18, and C19**

Vars	R-sq	Adj. R-sq	C-p	s	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	20	21	22	23	24	
4	70.9	70.1	71.4	1.2498			X	X					X	X													
5	75.5	74.7	39.7	1.1498			X	X					X	X									X				
6	77.9	77.0	24.5	1.0964			X	X					X	X			X						X				
7	79.2	78.2	17.0	1.0672	X		X	X					X	X			X						X				
8	80.0	78.9	13.1	1.05	X		X	X					X	X			X						X				X
9	81.0	79.8	7.7	1.0266	X		X			X			X	X			X					X		X		X	X
10	81.5	80.1	6.5	1.0181	X		X			X			X	X			X					X	X		X		X
11	81.7	80.3	6.8	1.0155	X		X			X			X	X			X					X	X		X	X	X
12	81.9	80.3	7.5	1.014	X		X			X	X		X	X			X					X	X	X	X		X

Note: Omitting C18 "Sq Inch", and C19 "Ln (sq inches)"

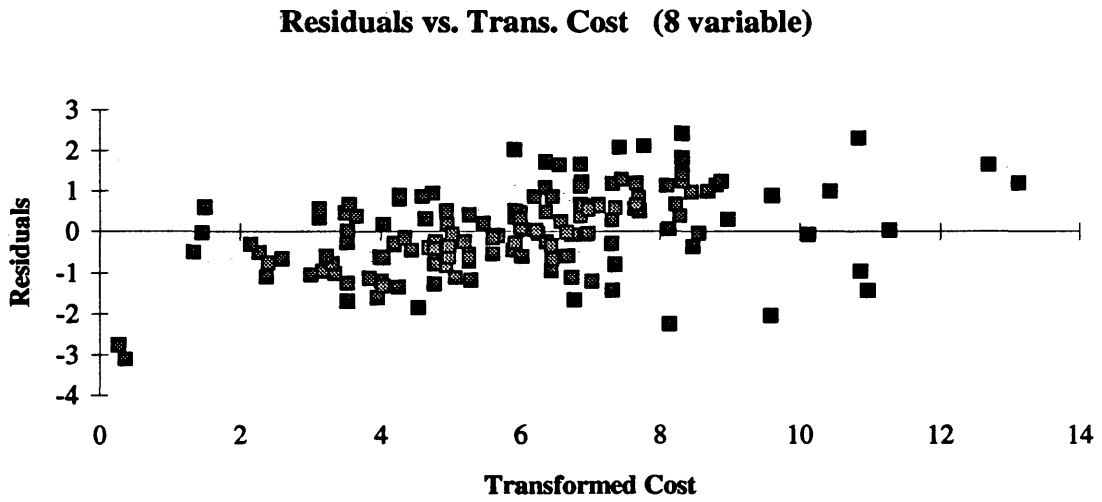
**Table 7 Best Subsets Regression Omitting C3, C4, C5, and C6**

Vars	R-sq	Adj. R-sq	C-p	s	1	2	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
4	73.8	73.1	65.5	1.1848					X	X								X		X				
5	77.1	76.3	42.2	1.1133					X	X			X					X		X				
6	78.7	77.8	31.3	1.0763					X	X			X					X		X				X
7	80.3	79.3	20.9	1.0393	X				X	X				X				X		X				X
8	81.1	80.0	16.6	1.0217	X				X	X			X		X			X		X				X
9	81.6	80.4	14.6	1.0115	X			X	X	X				X	X			X		X				X
10	82.0	80.7	13.1	1.0027	X			X	X	X				X	X	X		X		X				X
11	82.3	80.9	12.9	0.99834	X			X	X	X				X	X	X		X		X			X	X
12	82.6	81.1	12.5	0.99331	X	X		X	X	X				X	X			X		X		X	X	X

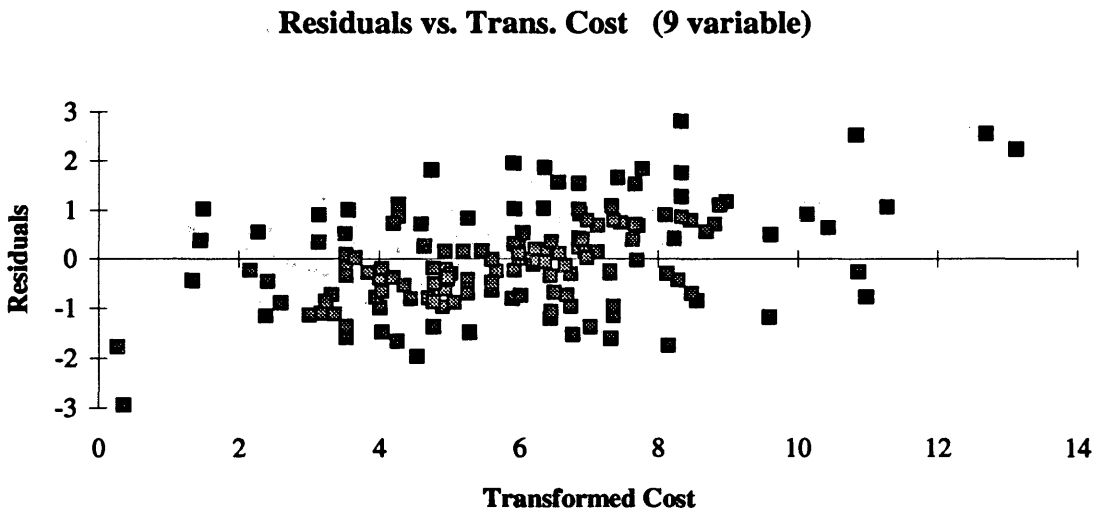
Note: Omitting C3 "Width", C4 "(Ln (Width))^2", C5 "Length", and C6 "Ln (length)"

Table 6 indicates the 9 variable model, but Table 7 does not give a definitive choice when considering the R<sup>2</sup> and the C<sub>p</sub> score. However, by re-tracing steps and looking at models generated in previous best subset regressions for any that do not contain both the Length/Width and Square Inches variables revealed the same 8 variable model defined in Tables 4 and 7. Although this model does not have an ideal C<sub>p</sub> score, the R<sup>2</sup> appeared to be very comparable to the 9 variable model highlighted in Table 6, and thus warranted a closer look.

A direct comparison of the 9 and 8 variable models using the residual plots from each model does not reveal any clear cut differences between the models. Refer to Figures 12 and 13 for examples as well as the data diskette for all of the plots associated with each model.



**Figure 12: Residual Plot of 8 Variable Model (Transformed)**



**Figure 13: Residual Plot of the 9 Variable Model (Transformed)**

The fitted values generated by both regressions were utilized in conjunction with the reverse transformation in an effort to evaluate each model's prediction error in terms of the actual fitted cost associated with each model.

Plots of the prediction errors and some descriptive statistics on the prediction error columns point towards the 8 variable model as the better of the two. This result also follows a rule of thumb when confronted with the choice of two comparable models. Simply stated, the rule of thumb would be to choose the simpler model of the two. Refer to the data diskette for the prediction error plots, reverse transformed data, and descriptive statistics for each model. The complete regression of the 8 variable model is as follows in Figure 14.

**The Regression Equation is:**

$$\text{Transcost} = 7.87 + 0.583 C1 - 0.0521 C9 + 0.000144 C10 - 0.0222 C13 + 0.00266 C15 + 0.0300 C18 - 1.50 C20 - 0.715 C24$$

<i>Predictor</i>	<i>Coef</i>	<i>Stdev</i>	<i>t-ratio</i>	<i>p</i>	<i>VIF</i>
Constant	7.8705	0.3887	20.25	0	
RJR (1) vs.Lundahl (0)	0.5831	0.169	3.45	0.001	1
Order Qty	-0.05212	0.004313	-12.08	0	8.5
(order qty)^2	0.00014	0.0000142	10.16	0	8.4
Smallest	-0.02221	0.005916	-3.75	0	1.3
Largest	0.00266	0.001056	2.52	0.013	1.2
Sq Inch	0.03004	0.001818	16.52	0	1.1
Two Layers	-1.5037	0.2303	-6.53	0	1.1
Copper 2	-0.7149	0.1966	-3.64	0	1.1

s = 1.022

R-sq = 81.1%

R-sq(adj) = 80.0%

**Analysis of Variance Tables:**

<i>SOURCE</i>	<i>DF</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>p</i>
Regression	8	630.903	78.863	75.55	0
Error	141	147.177	1.044		
Total	149	778.08			

<i>SOURCE</i>	<i>DF</i>	<i>SEQ SS</i>
RJR (1) vs.Lundahl (0)	1	18.562
Order Qty	1	107.502
(order qty)^2	1	62.192
Smallest	1	76.565
Largest	1	47.005
Sq Inch	1	266.631
Two Layers	1	38.653
Copper 2	1	13.794

Note: The Column indexes correspond to the variable names of the Predictor column in the first table.

**Figure 14: Regression Summary of the Final 8 Variable Model**

## 2.6 Resulting Model

Using the results generated by the regression summarized in Figure 14, the model can be substituted back into the form that is defined in equation 2.7. Recall equation 2.7.

$$\left( \frac{y_i^{\hat{\lambda}} - 1}{\hat{\lambda}} \right) = \beta_0 + \beta_1 X_{i,1} + \dots + \beta_{p-1} X_{i,p-1} + \epsilon_i \quad (2.7)$$

Using this result, the reverse transformation of the model can be conducted so that the result takes the following form.

$$\hat{Y} = \exp \left( \frac{\ln \left\{ \hat{\lambda} (\beta_0 + \beta_1 X_{i,1} + \dots + \beta_{p-1} X_{i,p-1} + \epsilon_i) \right\} + 1 \right)}{\hat{\lambda}} \quad (2.8)$$

By substituting the coefficients and column numbers, the model takes the resulting form.

$$\hat{Y} = \exp \left( \frac{\ln \left\{ \hat{\lambda} (7.87 + 0.583 C1 - .0521 C9 + .000144 C10 - .0222 C13 + .00266 C15 + .03 C18 - 1.5 C20 - .715 C24) \right\} + 1 \right)}{\hat{\lambda}} \quad (2.9)$$

Now by substituting in for  $\hat{\lambda} = 0.34746$  the final equation is defined below.

$$\hat{Y} = \exp \left[ \frac{\ln(3.7347 + 2.026 C1 - .01811 C9 + .0002015 C10 - .007718 C13 + .0009246 C15 + .01044 C18 - .5225 C20 - .2484 C24)}{0.34746} \right]$$

C1 = RJR (1) vs.Lundahl (0)

C9 = Order Qty

C10 = (order qty)^2

C13 = Smallest

C15 = Largest

C18 = Sq Inch

C20 = Two Layers

C24 = Copper 2

(2.10)

As a result, the 8 variable model appears to be a good choice. The true test of this model's utility would be to continually compare the actual price quotes for printed circuit boards against what the model generates to see how well it matches the actual quote. Further study could also be conducted in the area of determining the change of regression variance "s" as a result of the transformation process against models that were not transformed.

### Chapter 3

## CONCLUSION

In conclusion, this study has been successfully completed and achieved all of the goals set forth at the inception of the project. It has been shown that the Atkinson score statistic transformation can be a useful tool when analyzing data sets with regression analysis. Furthermore, it has been shown that using the  $C_p$  Criterion and implementing Zero-One indicator variables into the data set prove to be two very useful tools to consider when conducting regression analysis on large data sets. Recall that the  $C_p$  Criterion was used during the best subsets regression as a preliminary decision criterion for choosing the resulting model in conjunction with the standard criteria of  $R^2$  and estimated variance. Similarly, recall that zero-one indicator variables were used to consolidate the data into one large set by quantifying some of the qualitative aspects of the data. Finally, there is absolutely no substitute for seeing the data graphically to determine what kind of relationships exist between variables. All of these tools or steps in the analysis have resulted in a simple but effective model that is currently in use by Advanced Energy. This was the original reason for conducting this study.

Though it is not likely that this model will have a direct economic impact on Advanced Energy's profit and loss statement, it is safe to say that the model does give the Research

and Development Group another tool to consider when they embark on new design avenues. It provides the engineers with instant feedback on the potential costs that their designs will have on a final product. By having the power to maintain a perspective of where the design process is going in terms of cost, can only help the engineer stay within whatever constraints were imposed on the project when it was conceived. As stated earlier, the ability to maintain the level of leading edge technology while keeping costs of production in check is how Advanced Energy has continued to grow and lead the market in industrial power supply applications.

As Advanced Energy continues to solicit price quotes from its vendors, an area for further study concerning this model would be model validation and adjustment. In conjunction with this, further work could be conducted to determine what effect the transformation has on the regression variance. This would be particularly useful to see how the variance reacts when reverse transforming back after the model has been chosen. Furthermore, work could be conducted using different transformation techniques other than the parametric family of power transformations. As noted earlier, the given transformation technique is not a guaranteed method and there may be others available that produce more desirable results. Throughout this study, I have kept in mind that the model can in no way account for the price fluctuations associated with exterior market, economy, and political factors. It would be interesting to continually check the model and adjust it according to how these external factors change over time.

In closing, this work has been both challenging and rewarding. The challenge has come from starting a project and carrying it through to completion. The reward comes in seeing the concepts taught in applied statistics utilized within the realm of problem solving outside the academic world. This work has afforded me the opportunity to see, on a first hand basis, how statistics and operations research are important tools that benefit the business and manufacturing sectors when applied responsibly.

## REFERENCES CITED

- Atkinson, A.C. 1985. Plots, Transformations, and Regression: An Introduction to Graphical Methods of Diagnostic Regression Analysis - (Oxford Statistical Science Series). New York: Oxford University Press.
- Atkinson, A.C. and Lawrance, A.J. 1989. A Comparison of Asymptotically Equivalent Test Statistics for Regression Transformation. Biometrika 76 (2): pp.223-229.
- Box, G.E.P. and Cox, D.R. 1964 An Analysis of Transformations (with discussion). Journal of the Royal Statistical Society, Series B 26: pp.211-246.
- Carroll, R.J. and Rupert, D. 1988. Transformation and Weighting in Regression - (Monographs on Statistics and Applied Probability). New York: Chapman and Hall, Ltd.
- Cook, R. Dennis and Weisberg, Sanford 1982. Residuals and Influence in Regression - (Monographs on Statistics and Applied Probability). New York: Chapman and Hall, Ltd.
- Draper, N.R. and Smith, H. 1981. Applied Regression Analysis - (Wiley Series in Probability and Mathematical Statistics). New York: John Wiley and Sons, Inc.
- Lawrance, A.J. 1987. The Score Statistic for Regression Transformation. Biometrika 74 (2): pp.275-279.
- Miller, A.J. 1990. Subset Selection in Regression - (Monographs on Statistics and Applied Probability). New York: Chapman and Hall, Ltd.
- Neter, John, Wasserman, William, and Kutner, Michael H. 1990. Applied Linear Regression. Third Edition. Homewood, Ill.: Richard D. Irwin, Inc.
- Tukey, J.D. 1957. On the Comparative Anatomy of Transformations. Ann. Math. Statist. 28: pp.602-632.

Weisberg, Sanford 1980. Applied Linear Regression - (Wiley Series in Probability and Mathematical Statistics). New York: John Wiley and Sons, Inc.

**SELECTED BIBLIOGRAPHY**

- Aitchison, J. 1986. The Statistical Analysis of Compositional Data - (Monographs on Statistics and Applied Probability). New York: Chapman and Hall, Ltd.
- Atkinson, A.C. 1982. Regression Diagnostics, Transformations and Constructed Variables (with discussion). Journal of the Royal Statistical Society, Series B 44: pp.1-36.
- Bartlett, M.S. 1947. The Use of Transformations. Biometrics 3: pp.39-52.
- Box, G.E.P. and Hill, W.J. 1974. Correcting Inhomogeneity of Variance with Power Transformation Weighting. Technometrics 16: pp.385-389.
- Draper, N.R. and Cox, D.R. 1969. On Distributions and their Transformations to Normality. Journal of the Royal Statistical Society, Series B 31: 472-476
- Pankratz, Alan 1991. Forecasting with Dynamic Regression Models - (Wiley Series in Probability and Mathematical Statistics). New York: John Wiley and Sons, Inc.