AN EXPLORATION OF THE ROBUSTNESS OF TRADITIONAL REGRESSION ANALYSIS VERSUS ANALYSIS USING BACKPROPAGATION NETWORKS

by

Ina Samanta Markham

Dissertation submitted to the Faculty of the

Virginia Polytechnic Institute and State University

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in

Management Science

APPROVED:

Terry R. Rakes

Terry R. Rakes, Chairman

Ernest C. Houck

Ernest C. Houck

Laurence J. Moore

Loren Paul Rees

Loren P. Rees

Robert T. Sumichrast

April 6, 1992

Blacksburg, Virginia

AN EXPLORATION OF THE ROBUSTNESS OF TRADITIONAL REGRESSION ANALYSIS VERSUS ANALYSIS USING BACKPROPAGATION NETWORKS

by

Ina Samanta Markham

Terry R. Rakes, Chairman

Management Science

(ABSTRACT)

Research linking neural networks and statistics has been at two ends of a spectrum: either highly theoretical or application specific. This research attempts to bridge the gap on the spectrum by exploring the robustness of regression analysis and backpropagation networks in conducting data analysis. Robustness is viewed as the degree to which a technique is insensitive to abnormalities in data sets, such as violations of assumptions.

The central focus of regression analysis is the establishment of an equation that describes the relationship between the variables in a data set. This relationship is used primarily for the prediction of one variable based on the known values of the other variables. Certain assumptions have to be made regarding the data in order to obtain a tractable solution and the failure of one or more of these assumptions results in poor prediction.

The assumptions underlying linear regression that are used to characterize data sets in this research are characterized by: (a) sample size and error variance, (b) outliers, skewness, and kurtosis, (c) multicollinearity, and (d) nonlinearity and underspecification.

By using this characterization, the robustness of each technique is studied under what is, in effect, the relaxation of assumptions one at a time. The comparison between regression and backpropagation is made using the root mean square difference between the predicted output from each technique and the actual output.

Dedication

I would like to dedicate this research to my parents, Dr. Ajit M. Samanta and Mrs. Anjali Samanta, for teaching me the value of education.

Acknowledgements

I would like to thank my chairman, Dr. Terry R. Rakes, for the encouragement and support he has provided throughout the dissertation stage of my doctoral program. His enthusiasm about my topic was very rewarding especially in the early stages of my disseration.

I would like to express my appreciation to Dr. Ernest C. Houck, Dr. Laurence J. Moore, Dr. Loren P. Rees, and Dr. Robert T. Sumichrast for graciously serving as members on my dissertation committee.

I would like to thank Dr. Bernard W. Taylor, III, Head of the Department of Management Science, for providing me with excellent research facilities and monetary support for my doctoral work at Virginia Tech.

I would like to thank Sylvia Seavey, Tracy McCoy, and Teena Long for their help with GML and script commands during the writing phase of my dissertation.

I would like to express my heartfelt gratitude to my parents and my siblings for their love and support of me throughout my scholastic endeavors. Their pride in my work has been a source of encouragement to me. My thanks also go to my stepdaughter, Anna Markham, for her acceptance of all the times I was too busy with my research.

I am eternally grateful to my husband, Dr. Steven E. Markham, for his love and support and also for all the things he has done to make the dissertation stage of my doctoral work a pleasurable experience.

Table of Contents

Chapter 1: Introduction 1

1.1 Purpose and Justification 2

1.1.1 Purpose 2

1.1.2 Justification 3

1.2 Scope and Limitations 5

1.2.1 Scope 5

1.2.2 Limitations 6

1.3 Plan of Presentation 7

Chapter 2: Literature Review **9**

2.1 Neural Networks 9

2.2 Neural Networks and Statistics 12

Chapter 3: Methodology **18**

3.1 The Basic Regression Model 18

3.2 Proposed Neural Network Architecture 20

3.3 Characteristics of Data Sets 21

3.4 Data Generation 23

3.5 Design 23

3.5.1 Sample Size and Variance 24

3.5.2 Error Distribution 24

3.5.2.1 Effect of Outliers in Error Distribution 25

3.5.2.2. Effect of Skewness and Kurtosis 25

3.5.3 Multicollinearity 26

3.5.4. Nonlinearity and Underspecification 27

3.5.4.1 Nonlinearity 27

3.5.4.2 Underspecification 28

3.6 Criterion for Comparison 29

Chapter 4: Sample Size and Variance 30

4.1 Generation of Data 30

4.1.1 Generation of Samples 31

4.2 Regression Analysis 32

4.3 Neural Network Analysis 34

4.3.1 Determination of Network 36

4.3.1.1 Determination of Learning Count for Sample Size 20 37

4.3.1.2 Identification of Best Network 38

4.3.1.3 Determination of Best Learning Count for Other Sample Sizes 38

4.3.2 The Neural Network Experiment 41

4.4 Comparison of Neural Networks and Regression 44

4.4.1 Analysis of ANOVA Assumptions 44

4.4.1.1 Test for Normality 44

4.4.1.2 Test for Homoscedasticity 46

4.4.2 Nonparametric Two-way Analysis 48

4.4.3 Graphical Analysis 49

4.4.4 Wilcoxon Signed-Ranks Test 55

4.5 Conclusions 58

Chapter 5: Outliers, Skewness, and Kurtosis 63

5.1 Error Term with Outliers 64

5.1.1 Generation of Samples 64

5.1.2 Regression Analysis 65

5.1.3 Neural Network Analysis 67

5.1.4 Comparison of Neural Networks and Regression 68

5.2 Skewed Error Distributions 72

5.2.1 Generation of Samples 75

5.2.2 Regression Analysis 75

5.2.3 Neural Network Analysis 77

5.2.4 Comparison of Neural Networks and Regression 79

5.3 Error Distributions with Kurtosis 81

5.3.1 Generation of Samples 84

5.3.2 Regression Analysis 84

5.3.3 Neural Network Analysis 86

5.3.4 Comparison of Neural Networks and Regression 88

5.4 Conclusions 90

Chapter 6: Multicollinearity 92

6.1 Generation of Data 93

6.2 Generation of Samples 94

6.3 Regression Analysis 94

6.4 Neural Network Analysis 99

6.5 Comparison of Neural Networks and Regression 102

6.6 Conclusions 109

Chapter 7: Nonlinearity and Underspecification **110**

7.1 Nonlinearity 111

7.1.1 Model with Quadratic Term 111

7.1.1.1 Regression Analysis 112

7.1.1.2 Neural Network Analysis 116

7.1.1.3 Comparison of Neural Networks and Regression 119

7.1.2 Exponential Model 123

7.1.2.1 Regression Analysis 123

7.1.2.2 Neural Network Analysis 125

7.1.2.3 Comparison of Neural Networks and Regression 127

7.2 Underspecification 128

7.2.1 Regression Analysis 129

7.2.2 Neural Network Analysis 133

7.2.3 Comparison of Neural Networks and Regression 137

7.3 Conclusions 148

Chapter 8: Summary and Conclusions **149**

8.1 Summary of Results 149

8.1.1 Sample Size and Variance 149

8.1.2 Outliers, Skewness, and Kurtosis 150

8.1.3 Multicollinearity 151

8.1.4 Nonlinearity and Underspecification 151

8.2 Conclusions 152

Bibliography **154**

Appendix A. Neural Network Basics **159**

Appendix B. SAS Program to Generate Normally Distributed Populations 164

Appendix C. SAS Program to Estimate Regression Parameters 165

Appendix D. SAS Program to Predict Using Regression 166

Appendix E. BASIC Program to Calculate Root Mean Square 167

Appendix F. Plot of Root Mean Square versus Learning Count 168

Appendix G. SAS Program to Calculate the RMS Difference 169

Appendix H. SAS Program to Test for Normality 170

Appendix I. MINITAB Commands to Test Homoscedasticity 171

Appendix J. NPSP Program for Two-way Layout with Interactions 172

Appendix K. SAS Program for Nonparametric One-way Layout 174

Appendix L. SAS Program to Generate Populations with Outliers in Error Distribution . . . 175

Appendix M. Normal Probability Plots for Moderate Outlier Case 176

Appendix N. SAS Program to Select Beta Distribution Parameters for Skewness and Kurtosis 177

Appendix O. SAS Program to Generate Populations with Skewed Error Distributions 178

Appendix P. SAS Program to Generate Populations in the Multicollinearity Cases 179

Appendix Q. SAS Program to Generate Populations with Quadratic Term 181

Appendix R. SAS Program to Generate Population for Exponential Model 182

Appendix S. SAS NLIN Procedure 183

Vita 184

List of Illustrations

Figure 1. Plot of Average RMS for Regression - Variance Curves 51

Figure 2. Plot of Average RMS for Backpropagation - Variance Curves 52

Figure 3. Plot of Average RMS Differences - Variance Curves 53

Figure 4. Comparison of the two techniques when there is positive correlation 103

Figure 5. Comparison of the two techniques when there is negative correlation 104

Figure 6. Comparison of the two models using Backpropagation - Positive Quadratic Term 138

Figure 7. Comparison of the two models using Regression - Positive Quadratic Term ... 139

Figure 8. Comparison of the two models using Backpropagation - Negative Quadratic Term 140

Figure 9. Comparison of the two models using Regression - Negative Quadratic Term ... 141

Figure 10. Comparison of the two techniques with the Underspecified Model - Positive Quadratic Term 142

Figure 11. Comparison of the two techniques with the Underspecified Model - Negative Quadratic Term 143

Figure 12. Mathematical model of a processing element in a neural network 160

Figure 13. A three-layer backpropagation neural network 162

List of Tables

Table 1. Experimental Design for each Sample Size and Variance Combination 33

Table 2. Root Mean Squares from Regression - Sample Size and Variance 35

Table 3. Summary of Networks and Best Learning Counts 39

Table 4. Summary of Networks and Average Root Mean Squares 40

Table 5. Summary of Sample Sizes and Best Learning Counts 42

Table 6. Root Mean Squares from Backpropagation Network - Sample Size and Variance . 43

Table 7. Root Mean Differences between Backpropagation and Regression - Sample Size and Variance 45

Table 8. Qualitative Summary of Results 59

Table 9. Selected R-square values from Regression 61

Table 10. Root Mean Squares from Regression - Outliers 66

Table 11. Root Mean Squares from Backpropagation Networks - Outliers 69

Table 12. Root Mean Square Differences - Outliers 71

Table 13. Beta Distribution Parameters and Skewness 74

Table 14. Root Mean Squares from Regression - Skewness 76

Table 15. Root Mean Squares from Backpropagation Networks - Skewness 78

Table 16. Root Mean Square Differences - Skewness 80

Table 17. Beta Distribution Parameters and Kurtosis 83

Table 18. Root Mean Squares from Regression - Kurtosis 85

Table 19. Root Mean Squares from Backpropagation Networks - Kurtosis 87

Table 20. Root Mean Square Differences - Kurtosis 89

Table 21. Root Mean Squares from Regression - Multicollinearity 97

Table 22. Summary of Tests and Wilcoxon T Values 100

Table 23. Root Mean Squares from Backpropagation Networks - Multicollinearity 101

Table 24. Root Mean Square Differences - Multicollinearity 106

Table 25. Root Mean Squares from Regression - Positive Quadratic Term 114

Table 26. Root Mean Squares from Regression - Negative Quadratic Term 115

Table 27. Root Mean Squares from Backpropagation - Positive Quadratic Term 117

Table 28. Root Mean Squares from Backpropagation - Negative Quadratic Term 118

Table 29. Root Mean Square Differences - Positive Quadratic Term 120

Table 30. Root Mean Square Differences - Negative Quadratic Term 121

Table 31. Root Mean Squares for the Exponential Model 126

Table 32. Root Mean Squares from Regression - Without Positive Quadratic Term 130

Table 33. Root Mean Squares from Regression - Without Negative Quadratic Term 131

Table 34. Root Mean Squares from Backpropagation - Without Positive Quadratic Term . 134

Table 35. Root Mean Squares from Backpropagation - Without Negative Quadratic Term 135

Table 36. Root Mean Square Differences - Without Positive Quadratic Term 145

Table 37. Root Mean Square Differences - Without Negative Quadratic Term 146

Chapter 1: Introduction

The human characteristic of self-referential inquiry, i.e. the mind thinking of itself, has given rise to the field of artificial intelligence (AI). Scientists have modeled the thinking, remembering, and problem-solving capabilities of the human brain with considerable success, resulting in numerous AI breakthroughs, including neural computing. In traditional expert systems, knowledge is explicit in the form of rules. Neural computing is different in that neural networks generate their own rules by learning from examples. Neural networks provide an approach to computing which is closer to human perceptions. Some areas of applications where neural networks have been successful are natural language processing (Sejnowski and Rosenberg 1987), character recognition (Burr 1987), image compression (Cottrell, et al. 1987), and pattern recognition (Alkon, et al. 1990).

Statistics can be viewed as the art and science of inducing what a data set may be telling us. This is typically achieved by modeling or finding the mathematical relationship between several inputs and one or more outputs. Once such a relationship is found, based on the sample data, it is used to interpolate between or extrapolate beyond the sample data. Neural networks also lend themselves well to such applications.

This research investigates the robustness of regression analysis versus analyses using backpropagation networks using data sets of varying characteristics. Robustness is viewed as the degree to which a technique is insensitive to abnormalities in data sets, such as violations of certain basic assumptions. The characteristics of data sets which render them better suited to analysis using backpropagation networks or traditional regression analysis are determined.

Statistics should merge with neural networks in such a way that analysis of data is divided according to interfaces between substantive issues rather than interfaces between methodological schools of thought. Instead of applying regression analysis or neural networks to data across the board just because traditionally one has conducted analysis using one method or the other, it should first be determined which method is best suited for each type of data set based on the data set characteristics, and then the appropriate method should be applied.

1.1 Purpose and Justification

Practicing statisticians employ formal strategy in their analysis of data. To achieve an effective data analyzing strategy, a formal description of the choices, actions, and alternatives involving neural computing needs to be developed.

1.1.1 Purpose

The evolution in AI technology, especially in neural networks, has opened a new field of research integrating AI with statistics. Although neural networks have successfully contributed to a wide range of disciplines, such as linguistics, biosciences, psychology, and engineering, statistics is one of the few disciplines that has been shown to make a contribution back to neural networks in the sense that the concept of least squares has been used to explain the backpropagation network

(White, 1989a). Research linking neural networks and statistics has been at two ends of a spectrum: either highly theoretical or application specific.

The purpose of this research is to bridge the gap on the spectrum. Through force of habit and because there are few well-defined alternatives, the average researcher, regardless of discipline, uses traditional statistical techniques such as regression in analyzing data sets. This research explores the use of backpropagation network models for conducting data analyses. The questions addressed are:

1. Can backpropagation networks be successfully used for analyzing data in a manner similar to traditional regression analysis?

2. What characteristics of data sets make the application of backpropagation networks more appropriate than traditional regression analyses and vice versa?

3. Why do these characteristics lend themselves better to a particular method?

The results and recommendations of this research will aid researchers in selecting the strategy to follow when faced with different data sets. The choice between backpropagation networks and regression analysis should be decided on a case-by-case basis, based on the predictive power of each technique relative to the data set characteristics. The rationale for using predictive power as the basis for comparison is that, according to Myers (1986), the primary use of regression is the prediction of one variable based on the known value of one or more other variables.

1.1.2 Justification

The major motivation behind this endeavor is exploration. At the current stage of neural network evolution, there are some areas in which neural networks do better than any other technology. As the number of neural network researchers grow, it is hoped that their individual incremental contribution will add up to a revolution, providing better understanding of the capabilities of neural networks.

This research is important for a number of reasons:

1. It explores neural networks as an alternative to regression analysis in a way that can serve to prove or dispell claims often made in the research literature.

2. It develops a classification of types of data sets and aids the researcher in choice of technique and how to apply each technique correctly.

3. It opens up avenues of research involving neural networks and statistics.

Neural networks have proved capable of producing reasonable results in situations where input is noisy or incomplete. These factors cause difficulties and biases in results using traditional statistical techniques. Prior to the application of any statistical tool on data sets that violate the assumptions of regression, data cleaning has to be done. In fact, the issues in fitting a straight line to a bivariate data set, as in simple linear regression, are almost entirely issues of data cleaning. The model that is fitted must be tested to see if the assumptions are justified. For example, in regression analysis, one may first fit a straight line and then explore any departures from that line, such as curvatures or heteroscedasticity.

The use of neural networks does away with the essentially iterative circular process in statistical analysis. When trained with a subset of the data, a neural network "learns" what the data set may be telling us, and creates output accordingly.

1.2 Scope and Limitations

Since it would be impossible to consider all the different neural network models and their learning techniques, a subset of backpropagation networks is chosen in such a way that generalization is possible. Similarly, the specific statistical technique explored in the context of this research is the commonly-used regression analysis.

By considering only a fundamental treatment of regression and backpropagation, this research does not attempt to bias the reader toward either technique. In those cases where there is a clear consensus as to the regression method that should be used, this research has used the recommended method. For example, the Gauss-Newton procedure is used with the Marquadt option in the experiment on nonlinear regression. Following the advice given by Professor Raymond Myers, an expert on regression, ordinary least squares method is used with multicollinearity. The presence of other remedial methods in regression to counter data complications is recognized but the level of regression analysis used in this study is the generally accepted approach. Similarly, with the backpropagation network, the quasi-naive approach is taken, without testing a variety of different squashing functions or trying to experimently determine the cost function that would provide the absolute best results.

1.2.1 Scope

In this research, data sets are characterized by variability of factors such as sample size and error variance, distribution of error, multicollinearity, nonlinearity and model underspecification. Different backpropagation network models are developed for each characteristic analyzed but no attempt is made to integrate these into any kind of master system. The networks are described in detail so that the user can easily employ any one that is appropriate.

Many AI researchers, including Marvin Minsky and Seymour Papert, and practitioners, such as Larry Jacel of AT&T, have criticized neural networks as limited to solving ″toy problems″ as opposed to having a more general pattern-recognition ability. Nevertheless, Minsky and Papert counter that ″toy problems″ may be less of a limitation than a prototype. They propose that the scale of the ″toy problem″ may actually be the scale at which human intelligence operates. In the epilogue to Perceptrons (1988), Minsky and Papaert argue that the human brain is built up of many small neural networks. Each network solves a few interrelated ″toy problems″. They contend that the proper focus of research is not the search for universal principles but the search for what kinds of processing best serve which kinds of problems. Herbert Simon, a Nobel laureate, agrees with Minsky and Papert that since the brain is a hierarchy of systems, the best machine should be too.

The scope of this research is to identify part of such a hierarchy of systems for regression data analysis, leaving the building of the machine to future researchers.

1.2.2 Limitations

As stated earlier, one limitation of this research is the lack of any kind of master system to conduct the entire gambit in data analysis. Also, because there are so many statistical techniques, the only statistical technique considered is regression analysis. Emulating Myers′ (1986) description of regression analysis as a collection of techniques in data analysis, this research is a collection of networks and their relative performances in analyzing varying data sets.

A second limitation is that only the individual violations of regression assumptions are studied. The relevant parameters that represent the violations discussed in Chapter 3 are varied one at a time. To vary all or some of the parameters at the same time would result in an exceptionally large number of combinations. This would be far too time consuming to be practical. By varying one parameter at a time and holding the others constant, at least their individual effects may be isolated.

1.3 Plan of Presentation

To guide the reader through the remainder of this dissertation, the following preview of the dissertation chapters has been provided.

Chapter *Description*

One: is an introduction to the conceptual framework of the research and the significance of the research for linking backpropagation neural networks and regression analysis.

Two: is a review of neural networks literature. It also looks at the body of neural networks literature as it applies to statistics.

Three: presents the methodology which was employed to compare the performance of traditional regression analysis against the performance using backpropagation networks. The characteristics of data sets that were considered and their development are reviewed. The measure which was employed to compare the performances of the techniques is discussed.

Four: discusses the performances of backpropagation neural networks and regression analysis in the context of varying sample sizes and error variances.

Five: presents the comparison between backpropagation networks and regression analysis based on nonnormal distributions of the error term, such as outliers, skewness, and kurtosis.

Six: is a comparison of backpropagation networks and regression analysis in the presence of varying degrees of multicollinearity.

Seven: is a comparison of backpropagation networks and regression analysis in the presence of nonlinearity and model underspecification.

Eight: presents a summary and the final conclusions. It looks at the implications of the research findings and provides recommendations for future application.

Chapter 2: Literature Review

Review of related literature is divided into two categories:

1. Neural Networks
2. Neural Networks and Statistics

2.1 Neural Networks

Ancient Greek philosophers, such as Plato (427-347 B.C.) and Aristotle (384-322 B.C.), were the first to suggest theoretical explanations of the brain and thinking processes. ″Neural computers″ belong to a class of cybernetic machines which have a much longer history than is generally known. Heron the Alexandrian was the first to build one such machine, the hydraulic automata, around 100 B.C. In the 18th century, mechanical devices for performing conceptual information processing - such as the slide rule for demonstrating syllogisms - were devised. Other mechanisms for logic operations were introduced in the nineteenth century.

W. S. McCulloch and W. A. Pitts (1943) were the first to conceive the fundamentals of neural computing in the early 1940's. Analytical neural modeling has been pursued in connection with psychological theories and neurophysiological research. Psychologists were also developing models of human learning. The most oustanding of these models was developed by Hebb (1949). Hebb's model proposed a learning law that is viewed as the start of artificial neural network training algorithms. Researchers in the 1950's and the 1960's combined these insights to build the first artificial neural network. Farley and Clark (1954) set up models for adaptive stimulus-response relations in random networks. Further elaborations of these theories were carried out by Rosenblatt (1958), Widrow and Hoff (1960), Caianiello (1961), Steinbuch (1961), and Minsky and Papert (1969). These and other researchers worked to develop networks, consisting of a single layer of neurons, called perceptrons. The burst of activity and optimism died down as it was discovered that the perceptrons were incapable of solving many simple problems including the exclusive-or problem (see Minsky and Papert for detailed explanation).

With the exception of a few scientists like Kohonen, Grossberg and Anderson, most researchers let neural networks lapse into obscurity for nearly two decades. In the 1970's and early 1980's a theoretical foundation emerged, upon which multilayer networks were constructed. Independent research efforts by Werbos (1974), Parker (1982), and Rumelhart, et al. (1986), resulted in backpropagation, a systematic training algorithm for multilayer networks that overcomes the limitations presented by Minsky and Papert (1969). The problems encountered by these early backpropagation networks were convergence and local minima.

Since then, researchers have improved and extended the basic backpropagation algorithm. Parker (1987) devised a second-order backpropagation method which improves the speed of convergence of the backpropagation algorithm. Parker's method uses second derivatives to produce a more accurate estimate of the correct weight change. A method for improving training characteristics of backpropagation networks was described by Stornetta and Huberman (1987).

Almeida (1987) and Pineda (1988a) have shown that learning can occur very rapidly in systems where backpropagation is applied to recurrent networks, i.e. networks whose outputs feedback to inputs.

Despite some existing limitations, backpropagation has dramatically expanded the range of problems to which neural networks can be applied and it has generated many successful demonstrations of its power. In the past few years, there has been an explosive increase in research activity in neural networks. Theory has been translated into application and neural network technology has been commercialized. NEC in Japan has applied backpropagation to a new optical-character recognition system, improving accuracy to over 99%. Burr (1987) used a backpropagation network to recognize handwritten English words. He reported an accuracy of 99.7%. A system converting printed English text into highly intelligible speech called NetTalk, developed by Sejnowski and Rosenberg (1987), has produced spectacular success.

Different neural network models were evaluated according to their performance in classifying handwritten digits by Guyon, et al. (1989), where classification was based on discriminant functions. The network architectures tested include layered networks with one or several layers of adaptive connections, fully connected recursive networks, and ad hoc networks with no adaptive connections. The multilayered networks, especially the three-layered network of adaptive units, fully connected between layers, outperformed all others. This particular network had a sigmoidal response function and was trained with the backpropagation algorithm.

The human brain continuously records new memories while ensuring that existing memories are not erased or corrupted in the process. Conventional neural networks faced this stability-plasticity dilemma. For example, in a backpropagation network, if a fully trained network must learn a new training set, it may disrupt the weights to the extent that complete retraining is necessary. Grossberg (1987) pointed out that in a real-world case, the network will be exposed to an environment that is constantly changing. In such a situation, the backpropagation network will continuously modify weights, never "learning" satisfactorily. His research, along with Carpenter (1987), resulted in the adaptive resonance theory (ART). ART networks classify inputs according to categories. If an input resembles a stored pattern category within a specified tolerance, the stored pattern is modified to make it more like the input. If an input does not match any stored pattern within the tolerance specified, a new category is created by storing a pattern like the input. Since it is still in its infancy, and its mathematics are complicated, ART applications are more sparse than those of the conventional neural networks like backpropagation.

A brief discussion of neural network basics is presented in Appendix A for the uninitiated reader.

2.2 Neural Networks and Statistics

Coverage of statistical techniques in the literature is extensive and a search of this literature was performed to identify the most widely used technique in statistical data analysis as regression analysis. In the past 30 years, practical data analysts as well as statistical theorists have contributed to advancements in this area. Myers (1986) describes regression analysis as ″a collection of statistical techniques that serve as a basis for drawing inferences about relationships among quantities in a scientific system″.

Sir Francis Galton first introduced the term ″regression″ in 1885 while demonstrating that offspring do not tend toward the size of parents but rather toward the average as compared to the parents. Since then, researchers have been faced with the question of what effect variables have on one another.

The central focus of regression analysis is the establishment of an equation that describes the relationship between the variables in a data set. The primary use of regression analysis is the prediction of one variable based on the known value of one or more other variables. When a single variable is used to estimate the value of an unknown variable, the method is referred to as simple regression analysis. Simple linear regression (SLR) is simple regression where the relationship between the known and unknown variables is linear.

In order to apply statistics to a real situation, assumptions must first be made about the situation and the phenomena involved. For example, in SLR, the assumptions of a linear relationship between the regressor and response variables, and the normality of errors are critical. The assumptions are made to obtain a tractable solution, even though usually the assumptions must be tested ex post.

The failure of one or more of the underlying assumptions results in difficulties with statistical techniques such as regression analysis. There are alternatives to the standard methodology when assumptions are violated but the process of first being able to identify the violation before analysis can be a lengthy and tedious task for most decision-makers. Hocking and Pendleton (1983), discuss the alternatives and point out that the solutions should be applied with discretion and only after data has been carefully scrutinized. According to Myers (1986) most data analysts overreact to a violation and overdo manipulations such as data transformations. It takes experience and sophistication to make the statistical strategies successfully accomplish their purpose of accomodating or countering the assumption violations. For example, Deaton, et al. (1983), provide a guideline for situations with heterogeneous variances and the use of weighted least squares - i.e. estimated weights should not be used unless they are each based on a sample size of approximately nine. A less sophisticated analyst will be tempted to use weighted least squares every time (Myers 1986).

One of the biggest dilemmas facing the regression analyst is how to cope with individual data points that do not fit the trend set by the rest of the data. The analyst will often succumb to the temptation of eliminating such data points in order to enhance the quality of fit. This is the appropriate strategy only if the outlier is a result of something external to the system being studied. In all other cases, further experimentation in the region where the outliers occur is required. Andrews, et al. (1978), provide a good discussion on the treatment of outliers.

Statistical packages, such as SAS, SPSS, and Minitab, are widely used by expert statisticians and novices alike. According to Gale and Pregibon (1985), most of the packages are abused due to ignorance of statistics on the part of novices. Such abuse actually provides an opportunity for the introduction of AI techniques in statistics that will make data analysis effective for the user who has a limited knowledge of statistics.

The widespread success of neural networks in addressing problems in other fields has suggested that they might prove useful in statistics. The stochastic approximation method was first proposed by Herbert Robbins and Sutton Monro (1951). Their concern was finding the solution to the equations E(m(Z_t,θ)), where E is the mathematical expectation of the random quantity m(Z_t,θ) and t represents time. The randomness of m(Z_t,θ) comes from Z_t, θ being chosen to make E(m(Z_t,θ)) equal zero and m is a freely chosen function. The solution of E(m(Z_t,θ)) = 0 is somewhat trivial if the probability distribution of Z_t is known. Robbins and Monro were interested in the case when the probability distribution of Z_t is unknown, and they proved that θ converges to a solution, θ', as t increases, thus providing a solution to E(m(Z_t,θ)) = 0. White (1989c) demonstrates that the backpropagation algorithm is nothing but a special case of the stochastic approximation method. In backpropagation, the weights are adjusted in response to errors in hitting the target. The ideal situation would be an error of 0. The errors are nothing but functions of the output, input and weights. If Z_t is viewed as consisting of input and output, and θ are the weights, then m(Z_t,θ) can represent backpropagation. The volume of literature in statistics and engineering using the stochastic approximation method has provided valuable insight into the advantages and disadvantages of backpropagation, by casting network learning as a statistical estimation problem and by affording the opportunity for improvements in network learning methods through the application of modern statistical theory, which can be generalized to other neural network learning procedures. This research uses the fact that both least squares method and backpropagation networks are based on the convergence of the expected value of the squared deviations (error) to zero. Starting from this common platform the relative performances of the two techniques under varying data conditions are studied.

Gale and Pregibon (1982) built a Regression Expert System (REX) which allows novices to perform regression analyses. REX detects and corrects violations of assumptions made by standard techniques. According to Gale and Pregibon (1985), REX was found to be weak in the interpretation of results and tutoring of novice statisticians.

A generalization of backpropagation to recurrent systems was derived by Werbos (1988) and applied to a model of natural gas markets. The variables to be predicted had to be predicted as functions of their own values at a previous time. When ordinary regression is used to estimate a model which predicts variables at time t + 1 as a function of time t, the forecasts tend to deteriorate, due to cumulative error effects. The generalized backpropagation consisted of three distinct backpropagation networks, each with different output evaluation components, and exploited the phenomenon of cumulative error.

Widrow and Winter (1988) describe the application of adaptive neural networks to statistical prediction. The input signal is delayed by Δ time units before being fed to the adaptive filter. The undelayed input serves as the desired output for the adaptive filter. The filter weights adapt to produce a best least squares estimate of the present input signal. The optimal weights are then copied into a "slave filter" which receives as input the undelayed signal and produces as output the best least squares prediction of the input Δ time units in the future. In this way, Widrow and Winter (1988), have shown that the future values of time-correlated signals can be estimated from present and past input samples.

White (1989c) provides a theoretical explanation of the similarity between backpropagation and nonlinear regression and how these are alternatives to solving the ordinary least squares problem. Least squares regression is a method of fitting a curve to a sample. A family of regression curves is specified and the member best fitting the data is selected, where quality of fit is measured in terms of total squared error. The function $f(x,\theta)$ can be used to define a member of the family of regression curves, with x representing explanatory variables and θ representing the parameters. This representation is identical to the output function of a backpropagation network, where x is the vector of inputs and θ is the vector of all weights (i.e. weights from input to hidden layer as well as weights from hidden layer to output layer). Thus, the network output function can be defined in terms of inputs and weights. White demonstrates that the weights $\theta^{\times}$ provide a network output which also minimizes the expected squared error. When f is nonlinear as a function of θ, the method of picking θ to minimize total squared error is called nonlinear least squares. The nonlinear least squares estimator, θ_n tends to $\theta^{\times}$ as n become large, since the law of large numbers of statistics guarantees that average squared error converges to expected square error. White thus demonstrates that neural network learning methods have a lot to offer the field of statistics.

In their research on training of artificial neural networks, Wang and Malakooti (1989) explain the statistical implication of artificial neural networks by proving that a learning rule which achieves trainability is statistically equivalent to a maximum likelihood estimator.

Shea and Lin (1989) compare the application of a neural network with discriminant analysis in the context of explosives detection in airline baggage. They used a three-layer, fully-connected, feedforward backpropagation network, a counterpropagation network, and a backpropagation with shared weights network and found the three-layer backpropagation network performed the best, exceeding the performance of the standard statistical technique.

Hornik, et al. (1989), theoretically established that multilayer feedforward networks are universal approximators. Even with as few as one hidden layer, provided there are sufficient numbers of hidden units, these networks are capable of approximating any Borel measurable function. According to Rohatgi (1976), every countable set of real numbers is a Borel set. A real-valued function f of a real variable x is Borel measurable if the set $\{x : -\infty < f(x) \leq y\}$ is a Borel set for every real number y. According to Hornik, et al., any failure in applications will be due to inadequate learning, insufficient number of hidden units, or the presence of a stochastic relationship between input and output instead of a deterministic one. These issues are not addressed with respect to what is needed to attain a given degree of approximation.

A single hidden-layer feedforward network is shown by White (1989b) to give a better fit than ordinary least squares to the Henon map, $Y_t = 1 - 1.4Y_{t-1}^2 + .3Y_{t-2}$. Both techniques are compared with 3 models: (a)Model 1, inputs Y_{t-1} and Y_{t-2}, (b)Model 2, input Y_{t-1} only, and (c)Model 3, input Y_{t-2} only. The network fits well even when one of the required inputs is omitted.

As can be seen, research linking neural networks and statistics has been either highly theoretical or specific applications to particular problems. This research attempts to bridge the gap by comparing the relative performance of backpropagation networks and regression analysis under certain data conditions. These conditions characterize the failure of assumptions underlying SLR and cause complications when a researcher is using regression analysis. The assumptions of SLR and the exact characterization of the data sets considered in this research are discussed in the next chapter.

Chapter 3: Methodology

The purpose of this chapter is to outline the procedures and methods which have been developed to compare the robustness of regression analysis versus backpropagation networks. It should be pointed out that, although the regression results can be determined analytically, the same cannot be done with backpropagation, and therefore, an empirical approach is taken in this research. Six areas of methodology are addressed: an explanation of the basic regression model, the backpropagation networks employed in the study, the identification of important characteristics of data sets used for comparison of the techniques, the procedure employed in data generation, the design of experiments conducted in this research, and the criterion for comparison.

3.1 The Basic Regression Model

The statistical technique that is used in this research is regression analysis. Although this research recognizes that there are sophisticated methods in regression to counter most any data complications, the level of regression analysis in this study is that of the average researcher. The following SLR model is considered.

$$Y_i = \alpha + \beta X_i + \varepsilon_i$$

where X is the regressor variable, Y is the measured response variable, and α and β are the intercept and slope respectively. ε_i is the model error, to account for the fact that the model is not exact. Regression involves fitting the above model to a set of data (values of X and Y), i.e. estimating the regression coefficients and formulating the fitted regression model

$$\hat{Y}_i = a + bX_i .$$

This fitted model is an estimate of the functional relationship that describes the data. Researchers traditionally use the fitted regression model to make predictions at X values not in the data set.

The underlying assumptions in regression are

1. The X_i are nonrandom, observed with negligible error,

2. The ε_i are random variables having a normal distribution with mean zero and constant variance σ^2 (i.e. homogeneous variance), and

3. The ε_i are uncorrelated from observation to observation.

The formulation of a linear model in regression analysis is an oversimplification of what occurs in actual observed processes, but linear models are approximations that have proved to work well in the range of the data used to build them. When the data are of reasonable quality, i.e. the assumptions underlying SLR are not violated, a linear model is very informative.

3.2 Proposed Neural Network Architecture

There are numerous aspects to the design of a neural network that have to be taken into account. Appendix A provides a discussion of the fundamentals of a neural network.

The selection of neural paradigms is an important step in the development of any neural network application and is based on the comparison of application requirements to paradigm capabilities. The application requirements of interest are network size (the number of layers and the number of elements in each layer), type of output, associative memory classification (autoassociative or heteroassociative), training method, and time constraints. The type of neural network chosen for this study is the three-layer feedforward backpropagation network, since White (1989b) and Shea and Lin (1989) both find that this network performs the best in prediction. There is an input layer, one hidden layer, and the output layer. The network is trained according to the supervised method. Supervised training is accomplished by sequentially applying training pairs of inputs and desired output. Although the training time for the backpropagation network is slow, level of accuracy is high. This decision was arrived at after studying the literature on neural networks and various network applications cited in Chapter Two. For example, Shea and Lin (1989) found that the three-layer backpropagation network outperformed other neural networks in an application with discriminant analysis.

A major remaining methodological issue is the neural network representation scheme. Although the literature supports binary representation as being the most widely used, this research utilizes real numbers as input and output and utilizes the MinMax table in the NeuralWorks (Klimasauskas, et al., 1989b) backpropagation network implementation to scale the range of values and convert them to a binary scheme internally.

With the exceptions of the multicollinearity and nonlinearity cases, all the models under consideration in this research involve the prediction of Y from one X. The X values for all cases range between 0 and 99. The number of input processing elements (PE′s) is determined by the number of X′s in each case studied and there is one output PE for all cases. The number of PE′s in the hidden layer are determined through experimentation in Chapter 4. Subsequent chapters use this best number of hidden layer PE′s.

This is a general description of the neural network architecture. Details and specific modifications are provided in the respective chapters.

3.3 Characteristics of Data Sets

Empirical evidence of common occurrences in data sets as well as the basic assumptions of SLR (stated earlier) were taken into consideration in the identification of important data characteristics. Difficulties with regression analysis are usually a failure of one or more assumptions.

According to Myers (1986), sample size is an important consideration in regression analysis. When sample size is too small, adequate measures of error in regression results cannot be computed and model assumptions cannot be checked. Large sample sizes incur higher costs in data collection and so the researcher is often faced with determining a sample size that strikes an acceptable balance.

The underlying method for estimation of model coefficients in SLR is the method of least squares. The homogeneous variance assumption in SLR allows that, at X = X_i, Var(Y_i) = Var(ε_i) = σ^2. This has important implications for the variance properties of the least square estimators, which are

$$\mathrm{Var(b)} = \frac{\sigma^2}{S_{xx}}, \text{ and}$$

$$\text{Var(a)} = \sigma^2\left(\frac{1}{n} + \frac{\overline{X}^2}{S_{xx}}\right)$$

where $S_{xx} = \sum(X_i - \overline{X})^2$. As the variance of the error term increases, the variances of the estimators increase, thereby affecting the quality of prediction. Thus, variance of the error term is an important consideration in regression analysis.

Normality of the ε_i is necessary for the estimators to have the property of minimum variance among all unbiased estimators. The presence of outliers or nonnormality of the error term reduce this property to minimum variance of all linear unbiased estimators. The least squares procedure allows outliers to exert disproportionate influence on regression results.

In many real-world situations, there is more than one regressor variable. Models with more than one regressor that are linearly related to the regressand are referred to as multiple linear regression models. In such models a condition in the regressor variables, known as multicollinearity, may exist. Multicollinearity occurs when there are near linear dependencies among the regressors, i.e. the regressors move with one another, resulting in poor estimation of regression coefficients and therefore poor prediction.

While SLR is adequate for many situations, there are many areas of engineering and the sciences where the experimental situation requires the use of nonlinear models. These models may be nonlinear in the regression parameters, making the computation of parameter estimates by elementary matrix algebra (as in least squares) impossible. In fact, nonlinearity brings about complications in the development of least squares estimators.

Model misspecification results in poor prediction using regression analysis. Underspecifying refers to the failure to include some variables and this condition causes bias in prediction.

The result of these considerations is the characterization of data in this research according to the following: sample size and error variance, distribution of the error term, multicollinearity, and nonlinearity and underspecification. By using this characterization, the robustness of each technique is studied under what is, in effect, the relaxation of assumptions one at a time. A more complete discussion of each of these characteristics is provided later in the chapter, under the heading Design.

3.4 Data Generation

This research utilizes data generated through the Statistical Analysis System (SAS), instead of collected data. The advantage of generated data over real data is that the true values of the parameters are known, allowing for evaluation of the techniques employed in estimating these parameters.

In each of the experiments outlined under Design, the population is generated according to the characteristics studied. Detailed description of data generation is presented in each chapter. Some subsets of the population are used as samples for regression analysis as well as for the recall or learning capabilities of the backpropagation network. The training sets for the backpropagation networks are also subsets of the populations generated but are distinct from the recall sets. The regression analysis is conducted using SAS.

3.5 Design

The framework of this research is based on the data set characteristics mentioned earlier, i.e. sample size and variance, error distribution, multicollinearity, and nonlinearity and model underspecification. Each experiment is designed around a particular characteristic of data sets. The following sections describe each of the characteristics as well as the levels of analyses.

3.5.1 Sample Size and Variance

The primary purpose of the experiment on sample size and variance is to establish a baseline combination of these two factors at which regression and backpropagation perform at about the same level.

Sample size is important in regression analysis. According to Myers (1986), if sample size is too small, one cannot compute adequate measures of error in regression results, and there can be no basis for checking model assumptions.

The population data is generated according to the traditional SLR equation

$$Y_i = \alpha + \beta X_i + \varepsilon_i$$

where ε_i is the error term generated from a normal distribution with mean of 0 and variance levels of 25, 100, 225, and 400.

In order to test the robustness of regression analysis against backpropagation networks, experiments with varying sample sizes are conducted for each level of variance. Samples of size 20, 50, 100, 200, and 500 are considered, to encompass small, medium, and large sample sizes. The combination of sample size and variance yielding comparable predictions for both regression and backpropagation are used in the experiments that follow.

3.5.2 Error Distribution

The distribution of the error term is a critical factor in regression analysis. This research explores the performance of backpropagation networks and SLR based on the violation of the assumption of normal distribution of error. The following aberrations of the normal distribution are studied:

error distribution with outliers, skewed error distributions, and varying kurtosis of error distributions. Distributions with skewness and kurtosis are included to emulate the nonnormal situation depicted by Professor Richard G. Krutchkoff in his research on the comparison of two techniques.

3.5.2.1 Effect of Outliers in Error Distribution

According to Myers (1986), the normal distribution is not a very good model of the "noise" or error in most processes. Outliers are often present in the error distribution and least squares, the basis of regression analysis, allows outliers to exert disproportionate influence on the results. In order to generate the error term from a distribution with outliers, the model suggested by Werbos (1974) is used.

$$Y_i = \alpha + \beta X_i + \varepsilon_i$$

where ε_i is the error term. 95% of the time, ε_i is generated from a normal distribution with mean of 0. The variance level used for the error is the level at which both techniques perform the same in the prior variance/sample size experiments in Chapter 4. Using this level helps to remove any bias in the later designs that might be caused by level of variance. In 5% of the cases, for mild outliers, chosen at random, the ε_i is generated from a normal distribution with mean zero and standard deviation twice the basic standard deviation of the error. For extreme outliers, 5% of the cases are from a normal distribution with mean zero and standard deviation three times that of the non-outliers.

3.5.2.2. Effect of Skewness and Kurtosis

The data for this experiment are generated according to the parameters of Pearson distributions, i.e. the first four moments, where μ_1 represents the mean, $\mu_2 = \sigma^2$, and μ_3 and μ_4 are the third and fourth central moments. Skewness is then defined as

$$\beta_1 = \frac{{\mu_3}^2}{{\mu_2}^3}$$

and kurtosis is defined as

$$\beta_2 = \frac{\mu_4}{{\mu_2}^2} .$$

The normal distribution belongs to the family of Pearson distributions, with $\beta_1 = 0$ and $\beta_2 = 3$.

The effect of skewness in the error is studied for two cases: positive skewness, or $\sqrt{\beta_1} = 1$, and negative skewness, or $\sqrt{\beta_1} = -1$.

For the study of the effect of kurtosis, a very peaked error distribution with $\beta_2 = 4$ and a flatter distribution with $\beta_2 = 2$, is considered.

These values for skewness and kurtosis were emulated from the discourse provided by Professor Krutchkoff on comparison of two techniques.

3.5.3 Multicollinearity

Multicollinearity exists when the regressor variables are not independent, i.e. near linear dependencies exist between regressor variables. A regression coefficient is a rate of change or partial derivative of the response with respect to a regressor variable. When multicollinearity exists, the regressors move with one another as well as with the response variable. Multicollinearity prohibits precise statistical inference. It lowers the precision in the estimation of a regression coefficient by inflating the variance of the coefficient. Instability of the regression coefficient may effect the quality of fit and prediction of the regression model.

In order to study the effect of multicollinearity, the following linear model with two regressor variables is considered

$$Y_i = \alpha + \beta X_1 + \gamma X_2 + \varepsilon_i.$$

Varying degrees of multicollinearities in the data set are studied by generating data sets with correlation of -.1, -.5, -.9, 0, .1, .5, and .9 between the two regressor variables. A detailed description of how this is accomplished is in Chapter 6.

3.5.4. Nonlinearity and Underspecification

The presence of nonlinearity in a regression model causes complications that render the use of OLS impossible and model underspecification results in poor prediction. Both these conditions require special treatment under regression analysis.

3.5.4.1 Nonlinearity

While SLR is adequate for many situations, there are many areas of engineering and the sciences where the experimental situation requires the use of nonlinear models. Although nonlinear regression (NLR) refers to the nonlinearity in the regression parameters, this research uses the term ″nonlinear″ to include a model with a quadratic term, which is linear in the regression coefficients but nonlinear in the relationship between the regressor variables, as well as an exponential model, which is a true nonlinear model.

Model with Quadratic Term

The following model with a quadratic term is studied.

$$Y_i = \alpha + \beta X_i + \theta X_i^2 + \varepsilon_i.$$

The X and X^2 are the regressor variables, Y is the measured response variable, and α is the intercept. β and θ are the slope coefficients and ε_i is the model error. The θ values used are -10, -5, -1, -.5, -.2, .2, .5, 1, 5, and 10.

Exponential Model

As an example of nonlinear regression (NLR) models the following exponential model is considered.

$$Y_i = \alpha e^{\beta X_i} + \varepsilon_i$$

where X is the regressor variable, Y is the response variable, and ε_i is the model error. α and β are the regression coefficients.

3.5.4.2 Underspecification

Model underspecification results in poor prediction using regression analysis. Underspecifying refers to the failure to include some variables and this condition causes bias in prediction.

In order to study the effect of underspecification, the population with the relationship

$$Y_i = \alpha + \beta X_i + \theta X_i^2 + \varepsilon_i$$

is considered, where X and X^2 are the regressor variables, Y is the measured response variable, and α is the intercept. β and θ are the slope coefficients and ε_i is the model error.

The effect of failing to include the quadratic term, when it has a positive as well as a negative coefficient, is studied separately. For the positive quadratic case, θ values of .2, .5, 1, 5, and 10 are explored and for the negative quadratic case, θ values of -10, -5, -1, -.5, and -.2 are considered.

3.6 Criterion for Comparison

When comparing two or more procedures, it is crucial to identify the criterion for comparison. A very useful criterion in statistics is mean squared error (MSE), the expected value of the square of the difference between the estimator and the parameter. This research uses two techniques to estimate Y, i.e. regression analysis and backpropagation networks. Since the population value of Y is known, the two estimators are compared to the true Y value using the principle of MSE. In order to have the magnitude of the error in the same "units" as the observations, the root of the squared differences is used. The predicted output, $\hat{Y}$, from each technique is compared to the corresponding actual output, Y, and the root mean square difference (RMS) for each technique is calculated as

RMS $= \dfrac{\sqrt{\sum(Y-\hat{Y})^2}}{n}$, where n is sample size.

By comparing the procedures in this way, it is possible to determine whether the two prediction techniques perform comparably, one technique is better than the other under certain conditions, or if one is uniformly better than the other.

Chapter 4: Sample Size and Variance

This chapter compares the predictive abilities of simple linear regression and neural networks in the presence of two factors, sample size and variance of the error term, in an attempt to determine the combination of these factors such that both regression and backpropagation perform at the same level. This combination is to be used as the baseline for subsequent experiments. In order to observe the performance of the two techniques over a range from small to large samples, sample sizes of 20, 50, 100, 200, and 500 are considered. The interaction of sample size with variance of the error term is hypothesized to have some effect on prediction. Variances of 25, 100, 225, and 400 were selected after preliminary experimentation with variance levels so as to provide a range where effects would be observed. A two by two factorial design is used in this chapter.

4.1 Generation of Data

The populations used in this chapter were generated according to the form

$$Y_i = \alpha + \beta X_i + \varepsilon_i$$

where X is the regressor variable, Y is the measured response variable, and α and β are the intercept and slope respectively. The ε_i is the model error, to account for the fact that the model is not exact. The ε_i are random variables from a normal distribution with mean zero and variances of 25, 100, 225, and 400. The X_i are random variables generated from a normal distribution of mean 50 and variance of 225 to obtain X values within the range 0 to 99. Although nonrandomness of X is an assumption in SLR, in practice, according to Myers (1986), the X_i do experience some random variation. The important assumption is that any random variation in X is neglible compared to the range in which it is measured. In order to insure that Y values also ranged from 0 to 99, a β value of 0.8 is chosen and an α of 10 is chosen. A distinct population of 10,000 values was generated, using SAS, for each level of variance. Appendix B is a sample of the SAS program used to generate the populations.

4.1.1 Generation of Samples

From each population, varying sample sizes were generated. For each combination of sample size and variance, five sets of data, i.e. X and Y values, were taken from the appropriate population and designated as training sets. In addition, four other distinct sets of data were taken from each population to be used as recall sets. Thus, for each combination of variance and sample size, nine distinct data sets were obtained from the appropriate population. Five training sets were used to make sure that any one data set with atypical values was not being used to train the network, thus creating a bias. Similarly the four recall sets insure that there is no bias in the predictive ability.

The training sets were used to train the neural networks as well as estimate α and β in regression. The recall sets were used to determine how well the two techniques predict the output given the information they receive. Each of the four recall sets was recalled on each of the five training sets, resulting in 20 output sets for each combination of sample size and variance, as shown in Table 1 on page 33. Although such a design may result in some interdependence among the observations, it was expected to be controlled by using ANOVA.

For each combination of sample size and variance, the predicted output, $\hat{Y}$, from each technique was compared to the corresponding actual output, Y, for each observation in each of the 20 output sets and a root mean square (RMS) was calculated for each of the 20 output sets, using the formula

RMS $= \frac{\sqrt{\sum (Y - \hat{Y})^2}}{n}$, where n is sample size.

Thus, for each sample size and variance combination, 20 RMS values were calculated for each technique.

4.2 Regression Analysis

Regression involves fitting the model in Section 4.1 to a set of data (values of X and Y), i.e. estimating the regression coefficients and formulating the fitted regression model

$$\hat{Y}_i = a + bX_i .$$

This fitted model is an estimate of the functional relationship that describes the data. Researchers traditionally use the fitted regression model to make predictions at X values.

For each sample size and variance combination, the five training sets were used to estimate the regression parameters α and β. Appendix C is the SAS program used for this purpose. The five pairs

Table 1. Experimental Design for each Sample Size and Variance Combination

Recall	Training Sets				
Sets	A	B	C	D	E
1					
2					
3					
4					

of estimated α and β values were then used along with X values from each of the four recall sets to predict $\hat{Y}$ values using another SAS program (see Appendix D) based on the fitted model

$$\hat{Y}_i = a + bX_i .$$

Thus, for each sample size and variance combination, 20 output sets were created using regression. Each of these 20 output files with X and $\hat{Y}$, along with the appropriate recall files which have the corresponding Y values, were evaluated by a BASIC program (Appendix E) that calculates the RMS for each output set.

Thus, for each combination of sample size and variance, 20 RMS values were obtained, resulting in 400 RMS values since there are four variance levels and five sample sizes. The results from regression are summarized in Table 2 on page 35. The columns represent the five levels of sample size and the rows represent the four variance levels.

4.3 Neural Network Analysis

The selection of neural paradigms is an important step in the development of any neural network application. The application requirements of interest are network type and size, representation scheme, training method, and learning count.

The networks were trained according to the supervised method. Supervised training is accomplished by sequentially applying training pairs of inputs and desired output. The network type chosen for this study is the backpropagation network. Although the training time for the backpropagation network is slow, level of accuracy is high. This decision was arrived at after studying the literature on neural networks and various network applications cited in Chapter Two. For ex-

Table 2. Root Mean Squares from Regression - Sample Size and Variance

σ^2	Sample Sizes									
	20		50		100		200		500	
25	0.23624	0.22400	0.20183	0.20117	0.13742	0.11236	0.07916	0.06856	0.02230	0.01765
	0.23155	0.33247	0.20103	0.20198	0.12934	0.11675	0.07494	0.07650	0.00987	0.01342
	0.25210	0.19964	0.20096	0.20084	0.10863	0.10098	0.07308	0.06883	0.02045	0.01984
	0.18751	0.30401	0.20036	0.20187	0.10653	0.11005	0.07653	0.07712	0.01730	0.00994
	0.23529	0.20054	0.20179	0.20021	0.11279	0.11184	0.07184	0.06855	0.02174	0.01875
	0.19781	0.32402	0.20105	0.20064	0.11817	0.11923	0.07502	0.07657	0.17990	0.01684
	0.27808	0.18974	0.20005	0.20038	0.12140	0.12076	0.07240	0.06979	0.02009	0.02165
	0.21632	0.25403	0.20125	0.20032	0.12105	0.12369	0.07470	0.07711	0.01932	0.02735
	0.24781	0.20210	0.20134	0.20092	0.10889	0.12164	0.07203	0.06843	0.02186	0.02294
	0.18924	0.31493	0.20148	0.20176	0.12058	0.11896	0.07533	0.07658	0.01872	0.01908
100	2.57464	2.07191	1.59508	1.60160	1.10997	1.92825	0 69117	0.67232	0.44098	0.44182
	2.65369	3.22178	1.14117	1.56287	1.77370	0.98430	0.71624	0.68740	0.45419	0.45573
	2.60338	1.58139	1.42351	1.20896	1.00325	1.69432	0.75941	0.90170	0.43214	0.43017
	2.54583	2.71164	1.22572	1.31927	1.52321	1.60100	0.67586	0.85307	0.44006	0.43921
	2.60091	1.81085	1.60054	1.38981	1.23681	1.53725	0.71284	0.88983	0.44125	0.44901
	2.49982	2.91111	1.51237	1.49869	1.64231	1.42106	0.68261	0.46310	0.44068	0.43975
	2.81141	1.62644	1.39283	1.46875	1.35217	1.48923	0.78602	0.95413	0.45121	0.44961
	2.57931	2.72939	1.35060	1.26945	1.71428	1.37211	0.70451	0.80439	0.44549	0.43065
	2.91905	1.66728	1.47761	1.40108	1.82750	1.32896	0.87471	0.80914	0.46009	0.45735
	2.69780	2.79805	1.29816	1.50124	1.58435	1.52008	0.71608	0.82335	0.43281	0.43276
225	2.57864	2.16091	1.62863	1.23526	0.94057	0.98538	0.69517	0.68232	0.44411	0.45435
	2.56329	3.18287	1.68119	1.60321	0.98923	0.98010	0.72624	0.70923	0.45789	0.43226
	2.72341	2.11967	1.12647	1.53216	0.97216	0.96658	0.61423	0.66921	0.43257	0.42218
	2.32476	2.58743	1.48737	1.61486	0.98919	0.97223	0.70783	0.71246	0.41077	0.40928
	2.56834	2.10984	1.23784	1.48836	0.81421	0.90247	0.68912	0.67536	0.45863	0.41965
	2.43578	2.37216	1.58921	1.51043	0.82973	0.88632	0.69524	0.69910	0.40287	0.45926
	2.41386	2.09541	1.39847	1.27066	0.87265	0.91486	0.67428	0.66361	0.44692	0.40883
	2.20963	2.39654	1.41329	1.60981	0.85537	0.96431	0.68925	0.69064	0.43729	0.44398
	2.01235	2.16342	1.46592	1.39206	0.83487	0.90031	0.62039	0.65714	0.43786	0.41731
	2.33778	2.13728	1.40418	1.57926	0.89067	0.90205	0.67386	0.69895	0.46051	0.43172
400	3.40586	3.10787	2.44294	1.85288	1.41086	1.47807	1.03676	1.00840	0.66617	0.68153
	3.98054	4.83267	2.52178	2.08677	1.48384	1.47015	1.07436	1.03110	0.68684	0.70067
	3.62381	3.29317	2.10632	1.97236	1.40683	1.45372	1.88790	1.88325	0.65231	0.62386
	3.25018	4.00156	2.38712	2.19436	1.59218	1.55736	1.92765	1.93482	0.64320	0.69827
	3.26712	3.99643	2.21857	1.98745	1.40065	1.49207	1.12431	1.12746	0.64279	0.65836
	4.21687	4.21538	2.42729	2.37822	1.51392	1.52318	1.29863	1.23654	0.65009	0.69234
	3.10157	3.27185	2.23143	2.01327	1.41324	1.46673	1.20177	1.20031	0.67883	0.63927
	4.14768	4.62341	2.51863	2.03957	1.50321	1.48926	1.21005	1.20012	0.64028	0.67185
	3.32419	3.62439	2.40085	1.93750	1.42635	1.45938	1.08936	1.04326	0.67107	0.64129
	4.01760	4.43285	2.49399	2.16743	1.52176	1.49799	1.07839	1.05408	0.64816	0.68498

ample, Shea and Lin (1989) found that the three-layer backpropagation network outperformed other neural networks in an application with discriminant analysis.

The software used to implement the backpropagation paradigm is NeuralWorks professional II (1989b). The software was executed on a 386 personal computer running at 33 MHz.

A major remaining methodological issue is the neural network representation scheme. Although the literature supports binary representation as being very strong, coding may result in some loss of information. This research uses the X and Y values as real numbers and utilizes the MinMax table in NeuralWorks to facilitate learning and recall. The MinMax table scales data to continuous values between 0 and 1 by looking at the minimum and maximum values.

4.3.1 Determination of Network

The types of backpropagation networks investigated to determine the most appropriate one for this study were networks with one input layer processing element, one output layer processing element, and hidden layers with the following characteristics:

1. one hidden layer, one processing element, sigmoidal transfer function
2. one hidden layer, one processing element, sine transfer function
3. one hidden layer, one processing element, hyberbolic tangent function
4. one hidden layer, two processing elements, sigmoidal transfer function
5. one hidden layer, two processing elements, sine transfer function
6. one hidden layer, two processing elements, hyberbolic tangent function
7. two hidden layers, one processing element each, sigmoidal transfer function
8. two hidden layers, one processing element each, sine transfer function
9. two hidden layers, one processing element each, hyberbolic tangent function
10. two hidden layers, two processing elements each, sigmoidal transfer function

11. two hidden layers, two processing elements each, sine transfer function
12. two hidden layers, two processing elements each, hyberbolic tangent function

In order to determine the best network for this study, the following steps were required:

1. Determination of the best learning count (level of training) that provides the best prediction, for each type of network. Without loss of generality, a sample size of 20 and variance of 25 was used for the determination of the best learning count.
2. Identification of the best network using the best learning count for sample size 20 and variance 25.
3. Determination of the best learning counts for the remaining sample sizes using the best network.

4.3.1.1 Determination of Learning Count for Sample Size 20

Each of the above networks was trained for 50,000 presentations using the same training set (i.e. training set for sample size of 20 and variance of 25) and recalled on the same recall set after every 5,000 presentations. During recall, for each network, the input and the predicted output were written to a file. As with regression analysis, the neural network output files were evaluated by the same BASIC program to compare actual Y and $\hat{Y}$ and to calculate the RMS for the outputs from each network. For each network, the RMS associated with the learning counts up to 50,000 were plotted (Appendix F is a plot for one of these networks). A RMS value of 0 indicates that there is no difference between Y and $\hat{Y}$ and the larger the RMS, the poorer the prediction. The best learning count is the point where the plot flattens out, indicating that higher count does not reduce RMS.

The best learning count for each network was noted and the results indicate that the transfer function does not have an effect on the best learning count but the number of hidden layers and the number of processing elements in each hidden layer do have an effect. The networks (regardless of the type of transfer function) and the respective best counts (for sample size 20 and and variance 25) are summarized in Table 3 on page 39.

Thus, the backpropagation networks with two hidden layers and two processing elements in each hidden layer, regardless of the transfer function, performs the best when trained 60,000 times. All the other networks need to be trained only 20,000 times for their best prediction.

4.3.1.2 Identification of Best Network

All the networks mentioned in Section 4.3.1 (trained at their best levels according to Section 4.3.1.1) were recalled on all four recall sets for sample size of 20 and variance of 25. The RMS was calculated for each recall set. Then, for each network, the average RMS of the four appropriate recall sets was calculated. These average RMS values were compared to identify the best network for this study. The results are shown in Table 4 on page 40.

The backpropagation network with one hidden layer, two processing elements, and the sigmoidal function has the lowest average RMS. This indicates the best prediction values for Y. Thus, each subsequent experiment in this research uses a three-layer backpropagation network with two processing elements in the hidden layer, one processing element each in the input and output layers, and a sigmoidal transfer function.

4.3.1.3 Determination of Best Learning Count for Other Sample Sizes

The backpropagation network chosen in the previous section was used for the remaining levels of sample size (with variance 25), and the method for obtaining the best learning count in Section 4.3.1.1 was repeated. The results are summarized in Table 5 on page 42.

Table 3. Summary of Networks and Best Learning Counts

Network	Best Learning Count
One hidden layer, one processing element	20,000
One hidden layer, two processing elements	20,000
Two hidden layers, one processing element each	20,000
Two hidden layers, two processing elements each	60,000

Table 4. Summary of Networks and Average Root Mean Squares

Network	Average RMS
One hidden layer, one PE, sigmoidal	.99159
One hidden layer, one PE, sine	1.13193
One hidden layer, one PE, hyberbolic	.90658
One hidden layer, two PE, sigmoidal	.57594
One hidden layer, two PE, sine	.90079
One hidden layer, two PE, hyberbolic	.83170
Two hidden layers, one PE each, sigmoidal	.87082
Two hidden layers, one PE each, sine	1.06303
Two hidden layers, one PE each, hyperbolic	.97965
Two hidden layers, two PE each, sigmoidal	.75847
Two hidden layers, two PE each, sine	1.21573
Two hidden layers, two PE each, hyperbolic	1.06751

It should be stressed at this point that any level of learning count can be implemented for any sample size by the repeated presentation of the original sample (i.e. each input/output set in a sample of 100 is presented 350 times to get a learning count of 35,000).

In subsequent analyses of data, networks trained according to the appropriate learning count, based on sample size, are utilized.

4.3.2 The Neural Network Experiment

The experimental design is a two by two factorial design, with five levels of sample size and four levels of variance. Sample sizes of 20, 50, 100, 200, and 500 and variances of 25, 100, 225, and 400 are used. Within this framework, for each level of sample size and variance, there were five training sets and four recall sets. A three-layer backpropagation network with two processing elements in the hidden layer, one processing element each in the input and output layers and a sigmoidal transfer function is used. For each combination of sample size and variance, each of the five training sets resulted in a network which was then recalled on each of the four recall sets, resulting in 20 output sets. The training sets and the recall sets were of the appropriate sample size. For example, for sample size of 20 and variance 25, all recall and training sets were of size 20. The five trained networks (trained for 20,000 presentations) were recalled on each of the four recall sets, resulting in 20 output sets and the RMS for each set was calculated using the same BASIC program as regression (Appendix E). Thus, for sample size 20 and variance 25, there were 20 RMS values corresponding to the 20 RMS values obtained from regression analysis.

The experiment was repeated for each combination of sample size and variance. For each combination of sample size and variance 20 RMS values were obtained, resulting in 400 RMS values corresponding to the 400 RMS values obtained from regression. The neural network results are summarized in Table 6 on page 43.

Table 5. Summary of Sample Sizes and Best Learning Counts

Sample Size	Best Learning Count
20	20,000
50	25,000
100	35,000
200	40,000
500	50,000

Table 6. Root Mean Squares from Backpropagation Network - Sample Size and Variance

σ^2	Sample Sizes									
	20		50		100		200		500	
25	0.74877	1.04208	0.53216	0.54218	0.38790	0.34364	0.23979	0.18744	0.10329	0.11574
	0.81775	0.85497	0.53964	0.50036	0.37286	0.34937	0.25151	0.26316	0.05376	0.08937
	0.36926	0.53966	0.52143	0.52317	0.32987	0.32114	0.28888	0.25548	0.10986	0.09903
	0.36389	0.38807	0.50839	0.51294	0.32176	0.33298	0.31452	0.32526	0.09895	0.06561
	0.37553	0.63371	0.50227	0.52146	0.32067	0.33006	0.24454	0.19695	0.11214	0.09836
	0.42032	0.45609	0.52869	0.51732	0.32143	0.32010	0.26956	0.26290	0.19732	0.09987
	0.42567	0.73542	0.51298	0.51097	0.32735	0.31998	0.34740	0.19906	0.10639	0.11547
	0.47403	0.51703	0.52170	0.51906	0.33117	0.33947	0.24761	0.29156	0.13314	0.18376
	0.43120	0.79591	0.51063	0.51134	0.31067	0.32785	0.35797	0.24048	0.10522	0.13220
	0.51286	0.61686	0.53066	0.52814	0.32149	0.32864	0.30656	0.31179	0.10619	0.19852
100	2.84810	1.91418	1.58321	1.59372	1.20117	1.23128	0.58317	0.54969	0.26182	0.27176
	2.74225	3.08678	1.41235	1.49368	1.86735	1.01372	0.62237	0.63695	0.30435	0.31018
	2.52240	1.82308	1.54483	1.54216	1.01829	1.57218	0.58171	0.57813	0.26635	0.25243
	2.61720	2.94704	1.45070	1.51897	1.92536	1.35218	0.68347	0.54525	0.26726	0.26917
	2.81259	2.60013	1.59176	1.53725	1.05380	1.42775	0.62513	0.60368	0.30921	0.28008
	2.21321	2.98951	1.40613	1.50321	1.73864	1.58376	0.61514	0.41735	0.27217	0.27188
	2.79813	1.59531	1.48081	1.43376	1.27285	1.69216	0.53717	0.60208	0.30524	0.29178
	2.76372	2.86173	1.40867	1.53216	1.79365	1.43289	0.61268	0.57940	0.28112	0.26927
	2.92993	1.91235	1.40638	1.51068	1.31726	1.75338	0.72841	0.63218	0.32164	0.30987
	2.78341	2.93546	1.41096	1.50031	1.67724	1.37651	0.69103	0.61105	0.26126	0.26224
225	2.32726	2.01432	1.41321	1.09876	0.80522	0.82176	0.52923	0.49376	0.21963	0.23215
	2.61217	2.10966	1.42365	1.59217	0.89218	0.80715	0.52934	0.49876	0.24936	0.18542
	2.61322	2.09836	1.09056	1.33889	0.71214	0.74325	0.34683	0.39873	0.19961	0.18635
	2.04831	2.55326	1.30654	1.40973	0.78694	0.72631	0.50168	0.53418	0.17978	0.17163
	2.11327	2.00985	1.04361	1.30985	0.63358	0.69328	0.49725	0.48566	0.25611	0.18132
	2.31768	2.09745	1.39718	1.36674	0.69317	0.60951	0.49079	0.48265	0.17865	0.30112
	2.31926	2.10711	1.17752	1.09188	0.64219	0.70188	0.44321	0.41985	0.22721	0.17184
	2.39371	2.05328	1.30562	1.41387	0.61743	0.78226	0.47214	0.51923	0.21981	0.22964
	2.09876	2.27516	1.28369	1.18972	0.61153	0.71284	0.38991	0.39732	0.22865	0.18633
	2.00870	2.09360	1.29348	1.33241	0.72113	0.67341	0.50913	0.39772	0.22746	0.19697
400	3.39216	3.04365	2.23272	1.83926	1.21321	1.25132	0.83723	0.68265	0.42185	0.42273
	3.76817	4.51442	2.41264	2.06132	1.22846	1.27896	0.81253	0.88216	0.43926	0.48712
	3.43169	3.01965	2.08426	1.80054	1.18768	1.21678	1.60387	1.53218	0.33675	0.33188
	3.22321	3.91765	2.11832	1.92745	1.29325	1.28135	1.65163	1.74376	0.31023	0.53237
	3.01763	3.98521	2.09234	1.72147	1.39216	1.30083	0.91327	0.88627	0.33129	0.32487
	4.09388	4.04627	2.21143	1.91832	1.35263	1.38892	1.03176	1.09874	0.40165	0.55638
	3.00840	3.16472	2.01836	1.92316	1.18126	1.23212	0.89876	0.99895	0.44872	0.40113
	3.98765	4.51327	2.28657	1.98963	1.29386	1.27675	1.08173	0.91273	0.44312	0.49276
	3.25630	3.33768	2.14747	1.79964	1.15321	1.20817	0.83125	0.92174	0.51865	0.34328
	3.99273	4.23110	2.21210	1.90036	1.30134	1.28918	0.88734	0.79328	0.37672	0.54920

4.4 Comparison of Neural Networks and Regression

A SAS program (Appendix G) is used to calculate the difference between the RMS from neural network, RMS_N, and the RMS from regression, RMS_R. This difference is calculated as RMS_N - RMS_R for each observation in each combination of sample size and variance. A negative difference in the RMS indicates that neural networks have a lower RMS and therefore do a better job in predicting the Y while a positive difference implies that regression gives a better prediction of Y. The 400 RMS differences, corresponding to the 400 pairs of RMS values from regression and backpropagation, are summarized in Table 7 on page 45.

4.4.1 Analysis of ANOVA Assumptions

Due to its wide usage and overall comprehensibility, this study initially intended to use the two-way analysis of variance on its results. In order to use two-way ANOVA, the validity of the normality and homoscedasticity assumptions had to be determined.

4.4.1.1 Test for Normality

The RMS differences for each combination of sample size and variance (there were 20 such combinations and 20 differences for each combination) were tested to see whether they are from normal populations. The Kolmogorov-Smirnov goodnes-of fit test was used since it is a nonparametric procedure appropriate for small samples. The hypotheses were formulated as:

H_0 : RMS differences come from a normal distribution

H_1 : H_0 not true

Table 7. Root Mean Differences between Backpropagation and Regression - Sample Size and Variance

	Sample Sizes									
σ^2	20		50		100		200		500	
25	0.51253	0.81808	0.33033	0.34101	0.25048	0.23128	0.15881	0.11888	0.08099	0.09809
	0.58620	0.52250	0.33861	0.29838	0.24352	0.23262	0.17657	0.18666	0.04389	0.07595
	0.11716	0.34002	0.32047	0.32233	0.22124	0.22016	0.21580	0.18665	0.08941	0.07919
	0.17638	0.08406	0.30803	0.31107	0.21523	0.22293	0.23799	0.24814	0.08822	0.05567
	0.14024	0.43317	0.30048	0.32125	0.20788	0.21822	0.17270	0.12840	0.09040	0.07961
	0.22251	0.13207	0.32764	0.31668	0.20326	0.20087	0.19454	0.18633	0.17933	0.08303
	0.14759	0.54568	0.31293	0.31059	0.20595	0.19922	0.27500	0.12927	0.08630	0.09387
	0.25771	0.26300	0.32045	0.31874	0.21012	0.21578	0.17291	0.21445	0.11382	0.15641
	0.18339	0.59381	0.30929	0.31042	0.20178	0.20621	0.28594	0.17205	0.08336	0.10926
	0.32362	0.30193	0.32918	0.32638	0.20091	0.20968	0.23123	0.23521	0.08746	0.17944
100	0.27346	-0.1577	-0.0119	-0.0079	0.09120	-0.6969	-0.1080	-0.1226	-0.1792	-0.1701
	0.08856	-0.1350	0.27118	-0.0692	0.09365	0.02942	-0.0939	-0.0505	-0.1498	-0.1456
	-0.0809	0.24169	0.12132	0.33320	0.01504	-0.1221	-0.1777	-0.3236	-0.1658	-0.1778
	0.07137	0.23540	0.22498	0.19970	0.40215	-0.2488	0.00761	-0.3078	-0.1728	-0.1701
	0.21168	0.78928	-0.0088	0.14744	-0.1830	-0.1095	-0.0877	-0.2862	-0.1321	-0.1659
	-0.2866	0.07840	-0.1062	0.00452	0.09633	0.16270	-0.0675	-0.0458	-0.1685	-0.1679
	-0.0133	-0.0311	0.08798	-0.0349	-0.0793	0.20293	-0.2489	-0.3521	-0.1459	-0.1578
	0.18441	0.13234	0.05807	0.26271	0.07937	0.06078	-0.0918	-0.2249	-0.1644	-0.1614
	0.01088	0.24507	-0.0712	0.10960	-0.5102	0.42442	-0.1463	-0.1769	-0.1385	-0.1475
	0.08561	0.13741	0.1128	-0.0009	0.09289	-0.1436	-0.0251	-0.2123	-0.1716	-0.1705
225	-0.2474	-0.0576	-0.2154	-0.1365	-0.1354	-0.1636	-0.1659	-0.1886	-0.2245	-0.2222
	-0.0415	-0.1121	-0.2575	-0.0110	-0.0971	-0.1729	-0.1969	-0.2105	-0.2085	-0.2468
	-0.1102	-0.0213	-0.0359	-0.1933	-0.2602	-0.2233	-0.2674	-0.2705	-0.2329	-0.2358
	-0.2765	-0.0342	-0.1808	-0.2051	-0.2023	-0.2459	-0.2062	-0.1783	-0.2309	-0.2377
	-0.2507	-0.0124	-0.1942	-0.1785	-0.1806	-0.2092	-0.1919	-0.1897	-0.2025	-0.2383
	-0.1165	-0.2651	-0.1920	-0.1437	-0.1366	-0.2768	-0.2045	-0.2165	-0.2242	-0.1581
	-0.0202	-0.0421	-0.2209	-0.1788	-0.2305	-0.2129	-0.2311	-0.2438	-0.2197	-0.2369
	-0.1109	-0.1214	-0.1077	-0.1959	-0.2379	-0.1821	-0.2171	-0.1714	-0.2175	-0.2143
	-0.0037	-0.0699	-0.1823	-0.2024	-0.2233	-0.1875	-0.2305	-0.2598	-0.2092	-0.2309
	-0.2245	-0.1274	-0.1107	-0.2469	-0.1695	-0.2286	-0.1647	-0.3012	-0.2331	-0.2348
400	-0.0137	-0.0642	-0.2102	-0.0136	-0.1977	-0.2268	-0.1995	-0.3258	-0.2443	-0.2588
	-0.2124	-0.3183	-0.1091	-0.0255	-0.2554	-0.1912	-0.2618	-0.1489	-0.2476	-0.2136
	-0.1921	-0.2735	-0.0221	-0.1718	-0.2192	-0.2369	-0.2840	-0.3511	-0.3156	-0.2919
	-0.0269	-0.0839	-0.2688	-0.2669	-0.2989	-0.2761	-0.2760	-0.1911	-0.3329	-0.1659
	-0.2495	-0.0112	-0.1262	-0.2659	-0.0085	-0.1912	-0.2110	-0.2412	-0.3115	-0.3335
	-0.1229	-0.1691	-0.2159	-0.4599	-0.1613	-0.1343	-0.2669	-0.1378	-0.2484	-0.1359
	-0.0932	-0.1071	-0.2131	-0.0901	-0.2319	-0.2346	-0.3031	-0.2014	-0.2301	-0.2381
	-0.1603	-0.1101	-0.2321	-0.0499	-0.2094	-0.2125	-0.1283	-0.2874	-0.1972	-0.1791
	-0.0679	-0.2867	-0.2534	-0.1379	-0.2731	-0.2512	-0.2581	-0.1215	-0.1524	-0.2980
	-0.0249	-0.2018	-0.2819	-0.2671	-0.2204	-0.2088	-0.1911	-0.2608	-0.2714	-0.1358

The test statistic is D, the largest absolute difference between the sample cumulative probability distribution and the theoretical cumulative distribution. The null hypothesis is accepted if D < $D_{\alpha,\text{ number of observations}}$.

The Kolmogorov-Smirnov test was conducted at the .05 level of significance for each combination of sample size and variance (Appendix H contains the SAS program for the K-S test). For each of the 20 sets that were tested, the D values are non-significant, i.e. they are less than the critical value of .294. It is concluded that the populations of the RMS differences are all normally distributed.

4.4.1.2 Test for Homoscedasticity

The Bartlett test was used to test whether the 20 populations of RMS differences have equal variances. The assumptions underlying this test are that each of the populations is normal and that samples are randomly obtained and independent. Normality was tested in Section 4.4.1.1 and the manner in which the training and recall sets were generated insured randomness and independence.

The hypotheses for the Bartlett test are:

$H_0 : \sigma_1{}^2 = \sigma_2{}^2 = \ldots = \sigma_r{}^2$

H_1 : Not all $\sigma_i{}^2$ are equal

where r refers to the number of normal populations. If $s_i{}^2$ denotes the sample variance from the ith population and df_i denotes the degrees of freedom associated with the sample variance $s_i{}^2$, then the weighted arithmetic average of the sample variances using the associated degrees of freedom as weights is the mean square error

$$\mathrm{MSE} = \frac{1}{df_T} \sum df_i s_i^2$$

where $df_T = \sum df_i$

Similarly, the weighted geometric average of the s_i^2, denoted by GMSE, is

$$\mathrm{GMSE} = [(s_1^2)^{df_1}(s_2^2)^{df_2}...(s_r^2)^{df_r}]^{\frac{1}{df_T}}.$$

The following relation holds between these two averages : GMSE ≤ MSE and the two averages are equal if all the sample variances are equal. A value close to 1 for the ratio MSE/GMSE is evidence that the population variances are equal.

The Bartlett test uses the test statistic

$$\mathrm{B} = \frac{df_T}{C}(\log_e MSE - \log_e GMSE)$$

which has a χ^2 distribution with r-1 degrees of freedom. C is calculated as

$$\mathrm{C} = 1 + \frac{1}{3(r-1)}\left[(\sum \frac{1}{df_i}) - \frac{1}{df_T} \right].$$

The computational form of the test statistic is

$$\mathrm{B} = \frac{1}{C}[(df_T)\log_e MSE - \sum(df_i)\log_e s_i^2].$$

The decision rule is to reject H_0 if B is greater than the χ^2 at 1 - α level of significance and r - 1 degrees of freedom.

The standard deviations for each of the 20 sets of RMS differences were calculated and Minitab was used to perform the Bartlett test. Appendix I shows the necessary Minitab commands. The test statistic is calculated to be 504.459, which is greater than the critical value of 30.14 ($\chi^2_{.95,19}$). It may be concluded that the population variances are not equal and that a two-way ANOVA cannot be used. Instead, a nonparametric procedure will have to be used to analyze the data. No attempt was made to stabilize the variances through transformation since subsequent interpretation of transformed data could be inaccurate.

As an aside, not being able to use ANOVA may result in the lack of control regarding the interdependence in the observed data. Since tremendous amount of time and effort has been spent in generating the data, this researcher decided to proceed with the nonparametric analysis and then ascertain whether the experimental design made any difference to the results.

4.4.2 Nonparametric Two-way Analysis

In order to analyze the data, NPSP, a Nonparametric Statistical package developed by Professor Walter R. Pirie of the Department of Statistics at VPI&SU, was used. Most nonparametric analyses of two-way layouts do not deal with interactions. NPSP is a routine for two-way layout with interactions and is based on the research by McKean and Hettmansperger (1976). Their method is similar in spirit to least squares so that the results can be easily interpreted. They use the notion of dispersion of residuals based on a linear combination of the residuals introduced by Jaeckel (1972) and consider the reduction in dispersion just as the least squares method treats the reduction in the sum of squares. In NPSP, Pirie uses the method of general linear models by rank dispersion, where the dispersion function $D_2(e) = \sum e_i$ Rank (e_i) is minimized instead of $D_1(e) = \sum e_i^2$. NPSP also constructs a table similar to the ANOVA table.

In a two-way layout with equal number of observations per cell, assuming that there are k treatments or factor 1 (variance in this study) and b blocks or factor 2 (sample size in this study) and n > 1 observations per cell, Pirie developed an F test statistic for use in NPSP.

Appendix J is the listing for the NPSP analysis on the RMS differences between backpropagation networks and regression models. The test statistic for interactions is 8.6710, which is greater than the critical value of $F_{.95,12,380} = 1.75$. Thus, at an α level of .05, there are significant interactions between sample size and variance. Although the main effects are also significant, the presence of interaction masks the true main effects.

The NPSP analysis was also conducted separately on the RMS values from regression and backpropagation. The test statistic value for interactions is 9.0843 for regression and 8.8725 for backpropagation. In both cases, the test statistic is greater than the critical value of 1.75, implying the presence of significant interaction between sample size and variance.

According to Gibbons (1971), in analysis of variance, there is no method of identifying which populations differ most if the H_0 is rejected. All the tests are for tests which are collectively significant. The alternative is a pairwise test but conducting multiple pairwise comparisons result in the problem of unknown overall probabilities for Type I and Type II errors.

4.4.3 Graphical Analysis

In an attempt to isolate the nature of the interactions, the average RMS for each combination of sample size and variance are plotted against levels of sample size for regression, backpropagation and the difference between the two techniques. In Figure 1 on page 51 and Figure 2 on page 52, the average RMS is plotted against sample size, for regression and backpropagation respectively. The plot for mean RMS difference between the two techniques against sample size is in Figure 3 on page 53.

Both Figure 1 on page 51 and Figure 2 on page 52, where the four curves represent the four variance levels, indicate that

1. All four curves are monotonically decreasing and the lack of overall parallelism suggests the presence of interaction.
2. Since the curves do not cross each other, the plots suggest that the interaction is not extreme.

According to Figure 3 on page 53, where the average RMS difference between the two techniques are plotted against sample size, in addition to the above comments, the following observations may be made

1. The parallelism of the curves at variances of 225 and 400 suggests the absence of interaction between sample size and variance at these two levels of variance.
2. At smaller sample sizes of 20 and 50, all four curves are parallel, indicating no interaction.
3. Similarly, at large sample sizes of 200 and 500, the four curves are almost parallel, suggesting lack of interaction.
4. Between sample sizes of 50 and 200, the curves for variance levels of 25 and 100 are not parallel. This suggests the presence of interaction.

At the variance level of 25, regression appears to do better than the backpropagation networks, regardless of sample size. This changes at the higher variances. Backpropagation networks perform at a better rate than regression as sample size increases, although the rate of improvement is not equal at the four levels of variance. The rate of improvement of the backpropagation network over regression appears to slow down as variance increases from 25 to 400, with the rate being almost equal at variance levels of 225 and 400.

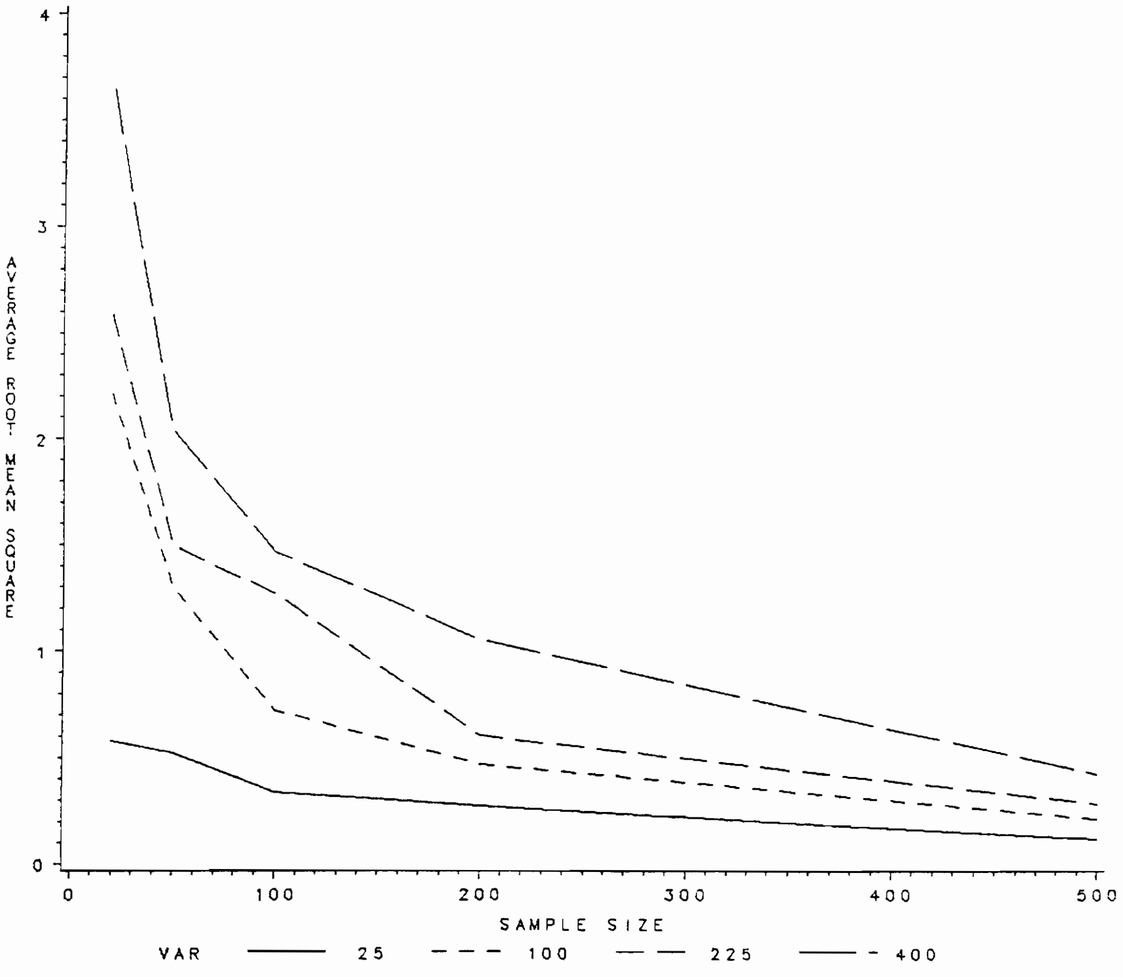


Figure 1. Plot of Average RMS for Regression - Variance Curves

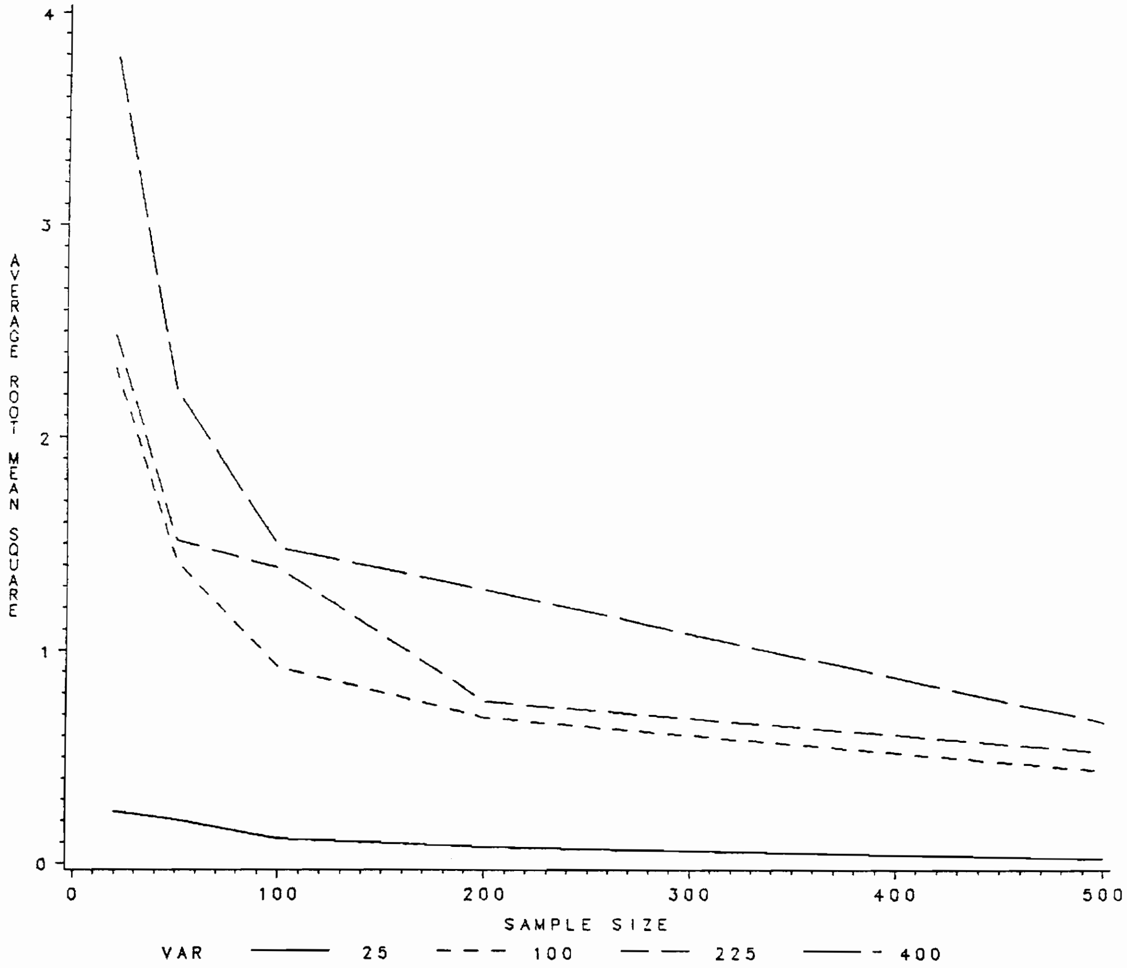


Figure 2. Plot of Average RMS for Backpropagation - Variance Curves

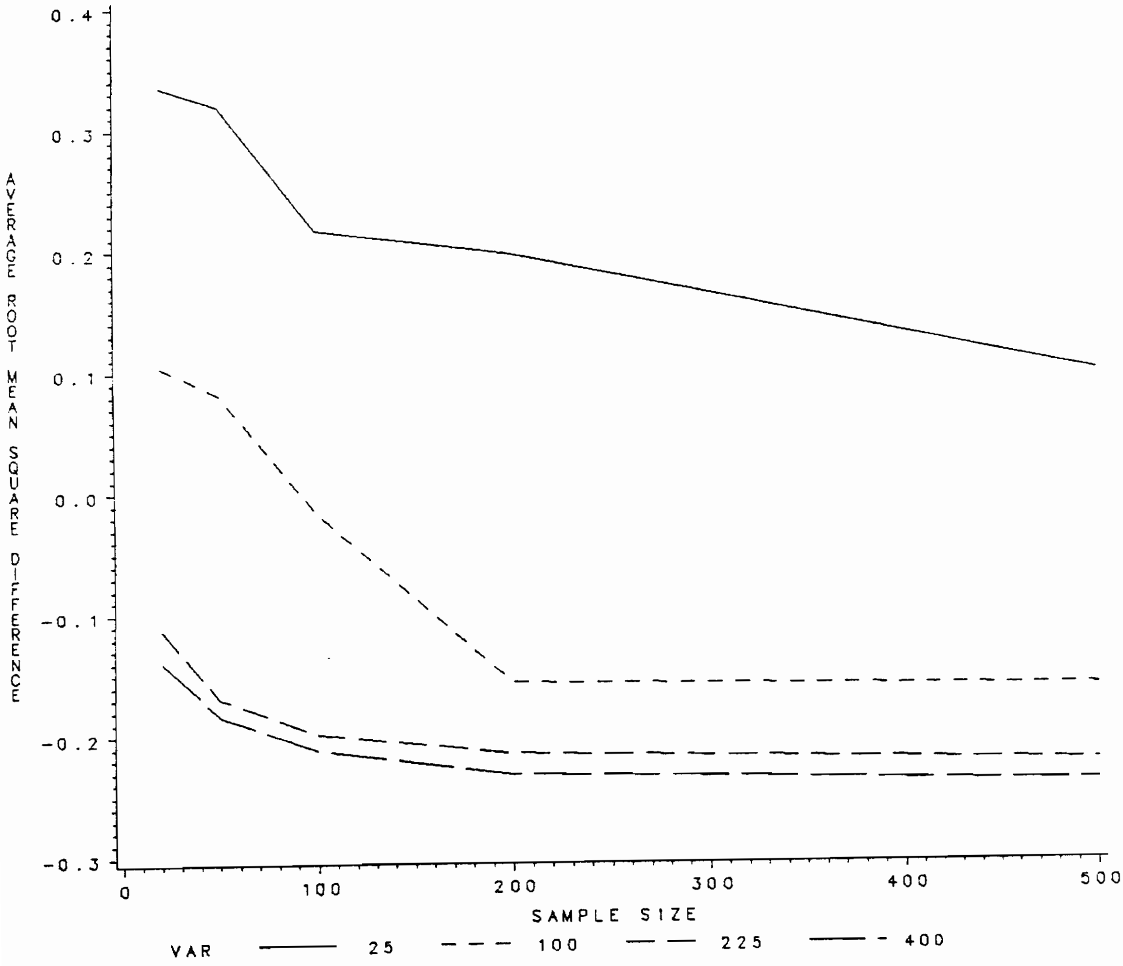


Figure 3. Plot of Average RMS Differences - Variance Curves

The interaction between sample size and variance is in effect between sample sizes of 50 and 200 for variance levels of 25 and 100. This means that as sample size increases from 50 to 200, when the variance of the error term is either 25 or 100, backpropagation networks appear to perform at a better rate than regression. At variance level of 100, although at sample sizes below 100, regression performs better, the two techniques predict at about the same level at sample size of 100 and the backpropagation network predicts more accurately than regression as sample size increases beyond 100. The graphical analysis raises the following questions:

1. Do regression and backpropagation perform significantly better with sample size of 500 than with sample size of 20, when variance is 25?
2. Do regression and backpropagation perform significantly better with sample size of 500 than with sample size of 20, when variance is 400?
3. Do regression and backpropagation perform significantly better when variance is 20 than when variance is 400, for sample size of 20?
4. Do regression and backpropagation perform significantly better when variance is 20 than when variance is 400, for sample size of 500?
5. Is the average RMS difference for sample size 100 and variance level 100 significantly different from 0?
6. For variance levels of 225 and 400, are the averages of the RMS differences at all sample sizes significantly less than 0?
7. Are the averages of the RMS differences for all sample sizes at variance of 25 significantly greater than 0?
8. At variance of 100, is the average RMS difference for sample sizes less than 100 significantly greater than 0?
9. At variance of 100, is the average RMS difference for sample sizes greater than 100 significantly less than 0?

Formal tests were required to answer these questions. Due to the absence of a nonparametric test for multiple pairwise comparisons, several one-sample tests were conducted, incurring the risk of compounding experiment-wise error. It was felt that these tests were necessary in order to be able to arrive at some sort of conclusion regarding the relative behavior of regression and backpropagation at different locations.

4.4.4 Wilcoxon Signed-Ranks Test

The Wilcoxon Signed-Ranks test is a nonparametric one-sample location problem, i.e. it is used to make inferences about the median of the population. The assumptions underlying this test are that the observations are independent and come from a continuous population symmetric about zero. The first assumption is valid for the data under consideration given the manner in which the data was generated and treated. The validity of the second assumption is insured by the normality of the underlying populations.

To perform the Wilcoxon test the following steps are required:

1. Set up the null hypotheses as $H_0 : \theta = 0$ where θ is the average RMS difference.
2. Alternatives can be $H_1 : \theta \neq 0$, or the one-tailed hypothesis $\theta < 0$ or $\theta > 0$.
3. Choose a level of significance α.
4. Rank all the RMS differences in the sample under consideration without regard to sign, giving the rank of 1 to the smallest difference.
5. After the ranking is completed, affix the sign of the RMS difference to each rank.
6. Let T^+ be the sum of all the positive ranks and T^- be the sum of all negative ranks. The test statistic T is the smaller of T^+ and T^-.
7. Reject the null hypothesis if T is $\leq$ the appropriate critical value from the Wilcoxon signed-ranks table.

At the risk of compounding the error, several individual tests are conducted in order to be able to arrive at some conclusions regarding the relative performances of backpropagation and regression under the conditions studied in this chapter.

For regression and backpropagation separately, the Wilcoxon signed-ranks test is carried out to determine whether the the average RMS at sample size of 500 is better than the average RMS at sample size of 20, when variance is 25. The lower one-tailed test is used where θ is the difference between the average RMS at size 500 and average RMS at size 20, so that a negative difference indicates that performance is better at sample size of 500. The test, conducted at an α level of .05, results in a test statistic value of 0 for both regression and backpropagation. Since the test statistics are 0, which is less than the critical value of 60 ($T_{.05,20}$), the decision is to reject the null hypothesis each time. At an α level of .05, the average RMS values for both techniques are significantly less than 0 for sample size of 500 than sample size of 20, at variance level of 25. Similar tests were carried out to determine whether the two techniques perform better with sample size of 500 than with sample size of 20 when variance is 400. Once again the test statistic values are 0, and the null hypothesis is rejected. At an α level of .05, the average RMS values for both techniques are significantly less than 0 for sample size of 500 than sample size of 20, at variance level of 400. This translates as both backpropagation networks and regression performing better with large sample size than with small sample size regardless of the variance level.

For each technique, the Wilcoxon signed-ranks test was also carried out to determine whether the the average RMS at variance of 225 is better than the average RMS at variance of 400, when sample size is 20 and 500. The lower one-tailed test was used where θ is the difference between the average RMS at variance 225 and average RMS at variance 400, for sample sizes of 20 and 500 separately. The tests, conducted at an α level of .05, result in test statistic values of 0 for both sample sizes for regression and backpropagation. Since the test statistics are 0, less than the critical value of 60 ($T_{.05,20}$), the decision is to reject the null hypothesis each time. At an α level of .05, the average RMS values for both techniques are significantly less than 0 for variance of 225 than variance of 400, at sample sizes of 20 as well as 500. Both backpropagation networks and regression perform better with smaller variance levels regardless of the sample size.

The two-tailed test at an α level of .10 was conducted for variance level 100 and sample size 100. The test statistic is 104, which is > the critical value of 60 ($T_{.10,20}$) and the decision is not to reject the null hypothesis. Thus the average RMS difference at variance 100 and sample size 100 appear not to be significantly different from 0, implying that, at this combination, the backpropagation network and regression predict at the same level of accuracy.

The lower one-tailed test at an α level of .05 was conducted for variance levels 225 and 400 at each sample size. The test statistics are 0, less than the critical value of 60 ($T_{.05,20}$) and the decision is to reject the null hypothesis each time. At an α level of .05, the average RMS differences are significantly less than 0 for all sample sizes at variance levels of 225 and 400. This translates as backpropagation networks performing better than regression at each sample size when variance is either 225 or 400.

The upper one-tailed test at an α level of .05 was conducted for variance level 25 across all sample sizes. The test statistics are again 0 and the decision is to reject the null hypothesis each time. The average RMS differences are significantly greater than 0 for all sample sizes at the variance level of 25. Thus, regression performs better than backpropagation networks for each sample size at variance of 25.

At the variance level of 100 and sample size of 20, the upper one-tailed test at an α level of .05 was conducted. The test statistic is calculated to be 52 and the decision is to reject the null hypothesis. Regression performs significantly better than the backpropagation network at variance 100 and sample size 20. At the same variance level but sample size of 500, the lower one-tailed test was conducted, resulting in the conclusion that backpropagation network outperforms regression.

Since the Wilcoxon signed-ranks test determined that the two techniques perform at the same level for the combination of sample size 100 and variance 100, subsequent analyses in this research use samples of size 100 where the variance of the error term is 100.

4.5 Conclusions

There are significant interactions between sample size and variance for regression analysis and backpropagation networks. Sample size and variance have an effect on the relative performances of regression and backpropagation, as determined by the RMS differences between the predictions made by the two techniques.

Both regression and backpropagation perform better when sample size is large as opposed to small, and at lower variance levels rather than higher variance levels.

Regression performs better than backpropagation for sample sizes between 20 and 500 at variance of 25. At variance of 100, regression outperforms backpropagation for sample sizes less than 100. For sample sizes greater than 100, at variance 100, backpropagation does better.

At sample size of 100 and variance of 100, regression and backpropagation perform at the same level.

Backpropagation networks perform better than regression at variance levels greater than 225, across all sample sizes.

In an effort to simplify the results and capture the essence of the conclusions, a qualitative summary is presented in Table 8 on page 59, where the sample size is categorized as small, medium or large and the variance is categorized as low, medium, or high. The R represents the fact that regression

Table 8. Qualitative Summary of Results

	Sample Size		
Variance	Small	Medium	Large
Low	R	R	R
Medium	R	S	B
High	B	B	B

performs better, the S stands for the two techniques performing the same, and the B represents backpropagation networks performing better.

As can be seen, regression appears to perform better at low variance levels while backpropagation performs better at high variance levels. For the medium variances, interaction between sample size and variance influences the performances of both techniques.

These results may seem counter intuitive. Since all the assumptions are satisfied, the OLS estimators of α and β have minimum variance among all unbiased estimators. It would seem that, while at low variance, regression outperforms backpropagation, as the variance increases, the two techniques would perform at about the same level.

Due to the fact that there is much about the working of backpropagation networks that is still lacking in theoretical explanation, the reason it performs better than regression at high variances can only be hypothesized. Some possible explanations for this counter intuitive result are attempted.

The experimental design shown in Table 1 on page 33 may have resulted in interdependence among the observations across the training sets for any one recall set. The effect is not clear since the analysis was conducted on the RMS differences and not the observed data, $\hat{Y}$. This seems to be a concern in theory but not in practice, since all the RMS differences for high variances are negative. This means that even if one considered only the recall sets for one of the training sets, and thus eliminated interdependence, the results appear the same.

The OLS estimates for the regression parameters were poor when variance was high. For example, in the case of sample size 20 and variance 400, the α and β were estimated to be -14.3692 and 0.6814 respectively. The true values for α and β were 10 and 0.8 respectively. Selected R^2 values from regression are shown in Table 9. The poor estimates of the parameters and the low R^2 values for high variances indicate that the data is so scattered that estimation is very difficult. In that sense, neither technique can perform well, and it may not be a significant finding that backpropagation outperforms regression.

The very manner in which backpropagation networks are used could be a possible explanation. While regression gets a ″snapshot″ look at the training data in order to estimate the parameters, backpropagation sees several replications of the data and is able to improve its ″understanding″ of the underlying function by adjusting and readjusting the weights. Although both are trained on the same sample size, backpropagation ″sees″ the data many more times.

In determining the learning count for each sample size, preliminary experimentation was conducted for the backpropagation. The learning count was determined to be the point where the RMS was lowest. Overtraining a network results in deterioration of performance, a condition refered to as the ″grandmother effect″. Knowing when to stop training may have given backpropagation an unfair edge over regression.

Finally, the backpropagation network may ″estimate″ the underlying function in a manner similar to biased estimation. This could explain why OLS estimates, with the minimum variance of all unbiased estimates, could not outperform backpropagation at high variances.

All of the above explanations are hypotheses and further research is needed to be able to arrive at more concrete explanations. The purpose of this chapter was to determine data characteristics that would provide a common point for regression and backpropagation so that subsequent experiments would not be biased one way or another. This has been accomplished and the point of reference is sample size of 100 and variance 100.

Table 9. Selected R-square values from Regression

	Sample Sizes				
σ^2	20	50	100	200	500
25	0.6645	0.6791	0.8692	0.9848	0.9946
100	0.5188	0.6428	0.7271	0.7508	0.7947
225	0.2844	0.3371	0.3634	0.4172	0.4892
400	0.2599	0.2765	0.3167	0.4027	0.4238

Chapter 5: Outliers, Skewness, and Kurtosis

This chapter investigates the performance of simple linear regression and backpropagation network under conditions when the error term is not normally distributed. Specifically, this chapter considers the cases of outliers, skewness, and kurtosis. A separate section deals with each case. Section 5.1 compares the two techniques when the error distribution is normal, has moderate outliers and extreme outliers. In Section 5.2, the comparison involves the normal error term and error distributions with positive and negative skewness. In the last section, backpropagation network and regression are compared when error distributions are peaked, normal, and flat. Based on the results of Chapter 4, sample size of 100 and variance of 100 were used as the standard, i.e. the normal distribution. From each population, five training sets of size 100 and four recall sets of size 100 were taken. The training sets were used to train networks as well as estimate α and β in regression. The recall sets were used to determine how well the two techniques predict the output given the training they receive. Each of the four recall sets were recalled on each of the five training sets, resulting in 20 output sets for each characteristic under consideration.

The predicted output, $\hat{Y}$, from each technique were compared to the corresponding actual output, Y, for each observation in each of the 20 output sets and a root mean square (RMS) was calculated for each of the 20 output sets for each characterization.

5.1 Error Term with Outliers

The populations were generated according to the form

$$Y_i = \alpha + \beta X_i + \varepsilon_i$$

where X is the regressor variable, Y is the measured response variable, and α and β are the intercept and slope respectively. The ε_i is the model error. The X_i are random variables generated from a normal distribution of mean 50 and variance of 225 to obtain X values within the range 0 to 99. A β value of 0.8 and α of 10 are chosen to be consistent with the previous chapter. The ε_i are random variables having a normal distribution with mean zero and variance of 100 for the normal case. In the case of moderate and extreme outliers, the ε_i were obtained from two distributions. For moderate outliers, 95 percent of the ε_i were generated from a normal distribution with mean 0 and variance 100 while 5 percent of the ε_i were generated from a normal distribution with mean 0 and variance 400. For extreme outliers, 5 percent of the ε_i were generated from a normal distribution with mean 0 and variance 900. A distinct population of 10,000 was generated using SAS for each level of outlier. Appendix L is a sample of the SAS program used to generate the populations.

Normal probability plots for the ε_i were studied to ascertain that the outliers did result in the error distribution being nonnormal. A plot that is linear suggests normality whereas departures from linearity suggests nonnormality. Appendix M shows the probability plots for the moderate outlier case.

5.1.1 Generation of Samples

From each population, five training sets of size 100 and four recall sets of size 100 were taken. The same procedure as in Chapter 4 was used. The training sets were used to train networks as well as estimate α and β in regression. The recall sets were used to determine how well the two techniques predict the output given the training they receive. Each of the four recall sets were recalled on each of the five training sets, resulting in 20 output sets for each level of outlier.

5.1.2 Regression Analysis

The five training sets for each level of outlier were used to estimate the regression parameters α and β Appendix C is the SAS program used for this purpose. The five sets of estimated α and β values, for each level of outlier, were then used along with X values from each of the appropriate recall sets to predict $\hat{Y}$ values using another SAS program (see Appendix D) based on the fitted model

$$\hat{Y}_i = a + bX_i .$$

Thus, for each level of outlier, 20 output sets were created using regression. Each of these 20 output files with X and $\hat{Y}$, along with the recall files which have the Y values, were evaluated by a BASIC program (Appendix E) that calculates the RMS for each output set for each level of outlier.

For each level of outlier, 20 RMS values were obtained. The results from regression are summarized in Table 10 on page 66.

A nonparametric one-way analysis of variance was conducted using the SAS routine NPAR1WAY (see Appendix K) to determine whether the average RMS values from regression analysis are equal for the normal, moderate outlier, and extreme outlier cases.

The hypotheses for the test were formulated as:

H_0 : The average RMS are equal for all three cases

Table 10. Root Mean Squares from Regression - Outliers

Normal		Mild Outlier		Extreme Outlier	
1.10997	1.92825	1.26102	1.45195	1.62731	1.89341
1.77370	0.98430	1.92258	1.62785	2.54932	2.13856
1.00325	1.69432	1.74852	1.38178	2.30673	1.79561
1.52321	1.60100	1.64731	1.83862	2.16568	2.43231
1.23681	1.53725	1.56998	1.37773	2.05790	1.78997
1.64231	1.42106	1.15402	1.66282	1.47819	2.18730
1.35217	1.48923	1.17605	1.48902	1.50889	1.94532
1.71428	1.37211	1.45088	1.70212	1.89192	2.24206
1.82750	1.32896	2.11690	1.94359	2.82014	2.57860
1.58435	1.52008	1.03440	1.29640	1.31148	1.67662

H_1 : At least one average RMS is different

The test was conducted at the .05 level of significance. The result of the Nonparametric One-way Anova is an F-value of 16.649 and a p-value of 0.0001, and therefore the null hypothesis is rejected. The presence of outliers does make a difference to the performance of regression.

In order to determine whether regression performs better when there are no outliers than when there are either moderate or extreme outliers, the Wilcoxon signed-ranks test was used. The null hypotheses were set up as $H_0 : \theta \geq 0$ where θ is the difference between average RMS from the normal case and the average RMS from either the moderate or the extreme outlier case. The alternatives were of the form $H_1 : \theta < 0$.

The test statistic value is 72 for the difference between normal and moderate outlier case, > the critical value of 60 ($T_{.05,20}$), and the decision is not to reject the null hypothesis. There is no difference in the performance for regression when there are no outliers and when there are moderate outliers.

In the test between normal and extreme outliers, the test statistic value is 14, which is < the critical value, and the decision is to reject the null hypothesis. The presence of extreme outliers does deteriorate the performance of regression.

5.1.3 Neural Network Analysis

The analysis in this chapter uses a three-layer backpropagation network with two processing elements in the hidden layer, one processing element each in the input and output layers and a sigmoidal transfer function. The networks were trained according to the supervised method for 35,000 times. The X and Y values were represented as real numbers and the MinMax table was utilized to facilitate learning and recall.

Each of the five training sets results in a network which was then recalled on each of the four recall sets, resulting in 20 output sets for each outlier level. The RMS for each output set was calculated using the same BASIC program as regression (Appendix E). Thus, for the moderate as well as the extreme outlier case, there were 20 RMS values corresponding to the 20 RMS values obtained from regression analysis. The RMS values for the normal case of sample size 100 and variance 100 were obtained from the previous chapter, for both regression and backpropagation network. The neural network results are summarized in Table 11 on page 69.

A nonparametric one-way analysis of variance was conducted using the SAS routine NPAR1WAY (see Appendix K) to determine whether the average RMS values from backpropagation are equal for the normal, moderate outliers, and extreme outliers cases.

The hypotheses for the test were formulated as:

H_0 : The average RMS are equal for all three cases

H_1 : At least one average RMS is different

The test was conducted at the .05 level of significance. The result of the nonparametric one-way ANOVA is an F-value of 0.307 and a p-value of 0.7366, and therefore the null hypothesis is not rejected. The presence of outliers does not make a difference in the performance of backpropagation.

5.1.4 Comparison of Neural Networks and Regression

A SAS program (Appendix G) is used to calculate the difference between the RMS from neural network, RMS_N, and the RMS from regression, RMS_R. This difference is calculated as RMS_N - RMS_R for each observation in each output set for each outlier case. A negative difference in the

Table 11. Root Mean Squares from Backpropagation Networks - Outliers

Normal		Mild Outlier		Extreme Outlier	
1.20117	1.23128	1.60450	1.33177	1.75343	1.51304
1.86735	1.01372	1.15368	1.41276	1.35607	1.58442
1.01829	1.57218	1.35078	1.07954	1.52979	1.29072
1.92536	1.35218	1.29375	1.18417	1.47952	1.38294
1.05380	1.42775	1.37163	1.28344	1.54817	1.47044
1.73864	1.58376	1.12252	1.50861	1.32860	1.66891
1.27285	1.69216	1.39617	1.21183	1.56980	1.40732
1.79365	1.43289	1.20775	1.51950	1.40373	1.67851
1.31726	1.75338	1.61984	1.96309	1.76695	1.18808
1.67724	1.37651	1.82327	1.65433	1.06484	0.91593

RMS indicates that neural networks have a lower RMS and therefore do a better job in predicting the Y while a positive difference implies that regression gives a better prediction of Y. The RMS differences are summarized in Table 12 on page 71.

A nonparametric one-way analysis of variance was conducted using the SAS routine NPAR1WAY (see Appendix K) to determine whether the average RMS differences are equal for all three levels - i.e. normal errors, moderate outliers, and extreme outliers.

The hypotheses for the test were formulated as:

H_0 : The average RMS differences are equal for all three
H_1 : At least one average RMS difference is different

The test was conducted at the .05 level of significance.

The result of the nonparametric one-way ANOVA is an F-value of 14.592 and a p-value of 0.0001, and therefore the null hypothesis is rejected. The presence of outliers does make a difference in the relative performances of regression and neural networks.

In Chapter 4, the Wilcoxon signed-ranks test was used to determine that for sample size of 100 and variance 100 in the normal case, both regression and neural networks performed at the same level. This test was also used to determine if the average difference between the two techniques is significantly different from 0 in the case of moderate outliers as well as in the case of extreme outliers. The null hypotheses were set up as $H_0 : \theta = 0$ where θ is the average RMS difference for moderate and extreme outliers. The alternatives were $H_1 : \theta \neq 0$. The test statistic values are 58 for moderate outliers and 3 for extreme outliers, and both are < the critical value of 60 ($T_{.10,20}$) and the decision is to reject the null hypothesis in both cases, at $\alpha = .10$. Thus the average RMS difference when there are any type of outliers is significantly different from 0, implying that the backpropagation network and regression do not predict at the same level of accuracy.

Table 12. Root Mean Square Differences - Outliers

Normal		Mild Outlier		Extreme Outlier	
0.09120	-0.6969	0.34348	-0.1202	0.12612	-0.3803
0.09365	0.02942	-0.7689	-0.2151	-1.1933	-0.5541
0.01504	-0.1221	-0.3977	-0.3022	-0.7769	-0.5049
0.40215	-0.2488	-0.3536	-0.6545	-0.6862	-1.0494
-0.1830	-0.1095	-0.1984	-0.0943	-0.5097	-0.3195
0.09633	0.16270	-0.0315	-0.1542	-0.1496	-0.5184
-0.0793	0.20293	0.22012	-0.2772	0.06091	-0.5380
0.07937	0.06078	-0.2431	-0.1826	-0.4882	-0.5636
-0.5102	0.42442	-0.4971	0.01950	-1.0532	-1.3905
0.09289	-0.1436	0.78887	0.35793	-0.2466	-0.7607

The lower one-tailed test at an α level of .05 was conducted for the both outlier cases. The test statistic values are less than the critical value of 60 ($T_{.05,20}$) and the decision is to reject the null hypothesis again. At an α level of .05, the average RMS differences are significantly less than 0 for both moderate and extreme outlier cases. Thus, backpropagation networks appear to perform better than regression when there are outliers.

5.2 Skewed Error Distributions

Skewness can be described as departure from symmetry of a frequency curve. In order to generate populations with error terms that are positively and negatively skewed, the beta distribution is used to represent the skewed error distribution. Skewness of a distribution is measured by

$$\beta_1 = \frac{\mu_3{}^2}{\mu_2{}^3}$$

where the μ_2 is variance and μ_3 is the third central moment. A positively skewed distribution with $\sqrt{\beta_1}$ = 1 and a negatively skewed distribution with $\sqrt{\beta_1}$ = -1 are used, emulating the research conducted by Professor Krutchkoff. The normal distribution is perfectly symmetrical with β_1 = 0.

A random variable is said to have a beta distribution with parameters α and β, $\alpha > 0$, $\beta > 0$, if its probability density function is

$$f(x) = \frac{x_{\alpha-1}(1-x)^{\beta-1}}{B(\alpha, \beta)}$$

if $0 < x < 1$ and $f(x) = 0$ otherwise. The coefficient of skewness for the beta distribution is

$$\frac{2(\beta - \alpha)(\alpha + \beta + 1)^{.5}}{(\alpha + \beta + 2)(\alpha\beta)^{.5}}.$$

Using these formula to obtain the appropriate skewness, the populations for this section were generated. The form for the population is

$$Y_i = \alpha + \beta X_i + \varepsilon_i$$

where X is the regressor variable, Y is the measured response variable, and α and β are the intercept and slope respectively. The ε_i is the model error. The X_i are random variables generated from a normal distribution of mean 50 and variance of 225 to obtain X values within the range 0 to 99. A β value of 0.8 and α of 10 are chosen to be consistent with the previous chapter. The ε_i are random variables with beta distribution. In order to determine the appropriate α and β values for the beta distribution that will result in skewness of + 1 and - 1, several combinations of the two beta distribution parameters, along with the formula for skewness were tried. Appendix N is the SAS program used for this purpose. The appropriate α and β values are tabulated in Table 13 on page 74.

The normal case (from Chapter 4) is compared with the case of positively skewed error distribution as well as negatively skewed error distribution. A distinct population of 10,000 was generated using SAS for each type of skewness. Appendix O is a sample of the SAS program used to generate the populations.

Normal probability plots for the ε_i were studied to ascertain that skewness did result in the error distribution being nonnormal. For both positive and negative skewness, the plots are nonlinear, suggesting nonnormality.

Table 13. Beta Distribution Parameters and Skewness

α	β	Skewness
1.5	0.5	-1
0.5	1.5	+1

5.2.1 Generation of Samples

From each population, five training sets of size 100 and four recall sets of size 100 were taken. The same procedure as in Chapter 4 was used. The training sets were used to train networks as well as estimate α and β in regression. The recall sets were used to determine how well the two techniques predict the output given the training they receive.

5.2.2 Regression Analysis

The five training sets for each type of skewness were used to estimate the regression parameters α and β Appendix C is the SAS program used for this purpose. The five sets of estimated α and β values for each type of skewness were then used along with X values from each of the appropriate recall sets to predict $\hat{Y}$ values using another SAS program (see Appendix D) based on the fitted model

$$\hat{Y}_i = a + bX_i .$$

Thus, for each type of skewness, 20 output sets were created using regression. Each of these output files with X and $\hat{Y}$, along with the recall file which has the Y values, were evaluated by a BASIC program (Appendix E) that calculates the RMS for each output set for type of skewness.

For each type of skewness, 20 RMS values were obtained. The results from regression are summarized in Table 14 on page 76.

A nonparametric one-way analysis of variance was conducted using the SAS routine NPAR1WAY (see Appendix K) to determine whether the average RMS values from regression analysis are equal for the normal, positively skewed, and negatively skewed cases.

Table 14. Root Mean Squares from Regression - Skewness

Normal		Negative Skewness		Positive Skewness	
1.10997	1.92825	1.10920	1.30297	1.86292	1.93531
1.77370	0.98430	1.78058	1.14511	1.24086	1.73710
1.00325	1.69432	1.60393	1.48148	1.33098	1.57046
1.52321	1.60100	1.50123	1.23175	1.16796	1.43802
1.23681	1.53725	1.42274	1.69537	1.77889	2.07087
1.64231	1.42106	1.00062	1.22765	1.28783	1.05146
1.35217	1.48923	1.02297	1.51697	1.65255	1.55019
1.71428	1.37211	1.30188	1.34059	1.85630	1.39830
1.82750	1.32896	1.97778	1.55684	1.47292	1.05595
1.58435	1.52008	0.87923	1.80190	1.56687	1.64547

The hypotheses for the test were formulated as:

H_0 : The average RMS are equal for all three cases

H_1 : At least one average RMS is different

The test was conducted at the .05 level of significance. The result of the nonparametric one-way ANOVA is an F-value of 0.607 and a p-value of 0.5483, and therefore the null hypothesis is not rejected. Skewness does not make a difference in the performance of regression.

5.2.3 Neural Network Analysis

A three-layer backpropagation network with two processing elements in the hidden layer, one processing element each in the input and output layers and a sigmoidal transfer function is used. The networks were trained according to the supervised method for 35,000 times. The X and Y values were represented as real numbers and the MinMax table was utilized to facilitate learning and recall.

Each of the five training sets resulted in a network which was then recalled on each of the four recall sets, resulting in 20 output sets for each type of skewness. The RMS for each output set was calculated using the same BASIC program as regression (Appendix E). Thus, for the positive as well as the negative skewness case, there were 20 RMS values corresponding to the 20 RMS values obtained from regression analysis. The RMS values for the normal case of sample size 100 and variance 100 were obtained from the previous chapter, for both regression and backpropagation network. The neural network results are summarized in Table 15 on page 78.

Table 15. Root Mean Squares from Backpropagation Networks - Skewness

Normal		Negative Skewness		Positive Skewness	
1.20117	1.23128	1.83507	1.54458	1.44602	1.62011
1.86735	1.01372	1.35489	1.63084	1.26489	1.72336
1.01829	1.57218	1.56482	1.27592	1.74099	1.92186
1.92536	1.35218	1.50407	1.38736	1.24266	1.91247
1.05380	1.42775	1.58703	1.49309	1.47564	1.40602
1.73864	1.58376	1.32169	1.73293	1.35926	1.53826
1.27285	1.69216	1.61317	1.41682	1.68231	1.69856
1.79365	1.43289	1.41248	1.74453	1.34933	1.73948
1.31726	1.75338	1.85140	1.15188	1.35357	1.33583
1.67724	1.37651	1.00296	0.82301	1.53766	1.71957

A nonparametric one-way analysis of variance was conducted using the SAS routine NPAR1WAY (see Appendix K) to determine whether the average RMS values from backpropagation are equal for the normal, positively skewed, and negatively skewed cases.

The hypotheses for the test were formulated as:

H_0 : The average RMS are equal for all three cases

H_1 : At least one average RMS is different

The test was conducted at the .05 level of significance. The result of the nonparametric one-way ANOVA is an F-value of 0.830 and a p-value of 0.4413, and therefore the null hypothesis is not rejected. Skewness does not make a difference in the performance of backpropagation.

5.2.4 Comparison of Neural Networks and Regression

A SAS program (Appendix G) is used to calculate the difference between the RMS from neural network, RMS_N, and the RMS from regression, RMS_R. This difference is calculated as RMS_N - RMS_R for each observation in each output set for each type of skewness. A negative difference in the RMS indicates that neural networks have a lower RMS and therefore do a better job in predicting the Y while a positive difference implies that regression gives a better prediction of Y. The RMS differences are summarized in Table 16 on page 80.

A nonparametric one-way analysis of variance was conducted using the SAS routine NPAR1WAY (see Appendix K) to determine whether the average RMS differences are equal for all three levels - i.e. normal errors, positively skewed errors and negatively skewed errors.

The hypotheses for the test were formulated as:

Table 16. Root Mean Square Differences - Skewness

Normal		Negative Skewness		Positive Skewness	
0.09120	-0.6969	0.72587	0.24161	-0.4169	-0.3152
0.09365	0.02942	-0.4257	0.14936	0.02403	-0.0137
0.01504	-0.1221	-0.0391	0.04417	0.41001	0.35140
0.40215	-0.2488	0.00284	-0.3080	0.07470	0.47445
-0.1830	-0.1095	0.16429	0.26544	-0.3033	-0.6649
0.09633	0.16270	0.32107	0.21596	0.07143	0.48680
-0.0793	0.20293	0.59020	0.07623	0.02976	0.14837
0.07937	0.06078	0.11060	0.18769	0.49303	0.34118
-0.5102	0.42442	-0.1264	-0.6500	-0.1194	0.27988
0.09289	-0.1436	0.12373	-0.3221	-0.0292	0.07410

H_0 : The average RMS differences are equal for all three

H_1 : At least one average RMS difference is different

The test was conducted at the .05 level of significance.

The result of the nonparametric one-way ANOVA is an F-value of 0.527 and a p-value of 0.5932, and therefore the null hypothesis cannot be rejected. Neither positive nor negative skewness in the error distribution makes a difference to the relative performances of regression and neural networks.

In Chapter 4, the Wilcoxon signed-ranks test was used to determine that for sample size of 100 and variance 100 in the normal case, both regression and neural networks perform at the same level. Since the mean differences for positive and negative skewness are not significantly different from the normal case, it may be concluded that both techniques appear to perform at the same level when the error distribution is skewed.

5.3 Error Distributions with Kurtosis

Kurtosis is described as the degree of flattening of a frequency curve. In order to generate populations with error terms that are highly peaked, i.e. leptokurtic, and flat or platykurtic, the beta distribution is used to represent the error distribution with the appropriate characteristic. Kurtosis of a distribution is measured by

$$\beta_2 = \frac{\mu_4}{\mu_2^2}$$

where the μ_2 is variance and μ_4 is the fourth central moment. A leptokurtic distribution with $\beta_2 > 3$ whereas a platykurtic distribution with $\beta_2 < 3$ are used, in the same manner as Professor Krutchkoff. The normal distribution is said to be mesokurtic with β_2 of 3.

The coefficient of kurtosis for the beta distribution is

$$\frac{3(\alpha+\beta)(\alpha+\beta+1)(\alpha+1)(2\beta-\alpha)}{(\alpha\beta)(\alpha+\beta+2)(\alpha+\beta+3)}+\frac{\alpha(\alpha-\beta)}{(\alpha+\beta)}.$$

The populations for this section were generated according to the form

$$\mathrm{Y_i} = \alpha + \beta \mathrm{X_i} + \varepsilon_\mathrm{i}$$

where X is the regressor variable, Y is the measured response variable, and α and β are the intercept and slope, respectively. The ε_i is the model error. The X_i are random variables generated from a normal distribution of mean 50 and variance of 225 to obtain X values within the range 0 to 99. A β value of 0.8 and α of 10 are chosen to be consistent with the previous chapter. The ε_i are random variables from a beta distribution. In order to determine the appropriate α and β values that will result in kurtosis of 2 and 4, several combinations of the two beta distribution parameters along with the formula for kurtosis were tried. Appendix N is the SAS program used for this purpose. The appropriate α and β values are tabulated in Table 17 on page 83.

The normal case (from Chapter 4) is compared with both cases of kurtosis. A distinct population of 10,000 was generated using SAS for each degree of kurtosis. Appendix O is a sample of the SAS program used to generate the populations.

Table 17. Beta Distribution Parameters and Kurtosis

α	β	Kurtosis
1.5	1.5	2
10000	10000	4

5.3.1 Generation of Samples

From each population, five training sets of size 100 and four recall sets of size 100 were taken. The same procedure as in Chapter 4 was used. The training sets were used to train networks as well as estimate α and β in regression. The recall sets were used to determine how well the two techniques predict the output given the training they receive.

5.3.2 Regression Analysis

The five training sets for each degree of kurtosis were used to estimate the regression parameters α and β. Appendix C is the SAS program used for this purpose. The five sets of estimated α and β values for each degree of kurtosis were then used along with X values from each of the appropriate recall sets to predict $\hat{Y}$ values using another SAS program (see Appendix D) based on the fitted model:

$$\hat{Y}_i = a + bX_i$$

Thus, for each degree of kurtosis, 20 output sets were created using regression. Each of these output files with X and $\hat{Y}$, along with the recall files which had the Y values, were evaluated by a BASIC program (Appendix E) that calculates the RMS for each output set for each degree of kurtosis.

For each degree of kurtosis, 20 RMS values were obtained. The results from regression are summarized in Table 18 on page 85.

A nonparametric one-way analysis of variance was conducted using the SAS routine NPAR1WAY (see Appendix K) to determine whether the average RMS values from regression analysis are equal for the normal, leptokurtic, and platykurtic cases.

Table 18. Root Mean Squares from Regression - Kurtosis

Normal		Platykurtic		Leptokurtic	
1.10997	1.92825	1.46615	1.61322	1.15267	0.99340
1.77370	0.98430	1.74856	1.63894	1.19252	1.69461
1.00325	1.69432	1.08169	1.66356	1.41713	2.11302
1.52321	1.60100	1.56594	1.32185	1.91086	1.66228
1.23681	1.53725	1.87845	1.58823	1.42014	1.29794
1.64231	1.42106	1.71756	1.63648	1.16788	1.63240
1.35217	1.48923	1.45726	1.54782	1.33010	1.49103
1.71428	1.37211	1.32382	1.74701	1.41908	0.99298
1.82750	1.32896	1.54299	1.42305	1.35762	1.46221
1.58435	1.52008	1.50035	1.41063	1.71536	1.53278

The hypotheses for the test were formulated as:

H_0 : The average RMS are equal for all three cases

H_1 : At least one average RMS is different

The test was conducted at the .05 level of significance. The result of the nonparametric one-way ANOVA is an F-value of 0.772 and a p-value of 0.4669, and therefore the null hypothesis is not rejected. Kurtosis does not make a difference in the performance of regression.

5.3.3 Neural Network Analysis

A three-layer backpropagation network with two processing elements in the hidden layer, one processing element each in the input and output layers and a sigmoidal transfer function is used. The networks were trained according to the supervised method for 35,000 times. The X and Y values were represented as real numbers and the MinMax table was utilized to facilitate learning and recall.

Each of the five training sets result in a network which was then recalled on each of the four recall sets, resulting in 20 output sets for each degree of kurtosis. The RMS for each output set was calculated using the same BASIC program as regression (Appendix E). Thus, for the leptokurtic as well as the platykurtic case, there were 20 RMS values corresponding to the 20 RMS values obtained from regression analysis. The RMS values for the normal case of sample size 100 and variance 100 were obtained from the previous chapter, for both regression and backpropagation network. The neural network results are summarized in Table 19 on page 87.

Table 19. Root Mean Squares from Backpropagation Networks - Kurtosis

Normal		Platykurtic		Leptokurtic	
1.20117	1.23128	1.62466	1.50820	1.30159	1.16113
1.86735	1.01372	1.54746	1.62819	1.62007	1.57828
1.01829	1.57218	1.51632	1.13523	1.43932	0.96625
1.92536	1.35218	1.58836	1.21084	1.50923	1.74633
1.05380	1.42775	1.25475	1.60276	1.45396	1.05160
1.73864	1.58376	1.48758	1.45648	1.39749	1.46877
1.27285	1.69216	1.11911	1.51221	1.32768	1.72197
1.79365	1.43289	1.38270	1.67200	1.16466	1.49065
1.31726	1.75338	1.46877	2.06205	1.75064	1.18811
1.67724	1.37651	1.26092	1.53805	1.22926	1.12420

A nonparametric one-way analysis of variance was conducted using the SAS routine NPAR1WAY (see Appendix K) to determine whether the average RMS values from backpropagation are equal for the normal, leptokurtic, and platykurtic cases.

The hypotheses for the test were formulated as:

H_0 : The average RMS are equal for all three cases
H_1 : At least one average RMS is different

The test was conducted at the .05 level of significance. The result of the nonparametric one-way ANOVA is an F-value of 0.850 and a p-value of 0.4329, and therefore the null hypothesis is not rejected. Kurtosis does not make a difference in the performance of backpropagation.

5.3.4 Comparison of Neural Networks and Regression

A SAS program (Appendix G) is used to calculate the difference between the RMS from neural network, RMS_N, and the RMS from regression, RMS_R. This difference is calculated as RMS_N - RMS_R for each observation in each output set for each degree of kurtosis. A negative difference in the RMS indicates that neural networks have lower RMS and therefore do a better job in predicting the Y while a positive difference implies that regression gives a better prediction of Y. The RMS differences are summarized in Table 20 on page 89.

A nonparametric one-way analysis of variance was conducted using the SAS routine NPAR1WAY (see Appendix K) to determine whether the average RMS differences are equal for all three levels - i.e. normal errors, leptokurtic errors and platykurtic errors.

The hypotheses for the test were formulated as:

Table 20. Root Mean Square Differences - Kurtosis

Normal		Platykurtic		Leptokurtic	
0.09120	-0.6969	0.15851	-0.2404	0.15252	-0.0314
0.09365	0.02942	-0.0658	-0.0108	0.62667	-0.1163
0.01504	-0.1221	0.43463	-0.5283	0.02219	-1.1468
0.40215	-0.2488	0.02242	-0.1110	-0.4016	0.08405
-0.1830	-0.1095	-0.6237	0.01453	0.03382	-0.2463
0.09633	0.16270	-0.2299	-0.1800	0.22961	-0.1636
-0.0793	0.20293	-0.3382	-0.0356	-0.0024	0.23094
0.07937	0.06078	0.05888	-0.0750	-0.2544	0.49767
-0.5102	0.42442	-0.0742	0.63900	0.39302	-0.2741
0.09289	-0.1436	-0.2394	0.12742	-0.4861	-0.4086

H_0 : The average RMS differences are equal for all three

H_1 : At least one average RMS difference is different

The test was conducted at the .05 level of significance.

The result of the nonparametric one-way ANOVA is an F-value of 0.142 and a p-value of 0.8683, and therefore the null hypothesis cannot be rejected. Neither flatness nor peakedness of the error distribution makes a difference in the relative performances of regression and neural networks.

In Chapter 4, the Wilcoxon signed-ranks test was used to determine that for sample size of 100 and variance 100 in the normal case, both regression and neural networks performed at the same level. Since the mean differences for platykurtic and leptokurtic distributions are not significantly different from the normal case, it is concluded that both techniques appear to perform at the same level whether the error distribution is flat or peaked.

5.4 Conclusions

The presence of moderate outliers does not affect the performance of regression. Extreme outliers cause the performance of regression to deteriorate. For backpropagation networks, the presence of outliers have no effect on performance. In fact, backpropagation appear to perform significantly better than regression in the presence of outliers.

Since the skewness and kurtosis experiments were carried out for extreme values that cover the range of possible values, it can be concluded that neither skewness nor kurtosis has any effect on the individual performances of regression or backpropagation.

Regression analysis and backpropagation networks perform at the same level of accuracy in predicting when the error term is skewed positively or negatively, or when the error term is flat or peaked.

Chapter 6: Multicollinearity

This chapter investigates the performance of multiple regression analysis and backpropagation network under conditions where there is simple correlation between the regressor variables. Multicollinearity exists when the regressor variables are not independent, i.e. near linear dependencies exist between regressor variables. In other words, multicollinearity is when the regressors move with one another as well as with the response variable. A regression coefficient is a rate of change or partial derivative of the response with respect to a regressor variable. When the X′s are moving with one another, it is difficult to get a precise estimate of Y because the ordinary least squares (OLS) procedure cannot clearly separate the rates of change of each X with Y.

Multicollinearity prohibits precise statistical inference. It lowers the precision in the estimation of regression coefficients by inflating the variances of the coefficients. Instability of the regression coefficients may effect the quality of fit and prediction of the regression model.

In this chapter, the relative performances of regression analysis and backpropagation network are explored in the presence of zero, low, moderate, and high multicollinearity of the data. To this end, correlation coefficients (between two regressor variables) of -.9, -.5, -.1, 0, .1, .5, and .9 are considered.

6.1 Generation of Data

In order to study the effect of multicollinearity, the following linear model with two regressor variables is considered

$$Y = \alpha + \beta X_1 + \gamma X_2 + \varepsilon_i$$

where X_1 and X_2 are the regressor variables, Y is the measured response variable, and α is the intercept. The β and γ are the slope coefficients and ε_i is the model error. The X_i are random variables generated from normal distributions. To obtain X values within the range 0 to 99, X_1 was generated with mean 50 and variance 225 while X_2 was generated with mean 45 and variance 225. In order to insure the correlation between the X's, X_1 was generated from a seed and then X_2 was generated from a linear combination, using the correlation coefficient, of the seeds for the two X's (see Appendix P for the SAS program). The population generated was then tested to check whether the correlation between the X's is what it is supposed to be. An α value of 10 is chosen to be consistent with the previous chapters while β and γ values of 5 and 7, respectively, are used to insure that both X's have a strong linear relationship with Y. The ε_i are random variables with a normal distribution with mean zero and variance of 100.

Varying degrees of multicollinearity are studied by generating data sets with correlation of -.9, -.5, -.1, 0, .1, .5, and .9 between the two regressor variables. A distinct population of 10,000 was generated using SAS for each level of multicollinearity.

6.2 Generation of Samples

From each population, five training sets of size 100 and four recall sets of size 100 were taken. The same procedure as in Chapter 4 was used. The training sets were used to train networks as well as estimate α, β and γ in regression. The recall sets were used to determine how well the two techniques predict the output given the training they receive. Each of the four recall sets was recalled on each of the five training sets, resulting in 20 output sets for each level of multicollonearity.

The predicted output, $\hat{Y}$, from each technique were compared to the corresponding actual output, Y, for each observation in each of the output sets combination and a root mean square (RMS) is calculated for each output set.

6.3 Regression Analysis

The effect of multicollinearity is instability of regression coefficients. The coefficients are very much dependent on the particular data set that generated them, resulting in OLS estimates that have very large variances.

According to Myers (1986), a researcher faced with multicollinearity in the data being studied has the following recourses:

1. Removal of one or more regressors. The presence of multicollinearity suggests that, in the case of k regressor variables, the actual model should involve fewer than k variables. The choice of variables for elimination is arrived upon by the researcher's knowledge of the phenomena being modeled or by diagnostics such as variance inflation factors and eigenvalues (Myers, 1986).

2. Transformation of the regressor variables. In the case where X_1 and X_2 are two regressors that are highly correlated, a new variable is defined as a function such as $X_1 + X_2$ or $\frac{X_1}{X_2}$. Such transformations reduce the dimensionality of the regressors and retain some of the informational content. One has to be careful of forming functions of regressor variables that do not make sense, e.g. add variables that are measured in different units.

3. Biased estimation of regression coefficients. OLS provides unbiased estimates that have the minimum variance of all linear unbiased estimators, but there is no upper bound on the variance of the estimators. Since multicollinearity can produce large variances, using OLS under these conditions results in a high price for unbiasedness. Biased estimation reduces the variance and stabilizes the regressor coefficients. Ridge regression is the most commonly used biased estimation technique. In ridge regression, the bias is incorporated by adding a small positive constant k, called the shrinkage parameter, to the main diagonal of the correlation matrix. The success of ridge regression in reducing the variance of the estimators lies in the choice of k. Much research has been conducted in procedures for choosing k (Draper et al., 1979).

4. Use OLS. The quality of prediction depends on where the X's are in the regressor space. If the X's are within the range where multicollinearity is characterized by the data, prediction will not be severely affected. In fact, any combination in the data range that reflect similar linear dependencies as those experienced in the data are likely to yield reasonably good prediction. According to Professor Raymond Myers, an expert in the field of regression analysis, if population multicollinearity is well depicted in the data, prediction will not be severely affected for X's in the data range. Unlike real life problems, where population multicollinearity is seldom well defined, this research uses data from populations that are generated by the researcher.

OLS method is used for this research. The five training sets for each level of multicollinearity were used to estimate the regression parameters α, β, and γ. The five sets of estimated α, β, γ values for each level of multicollinearity were then used along with X_1 and X_2 values from each of the appropriate recall sets to predict $\hat{Y}$ values using another SAS program based on the fitted model

$$\hat{Y}_i = a + bX_1 + gX_2 .$$

Thus, for each level of multicollinearity, 20 output sets were created using regression. Each of these output files with the X's and $\hat{Y}$, along with the recall files which have the Y values, were evaluated by a BASIC program (Appendix E) that calculates the RMS for each output set.

For each level of multicollinearity, 20 RMS values were obtained. The results from regression are summarized in Table 21 on page 97.

A nonparametric one-way analysis of variance is conducted using the SAS routine NPAR1WAY (see Appendix K) to determine whether the average RMS values from regression analysis are equal for all levels of multicollinearity.

The hypotheses for the test were formulated as:

H_0 : The average RMS are equal for all seven levels
H_1 : At least one average RMS is different

The test was conducted at the .05 level of significance. The result of the nonparametric one-way ANOVA is an F-value of 36.86 and a p-value of 0.0001, and therefore the null hypothesis is rejected.

Since there is a difference among the average RMS values, separate tests were conducted to determine whether there is a difference among the positively correlated sets and among the negatively correlated sets. Zero correlation is included in both sets.

The hypotheses for the test were formulated as:

Table 21. Root Mean Squares from Regression - Multicollinearity

$R_{X1,X2}$	Root Mean Squares				
-0.9	1.13035	1.09028	1.17891	1.19393	1.13096
	1.15256	1.24173	1.23836	1.25205	1.21055
	1.14542	1.18101	1.18731	1.17421	1.18649
	1.27006	1.13813	1.21170	1.12198	1.17785
-0.5	1.28684	1.21217	1.21197	1.28424	1.13807
	1.21990	1.18903	1.27428	1.14346	1.28029
	1.20230	1.11497	1.14017	1.18804	1.17703
	1.25696	1.27824	1.19756	1.15804	1.22931
-0.1	1.02142	1.06473	1.13718	1.23688	1.10737
	1.09668	1.25200	1.26901	1.13805	1.08458
	1.10022	1.22492	1.34606	1.11924	1.16807
	1.25075	1.24661	0.97093	1.25191	1.16629
0	1.06489	1.10142	1.03059	1.09040	1.08736
	1.09583	1.02402	1.07059	1.03360	1.10340
	1.10455	1.11628	1.08718	0.97928	1.01655
	1.11046	1.01682	1.05484	1.06680	1.07011
0.1	1.18471	1.17990	1.20335	1.28458	1.11233
	1.25974	1.07590	1.08111	1.13333	1.20024
	1.15571	1.25263	1.13225	1.15232	1.18740
	1.25044	1.24212	1.09983	1.23189	1.24450
0.5	1.20545	1.25647	1.29907	1.26638	1.28966
	1.26785	1.22838	1.21245	1.26902	1.22011
	1.26049	1.23629	1.20593	1.27345	1.23474
	1.26612	1.20828	1.29592	1.27081	1.29351
0.9	1.36487	1.38069	1.26504	1.37317	1.28392
	1.31528	1.20999	1.32291	1.28493	1.30816
	1.28929	1.36441	1.32323	1.39748	1.26779
	1.31396	1.30429	1.35834	1.28964	1.30043

H_0 : The average RMS are equal for positive correlation

H_1 : At least one average RMS is different

At α of .05, the F is 106.021 and the p-value is 0.0001, resulting in the rejection of the null hypothesis.

The test was repeated for negative correlation and the results of the nonparametric one-way ANOVA is an F of 19.556 and a p-value of 0.0001. The null hypothesis is again rejected.

For negative as well as positive correlation, the average RMS values are the not the same. Thus, multicollinearity does make a difference in the performance of regression.

A test to see if the average RMS values are the same for weak multicollinearity and no multicollinearity was also conducted. The hypotheses were formulated as:

H_0 : The average RMS values are equal for weak and zero multicollinearity

H_1 : At least one average RMS value is different

The test was conducted at the .05 level of significance. The result of the nonparametric one-way ANOVA is an F-value of 16.099 and p-value is 0.0001, and therefore the null hypothesis is rejected.

Weak multicollinearity does make a difference in the performance of regression.

In order to determine whether regression performs better when there is zero multicollinearity than when there is either a strong positive, strong negative, moderate positive, moderate negative, weak positive, or weak negative multicollinearity, separate Wilcoxon signed-ranks tests were used. The null hypotheses are set up as $H_0 : \theta \geq 0$ where θ is the average RMS difference between the levels of multicollinearity. The alternative was $H_1 : \theta < 0$.

The test statistic values for the separate tests are displayed in Table 22 on page 100. In each case, the test statistic value is < the critical value of 60 ($T_{.05,20}$) and the decision is to reject the null hypothesis at $\alpha = .05$.

Thus the average RMS when there is no multicollinearity is significantly less than 0, implying that regression predicts at a better level of accuracy when there is no correlation between the regressors than when there is any degree of multicollinearity.

6.4 Neural Network Analysis

The analysis in this chapter uses a three-layer backpropagation network with two processing elements in the hidden layer, one processing element in the output layer, two processing elements in the input layer, and a sigmoidal transfer function. The networks were trained according to the supervised method for and the best learning count (using the method outlined in Chapter 4) was determined to be 40,000. The X_1, X_2, and Y values were represented as real numbers and the MinMax table was utilized to facilitate learning and recall.

Each of the training sets results in a network that was then recalled on each of the four recall sets, resulting in 20 output sets for each multicollinearity level. The RMS for each output set was calculated using the same BASIC program as regression (Appendix E). Thus, for each multicollinearity level, there were 20 RMS values corresponding to the 20 RMS values obtained from regression analysis. The neural network results are summarized in Table 23 on page 101.

A nonparametric one-way analysis of variance was conducted using the SAS routine NPAR1WAY (see Appendix K) to determine whether the average RMS values from backpropagation are equal for all levels of multicollinearity.

Table 22. Summary of Tests and Wilcoxon T Values

Test	Wilcoxon T
0 better than .9	0
0 better than -.9	0
0 better than .5	0
0 better than -.5	0
0 better than .1	0
0 better than -.1	22

Table 23. Root Mean Squares from Backpropagation Networks - Multicollinearity

$R_{X1,X2}$	Root Mean Squares				
-0.9	1.05917	1.05997	0.99382	1.11677	1.00530
	1.04837	1.07880	0.98722	1.09075	1.03119
	1.03977	0.94824	0.99530	1.08807	1.07703
	1.13945	1.15623	1.04292	1.10649	1.06467
-0.5	1.05140	1.04048	1.07587	1.02956	1.13838
	1.07418	1.06040	1.10228	1.06579	1.07376
	1.08380	1.01326	1.12501	1.01929	1.08511
	0.94798	1.05590	1.03286	0.97812	1.11809
-0.1	1.05762	1.05459	1.10148	1.09184	1.02472
	1.09739	0.98053	1.08681	1.06044	1.07116
	1.06003	1.09373	1.03173	0.97238	1.06920
	1.06059	1.06751	1.07683	1.02068	1.03632
0	0.99630	1.01558	1.05147	1.05210	1.05017
	1.03894	0.98919	0.94668	1.06567	1.10445
	1.00983	1.08169	1.07818	1.05131	1.00550
	1.07160	1.02124	1.05665	1.05085	1.10383
0.1	1.01684	1.06168	1.02690	0.99505	1.01768
	1.02464	1.07902	1.11513	1.03965	1.04728
	1.08887	1.11034	1.17345	0.94174	1.09595
	1.02373	1.03327	1.11046	0.99227	1.05290
0.5	0.96818	1.03561	1.07333	1.10998	1.00749
	1.07453	1.05569	1.06705	1.03770	1.04191
	1.02827	0.99854	1.10242	1.02079	1.08485
	1.09766	1.11132	1.04636	1.01437	1.05486
0.9	1.10034	1.04648	1.08113	0.95899	1.01891
	1.07241	1.03217	1.01730	0.96567	1.01311
	1.01466	1.08598	1.01623	1.05333	1.00447
	1.04061	1.02134	1.08479	0.99394	1.14651

The hypotheses for the test were formulated as:

H_0 : The average RMS are equal for all seven levels

H_1 : At least one average RMS is different

The test was conducted at the .05 level of significance. The result of the nonparametric one-way ANOVA is an F-value of 0.651 and a p-value of 0.6892, and therefore the null hypothesis is not rejected.

Since there appears to be no difference among the average RMS values, separate tests were not required.

Thus backpropagation appears to predict at the same level of accuracy whether or not there is correlation between the regressors.

6.5 Comparison of Neural Networks and Regression

In order to compare the relative performances of the two techniques, the average RMS values for backpropagation network and regression analysis are plotted separately for the positive and negative correlation values, in Figure 4 and Figure 5 respectively, with zero being included in each set. In the positive as well as the negative case, it is apparent that the performance of regression deteriorates as the magnitude of the correlation is increased.

On the other hand, for both positive and negative correlation, backpropagation performs at the same level regardless of the magnitude of the correlation. Whether or not there is any multicollinearity, backpropagation seems to perform better than regression.

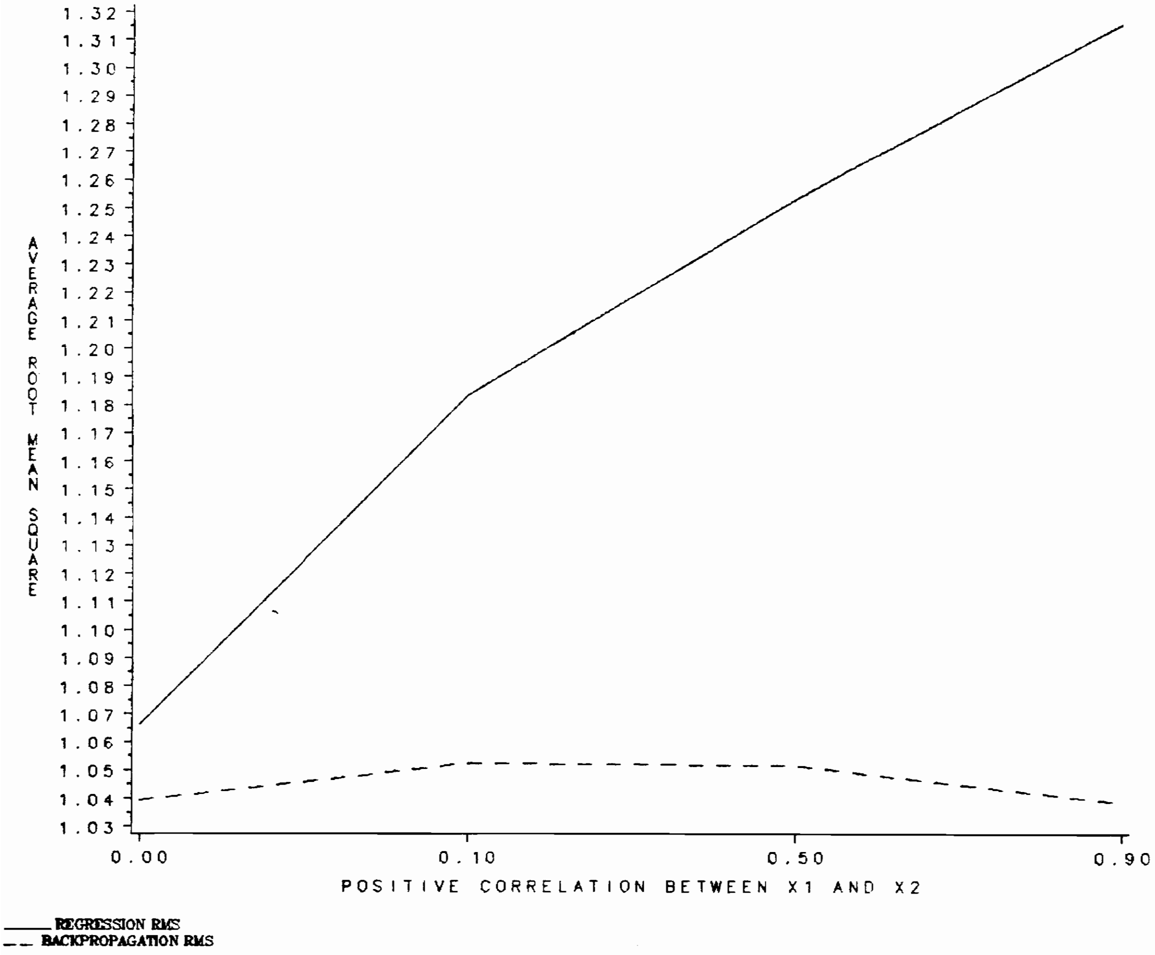


Figure 4. Comparison of the two techniques when there is positive correlation

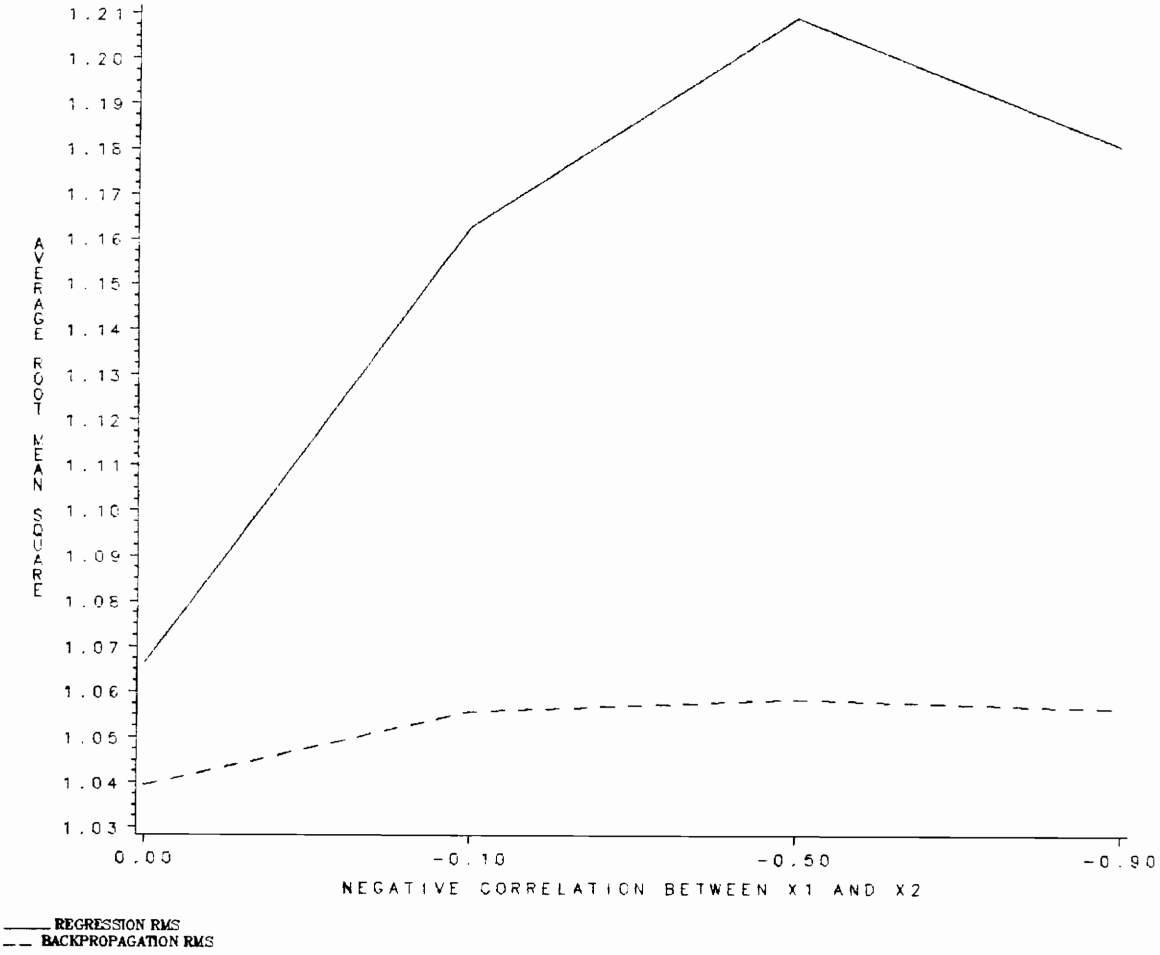


Figure 5. Comparison of the two techniques when there is negative correlation

A SAS program (Appendix G) is used to calculate the difference between the RMS from neural network, RMS_N, and the RMS from regression, RMS_R. This difference is calculated as RMS_N - RMS_R for each observation in each output set for each level of multicollinearity. A negative difference in the RMS indicates that neural networks have a lower RMS and therefore do a better job in predicting the Y while a positive difference implies that regression gives a better prediction of Y. The RMS differences are summarized in Table 24 on page 106.

A nonparametric one-way analysis of variance was conducted using the SAS routine NPAR1WAY (see Appendix K) to determine whether the average RMS differences are equal for all levels of multicollinearity - i.e. strong negative correlation, moderate negative correlation, weak negative correlation, no correlation, weak positive correlation, moderate positive correlation, and strong positive correlation.

The hypotheses for the test were formulated as:

H_0 : The average RMS differences are equal for all seven
H_1 : At least one average RMS difference is different

The test was conducted at the .05 level of significance.

The result of the nonparametric one-way ANOVA is an F-value of 21.443 and a p-value of 0.0001, and therefore the null hypothesis is rejected.

Since there appears to be a difference among the average RMS differences, separate tests were warranted to determine whether there is a difference among the positively correlated sets and among the negatively correlated sets. Zero correlation is included in both sets. The hypotheses for these two tests were formulated as:

H_0 : The average RMS differences are equal for all four

Table 24. Root Mean Square Differences - Multicollinearity

$R_{X1,X2}$	Root Mean Squares				
-0.9	-0.07118	-0.03031	-0.18509	-0.07716	-0.12566
	-0.10419	-0.16293	-0.25114	-0.16130	-0.17936
	-0.10565	-0.23277	-0.19201	-0.08614	-0.10946
	-0.13061	0.01810	-0.16878	-0.01549	-0.11318
-0.5	-0.23544	-0.17169	-0.13610	-0.25468	0.00031
	-0.14572	-0.12863	-0.17200	-0.07767	-0.20653
	-0.11850	-0.10171	-0.01516	-0.16875	-0.09192
	-0.30898	-0.22234	-0.16470	-0.17992	-0.11122
-0.1	0.03620	-0.01014	-0.03570	-0.14504	-0.08265
	0.00071	-0.27147	-0.18220	-0.07761	-0.01342
	-0.04019	-0.13119	-0.31433	-0.14686	-0.09887
	-0.19016	-0.17910	0.10590	-0.23123	-0.12997
0	-0.06859	-0.08584	0.02088	-0.03830	-0.03719
	-0.05689	-0.03483	-0.12391	0.03207	0.00105
	-0.09472	-0.03459	0.00900	0.07203	-0.01105
	-0.09286	0.00442	0.00181	-0.01595	0.03372
0.1	-0.16787	-0.11822	-0.17645	-0.28953	-0.09465
	-0.23510	0.00312	0.03402	-0.09368	-0.15296
	-0.06684	-0.14229	0.04120	-0.21058	-0.09145
	-0.22671	-0.20885	0.01063	-0.23962	-0.19160
0.5	-0.23727	-0.22086	-0.22574	-0.15640	-0.28217
	-0.19322	-0.17269	-0.14540	-0.23132	-0.17820
	-0.23222	-0.23775	-0.10351	-0.25266	-0.14989
	-0.16846	-0.09696	-0.24956	-0.25644	-0.23865
0.9	-0.26813	-0.33421	-0.18391	-0.41418	-0.26501
	-0.24287	-0.17782	-0.30561	-0.31926	-0.29505
	-0.27463	-0.27843	-0.30700	-0.34415	-0.26332
	-0.27335	-0.28295	-0.27355	-0.29570	-0.15932

H_1 : At least one average RMS difference is different

The tests were each conducted at the .05 level of significance.

The results of the nonparametric one-way ANOVA are F values of 49.986 and 9.317 for the positive and negative cases respectively. The p-value is 0.0001 for both, and therefore the null hypothesis is rejected in each case.

A test to see if the average RMS differences are the same for weak multicollinearity and no multicollinearity was also conducted. The hypotheses were formulated as:

H_0 : The average RMS differences are equal for weak and zero multicollinearity
H_1 : At least one average RMS difference is different

The test was conducted at the .05 level of significance.

The result of the nonparametric one-way ANOVA is an F-value of 7.648 and p-value of 0.0011, and therefore the null hypothesis is rejected in each case. The presence of multicollinearity does make a difference in the relative performances of regression and neural networks.

In order to determine whether there is a difference between regression and backpropagation when there is no multicollinearity, the Wilcoxon signed-ranks test was used. This test was used to determine if the average difference between the two techniques is significantly different from 0 in the case of zero collinearity. The null hypotheses were set up as $H_0 : \theta = 0$ where θ is the average RMS difference. The alternative was $H_1 : \theta \neq 0$. The test statistic value is 32 which is < the critical value of 60 ($T_{.10,20}$) and the decision is to reject the null hypothesis at $\alpha = .10$. Thus the average RMS difference when there is no multicollinearity is significantly different from 0, implying that the backpropagation network and regression do not predict at the same level of accuracy. This result is surprising as one would expect similar performances from both techniques. A possible explanation could be that the inherent nature of training in backpropagation results in its exposure to more data and therefore its apparently better performance.

The lower one-tailed test at an α level of .05 was conducted to determine if backpropagation performs better than regression at zero multicollinearity. The test statistic is once again < the critical value of 60 ($T_{.05,20}$) and the decision is to again reject the null hypothesis. At an α level of .05, the average RMS difference is significantly less than 0 in the case of zero multicollinearity. Thus, backpropagation networks appear to perform better than regression when there is no multicollinearity.

In order to determine whether there is a difference between regression and backpropagation when there is strong negative and strong positive multicollinearity, separate Wilcoxon signed-ranks tests were used. The null hypotheses were set up as $H_0 : \theta = 0$ where θ is the average RMS difference. The alternative was $H_1 : \theta \neq 0$. The test statistic values are 2 and 0 ,for the strong negative and positive cases respectively, and since both are < the critical value of 60 ($T_{.10,20}$), the decision is to reject the null hypothesis at $\alpha = .10$ for both cases. Thus the average RMS difference when there is either strong negative or strong positive multicollinearity is significantly different from 0, implying that the backpropagation network and regression do not predict at the same level of accuracy, for either case.

The lower one-tailed test at an α level of .05 was conducted to determine if backpropagation performs better than regression at strong negative or strong positive multicollinearity. The test statistics are once again < the critical value of 60 ($T_{.05,20}$) and the decision is to reject the null hypothesis. At an α level of .05, the average RMS difference is significantly less than 0 in both cases of strong multicollinearity.

Thus, backpropagation networks appear to perform better than regression when there is no multicollinearity, strong negative multicollinearity, or strong positive multicollinearity.

6.6 Conclusions

Under the conditions studied, regression analysis performs better based on the RMS differences when there is no multicollinearity than when there is any correlation between the two regressors. Multicollinearity does not affect the performance of the backpropagation network. In fact, backpropagation outperforms regression analysis whether there is multicollinearity or not, when there are two regressors in the model. A possible explanation for the result at zero collinearity could be that the inherent nature of training in backpropagation results in its exposure to more data and therefore its apparently better performance.

Chapter 7: Nonlinearity and Underspecification

This chapter investigates the performance of regression analysis and backpropagation networks under two conditions, nonlinearity and underspecification of the regressor variables.

While SLR is adequate for many situations, there are many areas of engineering and the sciences where the experimental situation requires the use of nonlinear models. These models are nonlinear in the regression parameters, making the computation of parameter estimates by elementary matrix algebra (as in least squares) impossible. In fact, nonlinearity brings about complications in the development of least squares estimators.

Model underspecification results in poor prediction using regression analysis. Underspecifying refers to the failure to include some necessary variables and this condition causes bias in prediction.

The presence of nonlinearity in a regression model causes complications that render the use of OLS impossible and model underspecification results in poor prediction. Both these conditions require special treatment under regression analysis. Analysis with neural networks is less sensitive to these problems, as the learning procedure "discovers" the nonlinear aspects and self-specifies the model. This chapter looks at the relative performance of regression analysis and backpropagation networks under these complications.

7.1 Nonlinearity

Although nonlinear regression (NLR) refers to the nonlinearity in the regression parameters, this research uses the term "nonlinear" to include a model that is nonlinear in the relationship between the regressor variables. Two models are studied: a true nonlinear model, the exponential model, as well as a model with a quadratic term which is linear in the regression coefficients but nonlinear in the relationship between the regressors.

7.1.1 Model with Quadratic Term

Consider the following model which includes a quadratic term

$$Y_i = \alpha + \beta X_i + \theta X_i^2 + \varepsilon_i \, .$$

The X and X^2 are the regressor variables, Y is the measured response variable, and α is the intercept. The β and θ are the slope coefficients and ε_i is the model error.

The X values were generated from a normal distribution with mean 50 and variance 225 and X^2 values were then calculated. (see Appendix Q for the SAS program). The mean and variance for X are chosen to insure that X has values ranging from 0 to 99. An α value of 10 and β value of 5 are used, to be consistent with the previous chapters and to insure a strong linear relationship between X and Y. In order to capture the complexity caused by the presence of a quadratic term, θ values of -10, -5, -1, -.5, -.2, .2, .5, 1, 5, and 10 are used. The θ values ranged from .2 and -.2 to fifty times that magnitude at 10 and -10. The ε_i are random variables with a normal distribution with mean zero and variance of 100, since it is determined in Chapter 4 that both regression and back-propagation perform at the same level for error variance of 100. A distinct population of 10,000 was generated using SAS for each value of θ.

From each population, five training sets of size 100 and four recall sets of size 100 were taken. The same procedure as in Chapter 4 was used. The training sets were used to train networks as well as estimate α, β and θ in regression. The recall sets were used to determine how well the two techniques predict the output given the training they receive. Each of the four recall sets were recalled on each of the five training sets, resulting in 20 output sets for each level of θ.

The predicted output, $\hat{Y}$, from each technique were compared to the corresponding actual output, Y, for each observation in each output set and a root mean square (RMS) was calculated for each output set.

7.1.1.1 Regression Analysis

The presence of X^2 implies that there is correlation between the two terms in the regression model. Since the population is defined with this relationship between the X and X^2, according to Myers (1986), prediction will not be severely affected for X values in the data range. Unlike real life problems, where the population is seldom well defined, this research uses data from populations that are generated by the researcher.

OLS method is used for this research. The five training sets for each level of θ were used to estimate the regression parameters α, β, and θ. The five sets of estimated α, β, θ values were then used along with X and X^2 values from each of the appropriate recall sets to predict $\hat{Y}$ values using another SAS program based on the fitted model

$$\hat{Y}_i = a + bX_i + cX^2_i .$$

Thus, for each level of θ, 20 output sets were created using regression. Each of these output files with the X's, X^2's, and $\hat{Y}$'s and the recall files which have the Y values, were evaluated by a BASIC program (Appendix E) that calculates the RMS for each output set.

Each combination of training set and recall set resulted in a set of generated outputs which were compared to the actual Y values (see Appendix E), to calculate an RMS value. For each value of θ, 20 RMS values were obtained. These are the results of 20 analyses involving the five training sets and the four recall sets. The results from regression for the positive and negative θ values are summarized separately in Table 25 on page 114 and Table 26 on page 115.

A nonparametric one-way analysis of variance was conducted using the SAS routine NPAR1WAY (see Appendix K) to determine whether the average RMS values from regression analysis are equal for all positive θ values.

The hypotheses for the test were formulated as:

H_0 : The average RMS are equal for positive θ values

H_1 : At least one average RMS is different

The test was conducted at the .05 level of significance.

The result of the nonparametric one-way ANOVA is an F-value of 0.011 and a p-value of 0.9997, and therefore the null hypothesis is not rejected. For positive θ values, all the average RMS values appear to be the same.

The test was repeated for the negative θ values and the results of the nonparametric one-way ANOVA is an F of 0.006 and a p-value of 0.9999. The null hypothesis again is not rejected. For negative θ values, all the average RMS values appear to be the same.

Table 25. Root Mean Squares from Regression - Positive Quadratic Term

θ	Root Mean Squares				
0.2	0.98786	1.02313	1.11006	1.05562	1.07791
	1.01017	1.05921	1.09455	1.04493	1.00942
	0.96810	1.06208	0.97216	1.02997	1.02293
	1.06934	1.14595	1.11394	0.94600	0.99439
0.5	0.98598	1.02125	1.10818	1.05374	1.07603
	1.00829	1.05733	1.09267	1.04305	1.00754
	0.96622	1.06020	0.97028	1.02809	1.02105
	1.06746	1.14407	1.11206	0.94412	0.99251
1	0.98905	1.02432	1.11125	1.05681	1.07910
	1.01136	1.06040	1.09574	1.04612	1.01061
	0.96929	1.06327	0.97335	1.03116	1.02412
	1.07053	1.14714	1.11513	0.94719	0.99558
5	0.98623	1.02150	1.10843	1.05399	1.07628
	1.00854	1.05758	1.09292	1.04330	1.00779
	0.96647	1.06045	0.97053	1.02834	1.02130
	1.06771	1.14432	1.11231	0.94437	0.99276
10	0.98778	1.02305	1.10998	1.05554	1.07783
	1.01009	1.05913	1.09447	1.04485	1.00934
	0.96802	1.06200	0.97208	1.02989	1.02285
	1.06926	1.14587	1.11386	0.94592	0.99431

Table 26. Root Mean Squares from Regression - Negative Quadratic Term

θ	Root Mean Squares				
-0.2	0.97277	1.01547	1.12073	1.05481	1.08180
	0.99978	1.05916	1.10195	1.04187	0.99887
	0.94884	1.06263	0.95376	1.02376	1.01523
	1.07142	1.16419	1.12542	0.92209	0.98068
-0.5	0.97155	1.01425	1.11951	1.05359	1.08058
	0.99856	1.05794	1.10073	1.04065	0.99765
	0.94762	1.06141	0.95254	1.02254	1.01401
	1.07020	1.16297	1.12420	0.92087	0.97946
-1	0.97467	1.01737	1.12263	1.05671	1.08370
	1.00168	1.06106	1.10385	1.04377	1.00077
	0.95074	1.06453	0.95566	1.02566	1.01713
	1.07332	1.16609	1.12732	0.92399	0.98258
-5	0.97286	1.01556	1.12082	1.05490	1.08189
	0.99987	1.05925	1.10204	1.04196	0.99896
	0.94893	1.06272	0.95385	1.02385	1.01532
	1.07151	1.16428	1.12551	0.92218	0.98077
-10	0.97356	1.01626	1.12152	1.05560	1.08259
	1.00057	1.05995	1.10274	1.04266	0.99966
	0.94963	1.06342	0.95455	1.02455	1.01602
	1.07221	1.16498	1.12621	0.92288	0.98147

Thus, when there is a quadratic term in the model, the magnitude and sign of the coefficient of the quadratic term has no effect on the predictive ability of regression analysis. In other words, whether the quadratic influence is minor or pronounced, positive or negative, as long as it is confined to the relationship between the regressor variables but strictly linear in the coefficients, the degree of the quadratic term has little effect on predictive power.

7.1.1.2 Neural Network Analysis

The analysis in this chapter uses a three-layer backpropagation network with two processing elements in the hidden layer, one processing element in the output layer, two processing elements in the input layer, and a sigmoidal transfer function. The networks were trained using the best learning count for two inputs (determined to be 40,000 in Chapter 6). The X, X_2, and Y values were represented as real numbers and the MinMax table was utilized to facilitate learning and recall.

Each of the five training sets results in a network which was then recalled on each of the four recall sets, resulting in 20 output sets for each value of θ. The RMS for each output set was calculated using the same BASIC program as regression (Appendix E). Thus, for each θ value, there were 20 RMS values corresponding to the 20 RMS values obtained earlier from regression analysis. The neural network results, for positive and negative θ values are summarized in Table 27 on page 117 and Table 28 on page 118.

A nonparametric one-way analysis of variance was conducted using the SAS routine NPAR1WAY (see Appendix K) to determine whether the average RMS values from backpropagation networks are equal for all positive θ values.

The hypotheses for the test were formulated as:

H_0 : The average RMS are equal for positive θ values

Table 27. Root Mean Squares from Backpropagation - Positive Quadratic Term

θ	Root Mean Squares				
0.2	1.07224	1.04586	1.02863	1.05370	1.04770
	1.02146	1.04218	1.03158	1.04972	1.04119
	1.02562	1.06297	1.05209	1.03426	1.03386
	1.06402	1.07373	1.01020	0.99667	0.98033
0.5	1.06324	1.03686	1.01963	1.04470	1.03870
	1.01246	1.03318	1.02258	1.04072	1.03219
	1.01662	1.05397	1.04309	1.02526	1.02486
	1.05502	1.06473	1.00120	0.98767	0.97133
1	1.06417	1.03779	1.02056	1.04563	1.03963
	1.01339	1.03411	1.02351	1.04165	1.03312
	1.01755	1.05490	1.04402	1.02619	1.02579
	1.05595	1.06566	1.00213	0.98860	0.97226
5	1.07213	1.04575	1.02852	1.05359	1.04759
	1.02135	1.04207	1.03147	1.04961	1.04108
	1.02551	1.06286	1.05198	1.03415	1.03375
	1.06391	1.07362	1.01009	0.99656	0.98022
10	1.06781	1.04143	1.02420	1.04927	1.04327
	1.01703	1.03775	1.02715	1.04529	1.03676
	1.02119	1.05854	1.04766	1.02983	1.02943
	1.05959	1.06930	1.00577	0.99224	0.97590

Table 28. Root Mean Squares from Backpropagation - Negative Quadratic Term

θ	Root Mean Squares				
-0.2	1.05817	1.04197	1.03142	1.04679	1.04312
	1.02702	1.03973	1.03323	1.04435	1.03912
	1.02957	1.05248	1.04581	1.03487	1.03463
	1.05313	1.05908	1.02011	1.01182	1.00180
-0.5	1.05075	1.03457	1.02400	1.03937	1.03570
	1.01960	1.03231	1.02581	1.03693	1.03170
	1.02215	1.04506	1.03839	1.02745	1.02721
	1.04571	1.05166	1.01269	1.00440	0.99438
-1	1.05182	1.03564	1.02507	1.04044	1.03677
	1.02067	1.03338	1.02688	1.03800	1.03277
	1.02322	1.04613	1.03946	1.02852	1.02828
	1.04678	1.05273	1.01376	1.00547	0.99545
-5	1.05840	1.04222	1.03165	1.04702	1.04335
	1.02725	1.03996	1.03346	1.04458	1.03935
	1.02980	1.05271	1.04604	1.03510	1.03486
	1.05336	1.05931	1.02034	1.01205	1.00203
-10	1.05923	1.04305	1.03248	1.04785	1.04418
	1.02808	1.04079	1.03429	1.04541	1.04018
	1.03063	1.05354	1.04687	1.03593	1.03569
	1.05419	1.06014	1.02117	1.01288	1.00286

H_1 : At least one average RMS is different

The test was conducted at the .05 level of significance.

The result of the nonparametric one-way ANOVA is an F-value of 0.630 and a p-value of 0.6420, and therefore the null hypothesis is not rejected. The average RMS values are not significantly different for all positive θ values.

The test was repeated for the negative θ values and the results of the nonparametric one-way ANOVA is an F of 1.515 and a p-value of 0.2041. The null hypothesis again is not rejected and the average RMS values are not significantly different for all negative θ values.

In other words, when there is a quadratic term in the model representing a nonlinearity in the relationship between input variables, the magnitude and sign of the coefficient of the quadratic term has no effect on the performance of backpropagation networks.

7.1.1.3 Comparison of Neural Networks and Regression

A SAS program (Appendix G) is used to calculate the difference between the RMS from neural network, RMS_N, and the RMS from regression, RMS_R. This difference is calculated as RMS_N - RMS_R for each observation in each value of θ. A negative difference in the RMS indicates that neural networks have a lower RMS and therefore do a better job in predicting Y while a positive difference implies that regression gives a better prediction of Y. The RMS differences for the positive and negative quadratic cases are summarized in Table 29 on page 120 and Table 30 on page 121.

Table 29. Root Mean Square Differences - Positive Quadratic Term

θ	Root Mean Squares				
0.2	0.08438	0.02273	-0.0814	-0.0019	-0.0302
	0.01129	-0.0173	-0.0629	0.00479	0.03177
	0.05752	0.00089	0.07993	0.00429	0.01093
	-0.0053	-0.0722	-0.1037	0.05067	-0.0141
0.5	0.07726	0.01561	-0.0885	-0.0090	-0.0373
	0.00417	-0.0241	-0.0700	-0.0023	0.02465
	0.05040	-0.0062	0.07281	-0.0028	0.00381
	-0.0124	-0.0793	-0.1108	0.04355	-0.0211
1	0.07512	0.01347	-0.0906	-0.0111	-0.0394
	0.00203	-0.0262	-0.0722	-0.0044	0.02251
	0.04826	-0.0083	0.07067	-0.0049	0.00167
	-0.0145	-0.0814	-0.1130	0.04141	-0.0233
5	0.08590	0.02425	-0.0799	-0.0004	-0.0286
	0.01281	-0.0155	-0.0614	0.00631	0.03329
	0.05904	0.00241	0.08145	0.00581	0.01245
	-0.0038	-0.0707	-0.1022	0.05219	-0.0125
10	0.08003	0.01838	-0.0857	-0.0062	-0.0345
	0.00694	-0.0213	-0.0673	0.00044	0.02742
	0.05317	-0.0034	0.07558	-0.0001	0.00658
	-0.0096	-0.0765	-0.1080	0.04632	-0.0184

Table 30. Root Mean Square Differences - Negative Quadratic Term

θ	Root Mean Squares				
-0.2	0.08540	0.02652	-0.0893	-0.0080	-0.0386
	0.02724	-0.0194	-0.0687	0.00248	0.04025
	0.08073	-0.0101	0.09205	0.01111	0.01940
	-0.0182	-0.1051	-0.1053	0.08973	0.02112
-0.5	0.07920	0.02032	-0.0955	-0.0142	-0.0448
	0.02104	-0.0256	-0.0749	-0.0037	0.03405
	0.07453	-0.0163	0.08585	0.00491	0.01320
	-0.0244	-0.1113	-0.1115	0.08353	0.01492
-1	0.07715	0.01827	-0.0975	-0.0162	-0.0469
	0.01899	-0.0276	-0.0769	-0.0057	0.03200
	0.07248	-0.0184	0.08380	0.00286	0.01115
	-0.0265	-0.1133	-0.1135	0.08148	0.01287
-5	0.08554	0.02666	-0.0891	-0.0078	-0.0385
	0.02738	-0.0192	-0.0685	0.00262	0.04039
	0.08087	-0.0100	0.09219	0.01125	0.01954
	-0.0181	-0.1049	-0.1057	0.08987	0.02126
-10	0.08567	0.02679	-0.0890	-0.0077	-0.0384
	0.02751	-0.0191	-0.0684	0.00275	0.04052
	0.08100	-0.0098	0.09232	0.01138	0.01967
	-0.0180	-0.1048	-0.1050	0.09000	0.02139

A nonparametric one-way analysis of variance was conducted using the SAS routine NPAR1WAY (see Appendix K) to determine whether the average RMS differences are equal for all positive θ values.

The hypotheses for the test were formulated as:

H_0 : The average RMS differences are equal for positive θ values

H_1 : At least one average RMS difference is different

The test was conducted at the .05 level of significance.

The result of the nonparametric one-way ANOVA is an F-value of 0.162 and a p-value of 0.9570, and therefore the null hypothesis is not rejected. The relative performances of both techniques appear to remain the same for all positive θ values.

The test was repeated for the negative θ values and the results of the nonparametric one-way ANOVA is an F of 0.089 and a p-value of 0.9856. The null hypothesis again is not rejected. The relative performances of both techniques remain the same for all negative θ values.

The two-sided test was also conducted for the combined positive and negative θ values. The test results in an F value of 0.0136 and a p-value of 0.9986, implying again that the null hypothesis cannot be rejected.

All three tests conclude that the average RMS differences are equal. In other words, the magnitude and sign of the coefficient of the quadratic term have no effect on the relative performances of backpropagation networks and regression analysis.

7.1.2 Exponential Model

In many areas of the physical, chemical, and biological sciences, it is more common to use non-linear models. As an example of nonlinear regression (NLR) models, consider the model where

$$Y_i = \alpha e^{\beta X_i} + \varepsilon_i$$

defines the relationship between the regressor and response variables. The X were generated from a normal distribution with mean 50 and variance 225 in order to have X values ranging from 0 to 99. The values for α and β are 10 and 5 respectively, to be consistent with the previous chapters. The ε_i are random variables generated from a normal distribution with mean 0 and variance 100, since variance of 100 is determined in Chapter 4 to be the point where regression and backpropagation performed equally well. A population of 10,000 was generated using the SAS program in Appendix R.

7.1.2.1 Regression Analysis

Since OLS is easy to use, a data analyst may be tempted to transform the model to induce linearity. The exponential model can be transformed, by taking natural logarithms, into the alternative model

$$\ln Y_i = \ln \alpha + \beta X_i + \varepsilon_i .$$

In this model, it appears that SLR can be applied by regressing ln Y against X. The least squares procedure on this linearized form will not give the same parameter estimates as the original model. This is because, in the nonlinear model, least squares implies the minimization of the sum of squares of residuals on Y, but in the transformed model, the sum of squares of residuals on ln Y are being minimized. The error structure for the two models are different. In the nonlinear model the error structure is additive while the transformed model has a multiplicative error structure. Certain assumptions are made regarding the ε_i in the original model, namely that they are normal with mean 0 and common variance 100. These assumptions are not valid for the transformed model and therefore, it does not make sense to use the transformation. For the exponential model being studied there are no other transformations that would linearize the model and still maintain the basic assumptions.

This research used the Gauss-Newton procedure for finding the least squares estimators in the nonlinear model. This procedure is iterative and requires starting estimates for the parameters. Consider the general formulation of the model

$$Y_i = f(X_i, \theta) + \varepsilon_i$$

where θ is the vector of parameters. The steps in the procedure are

1. Begin with a Taylor series expansion around the starting values of the parameters, θ_0, to get the linearized form

   $$Y_i - f(X_i, \theta_0) = \gamma_1 w_{1i} + \gamma_2 w_{2i} + \ldots + \gamma_q w_{qi}$$

   where w_{ji} represents the derivative of the nonlinear function with respect to the jth parameter, q represents the number of parameters, and $\gamma_j = \theta_j - \theta_{j,0}$. The w_j's play the role of the regressors in a linear regression structure and the γ_j's play the role of the regression coefficients.

2. The γ_j's are estimated by linear least squares and the θ values are updated.

3. The updated $\theta_{j,0}$'s replace the starting values.

4. The two previous steps are repeated until convergence is reached, i.e. the residual sum of squares and the parameter estimates no longer change with iterations.

The drawback with the Gauss-Newton procedure is that convergence is not guaranteed, since it is dependent on the starting values given. A popular modification to the Gauss-Newton method that guarantees convergence is developed by Marquadt (1963). Since there is collinearity due to the fact that the w's are derivatives of the same function, the least squares estimator of the γ_j is too large. The Marquadt procedure reduces the size of the refinement of γ_j in much the same way as the biased estimation technique, ridge regression.

The NLIN procedure in SAS with the Marquadt option (see Appendix S), was used on the training sets to estimate the parameters. In order to "guess" starting values, ln Y was plotted against X for each training set. The slope was used as the starting value for β and the antilog of the intercept was the starting value for α. The estimated α and β values were then used along with X values from each of the recall sets to predict $\hat{Y}$ values using another SAS program based on the fitted model

$$\hat{Y}_i = ae^{bX_i}.$$

Each combination of training set and recall set results in a set of generated output, $\hat{Y}$, which were compared to the actual Y values (see Appendix E) to calculate an RMS value.

These 20 RMS values obtained from nonlinear regression are summarized in Table 31 on page 126.

7.1.2.2 Neural Network Analysis

The analysis in this section uses a three-layer backpropagation network with two processing elements in the hidden layer, one processing element in the output layer, one processing element in the input layer, and a sigmoidal transfer function. The networks were trained using the best learning

Table 31. Root Mean Squares for the Exponential Model

Source	Root Mean Squares				
Regression	0.92759	0.99958	1.17702	1.06590	1.11139
	0.97312	1.07323	1.14536	1.04408	0.97160
	0.88725	1.07908	0.89556	1.01356	0.99918
	1.09390	1.25028	1.18493	0.84215	0.94935
Backprop	1.05517	1.04398	1.03667	1.04730	1.04476
	1.03363	1.04242	1.03792	1.04561	1.04199
	1.03539	1.05124	1.04662	1.03906	1.03889
	1.05168	1.05580	1.02885	1.02311	1.01618
Difference	0.12757	0.04440	-0.1403	-0.0185	-0.0666
	0.06050	-0.0308	-0.1074	0.00153	0.07039
	0.14813	-0.0278	0.15106	0.02550	0.03971
	-0.0422	-0.1944	-0.1560	0.18095	0.07524

count of 35,000 (based on the results in Chapter 4). The X and Y values were represented as real numbers and the MinMax table was utilized to facilitate learning and recall.

Each of the five training sets results in a network which was then recalled on each of the four recall sets, resulting in 20 output sets. The RMS for each output set was calculated using the same BASIC program as regression (Appendix E). Thus, there were 20 RMS values corresponding to the 20 RMS values obtained from regression analysis. The neural network results are in Table 31 on page 126.

7.1.2.3 Comparison of Neural Networks and Regression

A SAS program (Appendix G) is used to calculate the difference between the RMS from the neural network, RMS_N, and the RMS from regression, RMS_R. This difference is calculated as RMS_N - RMS_R. A negative difference in the RMS indicates that neural networks have a lower RMS and therefore do a better job in predicting the Y while a positive difference implies that regression gives a better prediction of Y. The RMS differences for the nonlinear case are summarized in Table 31 on page 126.

In order to determine whether backpropagation and regression analysis perform the same when the underlying relationship between the X and Y is nonlinear, the Wilcoxon signed-ranks test was used. This test was used to determine if the average difference between the two techniques is significantly different from 0. The null hypotheses were set up as $H_0 : \theta = 0$ where θ is the average RMS difference. The alternative was $H_1 : \theta \neq 0$. The test statistic value is 94; because the test statistic is $>$ the critical value of 60 ($T_{.10,20}$), the decision is not to reject the null hypothesis at $\alpha = .05$, implying that the average RMS difference is not significantly different than 0.

Thus, backpropagation networks and regression analysis perform equally when the underlying relationship is an exponential function.

7.2 Underspecification

Underspecifying refers to the failure to include some variables in a regression model and this condition causes bias in prediction. In order to study the effect of underspecification, the population with the relationship

$$Y_i = \alpha + \beta X_i + \theta X_i^2 + \varepsilon_i$$

is considered, where X and X^2 are the regressor variables, Y is the measured response variable, and α is the intercept. The β and θ are the slope coefficients and ε_i is the model error. This is the same model used in section 7.1.1. The X values were generated from a normal distribution with mean 50 and variance 225 and X^2 values were calculated. (see Appendix Q for the SAS program). An α value of 10 and β value of 5 are used. The θ values used are -10, -5, -1, -.5, -.2, .2, .5, 1, 5, and 10. The ε_i are random variables having a normal distribution with mean zero and variance of 100. A distinct population of 10,000 was generated using SAS for each value of θ.

The relative performance of backpropagation network and regression analysis when the model is underspecified, i.e. without the quadratic term, is explored. The regression and backpropagation analysis were conducted as if the model were

$$Y_i = \alpha + \beta X_i + \varepsilon_i.$$

The effect of failing to include the quadratic term, when it has a positive as well as a negative coefficient, is studied separately.

From each population, five training sets of size 100 and four recall sets of size 100 were taken. The same procedure as in Chapter 4 was used. The training sets were used to train networks as well as estimate α and β in regression (since θ is not known to exist). The recall sets were used to determine how well the two techniques predict the output given the training they receive. Each of the four recall sets were recalled on the five training sets, resulting in 20 output sets for each level of θ.

The predicted output, $\hat{Y}$, from each technique were compared to the corresponding actual output, Y, and a root mean square difference (RMS) was calculated for each output set.

7.2.1 Regression Analysis

The underspecifed model was treated as in the case of SLR with only one regressor, X. The training sets for each value of θ were used to estimate the regression parameters α and β, as if there is no quadratic term. The estimated α and β values were then used along with X values from each of the appropriate recall sets to predict $\hat{Y}$ values using another SAS program based on the fitted model

$$\hat{Y}_i = a + bX_i .$$

For each level of θ, 20 analyses involving the five training sets and four recall sets were conducted, resulting in 20 output sets. These output files with the X's and $\hat{Y}$, and the recall files which have the Y values, were evaluated by a BASIC program (Appendix E) that calculates the RMS for each output set.

For each value of θ, 20 RMS values were obtained. The results from regression excluding the quadratic term, for the positive and negative θ values, are summarized in Table 32 on page 130 and Table 33 on page 131.

Table 32. Root Mean Squares from Regression - Without Positive Quadratic Term

θ	Root Mean Squares				
0.2	5.86124	5.9057	6.01546	5.94675	5.97488
	5.88939	5.95129	5.99589	5.93326	5.88845
	5.83630	5.95491	5.84143	5.91439	5.90550
	5.96407	6.06076	6.02036	5.80841	5.86949
0.5	14.6201	14.7269	14.9902	14.8352	14.8928
	14.6877	14.8362	14.9432	14.7929	14.6854
	14.5603	14.8449	14.5726	14.7477	14.7263
	14.8668	15.0988	15.0019	14.4934	14.6399
1	28.4140	29.0818	30.7278	29.6970	30.1190
	28.8364	29.7651	30.4342	29.4946	28.8222
	28.0398	29.8193	28.1169	29.2115	29.0781
	29.9567	31.4074	30.8013	27.6215	28.5378
5	145.495	147.224	151.487	148.817	149.910
	146.589	148.994	150.726	148.293	146.552
	144.526	149.134	144.725	147.560	147.215
	149.490	153.247	151.677	143.443	145.815
10	289.264	293.672	304.533	297.731	300.516
	292.051	298.180	302.595	296.395	291.958
	286.795	298.538	287.303	294.527	293.646
	299.445	309.018	305.013	284.034	290.081

Table 33. Root Mean Squares from Regression - Without Negative Quadratic Term

θ	Root Mean Squares				
0.2	6.05607	6.10627	6.23002	6.15252	6.18425
	6.08782	6.15764	6.20794	6.13731	6.08676
	6.02794	6.16172	6.03373	6.11602	6.10599
	6.17205	6.28111	6.23554	5.99648	6.06537
0.5	14.8341	14.9381	15.1942	15.0338	15.0995
	14.8999	15.0444	15.1485	15.0023	14.8977
	14.7759	15.0528	14.7879	14.9582	14.9375
	15.0742	15.3000	15.2056	14.7108	14.8534
1	28.5937	29.2766	30.9600	29.9058	30.3374
	29.0256	29.9754	30.6597	29.6988	29.0112
	28.2110	30.0309	28.2898	29.4092	29.2728
	30.1714	31.6550	31.0351	27.7831	28.7202
5	145.749	147.455	151.660	149.027	150.105
	146.828	149.201	150.910	148.510	146.792
	144.793	149.339	144.990	147.786	147.446
	149.690	153.396	151.848	143.724	146.065
10	289.514	293.900	304.710	297.940	300.712
	292.288	298.387	302.781	296.611	292.195
	287.057	298.743	287.563	294.751	293.875
	299.646	309.173	305.192	284.309	290.327

A nonparametric one-way analysis of variance was conducted using the SAS routine NPAR1WAY (see Appendix K) to determine whether the average RMS values from regression are equal for all positive θ values when the model is underspecified.

The hypotheses for the test were formulated as:

H_0 : The average RMS are equal for positive θ values
H_1 : At least one average RMS is different

The test was conducted at the .05 level of significance.

The result of the nonparametric one-way ANOVA is an F-value of 29579.925 and a p-value of 0.0001, and therefore the null hypothesis is rejected. The average RMS values are not the same for all positive θ values.

The test was repeated for the negative θ values and the results of the nonparametric one-way ANOVA is an F of 29905.445 and a p-value of 0.0001. The null hypothesis again is rejected, implying that the average RMS values are not equal.

In other words, the magnitude of the coefficient of the quadratic term has an effect on how well regression predicts for both positive as well as negative signs. This was, of course, a very predictable conclusion, as the underspecified model is strictly linear and is less able to correctly predict as the degree of nonlinearity increases.

In order to determine whether regression performs better when the missing quadratic term is .2 than when it is 10 and -.2 rather than -10, the Wilcoxon signed-ranks test was used. This test was used first to determine if the difference between the average RMS at .2 and 10, is significantly less than 0. The null hypothesis was set up as $H_0 : \theta = 0$ where θ = Average RMS at .2 - Average RMS at 10. The alternative was $H_1 : \theta < 0$. The test statistic is 0 , i.e. < the critical value of 60

($T_{.05,20}$) and the decision is to reject the null hypothesis at $\alpha = .05$, implying that the average RMS is significantly less at θ of .2 than at θ of 10. The test was repeated for the difference between -.2 and -10, and the test statistic is once again 0. The null hypothesis is rejected at $\alpha = .05$. The average RMS is also significantly less at θ of -.2 than at θ of -10. Once again, this is a very predictable result.

Thus, regression analysis performs better when the magnitude of the excluded quadratic term is smaller regardless of the sign.

7.2.2 Neural Network Analysis

The analysis in this section uses a three-layer backpropagation network with two processing elements in the hidden layer, one processing element in the output layer, one processing element in the input layer, and a sigmoidal transfer function. The networks were trained using the best learning count of 35,000 (based on the results in Chapter 4). The X and Y values were represented as real numbers and the MinMax table was utilized to facilitate learning and recall.

Each of the five training sets results in a network that was then recalled on each of the four recall sets, resulting in 20 output sets for each value of θ. The RMS for each output set was calculated using the same BASIC program as regression (Appendix E). Thus, for each θ value, there were 20 RMS values corresponding to the 20 RMS values obtained from regression analysis. The neural network results, for positive and negative θ values, are summarized in Table 34 on page 134 and Table 35 on page 135.

A nonparametric one-way analysis of variance was conducted using the SAS routine NPAR1WAY (see Appendix K) to determine whether the average RMS values from backpropagation are equal for all positive θ values when the model is underspecified.

Table 34. Root Mean Squares from Backpropagation - Without Positive Quadratic Term

θ	Root Mean Squares				
0.2	1.88504	1.84563	1.81991	1.85733	1.84838
	1.80919	1.84014	1.82431	1.85139	1.83865
	1.81540	1.87118	1.85494	1.82831	1.82772
	1.87276	1.88725	1.79237	1.77217	1.74776
0.5	1.89303	1.86024	1.83882	1.86997	1.86252
	1.82991	1.85566	1.84249	1.86503	1.85442
	1.83508	1.88150	1.86798	1.84581	1.84532
	1.88281	1.89487	1.81591	1.79910	1.77879
1	2.10309	2.06813	2.04530	2.07851	2.07057
	2.03579	2.06325	2.04921	2.07324	2.06193
	2.04130	2.09080	2.07639	2.05275	2.05223
	2.09220	2.10506	2.02086	2.00294	1.98128
5	5.93238	5.88268	5.85032	5.89744	5.88614
	5.83671	5.87574	5.85577	5.88994	5.87387
	5.84454	5.91490	5.89441	5.86082	5.86007
	5.91689	5.93518	5.81548	5.79000	5.75921
10	7.70631	7.58292	7.50234	7.61956	7.59151
	7.46879	7.56571	7.51613	7.60095	7.56104
	7.48824	7.66293	7.61205	7.52865	7.52680
	7.66785	7.71325	7.41610	7.35284	7.27640

Table 35. Root Mean Squares from Backpropagation - Without Negative Quadratic Term

θ	Root Mean Squares				
0.2	1.82932	1.79777	1.77718	1.80714	1.79997
	1.76860	1.79338	1.78070	1.80238	1.79218
	1.77357	1.81823	1.80522	1.78390	1.78343
	1.81949	1.83109	1.75513	1.73896	1.71942
0.5	1.86761	1.83552	1.81456	1.84505	1.83775
	1.80583	1.83104	1.81814	1.84021	1.82983
	1.81089	1.85633	1.84309	1.82140	1.82092
	1.85761	1.86942	1.79213	1.77567	1.75579
1	2.10871	2.07230	2.04853	2.08311	2.07484
	2.03864	2.06723	2.05260	2.07762	2.06585
	2.04437	2.09591	2.08090	2.05629	2.05575
	2.09736	2.11075	2.02309	2.00443	1.98188
5	6.53829	6.10101	5.81547	6.23086	6.13148
	5.69659	6.04004	5.86434	6.16491	6.02351
	5.76549	6.38455	6.20426	5.90870	5.90216
	6.40201	6.56288	5.50987	5.28570	5.01482
10	7.82049	7.68692	7.59970	7.72658	7.69623
	7.56339	7.66830	7.61463	7.70644	7.66325
	7.58443	7.77353	7.71846	7.62818	7.62618
	7.77886	7.82800	7.50635	7.43788	7.35514

The hypotheses for the test were formulated as:

H_0 : The average RMS are equal for positive θ values

H_1 : At least one average RMS is different

The test was conducted at the .05 level of significance.

The result of the nonparametric one-way ANOVA is an F-value of 40823.849 and a p-value of 0.0001, and therefore the null hypothesis is rejected. The average RMS values are not the same for all positive θ values.

The test was repeated for the negative θ values and the results of the nonparametric one-way ANOVA is an F of 43521.793 and a p-value of 0.0001. The null hypothesis again is rejected, implying that, for negative θ values, the average RMS values are not equal.

In other words, the magnitude of the coefficient of the quadratic term has an effect on the performance of backpropagation networks for both positive as well as negative signs.

In order to determine whether backpropagation performs better when the missing quadratic term is .2 than when it is 10, and -.2 rather than -10, the Wilcoxon signed-ranks test was used. This test was used to determine if the difference between the average RMS at .2 and 10, is significantly less than 0. The null hypothesis was set up as $H_0 : \theta = 0$ where θ = Average RMS at .2 - Average RMS at 10. The alternative was $H_1 : \theta < 0$. The test statistic is 0 , i.e. < the critical value of 60 ($T_{.05,20}$), and the decision is to reject the null hypothesis at $\alpha = .05$, implying that the average RMS is significantly less at θ of .2 than at θ of 10. The test was repeated for the difference between θ values of -.2 and -10, and the test statistic is once again 0. The null hypothesis is rejected at α = .05. Average RMS is significantly less at θ of -.2 than θ of -10.

Thus, backpropagation networks performs better when the magnitude of the excluded quadratic term is smaller regardless of the sign. This result, while not surprising, is not as predictable as the related regression results, since neural networks are supposed to "discover" the underlying nonlinearities. The network may have, in fact, "discovered" the relationships for the various values of θ, but just not as precisely for major quadratic influence as for minor quadratic influence. The major question that remains is whether the neural network is better able to deal with the quadratic term than the underspecified regression model.

7.2.3 Comparison of Neural Networks and Regression

In order to compare the relative performances of the two techniques, the average RMS values for the backpropagation network and for regression analysis for the correctly specified model (as in Section 7.1.1) and the underspecified model are plotted separately for the positive and negative θ values.

Both Figure 6 on page 138 and Figure 7 on page 139 compare the average RMS values from the correctly specified model versus the underspecified model using the backpropagation network and regression analysis for positive quadratic term. The equivalent comparison for negative quadratic term are in Figure 8 on page 140 and Figure 9 on page 141. In the positive as well as the negative case, it is apparent that the performance of both techniques deteriorates for the underspecified model as the magnitude of the quadratic coefficient increases.

In Figure 10 on page 142, the average RMS values using the underspecified model are plotted for backpropagation versus regression for the case of missing positive quadratic term. A similar plot for the missing negative quadratic term is in Figure 11 on page 143. In both cases, backpropagation performs better than regression and it also deteriorates at a lesser rate as the magnitude of the missing quadratic coefficient increases.

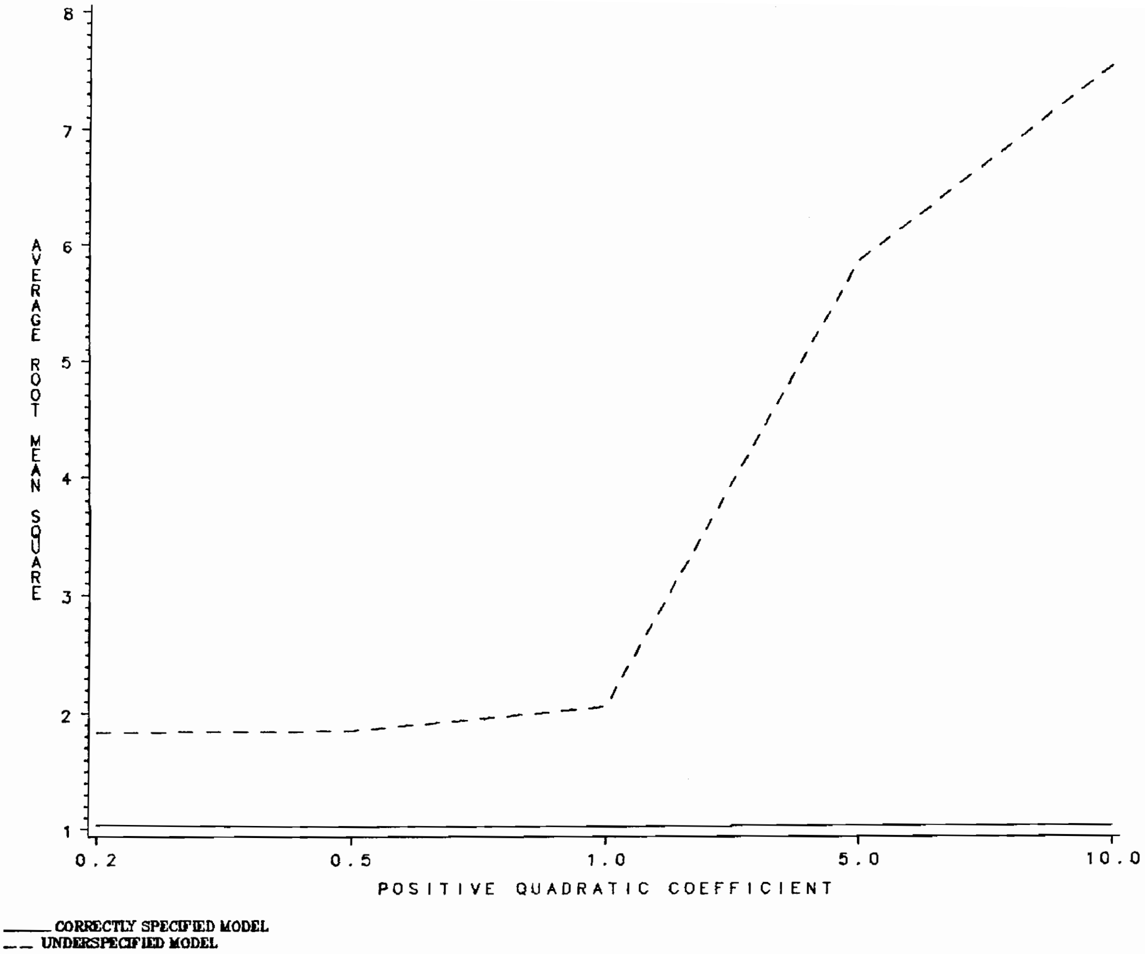


Figure 6. Comparison of the two models using Backpropagation - Positive Quadratic Term

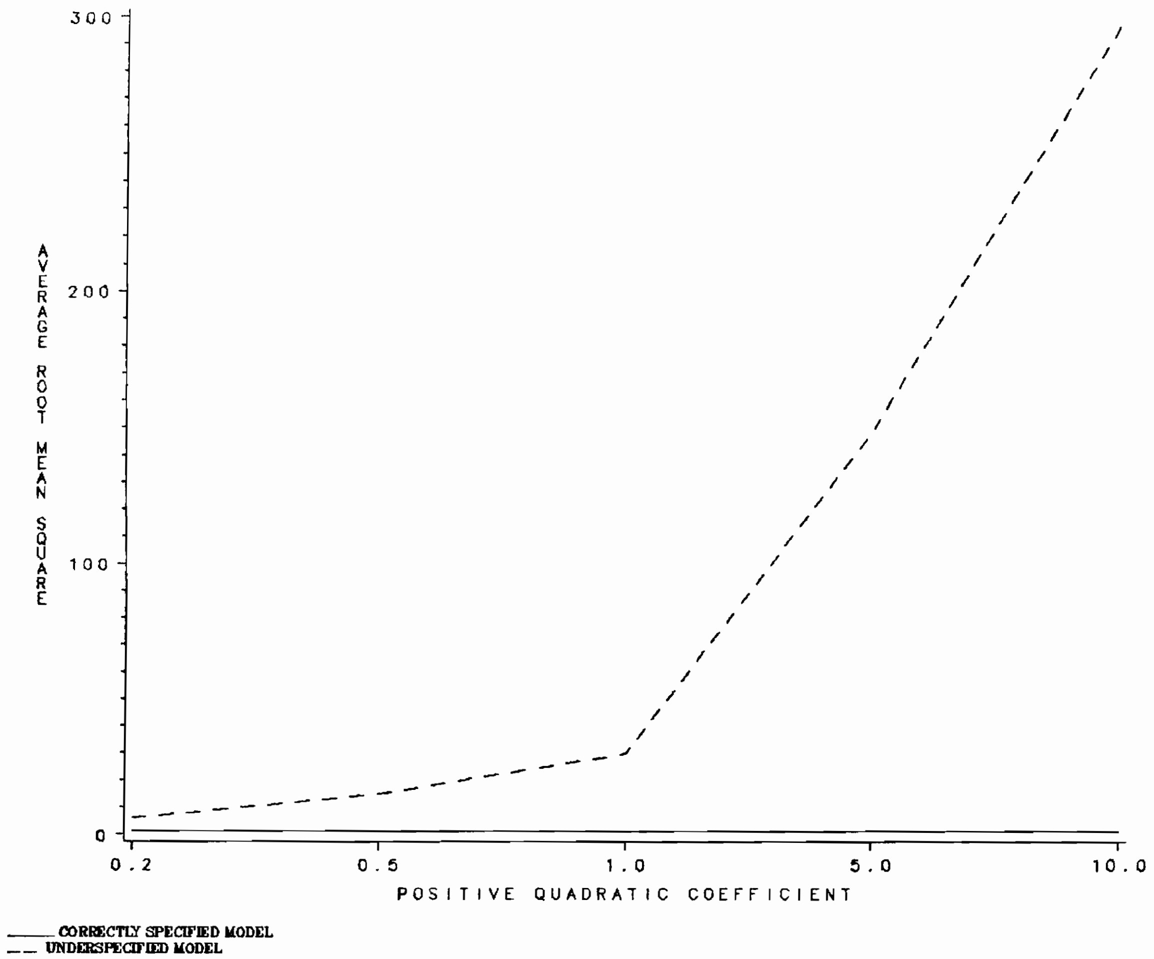


Figure 7. Comparison of the two models using Regression - Positive Quadratic Term

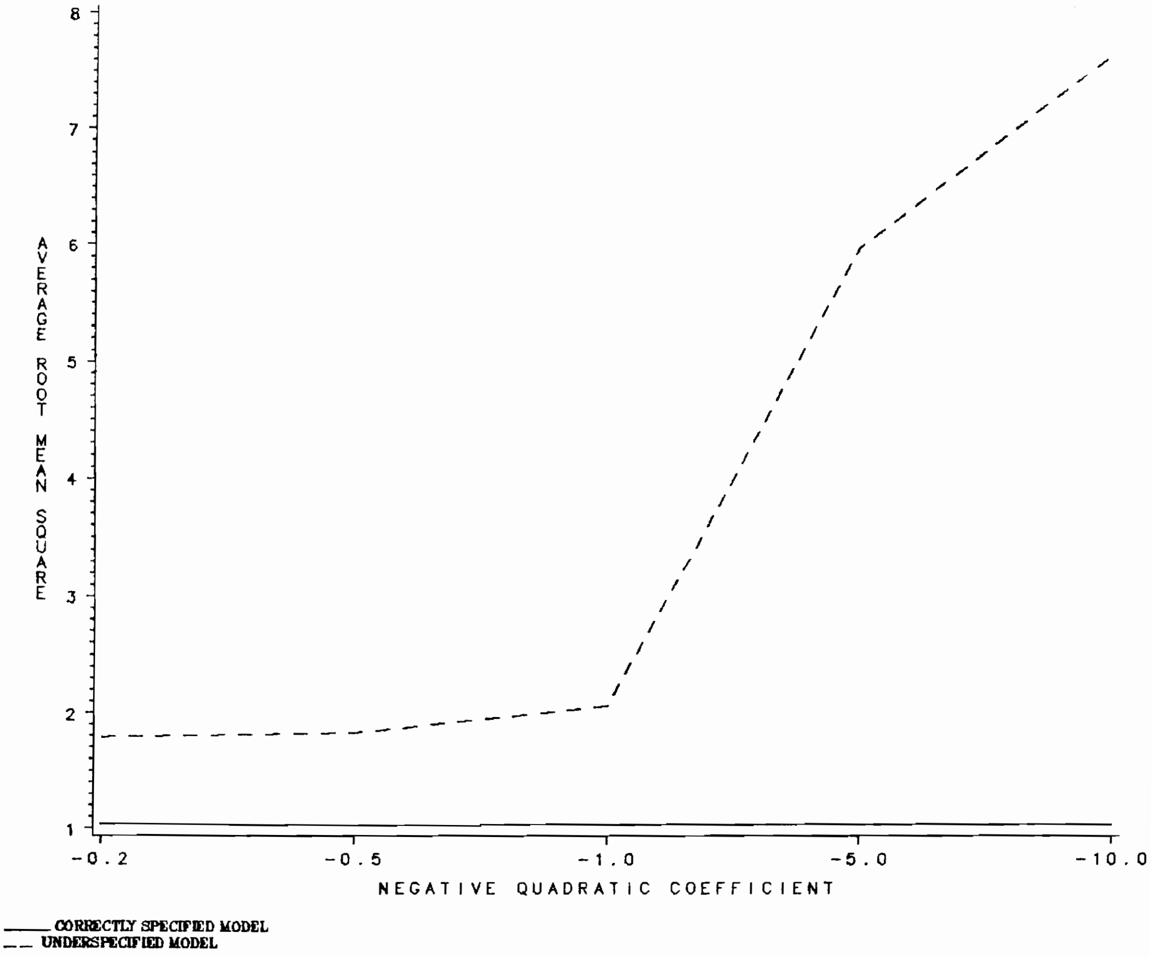


Figure 8. Comparison of the two models using Backpropagation - Negative Quadratic Term

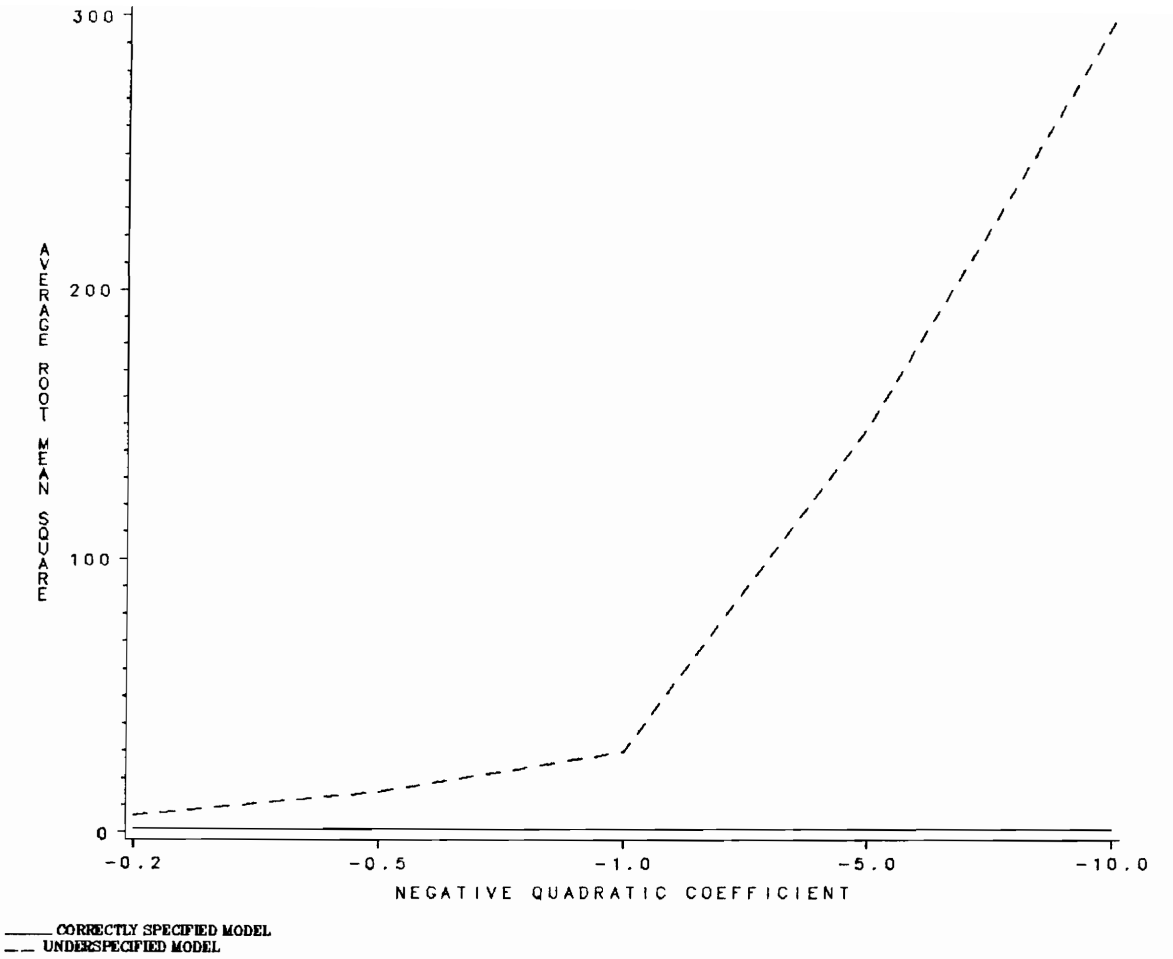


Figure 9. Comparison of the two models using Regression - Negative Quadratic Term

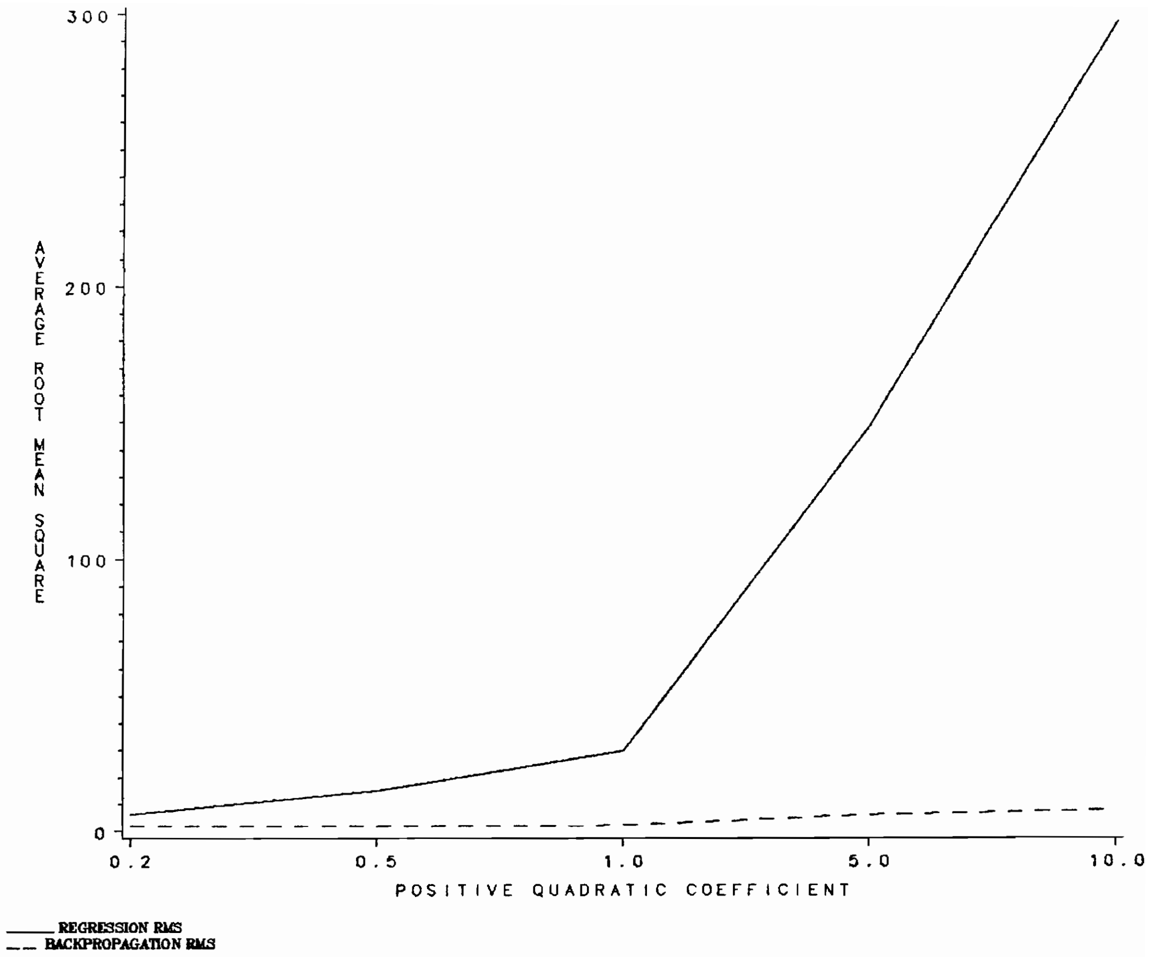


Figure 10. Comparison of the two techniques with the Underspecified Model - Positive Quadratic Term

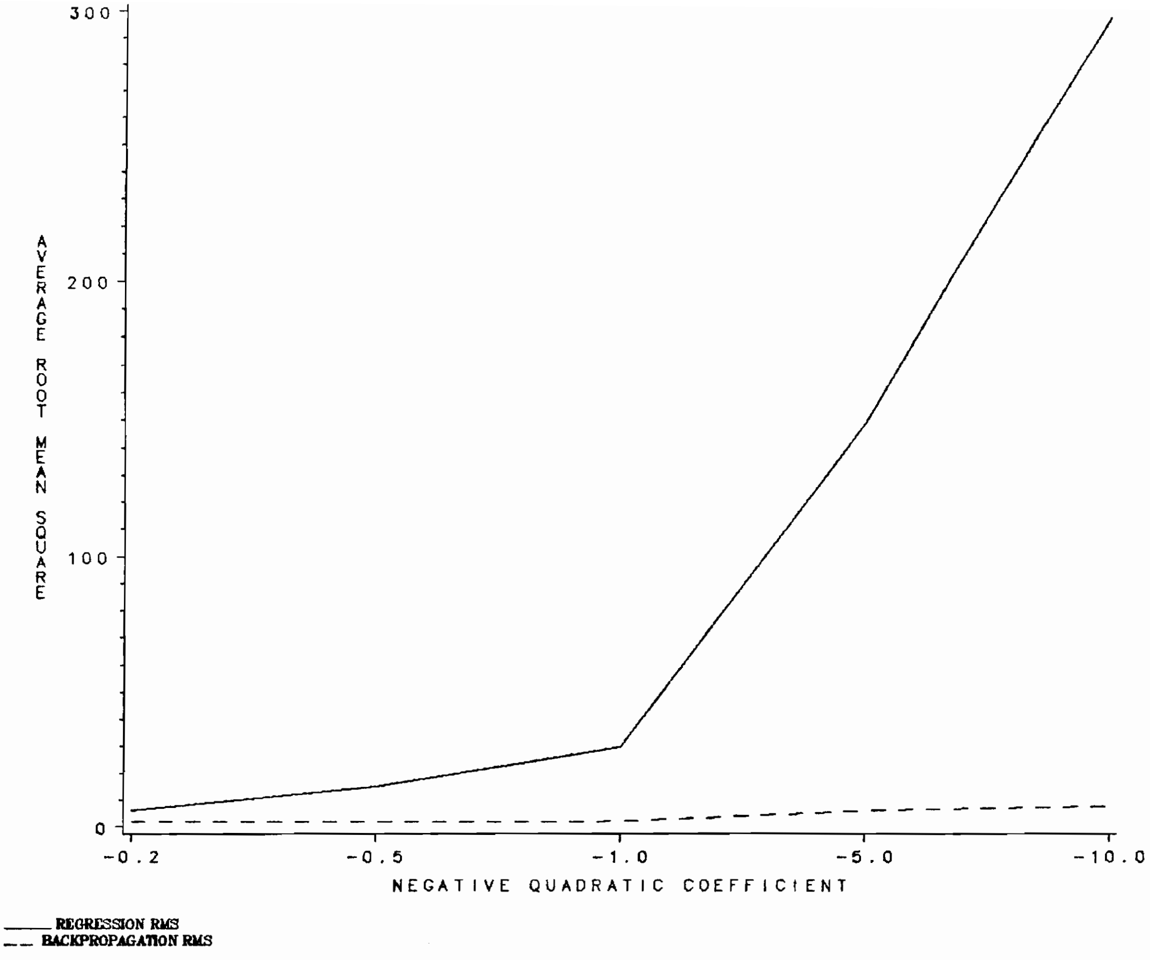


Figure 11. Comparison of the two techniques with the Underspecified Model - Negative Quadratic Term

A SAS program (Appendix G) is used to calculate the difference between the RMS from neural network, RMS_N, and the RMS from regression, RMS_R. This difference is calculated as RMS_N - RMS_R for each observation in each value of θ. A negative difference in the RMS indicates that neural networks have a lower RMS and therefore do a better job in predicting the Y while a positive difference implies that regression gives a better prediction of Y. The RMS differences for the positive and negative quadratic cases are summarized in Table 36 on page 145 and Table 37 on page 146.

A nonparametric one-way analysis of variance was conducted using the SAS routine NPAR1WAY (see Appendix K) to determine whether the average RMS differences are equal for all positive θ values when the model is underspecified.

The hypotheses for the test were formulated as:

H_0 : The average RMS differences are equal for positive θ values
H_1 : At least one average RMS difference is different

The test was conducted at the .05 level of significance.

The result of the nonparametric one-way ANOVA is an F-value of 28651.152 and a p-value of 0.0001, and therefore the null hypothesis is rejected. The average RMS differences are not the same for all positive θ values.

The test was repeated for the negative θ values and the results of the nonparametric one-way ANOVA is an F of 29206.181 and a p-value of 0.0001. The null hypothesis again is rejected, implying that the average RMS differences are not equal.

Table 36. Root Mean Square Differences - Without Positive Quadratic Term

θ	Root Mean Squares				
0.2	-3.9672	-4.0601	-4.1955	-4.0894	-4.1265
	-4.0802	-4.1111	-4.1715	-4.0818	-4.0498
	-4.0209	-4.0837	-3.9864	-4.0860	-4.0777
	-4.0913	-4.1735	-4.2279	-4.0362	-4.1217
0.5	-12.727	-12.866	-13.151	-12.955	-13.030
	-12.857	-12.980	-13.100	-12.927	-12.831
	-12.725	-12.963	-12.704	-12.901	-12.881
	-12.984	-13.204	-13.186	-12.694	-12.861
1	-26.311	-27.013	-28.682	-27.618	-28.048
	-26.800	-27.701	-28.385	-27.421	-26.760
	-25.998	-27.728	-26.040	-27.158	-27.025
	-27.864	-29.302	-28.780	-25.618	-26.556
5	-139.56	-141.34	-145.63	-142.92	-144.02
	-140.75	-143.11	-144.87	-142.40	-140.67
	-138.68	-143.21	-138.83	-141.69	-141.35
	-143.57	-147.31	-145.86	-137.65	-140.05
10	-281.55	-286.08	-297.03	-290.11	-292.92
	-284.58	-290.61	-295.07	-288.79	-284.39
	-279.30	-290.87	-279.69	-286.99	-286.11
	-291.77	-301.30	-297.60	-276.68	-282.80

Table 37. Root Mean Square Differences - Without Negative Quadratic Term

θ	Root Mean Squares				
-0.2	-4.2267	-4.3085	-4.4528	-4.3453	-4.3842
	-4.3192	-4.3642	-4.4272	-4.3349	-4.2945
	-4.2543	-4.3434	-4.2285	-4.3321	-4.3225
	-4.3525	-4.4500	-4.4804	-4.2575	-4.3459
-0.5	-12.966	-13.102	-13.379	-13.188	-13.261
	-13.094	-13.213	-13.330	-13.162	-13.067
	-12.965	-13.196	-12.944	-13.136	-13.116
	-13.216	-13.430	-13.413	-12.935	-13.097
-1	-26.485	-27.204	-28.911	-27.822	-28.262
	-26.987	-27.908	-28.607	-27.621	-26.945
	-26.166	-27.935	-26.208	-27.352	-27.217
	-28.074	-29.544	-29.012	-25.778	-26.738
-5	-139.21	-141.35	-145.84	-142.79	-143.97
	-141.13	-143.16	-145.04	-142.34	-140.76
	-139.02	-142.95	-138.78	-141.87	-141.54
	-143.28	-146.83	-146.33	-138.43	-141.05
-10	-281.69	-286.21	-297.11	-290.21	-293.01
	-284.72	-290.71	-295.16	-288.90	-284.53
	-279.47	-290.97	-279.84	-287.12	-286.24
	-291.86	-301.34	-297.68	-276.87	-282.97

In other words, the magnitude of the coefficient of the quadratic term has an effect on the relative performance of backpropagation networks and regression analysis for both positive as well as negative signs.

In order to determine whether backpropagation performs better than regression when there is either a positive or a negative quadratic term, the Wilcoxon signed-ranks test was used. This test was used to determine if the average difference between the two techniques is significantly less than 0 when θ values are -.2, -10, .2, and 10. The null hypotheses were set up as $H_0 : \theta = 0$ where θ is the average RMS difference. The alternative was $H_1 : \theta < 0$ in each case. The test statistic values in each case is 0, i.e. < the critical value of 60 ($T_{.05,20}$) and the decision is to reject the null hypothesis at $\alpha = .05$, implying that the average RMS difference is significantly less than 0.

Thus, backpropagation networks perform better than regression analysis when a quadratic term has been excluded, regardless of the size and sign of the coefficient of the quadratic term.

The Wilcoxon signed-ranks test was also conducted to determine whether the relative performance of backpropagation is better at larger θ values than at smaller θ values, for both the positive and the negative underspecified cases. The null hypothesis was set up as $H_0 : \theta = 0$ where $\theta =$ Average of (RMS difference at θ of 10 - RMS difference at θ of .2). A similar hypothesis for the negative case was also set up. The alternative was $H_1 : \theta < 0$ in each case. The test statistic value in each case is 0, i.e. < the critical value of 60 ($T_{.05,20}$) and the decision is to reject the null hypothesis at $\alpha = .05$, implying that the average differences between the RMS differences at θ of 10 and .2, and between θ of -10 and -.2 are significantly less than 0.

Backpropagation networks perform relatively better than regression analysis at larger values of θ than at smaller values of θ, regardless of the sign. Again, these results are consistent with statements in the literature that extoll the superiority of neural networks over regression in the presence of increasing nonlinearities.

7.3 Conclusions

Optimally specified regression models and backpropagation networks seem to perform at the same level of accuracy in predicting when there is a quadratic term in the model or the model is exponential. In the case of the model with a quadratic term, neither the magnitude nor the sign of the quadratic coefficient seems to have any effect on the individual or relative performances of the two techniques.

The results change dramatically when the model is underspecified, i.e. the quadratic term is left out. Both regression and backpropagation perform better when the magnitude of the missing quadratic coefficient is smaller than when it is larger, regardless of the sign. The magnitude of the quadratic coefficient also has an effect on the relative performance of the two techniques.

Backpropagation networks perform significantly better than regression analysis when a quadratic term has been excluded, regardless of the size and sign of the coefficient of the quadratic term. The relative performance of backpropagation networks improve significantly as the magnitude of the missing quadratic coefficient increases, regardless of the sign.

It seems reasonable to conclude that regression analysis is able to perform satisfactorily when the model is properly specified but when the model is incorrect, as is often the case, the prediction using regression is biased. This supports the theory put forth by Hecht-Nielsen (1989) that, in high-dimensional spaces (input dimensions greater than 3 to 10), regression techniques often fail to produce an appropriate approximation. Backpropagation networks do not seem to have this problem and therefore, at least in the underspecified quadratic case studied in this chapter, outperformed regression. It would be presumptuous, without further exploration of misspecified models, to generalize that neural networks perform better than regression whenever the underlying function is incorrectly approximated.

Chapter 8: Summary and Conclusions

This chapter summarizes the findings of the research conducted and states the resulting conclusions.

8.1 Summary of Results

The following sections summarize the results, in terms of the factors considered in the research.

8.1.1 Sample Size and Variance

There is significant interaction between sample size and variance for regression analysis and backpropagation networks. Sample size and variance appear to have an effect on the relative performances of regression and backpropagation.

Both regression and backpropagation perform better when sample size is large as opposed to small, and at lower variance levels rather than higher variance levels.

Regression performs better than backpropagation for all sample sizes at variance of 25. At a variance of 100, regression outperforms backpropagation for sample sizes less than 100. For sample sizes greater than 100, at variance 100, backpropagation does better.

At sample sizes of 100 and variance of 100, regression and backpropagation perform at the same level.

Backpropagation networks perform better than regression at variance levels greater than 225, across all sample sizes.

8.1.2 Outliers, Skewness, and Kurtosis

The presence of moderate outliers does not affect the performance of regression. Extreme outliers cause the performance of regression to deteriorate. For backpropagation networks, the presence of outliers has no effect on performance. In fact, backpropagation appear to perform significantly better than regression in the presence of outliers.

Neither skewness nor kurtosis has any effect on the individual performances of regression or backpropagation.

Regression analysis and backpropagation networks perform at the same level of accuracy in predicting when the error term is skewed positively or negatively, or when the error term is flat or peaked.

8.1.3 Multicollinearity

Regression analysis predicted more accurately when there is no multicollinearity than when there is any correlation between the two regressors. Multicollinearity does not affect the performance of the backpropagation network. In fact, backpropagation outperforms regression analysis whether there is multicollinearity or not, when there are two regressors in the model.

8.1.4 Nonlinearity and Underspecification

Optimally specified regression models and backpropagation networks seem to perform at the same level of accuracy in predicting when there is a quadratic term in the model or the model is exponential. In the case of the model with a quadratic term, neither the magnitude nor the sign of the quadratic coefficient has any effect on the individual or relative performances of the two techniques.

The results change dramatically when the model is underspecified, i.e. the quadratic term is left out. Both regression and backpropagation perform better when the magnitude of the missing quadratic coefficient was smaller than when it is larger, regardless of the sign. The magnitude of the quadratic coefficient also has an effect on the relative performance of the two techniques.

Backpropagation networks perform significantly better than regression analysis when a quadratic term has been excluded, regardless of the size and sign of the coefficient of the quadratic term. The relative performance of backpropagation network improves significantly as the magnitude of the missing quadratic coefficient increases, regardless of the sign.

8.2 Conclusions

Both techniques appear to perform at the same level under the following circumstances: when sample size is 100 and error variance is 100, in the presence of skewness or kurtosis of the error distribution, or when the underlying function is nonlinear and correctly specified.

It is not surprising to find that the two techniques are comparable when there is skewness or kurtosis of the error distribution, since regression analysis uses the reasonably robust method of ordinary least squares for estimation and neural networks can self-identify the nature of the relationships in a data set.

In the case of model specification, it appears that regression analysis is able to perform satisfactorily when the model is properly specified, but when the model is incorrect, as is often the case, the prediction using regression is biased. This supports the theory put forth by Hecht-Nielsen (1989) that, in high-dimensional spaces (input dimensions greater than 3 to 10), regression techniques often fail to produce an appropriate approximation. Backpropagation networks also seem to suffer from reduced information in the inputs, but not as much as regression, and therefore, at least in the underspecified quadratic case studied in this research, outperforms regression. It would be presumptuous, without further exploration of misspecified models, to generalize that neural networks perform better than regression whenever the underlying function is incorrectly approximated.

It is well-established that regression analysis performs quite well, as long as the assumptions are not violated, for sample sizes larger than 30. Assuming all other things are equal, an increase in sample size decreases the prediction variance in properly specified models. On the other hand, an increase in the variance of the error term increases the variances of the least square estimators, resulting in poor prediction. The interplay between sample size and variance leads to the conclusion that, for both techniques, performance is affected in a conflicting manner by the two factors. Increased sample size improves performance whereas increased variance deteriorates performance. Due to the presence of interaction between these two factors, it is not possible to determine the exact extent of the effect of sample size and variance. Nevertheless, it is concluded that backpropagation outperforms regression except when variance was 25, and when sample size is less than 100 for variance of 100. The relatively poor performance of backpropagation is not so much due to its ineffectiveness, but due to the effectiveness of regression analysis at low levels of error variance.

The results in the presence of outliers are no surprise. The least squares procedure allows outliers to exert disproportionate influence on regression results. Backpropagation networks do not have this handicap since they self-identify the underlying relationship.

The presence of multicollinearity affects parameter estimation in regression and renders poor prediction. In fact, the understanding and diagnosis of multicollinearity is imperative to anyone using regression analysis. The very nature of backpropagation networks reduces the difficulties associated with the detection and combating of multicollinearity. Although multicollinearity affects the performance of backpropagation networks, it is nowhere as damaging as it is to regression.

The purpose of this research is to explore the robustness of regression analysis and backpropagation networks. While it can be concluded that regression analysis is a reasonably robust technique, backpropagation networks have demonstrated superior performance in three notable instances: model underspecification, multicollinearity, and high error variances. It is suggested that researchers consider backpropagation networks for data which have these conditions or when the underlying function cannot be adequately approximated.

Bibliography

Alkon, D. L., Blackwell, K. T., Barbour, G. S., Rigler, A. K., and Vogl, T. P. "Pattern-Recognition by an Artificial Network derived from Biologic Neuronal Systems." Biological Cybernetics, 62, 1990: 363-376.

Almeida, L. B. "Neural Computers." Proceedings of the NATO ARW on Neural Computers, Dusseldorf. Heidelberg: Springer-Verlag, 1987.

✓ Anderson, J. A. "Categorization in a Neural Network." Organization of Neural Networks: Structure and Models. W. von Seelen, G. Shaw, U. M. Leinhos (ed.). Weinheim, Germany: VCH Verlagsgesellschaft, 1988.

Andrews, D. F., and Pregibon, D. "Finding outliers that matter." Journal of the Royal Statistical Society, Series B, 40, 1978: 85-93.

Bailey, D., and Thompson, D. "How to develop Neural-Network Applications." AI Expert, June 1990: 38-47.

Bruck, J., and Sanz, J. "A Study on Neural Networks." International Journal of Intelligent Systems, 3, 1988: 59-75.

Burr, D. J. "Experiments with a connectionist text reader." Proceedings of the IEEE International Conference on Neural Networks, 4, 1987: 717-724.

Caianiello, E. R. "Outline of a theory of thought-processes and thinking machines." Journal of Theoretical Biology, 2, 1961: 204-235.

Carpenter, G., and Grossberg, S. "A massively parallel architecture for a self-organizing neural pattern recognition machine." Computer Vision, Graphics, and Image Processing, 37, 1987: 54-115.

Caudill, M. "Neural Network Primer Part I." AI Expert, December 1987: 46-52.

Caudill, M. "Neural Network Primer Part II." AI Expert, February 1988: 55-61.

Caudill, M. "Neural Network Primer Part III." AI Expert, June 1988: 53-59.

Caudill, M. "Neural Network Primer Part IV." AI Expert, August 1988: 61-67.

Caudill, M. "Neural Network Primer Part V." AI Expert, November 1988: 57-65.

Caudill, M. "Neural Network Primer Part VI." AI Expert, February 1989: 61-67.

Caudill, M. "Neural Network Primer Part VII." AI Expert, May 1989: 51-58.

Caudill, M. "Neural Network Primer Part VIII." AI Expert, August 1989: 61-67.

Caudill, M. "Using Neural Networks: Representing Knowledge Part I." AI Expert, December 1989: 34-41.

Cottrell, G. W., Munro, P., and Zipser, D. "Image Compression by backpropagation: An example of extensional programming." Advances in cognitive science, volume 3. Norwood, NJ: Ablex, 1987.

Deaton, M. L., Reynolds, Jr., M. R., and Myers, R. H. "Estimation and hypothesis testing in regression in the presence of nonhomogeneous error variances." Communications in Statistics, B12 (1), 1983: 45-66.

Dietforide, T. G., and Michalski, R. S. "Discovering Patterns in Sequence of Events." AI, 25 (2), February 1985: 187-232.

Draper, N. R., and Van Nostrand, R. C. "Ridge regression and James Stein estimators: Review and comments." Technometrics, 21, 1979: 451-466.

Farley, B. G., and Clark, W. A. "Simulation of self-organizing systems by digital computer." Institute of Radio Engineers - Transactions of Professional Group of Information Theory, PFIT-4, 1954: 76-84.

Gale, W. A. Artificial Intelligence & Statistics. Reading, Massachussetts: Addison-Wesley Publishing Co., 1986.

Gale, W. A., and Pregibon, D. "An Expert System for Regression Analysis." Computer Science and Statistics Proceedings of 14th Symposium on the Interface. K. W. Heiner (ed.). New York: Springer-Verlag, 1982.

Gale, W. A., and Pregibon, D. "AI Research in Statistics." AI Magazine, 5 (4), 1985: 72-75.

Gibbons, J. D. Nonparametric Statistical Inference. New York: McGraw-Hill Book Company, 1971.

Grossberg, S. "Competitive Learning: From Interactive Activation to Adaptive Resonance." Cognitive Science, 11, 1987: 23-63.

✓ Guiver, J. P., and Klimasauskas, C. C. NeuralWorks, Networks II. Pittsburg, Pennsylvania: NeuralWare, Inc., 1989.

Gutierrez, M., Wang, J., and Grondin, R. "Estimating Hidden Unit Number for Two-layer Perceptrons." Proceedings of the IEEE International Conference on Neural Networks, I, 1989: 677-681.

Guyon, I., Poujaud, I., Personnaz, L., Dreyfus, G., Denker, J., and Le Cun, Y. "Comparing Different Neural Network Architectures for Classifying Handwritten Digits." Proceedings of the IEEE International Conference on Neural Networks, II, 1989: 127-132.

Hebb, D. O. Organization of Behavior. New York: Wiley, 1949.

Hecht-Nielsen, R. Neurocomputing. Reading, Massachusetts: Addison-Wesley Publishing Co., 1989.

Hettmansperger, T. P., and McKean, J. W. "Statistical Inference Based on Ranks." Psychometrika, 43, 1978: 69-79.

Hocking, R. R., and Pendleton, O. J. "The Regression Dilemma." Communications in Statistics, A12 (5), 1983: 497-527.

Hollander, M., and Wolfe, D. A. Nonparametric Statistical Methods. New York: Wiley, 1973.

Hornik, K., Stinchcombe, M., and White, H. "Multilayer Feedforward Networks are Universal Approximators." Neural Networks, 2, 1989: 359-366.

Jeffrey, W. "Neural Network Processing as a Tool for Function Optimization." Neural Network for Computing Snowbird Utah, AIP Proceedings, 151, 1986: 241

Klimasauskas, C. C. NeuralWorks, An Introduction to Neural Computing. Pittsburg, Pennsylvania: NeuralWare, Inc., 1989.

Klimasauskas, C. C., and Guiver, J. P. NeuralWorks, Networks I. Pittsburg, Pennsylvania: NeuralWare, Inc., 1989a.

Klimasauskas, C. C., Guiver, J. P., and Pelton, G. NeuralWorks Professional II and NeuralWorks Explorer. Pittsburg, Pennsylvania: NeuralWare, Inc., 1989b.

Kohonen, T. "An Introduction to Neural Computing." Neural Networks, 1, 1988: 3-16.

Krutchkoff, Richard G. Stochastic Simulation (class notes).

Marquadt, D. W. "An algorithm for least squares estimation of nonlinear parameters." Journal of the Society of Industrial and Applied Mathematics, 2, 1963: 431-441.

McCulloch, W. S., and Pitts, W. A. "A logical calculus of the ideas imminent in nervous activity." Bulletin of Mathematics and Biophysics, 5, 1943: 115-133.

McKean, J. W., and Hettmansperger, T. P. "Test of Hypotheses based on ranks in the linear model." Communications in Statistics - Theoretical Methods, A5 (8), 1976: 693-709.

Minsky, M., and Papert, S. Perceptrons. Cambridge, Massachussetts: MIT Press, 1969.

Minsky, M., and Papert, S. Perceptrons, Expanded Edition. Cambridge, Massachussetts: MIT Press, 1988.

Myers, R. H. Classical and Modern Regression with Applications. Boston, Massachussetts: Duxbury Press, 1986.

Neter, J., Wasserman, W., and Kutner, M. H. Applied Statistical Models. Homewood, Illinois: Irwin, 1985.

Pao, Y. "Functional Link Nets: Removing Hidden Layers." AI Expert, April 1989: 60-68.

Parker, D. B. "Learning logic." Invention Report S81-64, File 1, Stanford University, California, Office of Technology Licensing, 1982.

Parker, D. B. "Second order backpropagation: Implementing an optimal O(n) approximation to Newton's method as an artificial neural network." Manuscript submitted for publication, 1987.

Pineda, F. J. "Generalization of backpropagation to recurrent and higher order networks." Neural Information Processing Systems, 1988a: 602-611.

Pineda, F. J. "Dynamics and Architecture for Neural Computation." Journal of Complexity, 4, 1988b: 216-245.

Robbins, H., and Monro, S. "A Stochastic Approximation Method." Annals of Mathematical Statistics, 22, 1951: 400-407.

Rohatgi, V. K. An Introduction to Probability Theory and Mathematical Statistics. New York: Wiley, 1976.

Rosenblatt, F. "The perceptron: A probabilistic model for information storage and organization in the brain." Psychoanalytic Review, 65, 1958: 386-408.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. "Learning internal representations by error propagation." Parallel Distributed Processing, 1, 1986: 318-362.

Schor, R. H. "Design and Fitting of Neural Network Transfer Functions." Biological Cybernetics, 51, 1985: 357-362.

Schwartz, T. J. "8 Parables of Neural Networks." AI Expert, December 1989: 54-59.

Sejnowski, T. J., and Rosenberg, C. R. "Parallel networks that learn to pronounce English text." Complex Systems, 1, 1987: 145-168.

Shea, P. M., and Lin, V. "Detection of Explosives in Checked Airline Baggage Using an Artificial Neural System." Proceedings of the IEEE International Conference on Neural Networks, II, 1989: 31-34.

Smith, F. W. "A Trainable Nonlinear Function Generator." IEEE Transactions on Automatic Control, AC-11, April 1966: 212-218.

SAS Institute, Inc. SAS User's Guide: Statistics. Cary, North Carolina: SAS Institute, Inc.

Steinbuch, K. "Die Lernmatrix." Kybernetik, 1, 1961: 36-45.

Stornetta, W. S., and Huberman, B. A. "An improved three-layer, backpropagation algorithm." Proceedings of the IEEE International Conference on Neural Networks, 1987.

Toulouse, G., Changeux, J., Dehaene, S., and Nadal, J. "Neural Networks and Models for Learning and Memory." Organization of Neural Networks: Structure and Models. W. von Seelen, G. Shaw, U. M. Leinhos (ed.). Weinheim, Germany: VCH Verlagsgesellschaft, 1988.

Von Lehmen, A., Paek, E. G., Liao, P. F., Marrakchi, A., and Patel, J. S. "Factors Influencing Learning by Backpropagation." Proceedings of the IEEE International Conference on Neural Networks, I, 1988: 335-339.

Wang, J., and Malakooti, B. "On Training of Artificial Neural Networks." Proceedings of the IEEE International Conference on Neural Networks, II, 1989: 387-393.

Wasserman, P. D. Neural Computing. New York, New York: Van Nostrand Reinhold, 1989.

Werbos, P. J. "Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences." Ph.D. Dissertation, Harvard University, 1974.

Werbos, P. J. "Generalization of Backpropogation with Application to a Recurrent Gas Market Model." Neural Networks, 1, 1988: 339-356.

White, H. "Consequences and Detection of Misspecified Nonlinear Regression Models." Journal of the American Statistical Association, 76, 1981: 419-433.

White, H. "Learning in Artificial Neural Networks : A Statistical Perspective." Neural Computation, 1, 1989a:

White, H. "Some Asymptotic Results for Learning in Single Hidden-Layer Feedforward Network Models." Journal of the American Statistical Association, 84 (408), 1989b: 1003-1013.

White, H. "Neural Networks: Learning and Statistics." AI Expert, December 1989c: 48-52.

Widrow, B., and Hoff, M. E. "Adaptive Switching Circuits." WESCON convention, record Part IV, 1960: 96-104.

Widrow, B., and Winter, R. "Neural Nets for Adaptive Filtering and Adaptive Pattern Recognition." Computer, March 1988: 25-39.

Appendix A. Neural Network Basics

Neural networks are designed to mimic parts of the brain. The brain is made of many neurons comprised of dendrites, axons, and a nucleus. These neurons are interconnected through connection points called synapses. A neuron receives stimuli from other neurons through its dendrites at the synaptic gaps. The stimulus is weighted by the chemical ″strength″ of the synaptic gap. This weight represents the stored memory of the brain and incorporates all ″learned″ experiences. The weighted stimuli are combined by the nucleus to produce a response released through its axon, which then serves as stimuli to other neurons.

Neural networks are made up of interconnected processing elements (PE), which are the equivalent of neurons. As indicated in Figure 12 on page 160, the PE has inputs, denoted by X′s, coming from other PE′s or from the output of this PE looping back to itself. A weight of W_{ij} is assigned to each input and the weighted inputs are then combined by the PE, usually by a summation function, $I_i = \sum_j W_{ij}X_j$. The PE then transforms the weighted input to generate an output, Y. Typically, a sigmoid function, such as $Y = 1/(1 + \exp(-I_i))$, is used for the transformation. The PE learns by modifying its weights, the W_{ij}.

The two main phases in network operation are learning and recall. The process of modifying weights in response to sets of input and desired outputs is called learning. The learning rule of a

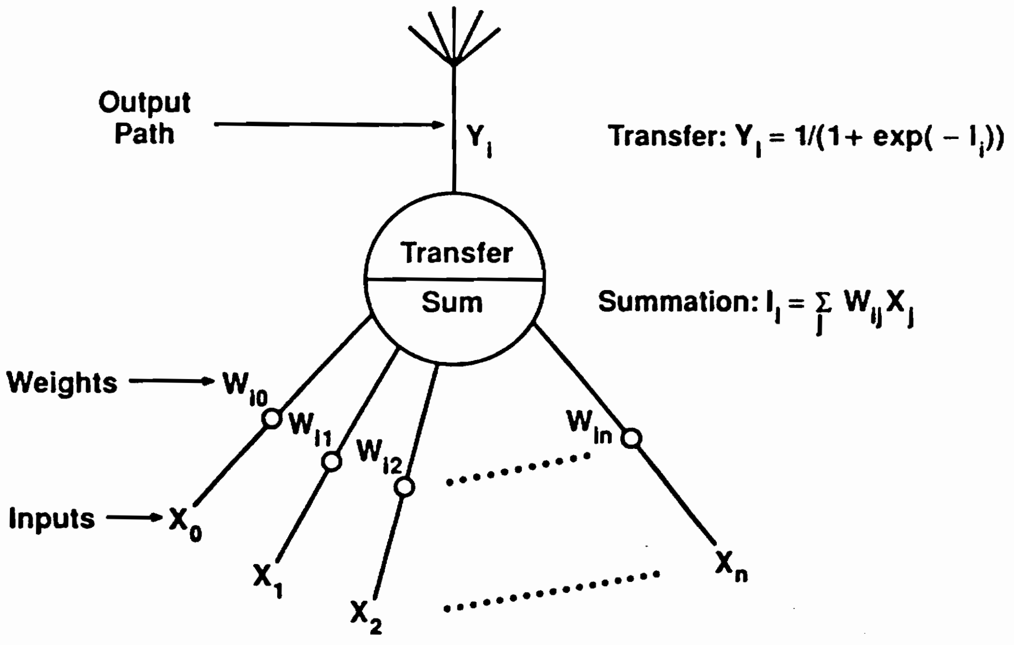


Figure 12. Mathematical model of a processing element in a neural network

network specifies how weights adapt in reponse to a learning example. Recall refers to how the network processes a stimuli presented at its input layer and creates a response at its output layer.

The structure of a neural network is defined by the interconnection architecture between the PE's, the transfer function, and the training laws. There are many possible network configurations of PE's, the most common one being the backpropagation network. The architecture of a backpropagation network, similar to the one that will be used in this research, is shown in Figure 13 on page 162. There is an input layer, one or more hidden layers, and an output layer of PE's. Backpropagation networks are used for supervised learning, where training data comprised of a set of inputs and the corresponding set of outputs are used to teach the network.

When input is presented to a neural network, by way of the input layer PE's, each input is multiplied by the weight W_{ij} connecting that input to the next PE in the middle layer. The weighted input is then presented to the hidden layer. In the case shown in Figure 13 on page 162, the three PE's in the input layer are presented with input of $X_1 = X_3 = 1$, and $X_2 = 0$. The input layer of a backpropagation network acts as buffer and no processing is done in this layer. The output of this first layer is the same as its input. The output is then transmitted to the two PE's in the hidden layer, attenuated by the weight along the paths indicated in Figure 13 on page 162. For example, the output from node 1 is multiplied by the weight W_{41}, i.e. 0.5, before being presented as input to node 4. Thus node 4 receives three weighted inputs, one from each input layer PE, namely 0.5, 0.0, and 0.4. The weighted inputs are summed at each PE in the hidden layer and then modified by the transfer function in the PE. With respect to node 4, the weighted input sum of 0.9 is transformed by the sigmoid function to yield the output of the hidden layer, $Y = 1/(1 + \exp(-0.9)) \simeq 0.71$. The output of the hidden layer is then presented as input to the output layer. The same weighting and transformation is carried out on the input to the output layer. The output of each PE in the output layer, in this example 0.61 and 0.65, is compared to the desired output from the training set, shown as $d_6 = 1$ and $d_7 = 0$. The error is computed as the difference between desired and actual output for each PE.

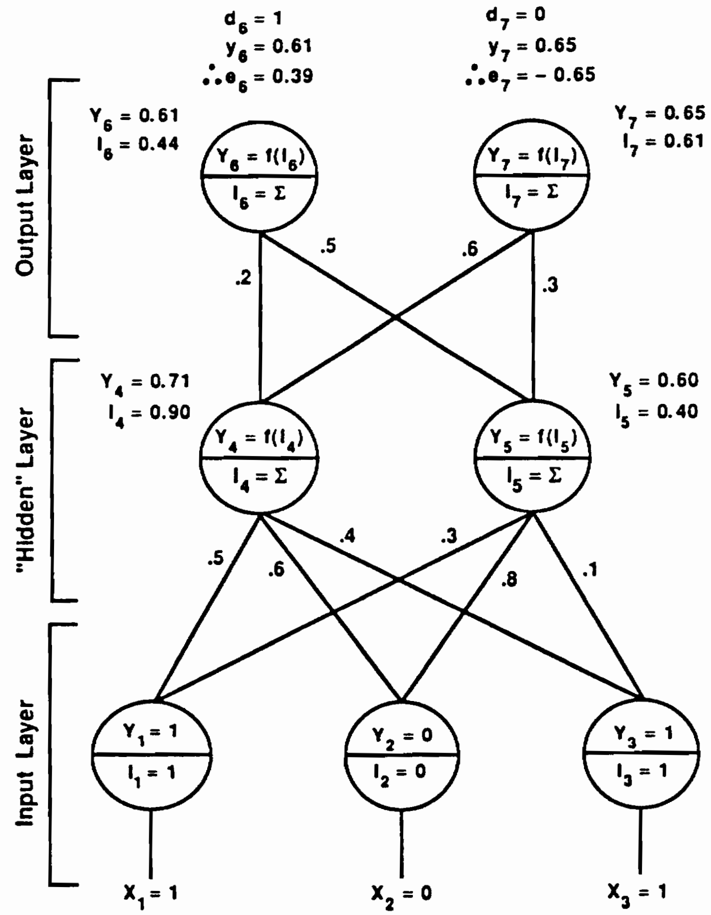


Figure 13. A three-layer backpropagation neural network

The error at each PE is propagated backwards through the network and weight changes are made throughout according to an algorithm to minimize the error. It is this process that gives the network its name. As each training set is presented to the network, the process is repeated with the weights being further modified each time.

The entire training data is typically presented to the network thousands of times before the network is trained, i.e. the error is reduced and stabilized. At this point input data never seen by the network can be presented to observe the output generated. For additional information on neural networks and the backpropagation network the reader is referred to Klimasauskas, et al. (1989b) and Wasserman (1989).

Appendix B. SAS Program to Generate Normally Distributed Populations

```
OPTIONS NODATE LS = 79;
TITLE 'GENERATING NORMALLY DISTRIBUTED POPULATION ';
TITLE2 'WHERE VARIANCE IS 100';
CMS FI OUT DISK VAR OUT A1;
DATA NORM; N = 10000;
        SEED4Z1 = 870724; SEED4Z2 = 890505;
        EM = 0; ES = 10; XM = 50; XS = 15;
        ALPHA = 10; BETA = .8;
    DO I = 1 TO N;
        CALL RANNOR(SEED4Z1, Z1); CALL RANNOR(SEED4Z2, Z2);
        E = EM + ES*Z1; X = XM + XS*Z2;
        Y = ALPHA + BETA*X + E;
        OUTPUT;
        FILE OUT; PUT X Y;
    END;
```

Appendix C. SAS Program to Estimate Regression Parameters

```
OPTIONS NODATE LS=79;
TITLE 'ESTIMATION OF REGRESSION PARAMETERS';
CMS FI INP DISK TV2020 NNA A1;
DATA; INFILE INP;
   INPUT X Y;
PROC GLM; MODEL Y=X/P CLM  ALPHA=.05;
```

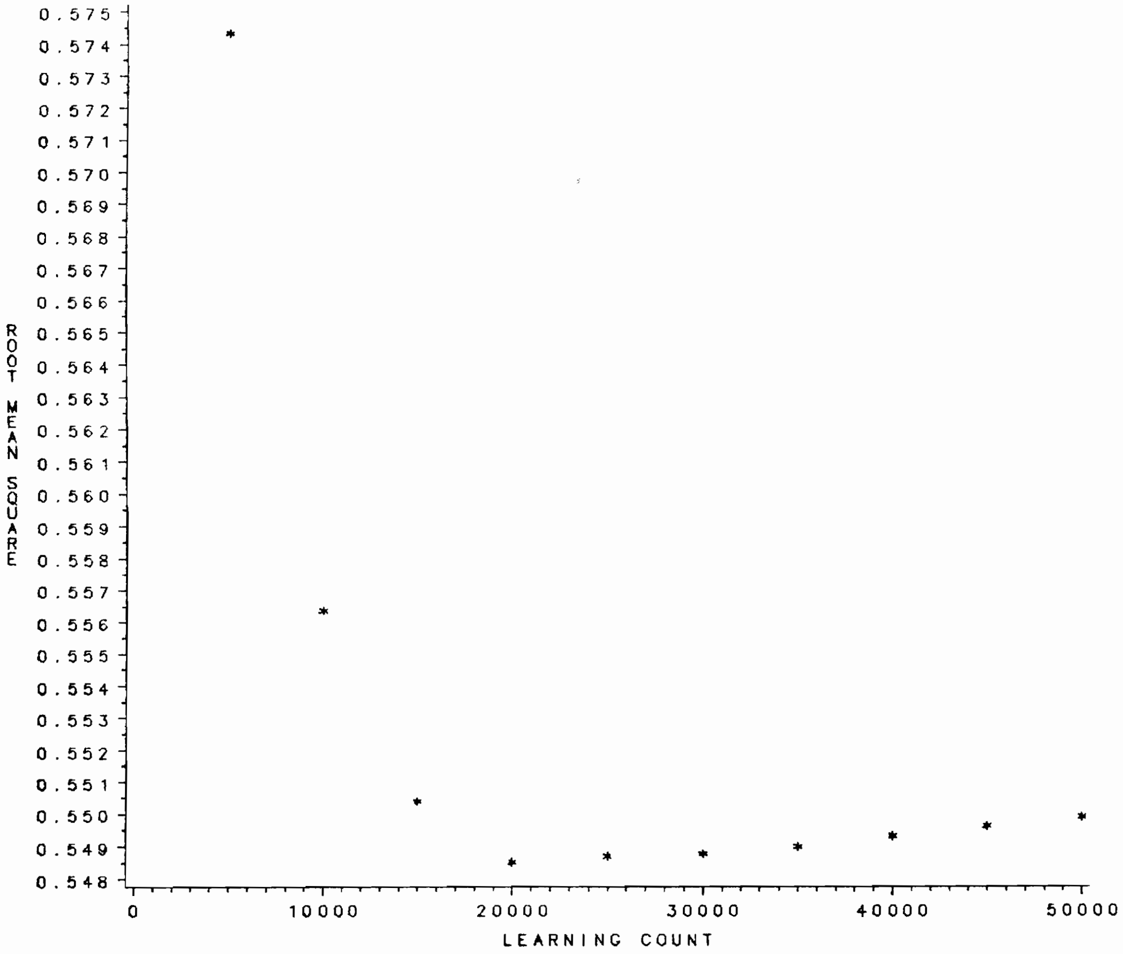
Appendix D. SAS Program to Predict Using Regression

```
OPTIONS NODATE LS=79;
TITLE 'PREDICTION USING PREDETERMINED REGRESSION PARAMETERS';
CMS FI INP DISK V2020   NNA A1;
CMS FI OUT DISK V2020   SAS A1;
DATA; INFILE INP;
   INPUT X Y;
    YHAT =  -3.085710027 + (0.995556650 * X);
   OUTPUT;
    FILE OUT; PUT X YHAT;
```

Appendix E. BASIC Program to Calculate Root Mean Square

```
20 CLS
30 COLOR 15,1
40 OPEN ″V2020.NNA″ FOR INPUT AS #1
60 OPEN ″V2020.NNR″ FOR INPUT AS #2
90 SUMSQ = 0
100 FOR I = 1 TO 20
175 INPUT #1,X,Y
190 INPUT #2,XH,YH
260 DIF = Y-YH
265 PRINT DIF
280 SUMSQ = SUMSQ + (DIF**2)
285 PRINT SUMSQ
290 NEXT I
320 MSQ = (SUMSQ**.5)/20
340 PRINT USING ″THE RMS DIFFERENCE IS : ##.#####″;MSQ
```

Appendix F. Plot of Root Mean Square versus Learning Count



Appendix G. SAS Program to Calculate the RMS Difference

```
OPTIONS NODATE LS=79;
TITLE 'CALCULATE THE DIFFERENCE BETWEEN NN AND REGRESSION';
CMS FI INP DISK V10500 DATA A1;
CMS FI OUT DISK D10500 DATA A1;
DATA; INFILE INP;
   INPUT VAR SIZE NN REG;
    DIFF  =  NN - REG;
   OUTPUT;
    FILE OUT; PUT VAR SIZE DIFF;
```

Appendix H. SAS Program to Test for Normality

```
OPTIONS NODATE LS=79;
TITLE 'KOLMOGOROV-SMIRNOV TEST FOR NORMALITY';
CMS FI INP DISK D5200 DATA A1;
DATA KS; INFILE INP;
    INPUT VAR SIZE DIFF;
PROC SORT DATA=KS;
BY DIFF;
DATA TEST; SET KS;
COUNT+1;
SX = COUNT/20;
FX = PROBNORM((DIFF - .1963765)/.0464366);
D = ABS(SX - FX);
PROC PRINT;
PROC MEANS MAX;
VAR D;
```

Appendix I. MINITAB Commands to Test Homoscedasticity

```
MTB > READ C1-C3
MTB > END
MTB > LET C4 = C3**2
MTB > LOG OF C4, PUT IN C5
MTB > LET C6 = 19*C4
MTB > LET C7 = 19*C5
MTB > NAME C1='VAR' C2='SIZE' C3='S' C4='SSQ' C5='LOG SSQ' C6='DFSSQ'
MTB > NAME C7='DFLOGSSQ'
MTB > PRINT C1-C7
MTB > LET K1 = SUM(C6)
MTB > LET K2 = SUM(C7)
MTB > LET K3 = K1/380
MTB > LOG OF K3, PUT IN K4
MTB > LET K5 = 380*K4
MTB > LET K6 = K5 - K2
MTB > NOTE    K7 = C   AND   K8 = B
MTB > LET K7 = 1 + (1/(3*19))*((20/19)-(1/380))
MTB > LET K8 = (1/K7)*K6
MTB > PRINT K1-K8
K1      5.79893
K2      -2103.10
K3      0.0152603
K4      -4.1825
K5      -1589.35
K6      513.752
K7      1.01842
K8      504.459
MTB > SAVE 'BARTLETT'
```

Appendix J. NPSP Program for Two-way Layout with Interactions

```
//B0392INA JOB 35A39,INA,TIME=3,REGION=2048K
/*JOBPARM LINES=20,ACCTPG
/*ROUTE PRINT VTVM1.INA
/*PRIORITY STANDARD
// EXEC PGM=NPSP
//STEPLIB DD DSN=A51240.NPSP.LOAD,DISP=SHR
//FT05F001 DD DDNAME=SYSIN
//FT06F001 DD SYSOUT=A
NONPARAMETRIC 2WAY ANOVA: FACTOR 1 (TRT) VARIANCE,
FACTOR 2 (BLK) SIZE
TWOWAY 5   4   5  20
(4X,F8.5)
GLMR2WAY 1
FINISHED
::
```

```
HETTMANSPERGER-MCKEAN PROCEDURE FOR TWO-WAY ANALYSIS OF
VARIANCE WITH INTERACTION,
USING THE RANKS OF THE RESIDUALS


   ANOVA TABLE

   SOURCE                  REDUCTION    D.F.  MEAN REDUCTION   F-RATIO

   MODEL - INTERACTION        3.3973     12         0.2831       8.6710

   ERROR                                380         0.0326

COMPARE THE F-RATIO TO AN F-DISTRIBUTION WITH  12
AND 380 DEGREES OF FREEDOM.


   ANOVA TABLE

   SOURCE                  REDUCTION    D.F.  MEAN REDUCTION   F-RATIO

   MODEL - FACTOR I          40.2294     3         13.4098      321.9443
```

```
          (TREATMENTS)

   ERROR                                  392          0.0417

COMPARE THE F-RATIO TO AN F-DISTRIBUTION WITH  3
AND 392 DEGREES OF FREEDOM.


   ANOVA TABLE

   SOURCE                REDUCTION    D.F.  MEAN REDUCTION   F-RATIO

   MODEL - FACTOR II        7.7642    4          1.9410       46.6008
          (BLOCKS)

   ERROR                                  392          0.0417

COMPARE THE F-RATIO TO AN F-DISTRIBUTION WITH  4
AND 392 DEGREES OF FREEDOM.
```

Appendix K. SAS Program for Nonparametric One-way Layout

```
OPTIONS NODATE LS=79;
TITLE 'NONPARAMETRIC ONE-WAY ANOVA';
CMS FI INP DISK VAR20 DATA A1;
DATA RAW; INFILE INP;
    INPUT VAR SIZE DIFF;
PROC NPAR1WAY ANOVA;
    CLASS SIZE;
    VAR DIFF;
RUN;
```

Appendix L. SAS Program to Generate Populations with Outliers in Error Distribution

```
OPTIONS NODATE LS=79;
TITLE 'GENERATING POPULATION WHERE ERROR TERM IS DISTRIBUTED';
TITLE2 'WITH OUTLIERS - 95 PERCENT OF ERROR IS FROM A NORMAL';
TITLE3 'DISTRIBUTION WITH MEAN 0 AND VARIANCE 100, 5 PERCENT OF';
TITLE4 'TIME IT HAS A VARIANCE OF 900';
CMS FI OUT DISK OUTL30 DATA A1;
DATA A; N=500;
      SEED4U=724; SEED4Z=515; SEED4X=812;
      ME1=0; SE1=10; ME2=0; SE2=30;
      MX=50; SX=15;
      ALPHA=10; BETA=.8;
      P=0.05;
    DO I=1 TO N;
      CALL RANUNI(SEED4U, U);
      CALL RANNOR(SEED4Z, Z);
      CALL RANNOR(SEED4X, V);
      IF U>P THEN DO; E1=ME1+SE1*Z; E2=.; DIST=1; E=E1; END;
      ELSE        DO; E2=ME2+SE2*Z; E1=.; DIST=2; E=E2; END;
      X=MX+SX*V;
      Y=ALPHA+BETA*X+E;
      OUTPUT;
      FILE OUT; PUT X Y;
    END;
```

Appendix M. Normal Probability Plots for Moderate Outlier Case

```
 37.5+                                  *
    |                                **
    |                               **+
    |                            ****
    |                         *****
    |                      + *****
    |                  *******
    |                *****
 -2.5+              ******
    |         ******
    |      ******
    |   ******
    | ***+ +
    |+ *
    |*
    |
-42.5+ *
     +----+----+----+----+----+----+----+----+----+----+----+
       -2     -1      0      +1       +2
```

Appendix N. SAS Program to Select Beta Distribution Parameters for Skewness and Kurtosis

```
OPTIONS NODATE LS=79;
TITLE 'SELECTED SKEWNESSES AND KURTOSOSES IN BETA DISTRIBUTIONS';
TITLE2 'P IS ALPHA AND Q IS BETA';
DATA A; DO P=1.0    TO 5.5   BY 0.5;
      DO Q=1.0    TO 5.5   BY 0.5;
   M=P/(P+Q);
   V=P * Q * (P+Q)**(-2) * (P+Q+1)**(-1);
   S=SQRT(V);
   SKEWNESS=(2 * (Q-P) * SQRT(P+Q+1))/((P+Q+2) * SQRT(P*Q));
   KURTOSIS=((3 * (P+Q) * (P+Q+1) * (P+1) * ((2*Q)-P))/((P*Q) * (P+Q+2)
         * (P+Q+3))) + ((P * (P-Q))/(P+Q));
OUTPUT; END; END;
PROC SORT; BY SKEWNESS;
PROC PRINT; VAR SKEWNESS KURTOSIS P Q M V S;
```

Appendix O. SAS Program to Generate Populations with Skewed Error Distributions

```
OPTIONS NODATE LS=79;
TITLE 'RANDOM NUMBERS - BETA DISTRIBUTION - POSITIVE SKEWNESS';
CMS FI OUT DISK POSKW OUT A1;
DATA RANBET; N=5000; ALPHA1=0.5; ALPHA2=1.5;
    SEED4E1=12345; SEED4E2=67890; SEED4X=52396;
    MX=50; SX=15;
    ALPHA=10; BETA=.8;
    DO I=1 TO N;
    CALL RANGAM(SEED4E1, ALPHA1, E1);
    CALL RANGAM(SEED4E2, ALPHA2, E2);
    CALL RANNOR(SEED4X, V);
    E=E1/(E1+E2);
    X=MX+SX*V;
    Y=ALPHA+BETA*X+E;
    OUTPUT; FILE OUT; PUT X Y;
    END;
PROC UNIVARIATE PLOT NORMAL; VAR E;
```

Appendix P. SAS Program to Generate Populations in the Multicollinearity Cases

```
OPTIONS NODATE LS = 79;
TITLE 'GENERATING RANDOM NUMBERS FROM A BIVARIATE';
TITLE2 'NORMAL DISTRIBUTION';
CMS FI OUT DISK MULTI DATA A1;
DATA A; N = 10000;
     SEED4Z1 = 724; SEED4Z2 = 515; SEED4Z3 = 812;
     MX1 = 50; SX1 = 15; MX2 = 50; SX2 = 15; R = 0.9;
     ME = 0;  SE = 10;
     ALPHA = 10; BETA = .8; GAMMA = .7;
     P = 1-R;
   DO I = 1 TO N;
     CALL RANNOR(SEED4Z1, Z1);
     CALL RANNOR(SEED4Z2, Z2);
     CALL RANNOR(SEED4Z3, Z3);
     X1 = MX1 + SX1*Z1;
     X2 = MX2 + SX2*(R*Z1 + P*Z2);
     E = ME + SE*Z3;
     Y = ALPHA + BETA*X1 + GAMMA*X2 + E;
     OUTPUT;
     FILE OUT; PUT X1 X2 Y;
   END;

PROC UNIVARIATE PLOT NORMAL; VAR X1 X2;
PROC PLOT; PLOT X1*X2/ VPOS=20 HPOS=40;
PROC CORR; VAR X1; WITH X2;
PROC REG; MODEL Y=X1 X2/ VIF COLLINOINT;
       MODEL Y=X1;
       MODEL Y=X2;
```

Appendix Q. SAS Program to Generate Populations with Quadratic Term

```
OPTIONS NODATE LS=79;
TITLE 'GENERATING POPULATION WITH QUADRATIC TERM';
CMS FI OUT DISK QUAD DATA A1;
DATA A; N=10000;
      SEED4X=724; SEED4E=515; SEED4Z3=812;
      MX=50; SX=15;
      ME=0;  SE=10;
      ALPHA=10; BETA=5; THETA=-10;
    DO I=1 TO N;
      CALL RANNOR(SEED4X, Z1);
      CALL RANNOR(SEED4E, Z2);
      X=MX+SX*Z1;
      XSQR=X**2;
      E=ME+SE*Z2;
      Y=ALPHA+BETA*X+THETA*XSQR+E;
      OUTPUT;
FILE OUT; PUT X XSQR Y;
    END;
```

Appendix R. SAS Program to Generate Population for Exponential Model

```
OPTIONS NODATE LS=79;
TITLE 'GENERATING FOR EXPONENTIAL MODEL';
CMS FI OUT DISK EXPO DATA A1;
DATA A; N=10000;
    SEED4X=724; SEED4E=515; SEED4Z3=812;
    MX=6; SX=2;
    ME=0;  SE=10;
    ALPHA=.5; BETA=.8;
  DO I=1 TO N;
    CALL RANNOR(SEED4X, Z1);
    CALL RANNOR(SEED4E, Z2);
    X=MX+SX*Z1;
    E=ME+SE*Z2;
    TEMP=EXP(BETA*X);
    Y=ALPHA*TEMP+E;
    OUTPUT;
FILE OUT; PUT X Y;
  END;
```

Appendix S. SAS NLIN Procedure

```
OPTIONS NODATE LS=79;
TITLE 'NONLINEAR REGRESSION - MARQUADT';
CMS FI INP DISK TNLR NNA A1;
DATA; INFILE INP;
   INPUT X Y;
PROC NLIN METHOD=MARQUADT;
PARMS ALPHA=.5 BETA=.8;
TEMP=EXP(BETA*X);
MODEL Y=ALPHA*TEMP;
DER.ALPHA=TEMP;
DER.BETA=ALPHA*X*TEMP;
OUTPUT OUT=STAT P=YHAT SSE=SSE PARMS=ALPHA BETA;
PROC PLOT; PLOT Y*X='X' YHAT*X='*'/OVERLAY;
```

Vita

The author was born in Mentakab, Malaysia on November 24, 1955 and she graduated from the Dato Ahmad High School in Lenggong, Malaysia, in 1973. She attended Lady Brabourne College, in the Calcutta University system, in Calcutta, India, and graduated (with honors) with a Bachelor of Arts degree in Economics in 1979. The author then entered the master's program in Economics at Calcutta University and graduated with a Master of Arts degree in Economics in 1982.

During the last semester of her master's work, the author was employed by Robson, Black & Ghosh, (Management Consultants) in Calcutta, India. She worked as a programmer analyst for the next three years. Following a summer at home in Malaysia, the author entered the M.B.A. program at Virginia Tech and received a Master's degree in Business Administration in 1986.

In 1987, the author entered the doctorate program in Management Science at Virginia Tech. She worked as an instructor in the Management Science department from Spring 1988 to Spring 1991. In the summer of 1989 she married Steven Edward Markham, Professor of Management. The author received a Doctor of Philosophy in Management Science in May 1992.

Ina Markham