UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

MODEL SELECTION USING AN INFORMATION THEORY APPROACH

A DISSERTATION

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

Degree of

DOCTOR OF PHILOSOPHY

By

ALI SALEH SHAQLAIH
Norman, Oklahoma
2010

Downloaded from SHAREOK. Persistent Handle URL: https://hdl.handle.net/11244/318482

MODEL SELECTION USING AN INFORMATION THEORY APPROACH

A DISSERTATION APPROVED FOR THE
DEPARTMENT OF MATHEMATICS

BY

Dr. Luther White, Chair

Dr. Kevin Grasse

Dr. Semion Gutman

Dr. Marilyn Breen

Dr. Musharraf Zaman

© Copyright by ALI SALEH SHAQLAIH 2010
All Rights Reserved.

DEDICATION

to

My father: Saleh,
and
The soul of my mother: Jazia.

Acknowledgements

First, I wish to express my gratitude to my research advisor, Professor Luther White, for teaching me a great deal of Mathematics, for encouraging me to explore and work in Mathematics, for his guidance, kindness and unconditional support during all these years of graduate school. I am grateful for having the opportunity to work with him.

I also wish to thank my doctoral committee members Dr. Musharraf Zaman, Dr. Kevin Grasse, Dr. Semion Gutman and Dr. Marilyn Breen for their valuable notes and helps in various ways and for serving in my committee.

Thanks are also to my wife, my kids, my brothers, all members of my family and all my friends for all kinds of support.

Contents

1 Introduction **1**

2 Application to Statistical Models **8**

2.1 Review of Literature for Resilient Modulus 8

2.2 Formulation of Models 10

2.3 The Models 11

2.3.1 Stress-Based Model 11

2.3.2 Multiple Regression Model 12

2.3.3 polynomial Model 13

2.3.4 Factorial Model 13

2.4 R^2 Analysis 16

2.5 AIC Analysis 18

2.6 Stability 21

3 Application to Neural Network Models **27**

3.1 Introduction 27

3.2 Network Architectures 30

3.3 Learning Rules 32

3.3.1 Supervised learning 32

3.3.2 Reinforcement learning 32
3.3.3 Unsupervised learning 33
3.4 The Models 33
3.4.1 Linear neural network model-LNN 34
3.4.2 General Regression Neural Network Model-GRNN 34
3.4.3 Radial Basis Function Network model-RBFN 35
3.4.4 Multi-Layer Perceptrons Network Model-MLPN 35
3.5 R^2 Analysis 36
3.6 AIC Analysis 37
3.7 Stability 40

4 Application to Physics-Based Models 47
4.1 Introduction 47
4.2 The girders Models 49
4.3 Data Analysis 55

5 Conclusion and Future Work 60
6.1 Least Square Theory 66
6.2 Likelihood Theory 68
6.3 Models 70
6.4 Model parameters 71
6.5 The principle of Parsimony 72
6.6 Cross validation 74
6.6.1 Holdout method 74
6.6.2 K-fold cross validation 75
6.6.3 Leave-one-out cross validation 75
6.7 Bootstrapping 76

6.8 Bayesian Method 77
6.9 Mallows' Cp 77
6.10 R^2 Method 78
6.11 Null Hypothesis Testing 78
6.11.1 Null hypothesis Testing Procedure 78
6.11.2 Problems with Null Hypothesis Testing 79
6.12 Choosing the variables in a model 80
6.12.1 Forward Selection 81
6.12.2 Backward Elimination 81
6.12.3 Stepwise 81
6.12.4 Principal Component Analysis (PCA) 82
7.13 Kullback-Leibler Information 86
7.14 Akaike's Information Criterion 90
7.15 AIC_c: A second order improvement 94
7.16 Confidence Set of Models 95
7.17 Relative Importance of Variables 97
7.18 Model Averaging 98

List of Tables

2.1 R^2 values (development and evaluation data sets) 16

2.2 AIC values (development data set) 19

2.3 AIC values (evaluation data set) 20

2.4 AIC and R^2 Ranking (development data set) 22

2.5 AIC and R^2 ranking (evaluation data set) 22

2.6 AIC and R^2 Ranking (Development and Evaluation Data sets) . . 23

3.1 R^2 values (development and evaluation data sets) 37

3.2 AIC values (development data set) 38

3.3 AIC_c values (evaluation data set) 39

3.4 AIC and R^2 ranking (development data set) 41

3.5 AIC and R^2 ranking (evaluation data set) 42

3.6 AIC and R^2 Ranking (development and evaluation data sets) . . . 42

4.1 AIC values 58

List of Figures

2.1 Experimental and Predicted values for Evaluation Data set, Factorial Model 25
2.2 Experimental and Predicted values for Evaluation Data set, polynomial Model 26

3.1 Experimental and Predicted values for Evaluation Data set, MLPN-1 44
3.2 Experimental and Predicted values for Evaluation Data set, MLPN-2 45
3.3 Experimental and Predicted values for Evaluation Data set, MLPN-3 46

Abstract

In this thesis we use the information theoretic approach in selecting the best model among many candidate models. It is shown that the information theoretic approach is better than the standard R^2 approach in selecting models. We use Akaike Information Criteria (AIC) to select the best model for resilient modulus of a soil and for a girder. This approach is applied to statistical models, neural network models and physics based models. The information theory approach is compared with the R^2 approach and it is found that the information theoretic approach is more stable and gives better results. The notion of ranking stability is introduced and is used as one of the reasons that makes information theory approach better than the R^2 approach. Important results are captured and compared to the results of the R^2 method in two different data sets.

Chapter 1

Introduction

With the increase ease of data collection and the more and more need for prediction, there is a growing need for methods of model selection. Model selection is the task of selecting a model from a set of potential models, given data [15]. A fundamental problem in applications is how to interpret data in the context of models for the purpose of eventually making predictions. We consider the situation in which we are presented with data. This data is considered an output dependent on various input parameters. The goal is to determine a functional dependence of the output on these input parameters so that predictions can be made. Typically in these situations, a family of relations are obtained using the given data. In fact, different members of the family may match different data better than others. The problem is then, how to determine or select which relation among the family that is best. For that matter, how does one decide what "best" means? Of the many possible models that one may have, how can one even begin to choose the model? Which model is the best? What is meant by best? These questions are in the preview of model selection.

Given a set of data that represent some actual measurement of some parameters. The goal is to recover the information that applies more generally to the process, not just to the particular data set. In fact that is why modeling is important; it helps us to predict. The better we can predict using the model, the better the model is. In model selection, we try to rank models in the candidate set relative to each other from best model to second best to poor. One might be close to reality model with, for example, 300 parameters, but it would be difficult understand the model and apply it. Thus one should tolerate some inexactness to facilitate a simpler model that gives easier understanding of the phenomenon. It is important that the best model is selected from a set of models that we have in the set of candidate models to be appropriately simple and precise [6].

There are typically three steps to studying and hence modeling a physical system [31]. First, a set of parameters are introduced to describe the system. Secondly, a forward problem or model is developed from physical principles or by fitting of data. The model allows us to make predictions of measurements of observable parameters given underlying physical parameters that may or may not be observable. Thirdly, inverse modeling attempts to determine information on physical parameters from data and measurements of the system and observable parameters. Ideally, there is an interaction among these steps to produce a collection of relations that constitute a model family. Since the objective is to make predictions, it is critical to assess the predictive value for the different family members. Indeed, the exercise has little value if, when applied to new data, one does not know which model to use or to believe.

It is important to realize that when we are looking for the best model, we are

not modeling the data but rather we are trying to model the information in the data. If we are modeling the data we could fit a high order fourier series terms or high degree polynomial terms until the fit is perfect but then it will be a very complex model that we can not deal with. Data contains both information and noise. Fitting the data perfectly would include modeling the noise and this is counter to our objective in modeling. Over fitting is a poor strategy and under fitting also means getting a poor model that will not give enough information, therefore we need a model that has a good balance between the over fitting and the under fitting [7].

In this thesis we apply the information theoretic techniques to a class of models given a collection of data set. Our work is heavily influenced by that described by Burnham and Anderson [6] for Biological models. In a previous study, [10], some statistical and neural network models were developed to model resilient modulus of a soil and then R^2 values of each model where used to decide which model is the best among the candidate models. Taking R^2 as a strategy to decide the best model is a very weak approach and it has many shortages [6]. In this thesis, models considered are statistical models, neural network models and Physics based models. Statistical models and neural network model are used to model the resilient modulus of a soil. The Physics based model are to model a girder from a given experimental data.

In this work, we assume that a collection of models has been given and that there exists an abstract model 'truth' that is not known to us. Although we do not actually know the truth model, there are available data consisting of observations of the truth model but, as we said, accompanied with noise. The task of model selection is to assess the ability of models in our family not only to fit

observation but to capture truth model behavior. Certainly, one way of capturing the data is to use a least squares criterion to determine physical parameters to fit observations. This amounts to using the so-called R^2 criterion as a criterion for model selection. While this procedure may determine best fit to data, it is well documented [6] that this does not necessarily produce a model that maximizes information and for that matter is the most useful. Baysian method and Mallows' Cp, are also well known methods for selecting the best model from a set of candidate models. For completeness, we summarized some of the famous methods of selecting a model in appendix A.

For the purpose of prediction, it is desirable to have a criterion that assesses models that is stable with respect to different data sets. Information criteria, specifically the Akaike information criteria (AIC), are statistical procedures that have been developed for just such a purpose. These techniques have received considerable attention in the literature and provide a collection of tools with which families of models may be analyzed with regard to their utility for capturing information. The work, [4], provides an excellent resource for the application of these information theoretic methods and influenced this work greatly. Furthermore, these procedures lead to rankings of the models within the candidate family and ways of combining information from the different models in the family. This leads in term to strategies for the determination of ensembles from which predictions may be made.

There are many reasons that made us consider this problem of modeling and some of these are:

• The importance of the resilient modulus (MR) in the mechanistic analysis of

the pavement system. Among major advantages, MR accounts for cyclic nature of vehicular traffic loading an d inelastic behavior that are particularly important for subgrade soils. MR also became the fundamental parameter in the AASHTO design guide to describe subgrade soils [3].

• Existence of a large data set from Ebrahimi study [10]. Having this large important set of data that was collected from different counties in the state of Oklahoma allows us to do this analysis to further investigate the best model that describes the Resilient modulus.

• The shortages that the R^2 method has. Although the R^2 method is largely used, it has many shortages that makes it not a good a strategy to use in model selection.

• The stability in AIC ranking of models. Stability of ranking is essential concept in model selection as if the rank is stable in two different data set, it will make it more appropriate to use.

• With all shortages of R^2 techniques and the stability of the AIC techniques, we thought it would be appropriate to use the AIC techniques in choosing the best model of the girders models.

We did this work through three steps. First, used the R^2 method in selecting the best model out of the set of candidate models. Second, we used the AIC techniques in deciding the best model in the same set of candidate models. Thirdly, to see which method is more stable, we apply these two methods on two different data sets and compared the ranking of models in each method for each data set. As we will see, the AIC ranking of models is stable for different data sets and thats makes it a better approach. By stability here, we mean if a model is ranked good for a certain data set, it will be ranked good in a different data set. For

that, stability in ranking models is very crucial.

This thesis consists of five chapters. In Appendix A, we give a brief introduction to likelihood theory and least square theory and the use of these two theories in modeling and we summarized the most known methods in model selection and briefly states the shortages and the advantages of each method. In Appendix B we went over the information theory approach and in details we put the techniques of this approach.

Chapter two mainly discusses the application of the information theory approach on statistical models that model resilient modules of a soil. We ranked the models by the R^2 method and by AIC techniques in both the development data set and the evaluation data set and then showed that the AIC approach was stable where as the R^2 ranking is not.

Chapter three discusses the application of the information theory approach on neural network models that model resilient modulus of a soil. We ranked the models by the R^2 method and by AIC techniques in both the development data set and the evaluation data set and then showed that the AIC approach is stable where as the R^2 ranking is not.

In chapter four, we used the information theory approach to choose the best model for modeling a girder and we found that the best model was the model that has two parameters. This tells us that having more variables in a model does not always gives a better model. Since the information theory approach takes under consideration the number of parameters in the model, it is not always a plus to

add more parameters to the model.

The results of this work are:

• The AIC approach in model selection is stable with respect to different data sets for the same parameters and therefore it is better to use than the R^2 approach.

• In modeling the resilient modulus of a soil and among the four statistical models, the only good model that should be considered is the factorial model.

• In modeling the resilient modulus of a soil by neural network models, the best model is MLPN-1 and the second is MLPN-2. All other considered models are very weak and should not be considered.

• For modeling a girder, the best model was the model that has two parameters and other models can be considered.

Chapter 2

Application to Statistical Models

In this chapter, we apply the information theory approach on statistical models to model resilient modulus of a soil and then compare it with the R^2 approach and show that the Information theory approach is more stable.

2.1 Review of Literature for Resilient Modulus

Resilient Modulus (MR) is the fundamental material parameter for mechanistic analysis of a multi-layered pavement system. MR is a measure of the elastic modulus of subgrade soils at a given stress level, and is defined as the ratio of an applied deviatoric stress σ_d to the recoverable strain ϵ_r so MR= $\frac{\sigma_d}{\epsilon_r}$ [3]. The resilient modulus (MR) of roadbed soils is a necessary parameter in pavement design since it is an expression of the elastic properties of the roadbed. The MR of roadbed soils is dependent on the soil type, water content, dry density, particle gradation, Atterberg limits, and stress states [34]. The new Mechanistic-Empirical Pavement Design Guide (MEPDG) allows the MR value of the roadbed soils to be determined from several different sources. The specific source to be used depends on the hierarchy of the applicable design level, which depends on the

class of pavement being designed and the available resources to the agency[34]. There are three levels of design/ analysis according to the Mechanistic-Empirical Pavement Design Guide(MEPDG) [13]:

Level 1: Actual laboratory resilient modulus testing is conducted to characterize the subgrade soil.

Level 2: Resilient modulus values are determined from other soil properties using correlations.

Level 3: Typical resilient modulus values are used based on soil classification.

Correlating MR with routine soil properties, as recommended for a Level- 2 design, is motivated by the fact that the determination of MR from laboratory and/or in-situ testing may be expensive and time consuming for certain applications, particularly for small projects. Different models have been proposed in the past to estimate MR from other soil properties and stress state. A majority of the existing models are based on statistical correlations of laboratory and/or field data [11].

Statistical models such as polynomial model and linear model are used to predict MR. For subgrade soils, the deviatoric stress was found to be a more influential parameter than the confining pressure [13]. In all the literature of modeling the MR, using statistical models, the value of R^2 was taking as an evidence of how the model fits the data and therefore decides the best model according to the value of R^2. We will show that this approach is very weak. We will introduce the use of the information theory approach in deciding the best model and we will show it is a better approach.

The data that is used in this study is the data that was used in Ebrahimi study [10]. This data is a total of 98 bulk soil samples and were collected from 16 different counties in the State of Oklahoma. This data was divided into two sets of data, the first is the development data set which is the data used in developing the models and the second is the evaluation data set which will be used in evaluating the fitting of the models. In fact, we will compare the ranking of the models in both data sets by the two methods.

2.2 Formulation of Models

For the formulation of a set of candidate models one should look at the published literature in the field of study, results of manipulating experiments and personal experience. Development of the a priors set of candidate models should include a global model which is a model that has many parameters, includes all potentially relevant affects and reflect causal mechanisms thought likely, based on the science of the situation. The global model should also reflect the study designed and attributes of the system studied. Tukey (1980) argues for the need for deep thinking and early exploratory data analysis, and that the results of these activities lead to good scientific questions and confirmatory data analysis. In this study, we have four statistical models which are: The stress-based model, the polynomial model, the multiple regression model and the factorial model. The factorial model here is considered the global model in this study as it contains all the independent variables and all combinations of them. The set of candidate models is chosen based on the literature published in many studies [10].

There are many approaches that people have been using to choose the best

model from a set of candidate models. All these approaches try to find the model that best close to the truth. As we stated earlier, many studies use the R^2 approach but this approach has many defects and there are many examples that show that this approach fails to choose the best model. In this study we will take the information theory approach and compare it with R^2 approach. By using the information theoretic approach, we mean using the value of the AIC of each candidate model to rank the models. One should ensure that that the same data set is used for each model, in other words, the same observations must be used for each analysis.

2.3 The Models

The set of candidate statistical models in this study consists of four models namely, Stress-Based Model, Multiple Regression Model, polynomial Model, Factorial Model. In each model, the dependent variable, MR, is correlated with seven independent variables, namely bulk stress (θ), deviatoric stress (σ_d), moisture content (w), dry density (γ_d), plasticity index (PI), percent passing No. 200 sieve ($P200$), and unconfined compressive strength (Uc). Of the seven independent variables used here only two (θ and σ_d) are stress-related. The five parameters ($\omega, \sigma_d, PI, P200, andUc$) are determined from routine soil testing [10].

2.3.1 Stress-Based Model

In this model, bulk stress (θ) and deviatoric stress (σ_d) are used as the model parameters, and they are correlated with MR as: $MR/Pa = k_1(\theta/Pa)^{k_2}(\sigma_d/Pa)^{k_3}$, where Pa represents atmospheric pressure, and k_1, k_2, and k_3 are regression constants.

The regression constants k_1, k_2, and k_3 are correlated with the selected soil properties or parameters ω, σ_d, PI, $P200$, and Uc. The dry density, σ_d, is normalized with respect to density of water and the unconfined compressive strength, Uc, is normalized with respect to the atmospheric pressure, Pa [10].

After fitting this model, it was found that $k_1 = 0.08789 + 0.1773(Uc/Pa) + 0.005048PI - -0.3967P200 + 1.2652w$, $k_2 = 0.5074 - -0.01336PI + 2.3432w - -0.3868\gamma_d, k3 = - - 0.6612 + 0.1589(Uc/Pa) - -0.2254P200$.
The R^2 for this model was 0.3226. R^2 tells us that the fit of this model is poor.

2.3.2 Multiple Regression Model

Multiple regression model is widely used because it is simple and linear in its variables and that makes it easy to use. The general equation for a multiple regression model for the independent variables utilized here could be expressed by the following equation:
$MR/Pa = b0 + b1w + b2(\gamma_d/\gamma_w) + b3PI + b4P200 + b5(Uc/Pa) + b6(\sigma_d/Pa) + b7(\theta/Pa)$ where bi represents the regression constants [10].

The model was found to be: $MR/Pa = 1.8050 - -0.4904w - -0.5747\gamma_d + 0.008083PI - -0.5123P200 + 0.2191(Uc/Pa) - -0.6401(s_d/Pa) - -0.0009399(\gamma/Pa)$.
The R^2 value was 0.4357. This tells us that the fitting is better than the stress-based model but still not that good.

2.3.3 polynomial Model

A polynomial model includes the basic components of a multiple regression model with the addition of higher order effects for the independent variables. For the independent variables considered here, a second order polynomial model could be expressed as follows:

$MR/Pa = b_0 + b_1 w + b_2 w^2 + b_3(\gamma_d/\gamma_w) + b_4(\gamma_d/\gamma_w)^2 + b_5 PI + b_6 PI^2 + b_7 P200 + b8P200^2 + b_9(Uc/Pa) + b_{10}(Uc/Pa)^2 + b1_{11}(\sigma_d/Pa) + b_{12}(\frac{\sigma_d}{Pa})^2 + b_{13}b7(\theta/Pa) + b14b7(\theta/Pa)^2$ where b_i represents the regression coefficients or models parameters [10].

Using the same data the polynomial model was found to be:

$$MR/Pa = 15.8002 + 2.9994w - 7.4142w^2 - 18.3291(\gamma_d/\gamma_w) + 5.4596(\gamma_d/\gamma_w)^2 + 0.02191PI - 0.0003142PI^2 - 0.3705P200 - 0.009229P200^2 + 0.2628(Uc/Pa) - 0.01050(Uc/Pa)^2 - 2.0332(\sigma_d/Pa) + 1.62950(\sigma_d/Pa)^2 - 0.01181(\theta/Pa) + 0.004735(\gamma/Pa)^2.$$

The R^2 for this model was found to be 0.4858 which is clearly better than the previous two models.

2.3.4 Factorial Model

Similar to the polynomial model, a factorial model also includes the components of a multiple regression model. However, instead of considering higher order effects of the independent variables, it accounts for interactions among different variables in the model. A full-factorial regression model consists of all possible products of the independent variables. The general equation for a fractional factorial design with second degree of interaction can be expressed as follows:

$$MR/Pa = b_0 + b_1 w + b_2(\sigma_d/\gamma_d) + b_3 PI + b_4 P_{200} + b_5(Uc/Pa) + b_6(\sigma_d/Pa) + b7(\theta/Pa) + b8w(\sigma_d/\gamma_w) + b_9 wPI + b_{10} wP200 + b_{11} w(Uc/Pa) + b_{12} w(\sigma_d/Pa) + b_{13} w(\gamma/Pa) + b_{14}(\sigma_d/\gamma_d)PI + b_{15}(\sigma_d/\gamma_w)P200 + b16(\sigma_d/\gamma_w)(Uc/Pa) + ... + b_{26} Uc(\sigma_d/Pa) + b_{27}(Uc/Pa)(U_C/Pa) + b28(\gamma/Pa)(\sigma_d/Pa).$$

The factorial model was found to be:

$$MR/Pa = 13.2514795 - 438.31923w - 13.311426\gamma_d + 2.41669221PI + 27.2918109P200 - 35.722370Uc - 11.240229\sigma_d + 85.5626222\theta + 278.441637w\gamma_d - 16.168513wPI - 1.0992907\gamma_d PI + 14.4631546wP200 - 9.7399663\gamma_d P200 - 1.7766282PIP200 + 332.908859wUc + 23.0157253\gamma_d Uc - .01410921PIUc + 38.7543639P200Uc + 799.099567w\sigma_d + 31.3535709\gamma_d\sigma_d - 3.3563515PI\sigma_d - 69.037455 * P200\sigma_d - 47.303113 * Uc * \sigma_d - 279.93578 * w\gamma - 38.643251 * \gamma_d * \gamma - 1.0518368 * PI * \gamma - 84.485170 * P200 * \gamma - 29.795776 * Uc * \gamma - 106.57364 * \sigma_d * \gamma + 9.15134484 * w * \gamma_d * PI - 28.679440 * w * \gamma_d P200 + 21.1200089 * w * PI * P200 + .662040611 * \gamma_d PIP200 - 200.60637w\gamma_d * Uc + 2.75148494 * w * PI * Uc - .02351306 * \gamma_d * PI * Uc - 244.11931 * w * P200 * Uc - 25.359274 * \gamma_d * P200 * Uc - 1.2799791 * PI * P200 * Uc - 588.87015 * w * \gamma_d\sigma_d + 15.1565883 * w * PI * \sigma_d + .481801078 * \gamma_d * PI * \sigma\sigma_d - 218.18881 * w * P200\sigma_d + 10.1638379\gamma_d * P200 * \sigma_d + 2.36460454 * PI * P200 * \sigma_d - 45.492450 * w * Uc * \sigma_d + 12.7119527 * \gamma_d Uc * \sigma_d + 3.45466718 * PI * Uc * \sigma_d + 73.1347282 * P200 * Uc * \sigma_d + 115.117756 * w * \gamma_d * \gamma + 7.15819491 * w * PI * \gamma - .16922596 * \gamma_d PI * \gamma\gamma + 319.715817 * w * P200 * \gamma + 37.2140312 * \gamma_d P200 * \gamma + .456277589 * PI * P200\gamma + 87.1036047 * w * Uc * \gamma + 12.8403258\gamma_d * Uc\gamma + .470678587 * PI * Uc\gamma + 17.7135930 * P200 * Uc * \gamma + 117.516534 * w * \sigma_d\gamma + 45.1075781 * \gamma_d * \sigma_d\gamma + 4.48108267 * PI\sigma_d\gamma + 123.422308 * P200 * \sigma_d\gamma + 69.9906302 * Uc * \sigma_d\gamma - 11.825725 * w\gamma_d * PI * P200 - 2.1155135 * w * \gamma_d * PIUc + 149.381038 * w * \gamma_d * P200 * Uc + .461733179 * w * PI * P200 * Uc + .843999848\gamma_d * PI * P200 * Uc - 3.1294192 * w\gamma_d * PI\sigma_d + 277.350974 * w * \gamma_d * P200 *$$

$\sigma_d - 14.172917 * w * PI * P200\sigma_d + .525633666 * \gamma_d * PI * P200 * \sigma_d + 98.7215822 * w * \gamma_d Uc\sigma_d - 15.964519 * w * PI * Uc * \sigma_d - 1.1448209 * \gamma_d * PI * Uc\sigma_d - 169.47515 * w * P200 * Uc * \sigma_d - 22.778234 * \gamma_d P200Uc\sigma_d - 2.0574402 PI * P200Uc\sigma_d - 1.1780048 * w * \gamma_d * PI\gamma - 136.19640 w\gamma_d * P200 * \gamma - 5.5904934 * w * PI * P200 * \gamma + .582825521 * \gamma_d * PI * P200 * \gamma - 32.417984 * w\gamma_d * Uc\gamma - 4.4154766 * w * PIUc * \gamma - .03434428 * \gamma_d * PI * Uc * \gamma - 35.024894 * w * P200Uc\gamma - 5.1994910 * \gamma_d * P200 * Uc\gamma + .197336556 * PI * P200 * Uc * \gamma + 12.9337904 w\gamma_d\sigma_d\gamma - 14.637167 wPI\sigma_d\gamma - 1.4228429 * \gamma_d * PI\sigma_d\gamma - 175.26170 * w * P200 * \sigma_d * \gamma - 51.332213\gamma_d * P200\sigma_d * \gamma - 4.4923407 * PI * P200\sigma_d * \gamma - 150.22381 * w * Uc\sigma_d * \gamma - 29.716170\gamma_d * Uc * \sigma_d\gamma - 2.2982062 * PI * Uc * \sigma_d * \gamma - 70.820468 * P200 * Uc * \sigma_d * \gamma + 0.00000000 w\gamma_d * PI * P200 * Uc + 0.00000000 * w\gamma_d * PI * P200 * \sigma_d + 5.73119283 * w * \gamma_d * PI * Uc * \sigma_d + 0.00000000 * w * \gamma_d P200 * Uc\sigma_d + 9.80008993 * w * PI * P200 * Uc * \sigma_d - .05009087 * \gamma_d * PI * P200 * Uc * \sigma_d + 0.00000000 * w * \gamma_d PI * P200\gamma + 1.92867824 * w * \gamma_d PI * Uc\gamma + 0.00000000 * w\gamma_d * P200 * Uc\gamma + 1.30268945 * w * PI * P200Uc\gamma - .38745801 * \gamma_d * PI * P200 * Uc * \gamma + 3.36590974 * w * \gamma_d * PI * \sigma_d * \gamma + 0.00000000 w\gamma_d * P200 * \sigma_d * \gamma + 11.4488598 wPI * P200 * \sigma_d\gamma + 1.17782173\gamma_d * PI * P200\sigma_d\gamma + 31.7684321 w\gamma_d * Uc * \sigma_d\gamma + 5.85679607 * w * PIUc\sigma_d\gamma + .607338016\gamma_d PI * Uc * \sigma_d$*BS$+ 128.506165 w * P200Uc\sigma_d\gamma + 26.8574361\gamma_d * P200 * Uc\sigma_d * \gamma + 1.63623868 * PI * P200 * Uc * sd * \gamma + 0.00000000 w\gamma_d PIP200 * Uc * \sigma_d + 0.00000000 * w\gamma_d * PI * P200Uc\gamma + 0.00000000 * w\gamma_d PI * P200 * \sigma_d\gamma + 0.00000000 w\gamma_d * PIUc\sigma_d * \gamma + 0.00000000 * w\gamma_d P200Uc * \sigma_d\gamma + 0.00000000 wPI * P200 * Uc * sd\gamma + 0.00000000\gamma_d PI * P200 * Uc\sigma_d\gamma - 4.6975464 w\gamma_d * PI * P200Uc\sigma_d\theta.$

The R^2 value for this model was 0.6595 which is the best fitting among all the four models [10].

2.4 R^2 Analysis

The statistical models were developed, as we stated earlier using the development data set which was data collected from many counties in the state of Oklahoma. For fuhrer analysis, another data set was (the evaluation data set) used to see how appropriate the fit of these model was. The evaluation data set was data collected from Woodward County and Rogers County [10] and was not used to develop the models. This provides different views on the prediction quality of the models [22]. A comparison is made between the R^2 values of the development data set and the evaluation data set.

Given data $y_i = y_i(x)$ where y_i is the experimental value of the input x. Let $\hat{y}_i$ be the predicted value of y using the model $\hat{y}_i(x, \alpha_i)$ with α_i parameters. The coefficient of multiple determination, R^2, is calculated as: R^2=1 $- \frac{SSE}{SSY}$ where SSE=$\Sigma(y_i - \hat{y}_i)^2$ and SSY= $\Sigma(y_i - \bar{y})^2$ where $\bar{y}$ is the average value of y_i. Under the R^2 method, one selects the best model to be the model that its R^2 is largest [19].

The following table summarizes the values of R^2 of the statistical models for both the development and the evaluation data sets.

Table 2.1: R^2 values (development and evaluation data sets)

	R^2 (development data set)	R^2 (evaluation data set)
Factorial Model	.6595	.4021
Multiple regression Model	.4357	.5370
Polynomial Mode	.4858	.5523
Stress-Based Model	.3226	.3666

First if we look at the R^2 values computed by using the development data

set. The R^2 value of the factorial model was 0.6595 and that was the highest value among all the models. On the other hand, the R^2 value of the factorial model computed by using the evaluation data set was only 0.4021. This value is far below the R^2 value when the development data set is used (0.6595). Even though the overall R^2 value for the development data set of the factorial model was 0.6595, it dropped significantly to 0.4021 for the evaluation data set. The full factorial model considered here contains 128 terms in the function, it may be considered a complex function among the four statistical models. Therefore, it is possible that the factorial model over-fitted the development data set and caused a poor prediction in the evaluation data set. This is an indication that when developing models, it is imperative to generate the model with a large data set and take the number of estimated terms under consideration. Furthermore, any model should be evaluated by other data sets that were not used in the development of the model. According to R^2 values, it can not be concluded that the factorial model is a good model as one time the R^2 value was much larger than when using the evaluation data set. Further more, when the development data set is used, the factorial model has the largest R^2 value where as it is the 2nd smallest value of R^2 when the evaluation data set is used. This tells us that the R^2 values are not accurate scale in deciding the best model. Similarly, the polynomial model predicted the MR/Pa values with an R^2 value of 0.5523 when using the evaluation data set where as the R^2 was 0.4858 in the case of the development data set. Again there is a large difference in the values of R^2 when different data sets are used. Similar results, one can see also for the Stress-Based model and the Multiple Regression model.

By looking at the table above, we can clearly see that the order of the best model totally changes when we change the data and that makes it hard to know

which is really best model. For example when the development data set is used, the factorial model was best where as it was the third best when the evaluation data set is used. Furthermore, When using the evaluation data set the multiple regression model was the best where as it was the third best when using the development data set. In the actual situation, one only has the data and needs to know the best model that will fit the data and therefore one can not know from this analysis which model to use as the question now is which is the best model to use? Is it the factorial model or the multiple regression model. We are looking for a model that will work for a generic data in order to use it in the application and It is clear that the R^2 analysis doesn't give us the best model here and that is one of the reasons that makes AIC approach better to use.

2.5 AIC Analysis

In order to see the best model according to AIC, we need to evaluate the AIC, Δ, ω values for each model. Given data $y_i = y_i(x)$ where y_i is the experimental value of MR at the experimental value of the input x. Let $\hat{y}_i$ be the predicted value of the MR using the model $\hat{y}_i(x, \alpha_i)$ with α_i parameters, then the standard deviation of the model $\hat{y}_i(x, \alpha_i)$ is calculated as $\sigma = \sqrt{\frac{\Sigma(y_i - \hat{y}_i)^2}{n}}$ where n is the the data size. It is important to mention here that n is the number of observations so each reading is considered a value. Also, $AIC = -2log(l(\hat{\theta}|y)) + 2k$, where k is the number of the estimated parameters in the model including σ and the intercept and $log(l(\hat{\theta}|y)) = \frac{-n}{2} log(\sigma^2)$. Given our set of 4 models then AIC difference for a model j is calculated as: $\Delta_j = AIC_j - AIC_{min}$, where AIC_{min} is the smallest value of the AIC values for all the 4 models and Akaike's weight for a model j is defined to be $w_j = \frac{\exp(-1/2\Delta_j)}{\Sigma_{r=1}^{R} \exp(-1/2\Delta_r)}$.

For a 95% confidence set on the actual K-L best model, we sum the Akaike weights from largest to smallest until the sum is just ≥ 0.95, then the corresponding subset of models is a type of confidence set on the K-L best models.

The following table summarizes the AIC, Δ, ω values for each model when the development data set is used.

Table 2.2: AIC values (development data set)

	k	σ	$Log(l)$	AIC	Δ	ω
Multiple Regression	9	0.042082	2927.357	-5836.71	94.39	0
Polynomial	16	.049461	2778.07	-5524.14	406.96	0
Factorial	30	.039088	2995.552	-5931.1	0	1
Stress-Based	4	.059448	2608.132	-5202.26	722.84	0

If we look at the results in the table above, we notice that the best model according to AIC analysis is the factorial model which also was the best according to R^2. If we look to all the three models (Multiple regression, polynomial, stress based), all have a 0 value of Akaike weight and this says that these models have very poor fit compared to the factorial model. We can get the same results if we look at the Δ values as all the three models have very big value of Δ ($\Delta >> 10$) which as we stated before tells us that these models have very poor fitting compared to the factorial model and therefore these models should not be considered as good models. The polynomial model has Akaike weight of 0 which stated that the fitting of the polynomial is very poor compared to the fitting of the factorial model. Further analysis also shows that AIC suggests that the second best model after the factorial model is the multiple regression model then the polynomial model then the last is the stress-based model (really has very poor fit) where the second best model according to R^2 is the polynomial model.

For further analysis, we find the 95% confidence set of models and we see clearly that the only best model here is the factorial model as its Akaike's weight, ω, is 1 ($\omega > 0.95$) which makes the other models not good models and therefore should not be considered for prediction of MR. Also if we calculate the evidence ratios for all models, we can see that these ratios are 0 which says that all these models are not good fitted models compared to the factorial model.

In conclusion, according to AIC analysis, the only model in this set of candidate models that has appropriate fitting is the factorial model where as we can not conclude these things from the R^2 analysis.

Now, we do the AIC analysis using the evaluation data set. The following table summarizes the results according to AIC analysis using the evaluation data set

Table 2.3: AIC values (evaluation data set)

	k	AIC	Δ	ω
Multiple Regression:	9	-1951.05	422.3557	0
Polynomial:	16	-2335.49	37.918	0
Factorial:	30	-2373.41	0	1
Stress-Based:	4	-2104.93	268.4747	0

By looking at the results in the previous table, we clearly can see that the only good model is the factorial model. Further more the 95% confidence set of models consists only of the factorial model and hence we get the same results we we did when the development data set was used.

2.6 Stability

In the practical situation, one usually has the data and needs to fit models to this data and usually the best model is unknown. How can one make sure that the real best model is chosen for the population as whole and not only good for the sample (data set)? How can it be claimed that this is the best model for the real situation. To over come this problem, we examine the models on a different data set and see if the best model still best model in the other data set. This brings us to the idea of stability of model ranking.

By stability of the ranking here, we mean, if the set of candidate models are fitted to different data sets, the ranking of the models will stay the same. In other words for different data sets (of course for the same variables) the first best model will still be the first best model in the different data sets and the worst model will still be ranked as the worst model. In our study here, we are using two data sets, the development data set and the evaluation data set and hence the ranking of a method will be stable if that method gives the same rank for the models in the two data sets. Getting a stable ranking is very essential and necessary as in the practical situation as one usually has the data and if the ranking is not stable, then it will be misleading and will give wrong or at least not accurate expectations of the output. On the other hand, once one knows that the ranking of a certain method is stable, then when this ranking is used in any practical situation for prediction, it will give the right expectation and it will choose the real best model.

To choose the best model, we look at the ranking of models in the AIC method and the R^2 method using both data sets and compare the results. The following

table summarizes the ranking of models using the development data sets.

Table 2.4: AIC and R^2 Ranking (development data set)

	According to AIC	According to R^2
Factorial Model	1	1
Multiple regression Model	poor	3
Polynomial Model	poor	2
Stress-Based Model	poor	4

By looking at the previous table we first notice that by both methods the best model is the factorial model. According to R^2 method, the second best model is the polynomial model and the third is the multiple regression model. On the other hand, according to AIC, the only best model is the factorial method. All other models are considered poor models and they should not be considered for prediction. This clear cut that the AIC method provides is really essential to judge what are the good models. By the R^2 method, we don't have a cut point at by which we can distinguish between the good models and the bad models.

Now we look at the AIC and R^2 rankings using the evaluation data which are summarized in the following table.

Table 2.5: AIC and R^2 ranking (evaluation data set)

	According to AIC	According to R^2
Factorial Model	1	3
Multiple regression Model	poor	2
Polynomial Model	poor	1
Stress-Based Model	poor	4

By looking at the results in the table above, we can see that according to R^2 method, the best model is the polynomial model, the second best model is the

multiple regression model and the third best is the factorial model. On the other hand according to AIC, the only good model is the factorial model and all other models should not be considered for prediction. The question now, is how can we decide, which model is good to use? Here we need to look at the ranking stability. The following table summarizes the ranking of the models according to AIC analysis and R^2 analysis for the evaluation Data set and the development data set .

Table 2.6: AIC and R^2 Ranking (Development and Evaluation Data sets)

	AIC Ranking (Dev., Eva.)	R^2 Ranking (Dev., Eva.)
Factorial Model	(1,1)	(1,3)
Multiple regression Model	(poor, poor)	(3,2)
Polynomial Model	(poor, poor)	(2,1)
Stress-Based Model	(poor, poor)	(4,4)

The table above shows that according to AIC analysis, the factorial model is the best model and all other models are not good model to consider and this is in both data sets but according to the R^2 analysis we see that when using the evaluation data set (Eva.), the best model is the polynomial model then the second best model is the multiple regression model and the factorial model is ranked the third one. On the other hand when using the development data set (Dev.), the best model was the factorial model then the second best model was the polynomial model and the third was best model was the multiple regression model. This shows that the R^2 results in ranking the models is not stable as it gives totally different ranks to the models when different data sets are used. The rank of the models according to AIC analysis stays the same on both data sets. Further more, we still get the three models, namely the polynomial model, the multiple regression model and the stress-based model are all considered poor

models when using different subsets data sets. This shows us that the AIC ranking is more stable and more appropriate.

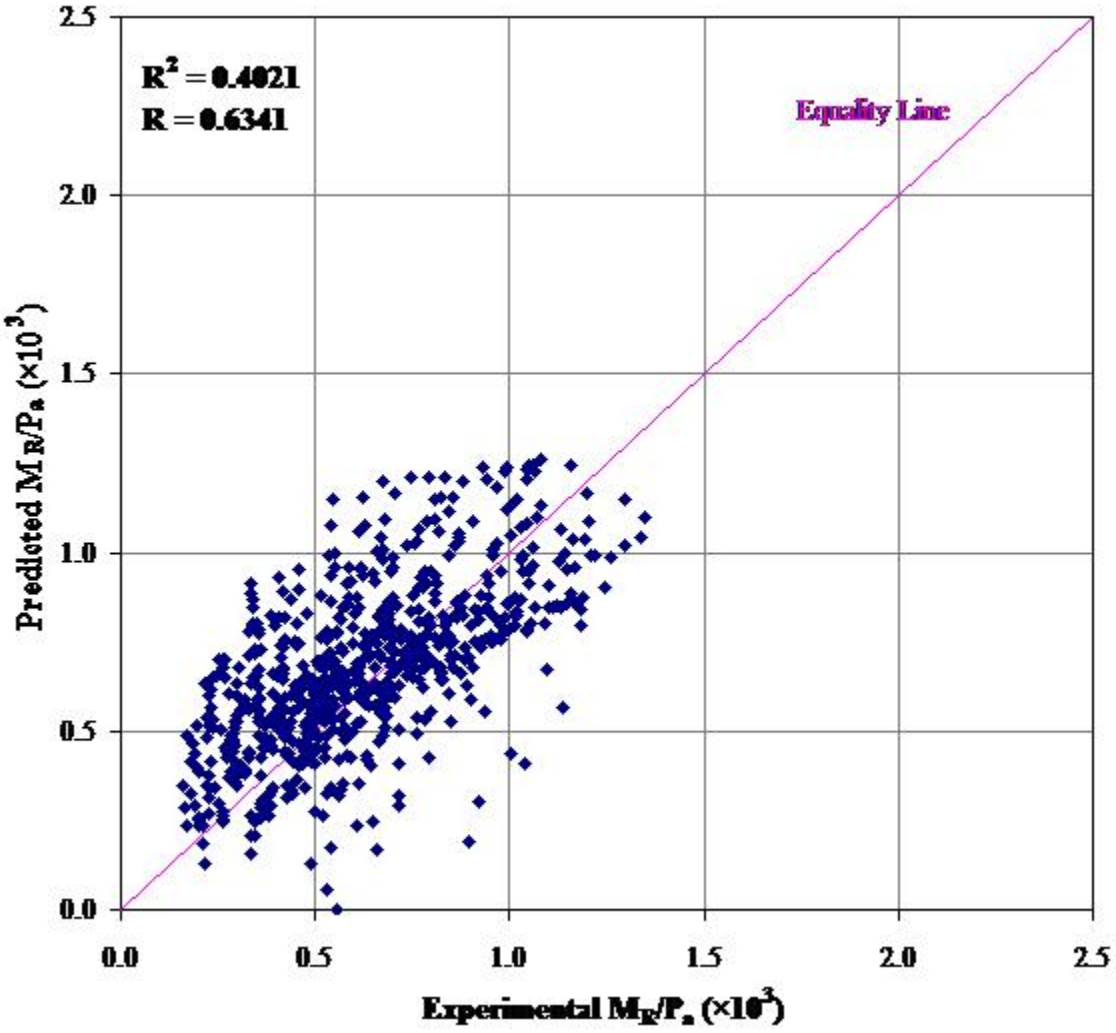


Figure 2.1: Experimental and Predicted values for Evaluation Data set, Factorial Model

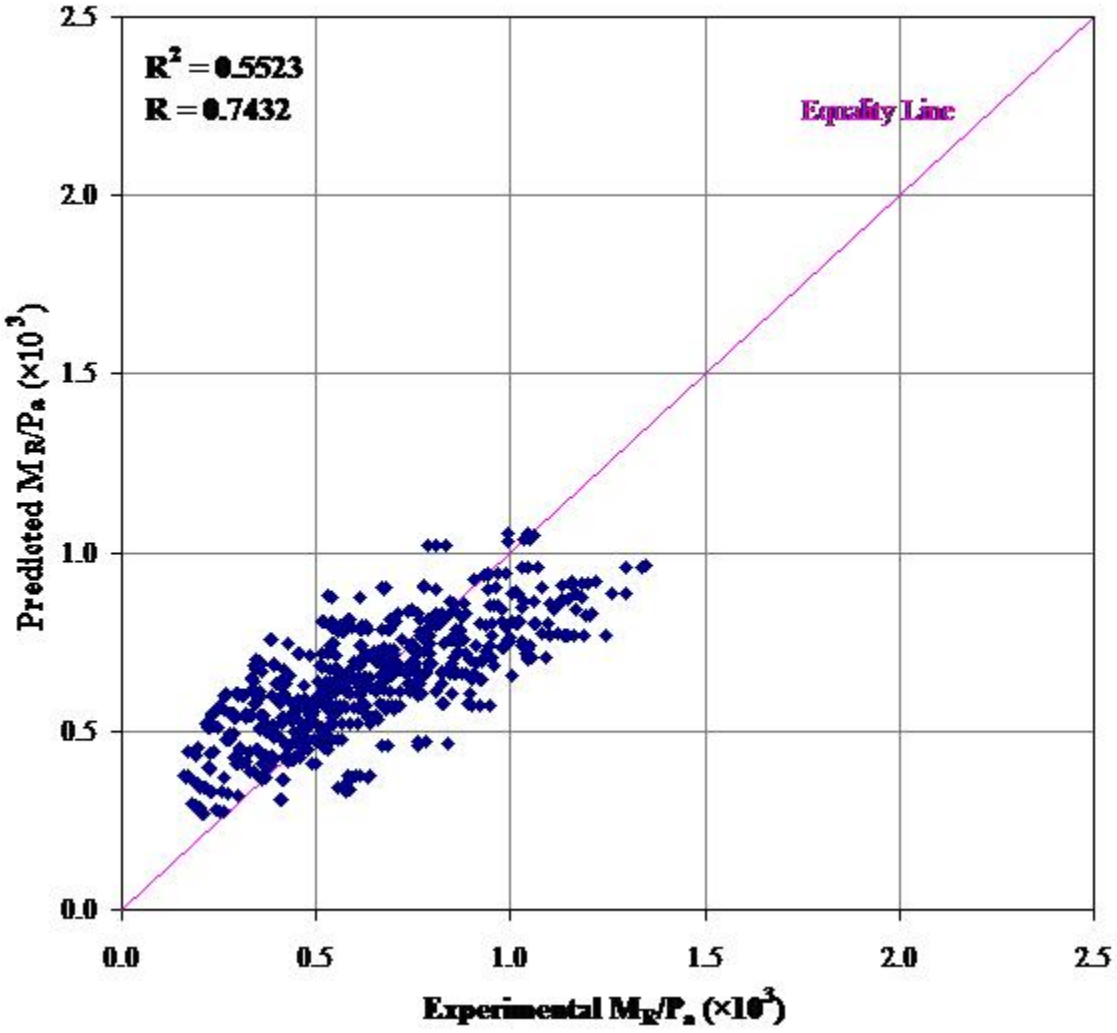


Figure 2.2: Experimental and Predicted values for Evaluation Data set, polynomial Model

Chapter 3

Application to Neural Network Models

An Artificial Neural Network (ANN) is an information processing paradigm that is inspired by the way biological nervous systems, such as the brain. The key element of this paradigm is the novel structure of the information processing system. It consists of a large number of highly interconnected processing elements called neurons working together to solve some problem [25]. In this chapter, we give a brief overview of the neural netowrk models then we introduce the models on which the AIC approach will be applied to.

3.1 Introduction

Neural network simulations appear to be a recent development, however this field was established before the advent of computers and has survived at least one major setback and several eras. An artificial neural network (ANN) is a tool that imitates the function of a biological neural network. The first artificial

neuron was produced in 1943 by the neurophysiologist Warren McCulloch and the logician Walter Pits but the technology available at that time did not allow them to do too much. Neural networks process information in a similar way the human brain does. The network is composed of a large number of highly interconnected processing elements (neurones) working in parallel to solve a specific problem. Neural networks learn by example. They cannot be programmed to perform a specific task. The examples must be selected carefully otherwise useful time is wasted or even worse the network might be functioning incorrectly. Neural networks and conventional algorithmic computers are not in competition but complement each other. [25].

Neural networks, with their remarkable ability to derive meaning from complicated data, can be used to extract patterns and detect trends that are too complex to be noticed by either humans or other computer techniques. A trained neural network can be thought of as an "expert" in the category of information it has been given to analyze. This expert can then be used to provide projections given new situations of interest and answer "what if" questions. Further more, an ability to learn how to do tasks based on the data given for training or initial experience. Also, an ANN can create its own representation of the information it receives during learning time. Neural network has a very wide use. They are used in signal processing, in control, in medicine and in pattern recognition [36].

We can think of Artificial Neural Network as another way of modeling the data. It has some numbers of layers, including an input layer, hidden layers and lastly output layer. There are many neural network models that one can generate. It provides some of the human characteristics of problem solving that are

difficult to simulate using any of the logical, analytical, or standard computing techniques [37]. The ANN has become an increasingly important tool due to its successes in many practical applications. One of the objectives of the ANN models is to find a function that can relate the input variable to the output variable. For example, in linear regression model, a function is obtained by changing the slope and intercept so that the function fits the data set. The same principle is applied to the ANN models. The ANN model is obtained by adjusting the weights between the processing elements. The ANN adjusts their weights by repeatedly presenting the input data to minimize the error between the historical and predicted output [12]. This phase is called "training" or "learning." The difference between an ANN model and a regression model is that a prior knowledge of the nature of the non-linearity is not required in ANN models. The degree of non-linearity of the ANN models can be changed easily by varying the number of hidden layers, number of nodes in each layer, and the transfer functions [**?**].

The architecture of ANN contains a number of simple, highly interconnected processing elements, known as "nodes". The main objective of the neural network approach is to find the weights through training a set of input data until the network reaches a minimum error. In the training process, a number of checks are performed in the network. After each epoch, the weights are adjusted and a sum of mean squared error between target and output values is calculated. The training process stops when the sum of mean squared error is minimized or falls within an acceptable range Different training algorithms can be used to train a network. The function of a training algorithm is to adjust the weights and thresholds using the training data set [28].

3.2 Network Architectures

There are different kinds of networks, some examples are:

- Single-input Neuron:

A single input neuron consists of a scalar input p which is multiplied by the scalar weight w to form wp then added to a bias b and then passed to the summer. The summer output n which is referred to as the net input goes into a transfer function f which produces the scalar neuron output a. The neuron output is calculated as $a = f(wp + b)$. As one can see, the output depends on the transfer function that one chooses [12].

- Multiple-input Neuron:

A neuron with R inputs, $p_1, p_2, ..., p_R$ are each weighted by corresponding elements $w_{1,1}, w_{1,2}, ..., w_{1,R}$ of the weight matrix W. The neuron output can be written as $a = f(WP + b)$ where p is $R \times 1$ vector and W is $1 \times R$.

- A layer of neurons:

A single layer network of S neurons is of the form $a = f(WP + b)$ where W is $S \times Rmatrix$ and a, b are vectors of length S. The layer consists of the weight matrix W, the summation and the multiplication operations, the bias vector b, the transfer function and the output vector.

- Multiple layers of Neurons:

Consider a network with several layers, each layer has its own matrix W, its own bias vector b, a net input vector n, and an output vector a, then we can write

this network as $a^3 = f^3(w^3 f^2(w^2 f^1(w^1 p + b^1) + b^2) + b^3)$.
A layer whose output is the network output is called an output layer. The other layers are called hidden layers. Multi layer networks are more powerful than a single layer networks [12].

The transfer function is very important in deciding the output of the network. There are many transfer function that are largely used such as:

- The linear function: $a = n$
- The Hard limit function: $a = \begin{cases} 0 & n < 0 \\ 1 & n \geq 0 \end{cases}$.
- The Symmetrical Hard Limit: $a = \begin{cases} -1 & n < 0 \\ 1 & n \geq 0 \end{cases}$.
- The Saturating Linear: $a = \begin{cases} 0 & n < 0 \\ n & 0 \leq n \leq 1 \\ 1 & n > 1 \end{cases}$.
- The Log-Sigmoid: $a = \frac{1}{1+e^{-}n}$

3.3 Learning Rules

By learning rule, it means the way to produce the weights and the bias of the network. Learning rules are sometimes referred to as training algorithms. The purpose of the learning rule is to train the network to perform some task. There are many types of neural network learning rules. They can be divided into three categories which are: supervised learning, unsupervised learning and reinforcement learning [12].

3.3.1 Supervised learning

In supervised learning the learning rule is provided with a set of examples (called training set) of a proper network behavior: $p_1, t_1, p_2, t_2, ..., p_r, t_r$, where p_i is an input to the network and t_i is the corresponding correct output (called target). This rule works as: while the inputs are applied to the network, the network output are compared to the targets. Then the learning rule is used to adjust the weights and the biases of the network so that it moves the network outputs closer to the target. This type of learning is largely used these days [12]

3.3.2 Reinforcement learning

This type of learning is similar to supervised learning, except that, instead of being provided with the correct output for each network input, the algorithm is only given a grade. The grade is a measure of the network performance over some sequence of inputs. This type of learning is currently much less common than supervised learning. Though, it appears in control system applications [12].

3.3.3 Unsupervised learning

In unsupervised learning, the weights and the biases are modified in response to network inputs only. There are no target outputs available. One can ask, how can we train a network if we don't know what is supposed to do? but, in fact most of these type of leanings perform some kinds of clustering operation. They learn to categorize the input patterns into a finite number of classes. This is useful in applications that are related to vector quantization.

3.4 The Models

The neural network models models that were generated to model resilient modulus are: Linear Network, General regression Neural network, Radial basis function network and multi-layer perceptron Network. In all these models the input layer consists of seven nodes, one node for each independent variable, namely moisture content (w), dry density (θ_d), plasticity index (PI), percent passing sieve No. 200 ($P200$), unconfined compressive strength (Uc), deviatoric stress (σ_d), and bulk stress (θ). The output layer consists of one node for the dependent variable, which is the MR. In applying the information theory approach on the neural network models, we treat the nodes as the estimated parameters since these nodes are estimated in the model.

The data that is used to generate the models is the same data that was used to generate the statistical models. As was stated earlier, this data is a total of 98 bulk soil samples and were collected from 16 different counties in the State of Oklahoma. The data was divided into two sets of data, the first is the development data set which is the data used in developing the models and the

second is the evaluation data set which will be used in evaluating the fitting of the models.

3.4.1 Linear neural network model-LNN

The Linear Network Model has only two layers, an input layer and an output layer. A linear model is typically represented using an $N \times N$ matrix and an $N \times 1$ bias vector. The weights correspond to the matrix, and the thresholds to the bias vector. When the network is executed, it effectively multiplies the input by the weight matrix then adds the bias vector [10].

A LNN model does not have any hidden layer and the number of the estimated parameters is 9. The R^2 value for the linear network model (using the development data set) was found to be 0.4.

3.4.2 General Regression Neural Network Model-GRNN

This model has four layers including the input layer, two hidden layers, and one output layer. The first hidden layer consists of the radial units. The number of nodes in the first hidden layer can be as many as the number of cases. The second hidden layer consists of units that help estimate the weighted average. The second hidden layer always has exactly one more node than the output layer. Since only one output is considered in the present study is (MR), the second hidden layer has two nodes.

The optimum number of nodes in the first hidden layer was determined using a trial and error approach. The second hidden layer had two nodes. From the trial and error approach, the best-fit GRNN model was found to have 1250 nodes in the first hidden layer, so the total estimated parameters is 1261. The R^2 value

for the GRNN model was 0.6015, which is significantly better than the Linear network model (0.4323).

3.4.3 Radial Basis Function Network model-RBFN

The radial basis function network uses an approach to divide the modeling space using hyper spheres. This model has three layers, namely input, hidden, and output layers. The hidden layer consists of radial units.

The RBFN model has one hidden layer. As in the case of the LN and GRNN models, a trial and error approach was used to determine the optimum number of nodes in the hidden layer. Following this approach, the optimum number of nodes in the hidden layer was found to be 100. The R^2 value of the RBFN model is 0.6284, which is slightly better than the GRNN model (0.6015) and much better than the LN model (0.4323).

3.4.4 Multi-Layer Perceptrons Network Model-MLPN

The multi-layer perceptrons network is one of the most popular network architectures in use these days [10]. The MLPN consists of an input layer, a number of hidden layers, and an output layer. In each of the hidden layers, the number of node can be varied. Due to the number of layers and the number of nodes in each layer, this model can adjust the architecture of the network based on the complexity of a problem. Three different models were used, which are with one hidden layer, with two hidden layers and with three hidden layers. Each of the nodes in the network performs a biased weighted sum of their inputs and passes this activation level through a transfer function to produce its output. The weights and biases in the network are adjusted using a training algorithm.

The number of hidden layers in the MLPN models can range from one to three. Here we have three MLPN models henceforth referred to as Multi-Layer Perceptrons Network-1, MLPN-1, Multi-Layer Perceptrons Network-2, MLPN-2, Multi-Layer Perceptrons Network-3 models, MLPN-3. The number of nodes in each of the three hidden layers was set at six nodes, based on the trial and error approach adopted. The estimated parameters in these models are: 15, 21, 27 respectively. The R^2 values of the Multi-Layer Perceptrons Network models were 0.5733, 0.5744, and 0.5587 for one, two and three hidden layers, respectively. These R^2 values indicate that all three Multi-Layer Perceptrons Network models are expected to better correlate the MR/Pa values than the linear network model (0.4323). However, the Multi-Layer Perceptrons Network models were worse than the General Regression Neural Network Model (0.6015) and the Radial Basis Function Network model (0.6284) [10].

3.5 R^2 Analysis

As was stated above, the neural network models were developed using data that was collected from many counties in the state of Oklahoma. For deeper analysis, another data set is used to evaluate the fit of the models. This data were not used in developing the models, we only use it to evaluate the models (the evaluation data set). A comparison is made between the R^2 values of the development data set and the evaluation data set. We use the same procedure as in chapter two, the table below summarizes the values of R^2 of the models using the development data set and the evaluation data set.

As we can see from the results in the above table, when the development data

Table 3.1: R^2 values (development and evaluation data sets)

	R^2 (development data set)	R^2(evaluation data set)
RBFN Model	.6284	.5557
GRNN Model	.6015	.4791
MLPN-2 Model	.5744	.6026
MLPN-1 Model	.5733	.6146
MLPN-3 Model	.5587	.5899
LNN Model	.4323	.5443

set is used, the best model is the Radial Basis Function Network model with R^2 value of 0.6284 which is a reasonably large value and tells us that the fit of this model to the development data is good. The second best model is the General Regression Neural Network Model with an R^2 value of 0.6015 which is considered a good fit to the data. The worst model is the Linear Network model with an R^2 value of 0.4323 which is a poor fit. We also notice that the models MLPN-1, MLPN-2 have very close values of R^2 (0.5733, 0.5744 respectively) which says that these two models have almost the same degree of goodness of fit to the data even though the number of nodes is different in the two models. When the evaluation data set is used the Multi-Layer Perceptrons Network-1 Model (MLPN-1) is the best model with an R^2 value of 0.6146. The second best model is MLPN-2 with an R^2 value of 0.6026 which still considered a good fit. The worst model is GRNN with an R^2 of 0.4791 which says this model has a poor fit.

3.6 AIC Analysis

Here we are using the AIC_c analysis and in order to apply this technique, we need to evaluate AIC_c, Δ, ω for each model. We are treating the total number of nodes in the models as the estimated parameters, so k will be the total number

of estimated nodes +1 (we add one for standard deviation (σ)). Also we use the AIC_c rather than AIC since in some models the ratio $(n/k) < 40$. As we stated in Appendix B, AIC_c=$AIC + ((2k(k+1))/(n-k-1))$ where n is the sample size and k is the number of the estimated parameters in the model [6]. It is also important to note here that n represents the number of the observations in the data.

The following table shows the results of the AIC_c calculations for the development data set:

Table 3.2: AIC values (development data set)

	k	AIC_c	Δ	ω
RBFN Model	109	-7608	56	0
GRNN Model	1261	-1785	5879	0
MLPN-2 Model	21	-7663.92	1.21	.353
MLPN-1 Model	15	-7665.13	0	.647
MLPN-3 Model	27	-7504	104	0
LNN Model	9	-7151	513	0

The analysis of AIC_c in the table above (Development data set) suggests that the Multi-Layer Perceptrons Network-1 Model is the best among all the set of candidate models. The second best model in the set of candidate models is Multi-Layer Perceptrons Network-2 Model. All other models are poor models and they should not be used (the value of $\Delta >> 10$) . If we find the confidence set of best models, we clearly can see this set consists only of the multilayer perceptron network-1 and Multi-Layer Perceptrons Network-2 Model as if their Akaike's weights are added, it will be 1 and that says that these are the only two models that should be considered. This means we should exclude the other models from the models of reasonable fit. Furthermore if we compute the evidence ratio for

the second best model namely Multi-Layer Perceptrons Network-2 Model, we find it to be .55 which says that the Multi-Layer Perceptrons Network-2 has considerable support to have a good fit.

We also calculated the AIC Values when the evaluation data set is used and we got the following results:

Table 3.3: AIC_c values (evaluation data set)

	k	AIC_C	Δ	ω
RBFN Model	109	-2216	271	0
GRNN Model	1261	14024	5879	0
MLPN-2 Model	21	-2401	86	.03
MLPN-1 Model	15	-2487	0	.97
MLPN-3 Model	27	-2344	143	0
LNN Model	9	-2357	130	0

By looking at the table above, we clearly can see that the best model is Multi-Layer Perceptrons Network-1 Model. The second best model is Multi-Layer Perceptrons Network-2. All other models are considered to have very poor fit as the Akaike weight of each one of them is 0. If we look at the set of confidence models, we see it consists only of the two models MLPN-1, MLPN-2.
We get valuable information here that we couldn't get from the R^2 analysis and it is that the second best fitting is the Multi-Layer Perceptrons Network-2 not the General Regression Neural Network. Second is that the General Regression Neural Network,Linear network, Multi-Layer Perceptrons Network-3 all have very poor approximations as their AIC_c is ($\Delta > 10$) which we couldn't get for the R^2 analysis . As we stated above, According to R^2 analysis of the development data set, the General Regression Neural Network is the second best model where as with the AIC analysis we see it is a poor a proximation as it has Akaike's weight

of 0. Also according to AIC_c analysis Multi-Layer Perceptrons Network-2 is the second best model where as it is the third in the R^2 analysis

3.7 Stability

To look at the stability of the ranking of each method, we look the rank of the models when the development data set is used and when the evaluation data set is used and see if the model has the same rank in both data sets. This is an important concepts in the practical situation as we need to be able to apply the model for situation in which we don't have data and therefore it is essential to have a stable ranking. To check the ranking stability in our study for the Neural Network Models, we compare the ranks in the AIC analysis with the ranks in the R^2 analysis. The table below summarizes the models from best to worse with both AIC and R^2 analysis for the development data set.

Table 3.4: AIC and R^2 ranking (development data set)

	According to AIC_c	According to R^2
RBFN Model	poor	1
GRNN Model	poor	2
MLPN-2 Model	2	3
MLPN-1 Model	1	4
MLPN-3 Model	poor	5
LNN Model	poor	6

If we look at the ranking of all the models according to R^2 and AIC for the development data set, we see that the best two models according to R^2 Method are the RBFN and GRNN where as according to AIC the best two models are MLPN-1 and MLPN-2. First we notice that the best models are not the same in the two methods. Furthermore, R^2 method doesn't give precisely which models have good models and which don't. In other word, using the R^2 method, we don't have a clear cut point by which we can decide if the model should be considered or not. For example, we know that the best two models are RBFN and GRNN but we don't know if MLPN-2 model (which is the third best) has a good fit or not. On the other hand when using the AIC analysis, we not only see that the best two models are MLPN-1, MLPN-2 but we know that all other models have poor fit and they should not be considered.

Now we look at the ranking of the models in the two methods when using the evaluation data set. The following table summarizes these ranks.

by looking at the table, we first notice that the best two models according to R^2 method are MLPN-1, MLPN-2 (which is different from when the development data set is used). Again, we don't have a clear cut point by which we can know which models should be considered or not. On the other hand the best two models according to AIC method are MLPN-1, MLPN-2 (which is the same

Table 3.5: AIC and R^2 ranking (evaluation data set)

	According to AIC_c	According to R^2
RBFN Model	poor	4
GRNN Model	poor	6
MLPN-2 Model	2	2
MLPN-1 Model	1	1
MLPN-3 Model	poor	3
LNN Model	poor	5

as when the development data set). Furthermore, according to AIC method, all other models are considered to have very poor fit and they can not be considered for any prediction.

We summarize the ranking of the models with respect to the two method and using the two data sets in the following table.

Table 3.6: AIC and R^2 Ranking (development and evaluation data sets)

	AIC Ranking (Dev.,Eva.)	R^2 Ranking (Dev.,Eva.)
RBFN Model	(poor, poor)	(1,4)
GRNN Model	(poor, poor)	(2,6)
MLPN-2 Model	(2,2)	(3,2)
MLPN-1 Model	(1,1)	(4,1)
MLPN-3 Model	(poor, poor)	(5,3)
LNN Model	(poor, poor)	(6,5)

First we notice that the R^2 ranking is not stable where as the AIC ranking is. In other words, according to R^2 method, we see that the best models are not the same if we look at the two different data sets. For example, in the development data set, the best model is RBFN where as this model is the fourth best in the evaluation data set. On the other hand when the evaluation data set is used the best model is MLPN-1 and this model is the fourth best in the development

data set. This non stability in the ranking makes this method very weak and sometimes misleading. On the other hand if we look at the ranking according to AIC we see it is stable and it gives the same best models in both data sets, namely the best model is MLPN-1 and the second best is MLPN-2. Furthermore all other models, according to AIC method have poor fit and they should not be used for prediction of MR. The non stability of the ranking in the R^2 analysis makes it hard to decide which one is the best model. On the other hand,the AIC_c ranking, as we saw, is stable and that helps us to choose the best model that works better for all data.

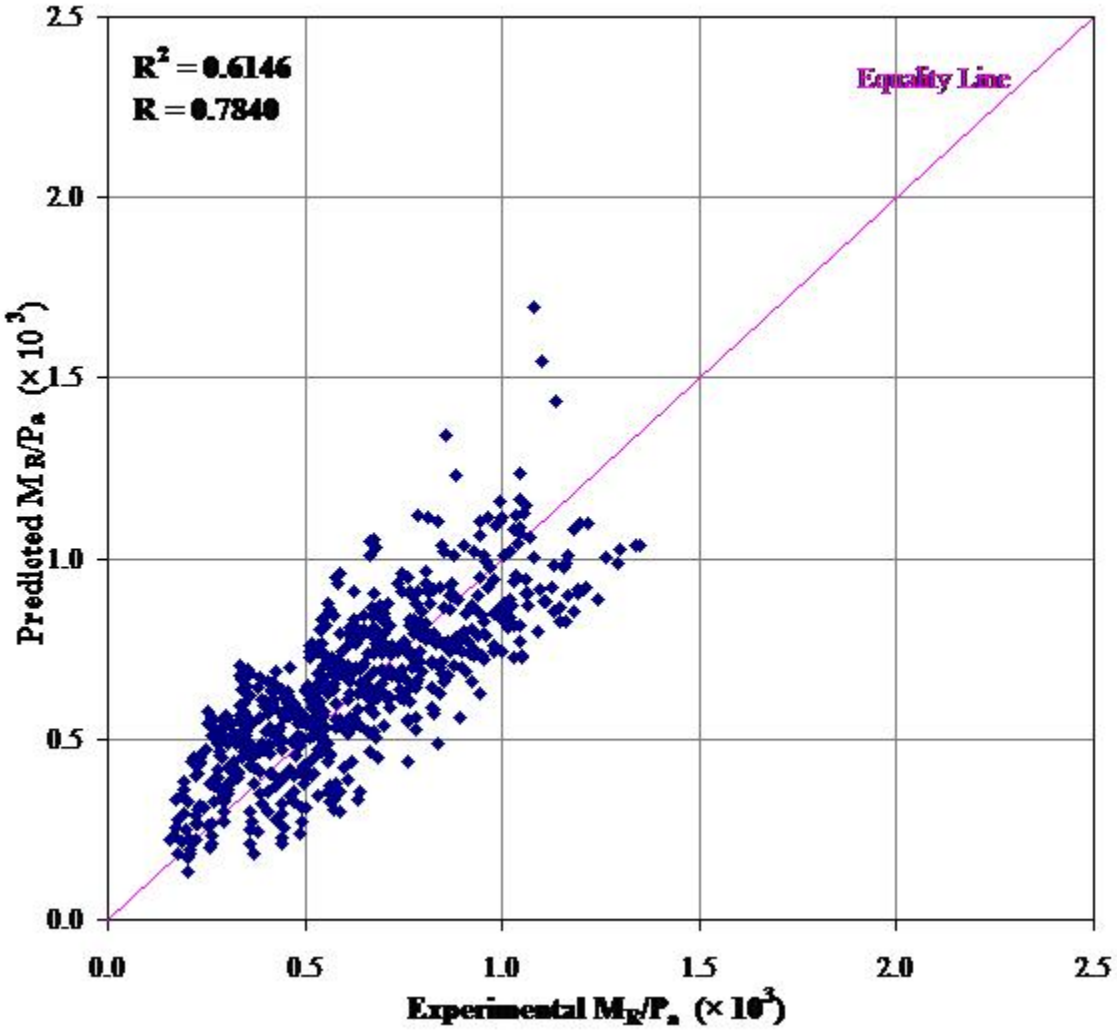


Figure 3.1: Experimental and Predicted values for Evaluation Data set, MLPN-1

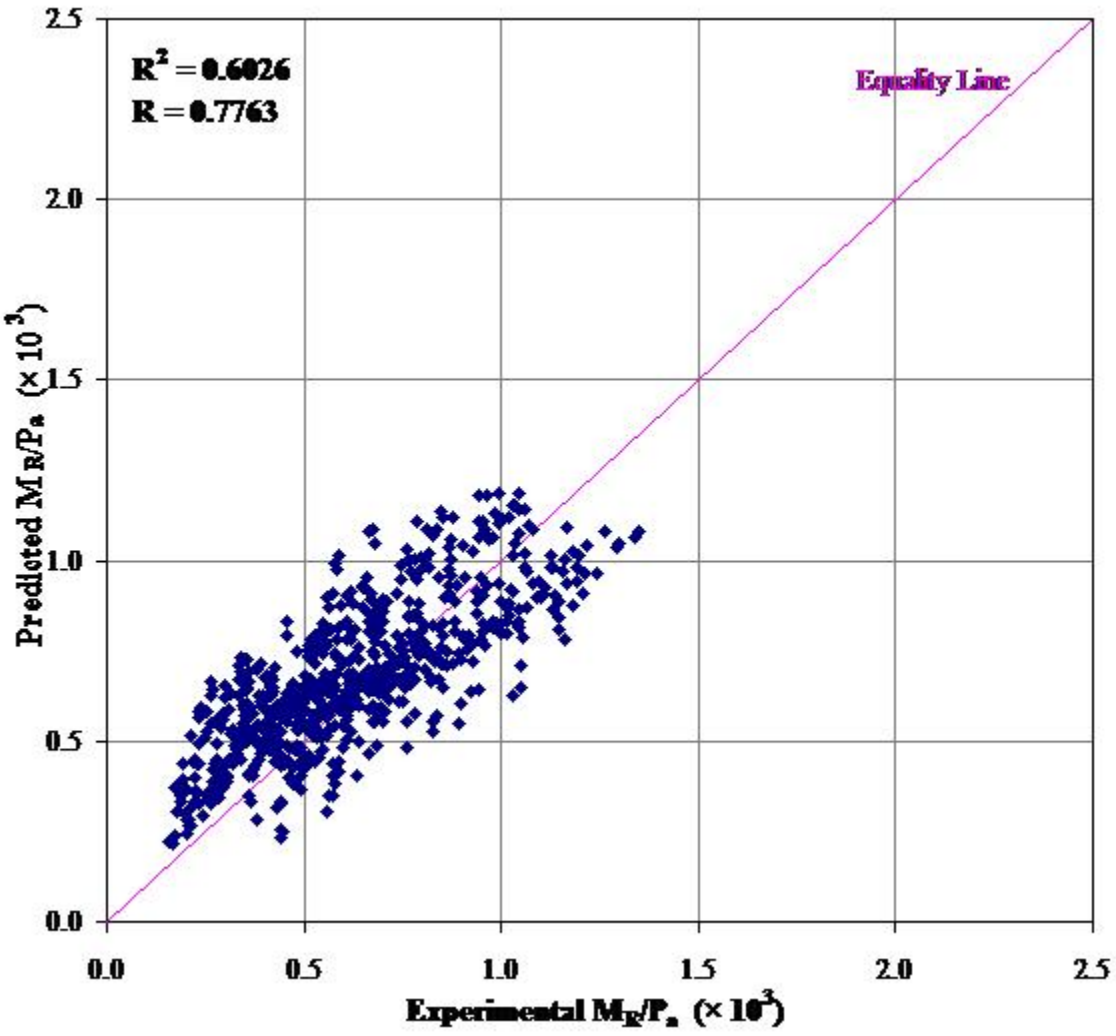


Figure 3.2: Experimental and Predicted values for Evaluation Data set, MLPN-2

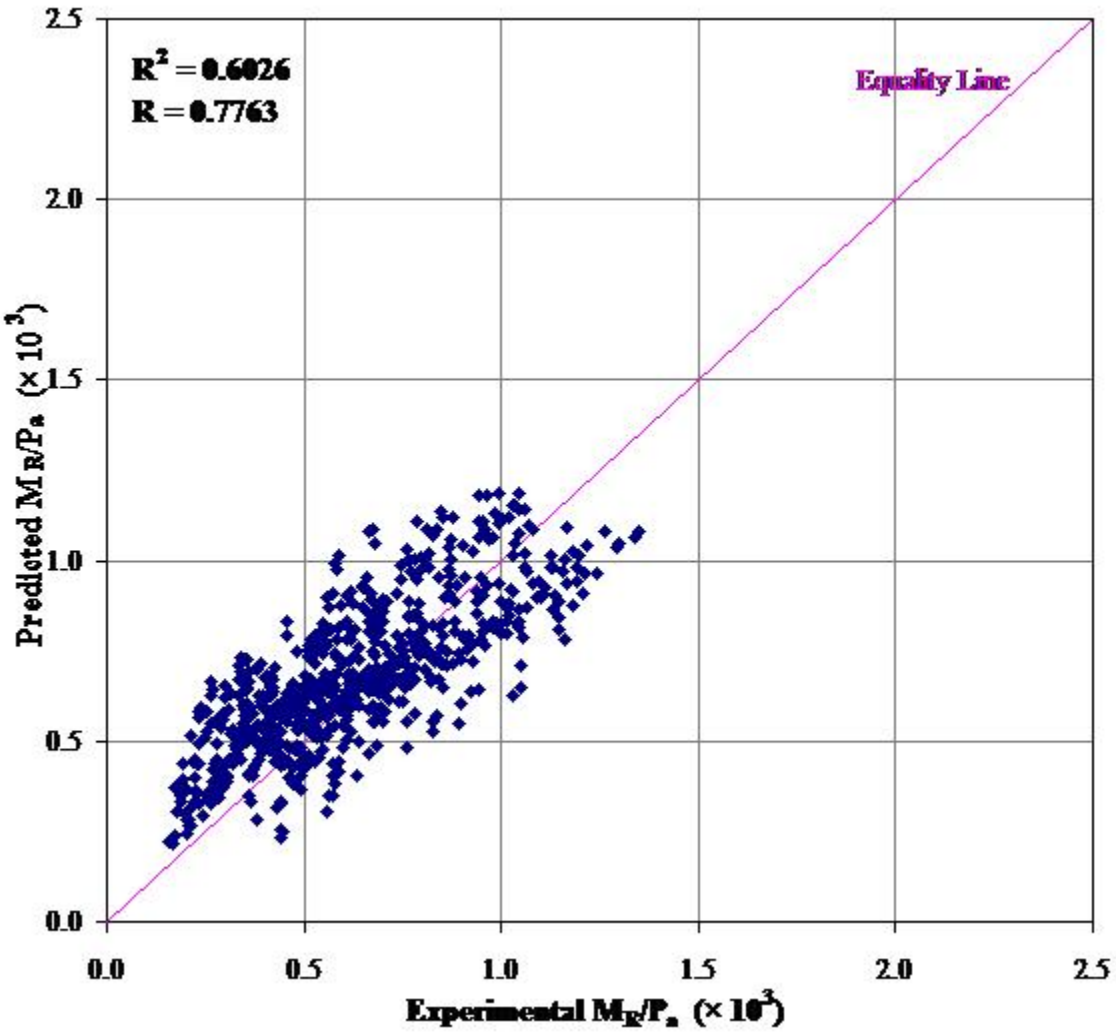


Figure 3.3: Experimental and Predicted values for Evaluation Data set, MLPN-3

Chapter 4

Application to Physics-Based Models

In this chapter we present the application of the information theory on physics models. This chapter relay on the work of Luther White and Jinsong Pei in their papers [35]. First we introduce the problem and then investigate the selection and ranking of models of a prestressed concrete girder with respect to data consisting of displacements that have been obtained from laboratory experiments. We use the information theoretic approach to select the best model. In fact we are looking for the optimal number of parameters that gives us the best model.

4.1 Introduction

For modeling a girder, consider a set Ω in $\mathbf{R^3}$ given by

$$\Omega = \{(x, y, z) : 0 < x < L, -k(z) < y < k(z), -h < z < h\}.$$

The function $k(z)$ will be given later. Displacements in the $x,\ y,$ and z directions are designated by $u,\ v,$ and w, respectively with

$$\mathbf{u} = \begin{bmatrix} u \\ v \\ w \end{bmatrix}$$

It is assumed that the material is isotropic and the small displacement gradient assumption applies [17]. The strains are expressed as $\epsilon_{11} = \frac{\partial u}{\partial x}$

$$\epsilon_{12} = \frac{1}{2}(\frac{\partial u}{\partial y} + \frac{\partial v}{\partial x})$$

$$\epsilon_{13} = \frac{1}{2}(\frac{\partial u}{\partial z} + \frac{\partial w}{\partial x})$$

$$\epsilon_{22} = \frac{\partial v}{\partial y}$$

$$\epsilon_{23} = \frac{1}{2}(\frac{\partial v}{\partial z} + \frac{\partial w}{\partial y})$$

$$\epsilon_{33} = \frac{\partial w}{\partial z}.$$

The stresses are expressed as

$\sigma_{11} = \frac{E}{(1+\mu)(1-2\mu)}[(1-\mu)\epsilon_{11} + \mu\epsilon_{22} + \mu\epsilon_{33}]$

$$\sigma_{12} = \frac{2E}{1+\mu}\epsilon_{12}$$

$$\sigma_{13} = \frac{2E}{1+\mu}\epsilon_{13}$$

$$\sigma_{22} = \frac{E}{(1+\mu)(1-2\mu)}[\mu\epsilon_{11} + (1-\mu)\epsilon_{22} + \mu\epsilon_{33}]$$

$$\sigma_{23} = \frac{2E}{1+\mu}\epsilon_{23}$$

$$\sigma_{33} = \frac{E}{(1+\mu)(1-2\mu)}[\mu\epsilon_{11} + \mu\epsilon_{22} + (1-\mu)\epsilon_{33}]$$

where E is Young's modulus and μ is Poisson's ratio [35]

4.2 The girders Models

We assume that the Poisson's ratio μ and the Young's modulus E are both functions of x. It is convenient to define the matrices

$$\widehat{b}_0 \;=\; \int_{-h}^{h} k(z) \begin{bmatrix} 1 & z \\ z & z^2 \end{bmatrix} dy$$

$$\widehat{b}_1 = \int_{-h}^{h} \begin{bmatrix} k(z) & 0 & \frac{k^3(z)}{3} \\ 0 & \frac{k^3(z)}{3} & 0 \\ \frac{k^3(z)}{3} & 0 & \frac{k^5(z)}{5} \end{bmatrix} dz$$

$$\widehat{b}_2 = \int_{-h}^{h} \begin{bmatrix} k(z) & 0 \\ 0 & \frac{4k^3(z)}{3} \end{bmatrix} dz.$$

Also, define the functions

$$\alpha(x) \;=\; \frac{2(1-\mu(x))}{(1+\mu(x))(1-2\mu(x))}$$

and

$$\beta(x) \;=\; \frac{4}{1+\mu(x)}.$$

The girder has a variable cross section that is described using numbers

$$h_i \text{ for } i = 1, ..., 6$$

such $h_i \;<\; h_{i+1}$, $h_1 \;=\; -h$, and $h_6 \;=\; h$ along with positive numbers

$$k_1,\ k_2,\ \text{and } k_3.$$

The function k giving the y boundary is given as follows

$$k(z) = \begin{cases} k_1, & h_1 < z < h_2 \\ k_1 + \frac{k_2-k_1}{h_3-h_2}(z-h_2), & h_2 < z < h_3 \\ k_2, & h_3 < z < h_4 \\ k_2 + \frac{k_3-k_2}{h_5-h_4}(z-h_4), & h_4 < z < h_5 \\ k_3, & h_5 < z < h_6 \\ 0, & \text{otherwise} \end{cases}$$

It is convenient to introduce the 5-tuples

$$\widehat{\alpha} \;=\; [0,\ \frac{k_2-k_1}{h_3-h_2},\ 0,\ \frac{k_3-k_2}{h_5-h_4},\ 0]$$

and

$$\kappa \;=\; [k_1,\; k_1 - \widehat{\alpha}(2)h_2,\; k_2,\; k_2 - \widehat{\alpha}(4)h_4,\; k_3]$$

It follows that the y boundary function may be written as

$$k(z) \;=\; \kappa(i) \;+\; \widehat{\alpha}(i)z, \text{ if } h_i \;<\; z \;<\; h_{i+1} \text{ for } i = 1, ..., 5$$

It is supposed that the sample girder used in the experiment has a spatial dependent Young's modulus with constant Poisson's ratio. The interval $(0, L)$ is partitioned into N_p subintervals (L_{k-1}, L_k) for $k = 1, ..., N_p$ of equal length with $L_0 = 0$ and $L_{N_p} = L$.

Define the characteristic functions

$$\Xi_k(x) = \begin{cases} 1, & x_{k-1} < x < x_k \\ 0, & \text{otherwise} \end{cases}$$

Set

$$E(x) \;=\; \sum_{k=1}^{N_p} E_k \Xi_k(x). \tag{4.1}$$

and define the parameter vector

$$q \;=\; [E_1 \; E_2 \; ... \; E_{N_p}]^T$$

External forcing is experimentally implemented by a vertical force applied to the girder centered at $x = L/2$. It is modeled in terms of the vector-valued function $\mathbf{f} \;=\; [f_1, f_2, f_3]^T$ where $f_1 \;=\; f_2 \;=\; 0$ and

$$f_3(x,y,z) = \begin{cases} f & \text{for } \xi_0 < x < \xi_1,\ -k(x) < y < k(x),\ -h < z < h \\ 0 & \text{otherwise,} \end{cases}$$

Setting

$$f_0 = 2f \int_{-h}^{h} k(z)dz$$

and

$$f_2 = \frac{2}{3} f \int_{-h}^{h} k^3(z)dz,$$

and then define the vector

$$F = \begin{bmatrix} 0 \\ 0 \\ f_0 \\ 0 \\ f_2 \end{bmatrix}$$

Spatial approximations are obtained using piecewise linear elements defined on a uniform mesh on $(0, L)$ [35]. Thus, partition the interval $(0, L)$ into N subintervals $[x_i, x_{i+1}]$ and define $M = N + 1$ functions $\{b_i \text{ for } i = 1, ..., M\}$ by

$$b_i(x) = \begin{cases} \frac{x - x_{i-2}}{x_{i-1} - x_{i-2}} & \text{for } x \in [x_{i-2}, x_{i-1}], \\ \frac{x_i - x}{x_i - x_{i-1}} & \text{for } x \in [x_{i-1}, x_i], \\ 0 & \text{otherwise,} \end{cases}$$

Also, define the vector-valued function $x \mapsto \mathbf{b}(x) = [b_1(x), ..., b_M(x)]^T$, and let $\mathbf{0}$ designate an M row-vector of zeros.

It is convenient to define the $5 \times 5M$ matrix valued function

$$x \mapsto B(x) \;=\; \begin{bmatrix} \mathbf{b}(x)^T & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{b}(x)^T & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{b}(x)^T & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{b}(x)^T & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{b}(x)^T \end{bmatrix}$$

Let $\mathbf{c}$ be a column vector $5M$ vector.

Finally, define the matrices

$$G_k = \int_0^L [\alpha \Xi_k(x) B_x(x)^T p_0^T \widehat{b}_0 p_0 B_x(x) \tag{4.2}$$

$$+\beta \Xi_k(x)(P_0 B(x) + P_1 B_x(x))^T \widehat{b}_1 (P_0 B(x) + P_1 B_x(x))$$

$$+\beta \Xi_k(x) B(x)^T P_2^T \widehat{b}_2 P_2 B(x)] dx.$$

where

$$p_0 \;=\; \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \end{bmatrix}$$

$$P_0 \;=\; \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

$$P_1 \;=\; \begin{bmatrix} 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

$$P_2 \;=\; \begin{bmatrix} 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

As for the external forcing term, define the vector

$$\mathbf{F} \;=\; \int_0^L F^T B(x) dx$$

and express the work as a function of $\mathbf{c}$

$$W(\mathbf{c}) \;=\; \mathbf{F}^T \mathbf{c}$$

In the experimental setup, the girder is supported by roller located at $x = 0$ and $x = L$ Boundary behavior is captured by penalizing the boundary displacements. Towards this end, introduce the matrices

$$K_0 \;=\; \begin{bmatrix} K_{01} & 0 & 0 & 0 & 0 \\ 0 & K_{02} & 0 & 0 & 0 \\ 0 & 0 & K_{03} & 0 & 0 \\ 0 & 0 & 0 & K_{04} & 0 \\ 0 & 0 & 0 & 0 & K_{05} \end{bmatrix},$$

$$K_L \;=\; \begin{bmatrix} K_{L1} & 0 & 0 & 0 & 0 \\ 0 & K_{L2} & 0 & 0 & 0 \\ 0 & 0 & K_{L3} & 0 & 0 \\ 0 & 0 & 0 & K_{L4} & 0 \\ 0 & 0 & 0 & 0 & K_{L5} \end{bmatrix},$$

along with

$$G^{(0)} \;=\; B(0)^T K_0 B(0)$$

and

$$G^{(L)} \;=\; B(L)^T K_L B(L).$$

The displacement vector is the solution of the set of linear equations

$$[\sum_{k=1}^{N_p} E_k G_k \;+\; G^{(0)} \;+\; G^{(L)}]\mathbf{c} \;=\; \mathbf{F} \tag{4.3}$$

Define the matrix

$$G(q) \;=\; \sum_{k=1}^{N_p} E_k G_k \tag{4.4}$$

so that the solution of (4.3), $\mathbf{c}$, is a function of q, $\mathbf{c} \;=\; \mathbf{c}(q)$ and

$$G(q)\mathbf{c}(\mathbf{q}) \;=\; \mathbf{F} \tag{4.5}$$

4.3 Data Analysis

Data consist of flexural measurements in which displacements resulting from various loadings of the girder are measured. The girder is composed of pre-stressed concrete and is 40 feet long, 3 feet deep and 1.5 feet wide. The values of $h_i,\ i = 1, ..., 6$ and $k_i,\ i = 1, 2, 3$ are given by

$$h_1 = -1.5,\ h_2 = -1.0,\ h_3 = -0.5,\ h_4 = 0.75,\ h_5 = 1.0,\ \text{and}\ h_6 = 1.5$$

and

$$k_1 = 0.75,\ k_2 = 0.25,\ \text{and}\ k_3 = 0.5$$

The load is applied at $x = 20$. Observations of displacements are taken at

$$x \ = \ 5, \ 10, \ 15, \ 20, \ 25, \ \text{and } 30 \text{ ft.}$$

For convenience we define an observation location vector that is an array containing the locations at which displacement measurements are made.

$$\mathbf{x}_o \ = \ \begin{bmatrix} 5 \\ 10 \\ 15 \\ 20 \\ 25 \\ 30 \end{bmatrix}.$$

Thus, we define the observation operator that amounts to the evaluation of a function $\phi \in H^1(0, L)$ at points contained in the vector $\mathbf{x}_o$ by

$$C\phi \ = < \delta_{\mathbf{x}_o}, \phi > = \ \begin{bmatrix} \phi(5) \\ \phi(10) \\ \phi(15) \\ \phi(20) \\ \phi(25) \\ \phi(30) \end{bmatrix}.$$

Rollers are located at locations $x = 0$ ft and $x = 35$ ft. The girder has been repaired at one end $x = 5$ ft. Data from an experiment consists of 100 7-tuples. The first component in each 7-tuple is the magnitude of the applied force. The remaining six numbers are the resulting displacements $\mathbf{z}$ measured at

the indicated locations along the girder. The measurements are made at the top surface of the girder.

To formulate the validation problem, it is assumed for ease that Poisson's ratio μ is a constant. Furthermore, the Young's modulus is a piecewise constant function defined on N_p subintervals each of length L/N_p. The values of the Young's modulus function are between 0 and $E_{max} \times 10^7$ where $E_{max} = 20$. A number N_D of the 7-tuples are selected as data. The admissible set Q_{ad} is defined to be a closed set in $\mathbb{R}^{N_p}$ given by

$$Q_{ad} = \{q = (E_1, ..., E_{N_p}) \in \mathbb{R}^{N_p} : 0 \leq E_i \leq E_{max}\}.$$

Given an admissible parameter vector $q \in Q_{ad}$, the approximating displacement vector $\mathbf{c}(q)$ is calculated from equation (4.4) [35]. The approximating displacement function is expressed as

$$\mathbf{V}^N(\mathbf{c}(q))(x) = B(x)\mathbf{c}(q)$$

The observation vector to be compared with data is

$$\xi(q) = C([0\ 0\ 1]PV^N(c(q))) = \begin{bmatrix} w_0(5) \\ w_0(10) \\ w_0(15) \\ w_0(20) \\ w_0(25) \\ w_0(30) \end{bmatrix}.$$

For applying the information theoretic approach in selecting the best model,

given the data $\mathbf{z}_j$ for $j = 1, ..., N_D$, and the 4 models which are models with 2 parameters, 3 parameters, 4 parameters and 6 parameters we need to evaluate the AIC, Δ, ω values for each model. AIC_c=$AIC + ((2k(k+1))/(n-k-1))$ where n is the sample size and AIC=$-2log(l(\hat{\theta}|y)) + 2k$, where k is the number of the estimated parameters in the model including σ and $log(l(\hat{\theta}|y)) = \frac{-n}{2}log(\sigma^2)$. It is important here to note that n represents the number of observations and therefore a vector of 7 entries is considered 7 observation. The reason we treated n this way is we think even if the vector sonsidered one element, it still consists of many observations and that affects the way the standard deviation is being calculated. Given our set of 4 models then AIC difference for a model j is calculated as: Δ_j =AIC_j-AIC_{min}, where AIC_{min} is the smallest value of the AIC values for all the 4 models and Akaike's Weight for a model j is defined to be w_j=$\frac{\exp(-1/2\Delta_j)}{\Sigma_{r=1}^{R}\exp(-1/2\Delta_r)}$. The following table summarizes the results of the AIC method.

Table 4.1: AIC values

Model	number of parameters	log(l)	AIC_c	Δ	ω	rank
1	2	19.04367	-37.9636	-2.86598E-5	.423473624	1
2	3	18.31454	-36.3791	1.58452	.19175469	3
3	4	19.0373	-37.6535	.310052632	.362654498	2
4	6	16.4813	-32.0594	5.904225806	.022117386	4

If we look at the table above, we see that the best model is the model that has two parameters. This model is ranked number 1 as itsΔ value is 0 and has the largest Akaike weight. The second best is the model that has four parameters and the third is the model with three parameters. To investigate further and to know which models can be considered and which should not, we look at the confidence set of models. The 95% confidence set of models consists of three models which are models 1, 2 and 3. We notice that model 4 (The model with 6 parameters)

is not included in the confidence set of models. This tells us this model should not be considered for modeling the girder. We see also that the ratio of evidence for model 3 is approximately 85% where as the ratio of evidence of model 2 is approximately 45% which says that model 3 is twice as best as model 2 with the fact that model 1 is the best. We can see from this analysis that having more parameters in the model doesn't always give better model and therefore if we are looking for the best model in our case then using two parameters will be best.

Chapter 5

Conclusion and Future Work

In this thesis, the information theory approach has been used as basis for model selection. We went over the information theoretic technique in model selection and the R^2 technique and we used these two techniques in selecting the best model among set of candidate models. We applied the technique on statistical models, neural network models and physics based models. We used the information theoretic approach to select the best statistical model for the resilient modulus of a soil, the best neural network model for resilient modulus of a soil and the best model for a girder. We introduced the stability of the ranking which is essential in model selection and we found that the information theoretic approach is more stable than the R^2 approach.

In chapter two, The information theory approach has been taken in deciding the best model to model (MR). Many reasons has been stated for why using AIC is better than using the R^2 approach. With the AIC approach, better results has been captured and more information about the models is known. For example, we saw that the best model in modeling the (MR) is the factorial model and

in this matter the R^2 and the AIC analysis agreed only when the development data set is used. Using the evaluation data set, shows that the R^2 approach didn't give a stable ranking and it conflicts the results when using the development data set. The AIC analysis suggested that the only statistical model that should be considered is the factorial model where as we couldn't get this result by the R^2 analysis as the value of the R^2 of the polynomial model is high but still not a good fitted model. Similarly, The stress based model and the multiple regression model are considered very poor models and they should not be considered for predicting MR. Further more, we saw that the AIC ranking is more stable and the model that ranked good according to AIC will be good for different subsets of the data. In particular, the AIC analysis, showed that the factorial model is the best model and in fact it is the only good model in the set of candidate models. Furthermore, the result we get from the information theory approach is stable and the ranking of the model stays the same in the development and the evaluation data sets which tells us that we can assure the best chosen model by AIC will be the best for the whole population not only in the data. In summary, Using the information theory approach is more appropriate and gives more information about the candidate models more than what R^2 does.

In chapter three, The information theory approach has been taken as a basis in in deciding the best neural network model to model (MR). Many reasons has been stated for why using AIC is better than using the R^2 approach. With the AIC approach, better results has been captured and more information about the models is known. In the Neural network models, the best model was Multi-Layer Perceptrons Network-1 Model and the second best model was Multi-Layer Perceptrons Network-2 Model and this ranking was stable and this result was the

same when the development data set or the evaluation data set is used. This notion of stability in this ranking is essential as by that we are sure that this model will fit what ever data used. The other four models represent a poor fit for the data even their R^2 values was considerably large. In summary, using the AIC is much better approach is deciding the best model and using the R^2 is sometimes misleading as we saw in the set of neural network models.

In chapter four, we used the information theoretic approach to choose the best model for a girder. We found that the model of two parameters is the best model to model the girder. The second best model is the model that has four parameters. This tells us that more parameters in the model is not always better.

Following the results from this thesis, We really support using the information theory approach for this kind of modeling rather than using the R^2 approach and testing hypothesis. we encourage using this information theoretic approach and investigate its stability with respect to different subsets of the data to prove the stability of the ranking.

For future work, more investigation is needed on the mathematical stability of the AIC as an approximation for minimizing the Kullbak-Leibler Information is needed. Furthermore some more research is also needed on how this AIC behaves if the number of the estimated parameters in the model is very close or larger than the data size. Another issue one needs to investigate more is how we consider the data size in the case if we are using the data as vectors.

Bibliography

[1] Akaike, H., (1973). Information theory as an extension of maximum likelihood principle. *Second International Symposium on Information theory*Akademiai Kiado,Budapest. PP 267-281.

[2] Akaike, H., (1981). Likelihood of a model and information criteria.*Journal of Econometrics*. pp3-14.

[3] American Association of State Highway and Transportation Officials(AASHTO)(1986). *Standard Specifications for Transportation Materials and methods of sampling and Testing*Washington, D.C.

[4] Anderson, D.(2008). *Model Based Inference in the life sciences: A primer on evidence*, Springer, NY.

[5] Anderson, D., Burnham, K., and Thompson, W. (2000). Null hypothesis Testing: Problems, Prevalence, and an altarnative. *Journal of Wildelife Management*Vol 64. PP 921-923.

[6] Burnham, D. and Anderson D.(2002). *Model selection and multimodelinference A pratical information-Theoritic Approach*,2nd edition,Springer, NY.

[7] Claeskens, G., and Hjort, N., (2009). *Model selection and model averaging*, 2nd edition, Cambridge University press, UK.

[8] Dia, S. and Zollars, J. (2002). Resilient Modulus of Minnesota Road Research Project Subgrade Soil. *Transportation Research Record* 1786, Transportation Research Board, National Research Council, Washington, D.C., pp. 20-28.

[9] Demuth H. and Beale M.(1996). *Neural Network Design* PWS.

[10] Ebrahimi, A. (2006). Regression and neural network modeling of resilint modulus based on routine soil propertries and stress states. *PhD Dissertation*, University of Oklahoma, Norman, OK.

[11] Farrar, M.J. and Turner, J.P. (1991). Resilient Modulus of Wyoming Subgrade Soils. *Mountain Planins Consortium Report No. 91-1*, The University of Wyoming, Luramie, Wyoming, USA.

[12] Haykin, W.L., (1994). *A neural Networks: A comprehensive Foundation*, Macmillan, College Publishing, New York.

[13] Hossain, S., (2009). Estimation of subgrade resilient modulus for Virginia Soil. *88th Transportation Research Board Annual Meeting* Washington D.C. January 2009.

[14] S.C. Hunter, *Mechanics of Continuous Media, 2nd ed.*, Ellis Horwood Limited, New York, 1983.

[15] Hurvich, C. M., and Tsai, C.L., (1989). Regression and time series model selection in small samples. *Biometrika* Vol 76. pp. 297-307.

[16] Larson, R. and Marx, M., (2007). An introduction to mathematical statistics and its applications. 3rd edition.

[17] Luenberger, D., G. *Optimization by Vector Space Methods*, Wiley, New York, 1969.

[18] May, R. and Witczah, M., (1981). Effective Granular Modulus to Model Pavement Response. *Transportation Research Record*. Transportation Research Board, National Research Council, Washington, D.C., pp. 1-9.

[19] Mendenhall, W. and Sincich T., (2003). A second course in statistics:regression analysis.6th ed.Pearson Education,Inc,New Jersy, USA.

[20] Meirovitch, L., (1967) *Analytical Methods in Vibrations*, McMillan, New York.

[21] Mendenhall, W., and Sincich, T., (2007). *Statistics for Engineerings and the Sciences*,5th ed,Pearson Prentice Hall, New Jeresy, USA.

[22] Myers, R., Montgomery, D., (2001). Generalized Linear Models: With Applications in Engineering and the Sciences, John Wiley and Sons Inc., New Jersey.

[23] Niederreiter, H., (1992). *Random Number Generation and Quasi-Monte Carlo Methods*, SIAM, Philadelphia, Pennsylvania, 1992.

[24] Park, H.I., Kweon, G.C., and Lee, S.R. (2006). Prediction of Resilient Modulus of Subgrade Soils and Subbase Materials Based on Artificial Neural Network.

[25] Patterson, D. (1999). *Artificial Neural Networks*Prentice Hall, Singapore.

[26] Schultz, M., (1973). *Spline Analysis*, Prentice-Hall, Englewood Cliffs, N.J.

[27] Shahin, M. A., Jaksa, M.B., and Maier, H.R. (2001). Artificial Neural Network Applications in Geotechnical Engineer. *Australian Geomechanics*, Vol. 36, No. 1, Australian Geomechanics Society, St. Ives, New South Wales, Australia, pp. 49-62.

[28] Shahin, M. A., Maier,H.R.,and Jaksa,M.B. (2004). Data Division for Developing Neural Networks Applied to Geotechnical Engineering. *Journal of Computing in Civil Engineering*,American society of Civil Engineering, Vol.18, No.2, PP.105-114.

[29] Shibata, R., (1983). A theoritical view of the use of AIC. *Time series Analysis:Theory and Practice* pp 237-244, Amsterdam, Holland.

[30] Solanki, P., Zaman, M. and Ebrahimi, A. (2009). Regression and Artificial Neural Modeling of Resilient Modulus of Subgrade Soils for Pavement Design Applications . *Soft Computing in Pavement and Geomechanical Systems: Recent Advances* (Edited by Gopalakrishnan, K., Ceylan, H. and Attoh-Okine, N.O.). Springer-Verlag, Series, Vol. 259, 2009, pp. 269-304.

[31] Tarantola, A., (2005). Inverse Problem Theory. *SIAM* Philadelphia, 2005.

[32] Tian, P., Zaman, M. M., and Laguros, J.G. (1998). Variation of Resilient Modulus of Aggregate Base and Its Influence on Pavement Performance. *Journal of Testing and Evaluation, JTEVA*, Vol. 26, No. 4, pp. 329-335.

[33] Tukey, J., (1980). We need both exploratory and confirmatory. *The American Statistician* Vol34, pp 23-25.

[34] Tyler, D., Baladi,G., Sessions, C., and Hiadar, S., (2009). Backcalculated and Laboratory Measured Resilient Modulus Values. *88th Transportation Research Board Annual Meeting* Washington D.C., January 2009.

[35] White, L., Jinsong, P. (2009). Deformation of narrow plates for modeling girders and bridges. *In preperation*

[36] Zaman, M., Solanki, P., Ebrahimi, A. and White. L. (2010). Neural Network Modeling of Resilient Modulus Using Routine Subgrade Soil Properties . *Int. J. of Geomechanics*, Vol. 10, No. 1, 2010.

[37] Zurda, J.M., (1992). Introduction to Artificial Neural Systems, West, St. Paul, Minnesota, USA.

Appendix A

Preliminaries

In this Appendix we give an introduction to modeling, least square theory and likelihood theory without exposing into the mathematical proofs. Estimation of model parameters and the principle of parsimony are explained briefly. We also summarized the most known methods for model selection.

6.1 Least Square Theory

Least square theory has been used a lot in modeling. To summarize the idea of this theory, lets assume that the dependent variable y is modeled as a function of the variable x. Lets take the very simple case, the linear model. The linear model is of the form $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, where the ϵ_i are the error terms often called the residuals. Under the least square theory, we want to estimate β_0 and β_1 that minimize the $\Sigma(\epsilon_i)^2$. In our example here, the parameter estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ minimizes the average squared error and defines a regression line that is best fit [15]

Example 1. Least square method

In the linear case, Given data $(x_1, y_1), ..., (x_N, y_N)$,

The error is $\Sigma(\epsilon_i)^2 = \sum_{i=1}^{N}(y_n-(ax_n+b))^2) = E(a,b)$. The goal is to find the values of a, b that minimize $E(a,b)$. To do that, we note that E is a function of two variables a, b so, we use calculus to find the minimum. We solve: $\frac{\partial E}{\partial a} = 0, \frac{\partial E}{\partial b} = 0$.

$$\frac{\partial E}{\partial a} = \sum_{n=1}^{n} 2(y_n - (ax_n + b)(-x_n) = \frac{\partial E}{\partial b} = \sum_{n=1}^{n} 2(y_n - (ax_n + b)(1) = 0 \quad (6.1)$$

We can rewrite equation (6.1) as:

$$(\sum_{i=1}^{n} x_n^2)a + (\sum_{i=1}^{n} x_n)b = (\sum_{1=1}^{n} x_n y_n), (\sum_{i=1}^{n} x_n))a + (\sum_{1=1}^{n} 1)b = (\sum_{i=1}^{n} y_n). \quad (6.2)$$

So we have the following matrix equation:

$$\begin{bmatrix} \sum_{n=1}^{N} x_n^2 & \sum_{n=1}^{N} x_n \\ \sum_{n=1}^{N} x_n & \sum_{n=1}^{N} \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} \sum_{n=1}^{N} x_n y_n \\ \sum_{n=1}^{N} y_n \end{bmatrix} \quad (6.3)$$

which implies:

$$\begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} \sum_{n=1}^{N} x_n^2 & \sum_{n=1}^{N} x_n \\ \sum_{n=1}^{N} x_n & \sum_{n=1}^{N} \end{bmatrix}^{-1} \begin{bmatrix} \sum_{n=1}^{N} x_n y_n \\ \sum_{n=1}^{N} y_n \end{bmatrix} \quad (6.4)$$

, provided that $M = \begin{bmatrix} \sum_{n=1}^{N} x_n^2 & \sum_{n=1}^{N} x_n \\ \sum_{n=1}^{N} x_n & \sum_{n=1}^{N} 1 \end{bmatrix}$ is invertible.

To see that M is invertible we find $detM$.

$detM = (\sum_{n=1}^{N} x_n^2 \sum_{n=1}^{N} 1) - (\sum_{n=1}^{N} x_n \sum_{n=1}^{N} x_n)$.

Since $\bar{x} = \frac{1}{N}\sum_{n=1}^{N} x_n$, we get:

$detM = N\sum_{n=1}^{N} x_n^2 - (N\bar{x})^2 = N\sum_{n=1}^{N}(x_n - \bar{x})^2$.

Therefore as long as not all x_n are equal, det M is not zero and therefore M is

invertible.

6.2 Likelihood Theory

Likelihood theory is much more general, far less used and easier to understand as a concept than least square Theory. In this section we briefly introduce the likelihood theory including some definitions, examples and theorems that we will need later.

Definition 6.1. Let $Y_1, ..., Y_n$ be n independent random variables with probability density functions (pdf)$f_i(y_i, \theta)$ depending on a vector-valued parameter θ. The joint density of n independent observations $y = (y_1, ..., y_n)^T$ is $f(y, \theta) = \prod_{i=1}^{n} f_i(y_i, \theta) = L(\theta, y)$. This expression is viewed as a function of the unknown parameter θ given the data y and is called the likelihood function.

Often we work at the natural logarithm of the likelihood function, the so-called log-likelihood function:
$logL(\theta, y) = \sum_{n=1}^{n} \log f_i(y_i, \theta)$.
A sensible way to estimate the parameter θ given the data y is to maximize the likelihood (or equivalently the log-likelihood) function by choosing the parameter value that makes the data actually observed as likely as possible.

Definition 6.2. The maximum-likelihood estimator (MLE) is defined as the value $\hat{\theta}$ such that $logL(\hat{\theta}, y) \geq logL(\theta, y)$ for all θ [16].

There are important distinctions between the terms probability and likelihood. Likelihood is relative or comparative and the likelihood values don't sum to 1. Both quantities are conditional on various things and both are useful in

inductive inference [16].

Example 2. The Log-Likelihood for the Geometric Distribution.
Consider a series of independent Bernoulli trials with common probability of success Π. The distribution of the number of failures Y_i before the first success has pdf equals to $P(Y_i = y_i) = (1 - \Pi)^{y_i}\Pi$, for $y_i = 0, 1, ...$
Direct calculation shows that $E(Y_i) = (1 - \Pi)/\Pi$. The log-likelihood function based on n observations can be written as $logL(\Pi, y) = \sum_{i=1}^{n}(\log(1-\Pi)+\log\Pi) = n(\hat{y}log(1 - \Pi) + log\Pi)$ where $\hat{y} = \sum_{n=1}^{n} y_i/n$ is the sample mean. The fact that the log-likelihood depends on the observations only through the sample mean shows that $\hat{y}$ is a sufficient statistic for the unknown probability Π. For example, the log-likelihood function for a sample of $n = 20$ observations from a geometric distribution when the observed sample mean is $\hat{y} = 3$.

There are many reasons that make the Likelihood theory important and worth of study. First, Likelihood and log-Likelihood functions form the general basis for deriving estimates of unknown parameters in the models of science hypothesis. Second, Model selection based on Kullback- Leiber information depends on the Likelihood theory. Third, The Maximum Likelihood Estimators are Normally distributed, have minimum variance and unbiased.

Definition 6.3. The score vector.
The first derivative of the log-likelihood function is called Fishers score function, and is denoted by $u(\theta) = \frac{\partial logL(\theta,y)}{\partial\theta}$.
Note that the score is a vector of first partial derivatives, one for each element of θ.

Example 3. The Score Function for the Geometric Distribution.

The score function for n observations from a geometric distribution is $u(\Pi) = \frac{dlogL}{d\Pi} = n(\frac{1}{\Pi} - \frac{\hat{y}}{1-\Pi})$. Setting $u(\Pi) = 0$ and solving for Π gives the maximum likelihood estimator $\hat{\Pi} = \frac{1}{1+\hat{y}}$.

Note that the MLE of the probability of success is the reciprocal of the number of trials. This result is intuitively reasonable as it is the longer it takes to get a success, the lower our estimate of the probability of success would be.

Suppose that in a sample of $n = 20$ observations we have obtained a sample mean of $\hat{y} = 3.$, then MLE of the probability of success would be $\hat{y} = 1/(1+3) = 0.25$.

6.3 Models

A model is a simplification or approximation of reality and hence does not reflect all of reality. While a model can never be "truth" a model might be ranked from very useful to useful, to some what useful to finally essentially useless[4]. Models are central to science as they allow a rigorous treatment and integration of Science hypothesis, Data, Statistical assumptions and estimates of unknown model parameters [6].

Model selection is the task of selecting a model from a set of potential models, given data. In a practical situation, we only have data and we want to use this data to predict the future or predict in another similar situation where we don't have data. In most cases, this is one of the fundamental tasks of scientific inquiry. Determining the principle behind a series of observations is often linked directly to a mathematical model predicting those observations. Models are central to science as they allow a rigorous treatment and integration of:

- Science hypothesis
- Data
- Statistical assumptions
- Estimate of unknown model parameters.

It is important to keep in mind that models are only approximation to full reality. Box(1979) said "...,all models are wrong, some are useful". In fact, we should think of the value of alternative models as better or worse, instead of right or wrong [4]. In the real world with real data, there is no valid concept of a model that is exactly true, representing full reality. If we had a true model, we would still have to estimate its many parameters. Some people view a true model must be considered infinite dimensional. Of course this is a useful concept and view but this is just another way of saying there is no valid way of a true model.Sometimes it is useful to think of realities a function f and let this f be infinite dimensional. Thus $f(x)$ represent conceptually the full truth and it is based on a very large number of parameters.

6.4 Model parameters

Estimating parameters is very important in mathematical modeling. Usually the data is used to approximate the parameters in the model. One of most important things that gives better estimated parameters is enlarging the sample size. If the sample size is small, the parameter estimates will have large variances and wide confidence intervals [**?**]. Usually data is used to both select a model and estimate the model parameters. There are two issues in this matter, one is which parameters should be included in the model and the second is how to approxi-

mate these parameters? For the first issue, the importance of the parameters in the data plays the role in deciding if this parameter will be in the model or not. There are many procedures that have been developed to estimate model parameters, the most three famous are: Least squares(LS), Maximum likelihood(ML), and Bayesian methods. We will summarize these methods later.

6.5 The principle of Parsimony

The principal of parsimony takes many forms and has many formulations in many areas ranging from philosophy, Physics and Mathematics. In mathematical modeling, Parsimony is the concept that a model should be as simple as possible with respect to the included variables, model structure, and number of parameters. Parsimony relates to under and over fitting models which in other words deals with the suitable number of parameters that should be included in the model [7]. Lets imagine fitting a model and imagine the fit is improved by a model with more parameters, then the question is where should one stop? Box and Jenkins (1970) suggested that the principle of parsimony should lead to a model with the smallest possible number of parameters for adequate of the data. Edwards(1970) says" ...too few parameters and the model will be so unrealistic as to make prediction unreliable, but too many parameters and the model will be so specific to the particular data set so to make prediction unreliable." So we want a proper trade off between under and over fitting. Either extreme will result will result in unreliable prediction. The fit of any model can be improved by increasing the number of parameters in the model, however too many parameters could lead to over fitting and to very complicated model that is difficult to use. The concept of parsimony has been important principle for several decades. This

notion appears to be simple, however it is very useful in modeling and statistical interference. We can think of parsimony as a function of the number of the estimated parameters in the model, say k. Given a fixed data, when k increases, squared bias decreases and that is good but the variance, or measure of uncertainty increases and that is of course is not good. That is there is a penalty or cost for adding more parameters. So the difficulty is what is a suitable k?

Parsimony is a conceptual goal because in the real world neither bias nor variance is known to the researcher analyzing the data. Parsimony is a desired characteristic of a model used for inference, and it is usually visualized as a suitable tradeoff between squared bias and variance of parameters estimators. Parsimony lies between the evils of under fitting and overfitting.In summary,parsimony represent a tradeoff between bias and variance as function of the dimension of the model. Of course we want to be close to the reality as much as we can and also we want the model to be as simple as possible. This agrees with what Albert Einstein have said" Everything should be made as simple as possible, but no simpler." In summary, in mathematical modeling, parsimony means only parameters that really matter ought to be included in a selected model [7].

In the practical situation, one is presented with data and usually many models are suggested. From these models, how one chooses the best model. Furthermore, what is meant by best here? Even in the same model, how can one decides the parameters that should be included in the model. Most selection model strategies work by assigning a certain score to each candidate model. In some cases there might be a clear best model, but in other cases these scores might reveal that there are several candidates that do almost as the best model. Many methods

has been used like Cross validation, the coefficient of multiple determination, bootstrapping and more. In this chapter we'll summarize some of these methods.

6.6 Cross validation

Cross validation is a model evaluation method that is largely known. It is used for model selection by choosing the model that has the smallest estimated error. It is better than the residual evaluation as in residual evaluations, they don't give an indication of how well the learner will do when it is asked to make new predictions for data it has not already seen. Cross validation overcomes this problem by not using the entire data set when training a learner. Some of the data is removed before training begins. Then when training is done, the data that was removed can be used to test the performance of the learned model on "new" data. There are many kinds of cross validation such as: holdout method, k-fold cross validation and leave-one-out cross validation.

6.6.1 Holdout method

This is the simplest kind of cross validation. The data is separated into two sets, called the training set and the testing set. Then the model is generated using the training set only. Then the model is used to predict the output values for the data in the testing set. Then some method is used like the least squares to find the mean absolute test set error which is used to evaluate the model. The advantage of this method is that it is fast and easy to compute, however, its evaluation can have a high variance. The evaluation may deponed heavily on which data end in the training set and which end in the testing set and therefore the evaluation heavily depends on how the splitting of the data is made.

6.6.2 K-fold cross validation

This method is improvement of the hand out method. The data set is divided into k subsets and the hold out method is repeated k times. Each time, one of the k subsets is used as the test set and the other $k-1$ subsets are put together to form a training set. Then the average error across all k-trails is computed. The advantage of this method is it doesn't matter much how the data is divided as every data pints will be in a test set exactly one time and will be in the training set $k-1$ times. The variance of the resulting estimate is reduced as k increased. The disadvantage of this method is the intensive calculation that are in the method as the algorithm has to be rerun k times which means it takes k times to make an evaluation. Improvement of this method also is to randomly divide the data into a test and training k different times. Doing this is good as one can independently choose how large each test set is and how many trials one averages over.

6.6.3 Leave-one-out cross validation

This is a special case of the k-fold validation taken with k is the number of the data points in the data set say, N. That means that N separate times and the function is trained on all data except for one point and a prediction is made for that point. As in the k-fold validation, the average error is computed and used to evaluate the model. Leave-one-out cross-validation often works well for estimating generalization error for continuous error functions such as the mean error but may perform poorly for discontinuous error functions such as the number of misclassified cases. In the case of discontinuity, k-fold cross validation is preferred. In general, cross validation method is very simple and empirical way

of comparing models. However, it has many shortages some of which are:

• The method can be time-consuming, since many training runs may be needed.

• Since not all the data is used for training, one might loose some information in the model.

• Over and under-fitting may occur when using cross validation.

6.7 Bootstrapping

In Bootstrapping, instead of repeatedly analyzing subsets of the data, one repeatedly analyze subsumable of the data. Each subsamable is a random sample with replacement from the full sample. Desponds on the data and the sample size, one can decide how many subsample to choose. Therefore, bootstrapping is a general approach to model selection based on building a sampling distribution for a statistic by resampling from the data at hand. Then one method is used on each sample.To be more clear, lets say we have data (x) with sample size n and 6 models that we need to see the best of them. We generate say 10,000 bootstrap data sets, each of size n and all derived by resampling the data with replacement. Then a criteria is used to decide which is best model in each bootstrap. After finding all the ranking for the models in the 10,000 bootstrap samples, then one find the bootstrap frequency for each model which is the number of times that a model was found to be best divided by 10,000 in our example. Then we can see which model has the largest relative frequency. As we can see, this is a good way to improve the criteria that is used to decide the best model but still it depends heavily on the criteria one chooses.

6.8 Bayesian Method

This method uses the rules of probability theory to select the best model. Prior probability distributions are used to describe the uncertainty in the unknowns. After observing the data, a distribution is assigned, then what is so called model indication is calculated for each model. The problems with this method is that it needs large data sets to be reasonable to use as the larger the data set the better the method. This comes from the probability theory that this method is built on. In more details, suppose that a set of k models are under consideration for data Y and that under the model M_k has density $p(Y \mid \theta_k, M_k)$ where θ_k is a vector of unknown parameters that indexes the members of the model M_k.Then a prior probability distribution $p(\theta_k \mid M_k)$ to the parameters of each model and also a prior probability is $p(M_k)$ is assigned to each model. In other words, this method has three stages which are, first the model M_k is generated from the probabilities $p(M_1), ...p(M_k)$, second the parameter vector θ_k is generated from $p(\theta_k \mid M_k)$,and third the data Y is generated from $p(Y \mid \theta_k, M_k)$. (Chipman, H. 2001)

6.9 Mallows' Cp

In this method, we want the model that minimize Cp, where $Cp = \frac{SSE}{(\hat{\sigma})^2} + 2P - n$, where p is the number of the estimated parameters in the model, n is the sample size and $(\hat{\sigma})^2$ is the mean square error of the model. The model with the lowest Cp value approximately equal to p is the most adequate model. Under a model not suffering from appreciable lack of fit, Cp has expectations nearly equal to p.

6.10 R^2 Method

This method is largely used to decide which model has best fit in a set of candidate models. Given data $y_i = y_i(x)$ where y_i is the experimental value of the input x. Let $\hat{y}_i$ be the predicted value y using the model $\hat{y}_i(x, \alpha_i)$ with α_i parameters. The coefficient of multiple determination, R^2, is calculated as: R^2=1 $- \frac{SSE}{SSY}$ where SSE=$\Sigma(y_i - \hat{y}_i)^2$ and SSY= $\Sigma(y_i - \bar{y})^2$ where $\bar{y}$ is the average value of y_i. Under the R^2 method, one selects the best model to be the model that its R^2 is largest[19]. Rencher and Pun (1980) found this approach to be very poor. There are many examples of models with large values of R^2 but these models represent poor approximation of the truth [6], [29].

6.11 Null Hypothesis Testing

This method is largely used and the dominant approach is to frame a question in two contrasting hypotheses, the first states that there is no difference between population parameters (usually called the null hypothesis) and the second represent either a unidirectional or bidirectional alternative (called the alternative hypothesis). These hypotheses correspond to different models. A substantional arbitrary level (α) usually is picked up to serve as a cut off for statistically significance versus statistically nonsignificant results.

6.11.1 Null hypothesis Testing Procedure

The first step of hypothesis testing is to convert the research question into null and alterative hypotheses. One usually starts with the null hypothesis (H_0). The null hypothesis is a claim of no difference. The opposing hypothesis is the alter-

native hypothesis (H_1). The alternative hypothesis is a claim of a difference in the population, and is the hypothesis the researcher often hopes to bolster. It is important to keep in mind that the null and alternative hypotheses reference population values, and not observed statistics.

The second step is to calculate a test statistic from the data. Large test statistics indicate data are far from expected, providing evidence against the null hypothesis and in favor of the alternative hypothesis.

The third step is to use p Value and write the conclusion. The test statistic is converted to a conditional probability called a $P-$value. The $P-$value answers the question If the null hypothesis were true, what is the probability of observing the current data or data that is more extreme? Small p values provide evidence against the null hypothesis because they say the observed data are unlikely when the null hypothesis is true.

6.11.2 Problems with Null Hypothesis Testing

Even this procedure is largely used in many areas but it has many problems. The fundamental problem in hypothesis testing is not it is wrong, but that it is uninformative in most cases and it has little use in variable or model selection. One curios problem in hypothesis testing is that most null hypothesis are assumed false in the way we state them, so rejecting the null hypothesis doesn't give solid information to us but rather it says we reject or can not reject the null hypothesis. A second problem is about the $p-$value is about events never occurred instead of being about a statement of evidence from an actual preserved event which is the data. Another problem is that the p value is explicitly conditional on the null hypothesis as it is computed based on the distribution of the test

statistic assuming the null hypothesis is true. Another problem is that the p value is dependent on the sample size and one can reject a null hypothesis with an enough large sample, even though if the true difference is trivially small. Another problem is that using fixed α-level to decide to reject or not reject the null hypothesis makes little sense as sample size increases. In this case even when the null hypothesis is true and sample size is infinite , a type I error (which is rejecting a null that is true) still occurs with probability α and this is not consistent as theoretically speaking α should go to zero as n goes to infinity. There is a common misuse of the $p-$value. The proper interpretation of the p value is based on the probability of the data given the null hypothesis and not the converse. Usually we can not accept or prove the null hypothesis, we only fail to reject it. The p value is not the probability that the null hypothesis is true as many people interpret it.

6.12 Choosing the variables in a model

In the practical situation, one knows what kind of model is needed but not sure what variables are good to include in the model and what variables should not be included. Again this is another situation where a trade off is needed. More variables means more accurate model but we don't need a very complicated model that is hard to deal with. Therefore, we trade some of the exactness to the simplicity. There are many methods that are used in deciding the good variables that should be in the model. Some of these methods are: Forward selection, backward elimination, stepwise and principal component analysis. We will briefly summarize in this section how each method works.

6.12.1 Forward Selection

This method is used when one knows the model to use but needs to decide the variables to include. Usually a criteria from the previous stated ones is used to decide. In this forward-selection technique begins with no variable in the model. For each of the independent variables, one adds variables to the model one at a time. At each step, each variable that is not already in the model is tested for inclusion in the model. It is tested by calculating the F statistic that reflects the variable contribution to the model if it is included. The p value for these F statistics are compared to the needed value for significant p. Thus we begin adding the most significant and continue adding variables until none of the remaining variables are significant.Thus variables are added one by one to the model until no remaining variable produces a significant F statistic. Once the variable is in the model, it stays.

6.12.2 Backward Elimination

The backward elimination method is used to decide the variables needed to be considered in a certain model. In this method, one begins by calculating the F statistics for a model, including all of the independent variables. Then the variables are deleted from the model one by one until the variables remaining in the model produce F statistics significant at the needed level. At each level the variable showing the smallest contribution to the model is deleted.

6.12.3 Stepwise

The stepwise method is a modification of the forward-selection technique and only differers in that the variables that are in the model don't necessarily stay in

the model. Variables are added one by one to the model and the F statistic for a variable to be added must be significant at the needed level. However, after the variable is added the method look at all the variables included already in the model and deletes any variable that doesn't produce an F statistic significant at the needed level. After this check is made on all the variables in the model, another variable is added and so on. This process ends only when none of the variables outside the model has a significant F statistic at the needed level or when the variable to be added to the model is the one just deleted from it.

6.12.4 Principal Component Analysis (PCA)

The principle component analysis is appropriate when one obtained measures on a number of observed variables and wish to develop a smaller number of artificial variables called the principal component that will account for most of the variance in the observed variables. From mathematical point view, PCA decomposes high-dimensional data into a low-dimensional subspace component and a noise component. A central issue here, is choosing the dimensionality of the subspace component so that all of the noise, but none of the signal is removed. To do this, the probability of the data for each possible subspace dimensionality is computed. For a given dimensionality, this requires integrating over all possible PCA decompositions (i.e. words over all subspaces).

Methodology of PCA

Suppose that $x_1, x_2, ...x_M$ are $N \times 1$ vectors. We summarize the steps to do PCA,

Step1: We find the average vector $\bar{x} = \frac{1}{M} \sum_{i=1}^{M} x_i$.

step2: We subtract the mean to get $\Phi_i = x_i - \bar{x}$.

step3: We form the matrix

$$A = \begin{bmatrix} \Phi_1 & \Phi_2 & ... & \Phi_N \end{bmatrix}$$

(Note A is an $N \times M$ matrix).

Step4: We find the sample covariance matrix $C = \frac{1}{M} \sum_{n=1}^{M} \Phi_n \Phi_n^T = AA^T$.

step5: We compute the eigenvalues of C:$\lambda_1 > \lambda_2 > ... > \lambda_N$.

step6: We compute the eigenvectors of C: $u_1, u_2, ..., u_N$.

step7: Since C is symmetric, $u_1, u_2, ..., u_N$ form a basis, so we can write $x - \bar{x} = b_1 u_1 + b_2 u_2 + ... + b_N u_N$.

step8: Then we only keep the terms corresponding to the K largest eigenvalues, so we have: $\hat{x} - \bar{x} = \sum_{i=1}^{K} b_i u_i$, where $K << N$ and hence the representation of $\hat{x} - \bar{x}$ into the basis $u_1, u_2, ..., u_k$ is thus

$$y = \begin{pmatrix} b_1 \\ b_2 \\ ... \\ b_K \end{pmatrix}$$

Geometric interpretation of PCA:

PCA projects the data along the direction where the data varies most. These directions are determined by the eigenvectors of the covariance matrix corresponding to the largest eigenvalues. Note that The magnitude of the eigenvalues corresponds to the variance of the data along the eigenvector directions.

How to choose the principle component:

The question now is how one decides what is K? In fact the criteria to find k is we need the smallest k that satisfies: $\frac{\sum_{i=1}^{K} \lambda_i}{\sum_{i=1}^{N} \lambda_i} >$ Threshold.

Dimensionality Reduction implies information Loss, and one usually wants to preserve as much information as possible. Therefore, we want to minimize the error, that is, minimize $\| x - \hat{x} \|$. The best low-dimensional space can be determined by the "best" eigenvectors of the covariance matrix of x (i.e., the eigenvectors corresponding to the "largest" eigenvalues). It can be shown that the low-dimensional basis based on principal components minimizes the reconstruction error: $e = \| x - \hat{x} \|$. It can also be shown that the error is equal to: $\frac{1}{2} \sum_{i=k+1}^{N} \lambda_i$. PCA is very important and is largely used in data analysis [16].

Appendix B

The Information Theory Approach

In the modeling problem, we always look for a model that is rich enough to explain the data and on the other hand simple enough to understand. Information theory is an integral part of almost any data analysis. When we receive something that decreases our uncertainty about the state of the world, it is called information [5]. Information can not be measured with instruments but can be defined in terms of probability distributions. In our practical life, researcher is usually presented with data, or some science hypotheses. Then a mathematical model is derived to well represent each of the hypothesis, then one can ask many questions:

- Given the data, which science hypothesis has the most empirical support?
- What is the rank of the hypotheses, given the data?
- What is the likelihood of one hypothesis versus another?
- How can we decide the best mode; among the candidate models?

These questions are in the core of model inference. As pointed out by [4],

there are three principles that regulate our ability to make inferences in the sciences:

- Simplicity and parsimony
- Several working hypothesis
- Strength of evidence

7.13 Kullback-Leibler Information

There are several ways of measuring closeness of a model f to a model g but the one that intimately linked to the maximum likelihood method is what is called Kullback-Leibler Information distance.

Definition 7.4. Kullback-Leibler information between models f and g is defined for continuous functions as: $I(f,g) = \int_{\Omega} f(x) \ln(\frac{f(x)}{g(x|\theta)})\, dx$, where ln denotes the natural logarithm and Ω denotes the space where the models are defined.

The notion $I(f,g)$ denotes the information lost when g is used to approximate f. A model g is a perfect model if $I(f,g) = 0$ where f is the true model. Sometimes $I(f,g)$ is called the distance from g to f [6].

The Kullback-Leibler distance can be conceptualized as a directed distance between the models f, g. It is also important to notice that it is not the distance (as defined in topology) as the distance from f to g is not the same as from g to f. The K-L distance between models is a fundamental quantity in science and information theory and it is logical basis for model selection in conjunction with likelihood inference.

Example 4. Let f be a gamma distribution with two parameters($\alpha = 4, \beta = 4$). Consider 4 models, g_i where $i = 0, ..., 4$, each with two parameters (See the table below).

Table (Values of K-L distances between distributions)

notation	Approximating model	$I(f, g_i)$	Rank
g_1	Weibull distribution ($\alpha = 2, \beta = 20$)	.04620	1
g_2	Lognormal distribution($\theta = 2, \sigma^2 = 2$)	.67235	3
g_3	Inverse Gaussian($\alpha = 16, \beta = 64$)	.06008	2
g_4	F distribution($\alpha = 4, \beta = 10$)	5.74555	4

In this example, the Weibull distribution is closest to f. In other words it loses the least information from f. The F distribution is relatively far from f.

Theorem 7.5. *let $I(f, g)$ be defined as above, then:*

a) $I(f, g) \geq 0$

b) $I(f, g) = 0$ if and only if $f = g$.

Proof

a. First, note that both $f(x)$ and $g(x)$ are valid probability distributions and hence satisfy $f(x) \geq 0, g(x) \geq 0$ and $\int_\Omega f(x)\, dx = 1, \int_\Omega g(x) dx = 1$.

Without loss of generality we assume that $f(x), g(x) > 0$.

Define a new function $h(x) = \frac{g(x)-f(x)}{f(x)}$;

thus $\frac{g(x)}{f(x)} = 1 + h(x)$.

And hence $-1 < h(x) < \infty$.

Now,

$$I(f,g) = \int_\Omega f(x) ln(\frac{f(x)}{g(x)})dx = -\int_\Omega f(x) log(\frac{g(x)}{f(x)})dx \tag{7.5}$$

Using the fact that $\int_\Omega f(x)h(x)dx = 0$, (7.1) becomes:

$$I(f,g) = \int_\Omega f(x)g(x)dx - \int_\Omega f(x) log(\frac{g(x)}{f(x)})dx.$$

$$= \int_\Omega f(x)g(x)dx - \int_\Omega f(x) log(1+h(x))dx = \int_\Omega f(x)[h(x) - log(1+h(x)]dx = \int_\Omega f(x)t(h(x))dx,$$

where $t(h(x)) = h(x) - log(1+h(x))$.

It is suffices to show that $t(h)$ is nonnegative.
To see this we notice that: $t(h) = h - log(1+h)$,
so $t'(h) = \frac{h}{1+h}$ and $t''(h) = \frac{1}{(1+h)^2}$.

The only critical number is 0 and since $t''(0) = 1 > 0$, therefore at $h = 0$, t has minimum and it is the only minimum and hence $t(h) \geq 0$ which proves a.

for part b,

It is obvious that if $f(x) = g(x)$ then $I(f,g) = 0$.

For the other direction, assume that $I(f,g) = 0$ then $\int_\Omega f(x)t(x)dx = 0$

which implies that $h(x) - log(1+h(x)) \equiv 0$ for all x.

Hence $h(x) = log(1 + h(x))$ and so $e^{h(x)} = 1 + h(x)$.

This implies that:

$$1 + h(x) + \sum_{i=2}^{\infty} \frac{1}{t!}[h(x)]^i = 1 + h(x),$$

$$\text{therefore } \sum_{i=2}^{\infty} \frac{1}{t!}[h(x)]^i = 0$$

Note that if $h(x) > 0$ then the latter above cant be true and this means: if $I(f, g) = 0$, then $h(x) \leq 0$ for all x.

Now if $h(x) < 0$ over any set of x values ϕ for which $\int_\phi f(x)dx > 0$

but this will imply $\int g(x)dx < \int f(x)dx$ which is a contradiction,

and hence $h(x)$=0 and that completes the proof.

Definition 7.6. The score vector of a model g is defined as:

$$u(x, \theta) = \frac{\partial}{\partial\theta} log(g(x \mid \theta_0))$$

and the information matrix function is defined as:

$I(\theta_0) = E_f \frac{\partial^2 log(g(x|\theta)}{\partial\theta^2}$, evaluated at $\theta = \theta_0$

Lemma 7.7. *If* θ_0 *satisfies* $\min_{\theta\in\Theta}[I(f, g)] = \int f(x)log(\frac{f(x)}{g(x|\theta_0)})dx$, *then:*

$$E_f[\frac{\partial}{\partial\theta} log(g(x \mid \theta_0))] = 0$$

proof

Since θ_0 minimizes [I(f,g)], then:

$\frac{\partial}{\partial\theta}\int f(x)ln(\frac{f(x)}{g(x|\theta_0)})dx = 0,$

then $\frac{\partial}{\partial\theta}\int f(x)log(f(x))dx - \frac{\partial}{\partial\theta}\int f(x)log(g(x \mid \theta))dx = 0,$

but since θ is not involved in $f(.)$, then the first term of the above is 0 and the second term can be written as:

$$\int f(x)\frac{\partial}{\partial\theta}log(g(x \mid \theta))dx \text{ at } \theta = \theta_0 = 0$$

and hence:

$$E_f[\frac{\partial}{\partial\theta}log(g(x \mid \theta_0))] = 0 \text{ and the proof is complete.}$$

7.14 Akaike's Information Criterion

Akaike introduced his information theoretic approach in a series of papers in the mid-1970s as a theoretical basis for model selection. Akeike's finding of a relation between the K-L information and the maximized log-likelihood has allowed major practical and theoretical advances in model selection and data analysis. In this section we introduce the AIC with its derivation and then talk about its use.

Theorem 7.8. *The estimate of the expected relative distance between the fitted model and the unknown true mechanism is* $AIC = -2log(l(\hat{\theta} \mid y)) + 2k$ *where* $log(l(\hat{\theta} \mid y))$ *is the numerical value of the log-likelihood at its maximum point and* k *is the number of the estimated parameters in the model.*

proof

We start by the K-L distance between two models f, g.

$$I(f, g(. \mid \theta_0)) = \int_\Omega f(x) \ln(\frac{f(x)}{g(x|\theta_0)})\, dx.$$

Note that $I(f, g)$ doesn't involve any data nor any value of x.

Given data y as a sample of $f(.)$, we find the MLE $\hat{\theta} = \hat{\theta}(y)$ and compute an estimate of $I(f, g(. \mid \theta_0))$ as:

$$I(f, g(. \mid \hat{\theta}(y))) = \int_\Omega f(x) \ln(\frac{f(x)}{g(x|\hat{\theta}(y))})\, dx.$$

So our goal here is to find $\hat{\theta_0}$ that minimizes $I(f, g)$.

On average the estimated value of $I(f, g)$ is $E_y[I(f, g(. \mid \hat{\theta}(y)))]$.

so our problem becomes select a model g to minimize $E_y[I(f, g(. \mid \hat{\theta}(y)))]$.

By logarithm properties:

$E_y[I(f, g(. \mid \hat{\theta}(y)))] = \int_\Omega f(x) \ln(f(x))\, dx - E_y[\int_\Omega f(x) \ln(g(x \mid \hat{\theta}(y)))\, dx$

$=$ Constant $- E_y E_x[\ln[g(x \mid \hat{\theta}(y))]]$.

So our problem becomes to maximize $E_y E_x[\ln[g(x \mid \hat{\theta}(y))]]$,

and therefore we need to maximize $T = \int_\Omega f(y)[\int_\Omega f(x) \ln(g(x \mid \hat{\theta}(y)))\, dx$.

T can be written as: $T = E_{\hat{\theta}} E_x[\ln[g(x \mid \hat{\theta})]]$,

where it is understood that MLE, $\hat{\theta}$, is based on sample y and the two expectations are for x and y (and hence $\hat{\theta}$) both with respect to truth f.

Now expand $\ln(g(x \mid \hat{\theta}))$ about θ_0,

so we have:

$\ln(g(x \mid \hat{\theta})) \approx ln(g(x \mid \theta_0)) + [\frac{\partial ln(g(x|\theta_0)}{\partial \theta}]'[\hat{\theta} - \theta_0] + \frac{1}{2}[\hat{\theta} - \theta_0]'[\frac{\partial^2 ln(g(x|\theta_0)}{\partial \theta^2}][\hat{\theta} - \theta_0]$.

take the expected value to get:

$$E_x[\ln(g(x \mid \hat{\theta}))] \approx E_x[\ln(g(x \mid \theta_0))] + E_x[\frac{\partial ln(g(x \mid \theta_0)}{\partial \theta}]'[\hat{\theta} - \theta_0] + \frac{1}{2}[\hat{\theta} - \theta_0]' E_x[\frac{\partial^2 ln(g(x \mid \theta_0)}{\partial \theta^2}][\hat{\theta} - \theta_0] \tag{7.6}$$

Note that E_x is used to mean E_f and $\hat{\theta}$ to mean $\hat{\theta}(y)$.

We recall that $E_x[\frac{\partial ln(g(x|\theta_0)}{\partial \theta}] = 0$ by the lemma above.

Denote $I(\theta)$ by $E_x[\frac{\partial^2 ln(g(x|\theta_0)}{\partial \theta^2}]$ to get:

$$E_x[\ln(g(x \mid \hat{\theta}))] \approx E_x[\ln(g(x \mid \theta_0))] - \frac{1}{2}[\hat{\theta} - \theta_0]' I(\theta)[\hat{\theta} - \theta_0] \tag{7.7}$$

Now we take the expectation with respect to $\hat{\theta}$ to get:

$$E_{\hat{\theta}} E_x[\ln(g(x \mid \hat{\theta}))] \approx E_x[\ln(g(x \mid \theta_0))] - \tfrac{1}{2} tr[I(\theta) E_{\hat{\theta}}[\hat{\theta} - \theta_0][\hat{\theta} - \theta_0]'$$

Let $T = E_{\hat{\theta}} E_x[\ln(g(x \mid \hat{\theta}))]$ and let $\Sigma = E_{\hat{\theta}}[\hat{\theta} - \theta_0][\hat{\theta} - \theta_0]'$

So now we have:

$$T = E_x[\ln(g(x \mid \theta_0))] - \frac{1}{2} tr[I(\theta)\Sigma] \tag{7.8}$$

Now, we expand $\ln(g(x \mid \theta_0))$ about $\hat{\theta}$ to get:

$$\ln(g(x \mid \theta_0)) \approx ln(g(x \mid \hat{\theta})) + [\tfrac{\partial ln(g(x|\hat{\theta})}{\partial \theta}]'[\theta_0 - \hat{\theta}] + \tfrac{1}{2}[\theta_0 - \hat{\theta}]'[\tfrac{\partial^2 ln(g(x|\hat{\theta})}{\partial \theta^2}][\theta_0 - \hat{\theta}]$$

by taking the expected value with respect to x we get:

$$E_x[\ln(g(x \mid \theta_0))] \approx E_x[ln(g(x \mid \hat{\theta}))] + E_x[\tfrac{\partial ln(g(x|\hat{\theta})}{\partial \theta}]'[\theta_0 - \hat{\theta}] + \tfrac{1}{2}[\theta_0 - \hat{\theta}]'[E_x \tfrac{\partial^2 ln(g(x|\hat{\theta})}{\partial \theta^2}][\theta_0 - \hat{\theta}]$$

Since $[\frac{\partial ln(g(x|\hat{\theta})}{\partial \theta}] = 0$, we get:

$$E_x[\ln(g(x \mid \theta_0))] \approx E_x[\ln(g(x \mid \hat{\theta}))] - \frac{1}{2} tr[E_x[\hat{I}(\hat{\theta})][\theta_0 - \hat{\theta}][\theta_0 - \hat{\theta}]'] \qquad (7.9)$$

We note that:

$E_x[\hat{I}(\hat{\theta})][\theta_0 - \hat{\theta}][\theta_0 - \hat{\theta}]' \approx [I(\theta_0)][E_x[\theta_0 - \hat{\theta}][\theta_0 - \hat{\theta}]']$

$= [I(\theta_0)][E_x[\hat{theta} - \theta_0][\hat{theta} - \theta_0]']$

$= [I(\theta_0)]\Sigma$

Now we substitute in (2.5) to get:

$$E_x[\ln(g(x \mid \theta_0))] \approx E_x[\ln(g(x \mid \hat{\theta}))] - \tfrac{1}{2} tr[I(\theta_0)\Sigma]$$

Now substitute in 7.4 to get:

$T \approx E_x[\ln(g(x \mid \hat{\theta}(x)))] - tr[I(\theta_0)\Sigma]$ So now we have:

$$\hat{T} \approx \ln(g(x \mid \hat{\theta}(x))) - \hat{tr}[I(\theta_0)\Sigma]$$

and the best model will be the one that has the largest value of $\hat{T}$.

For convention reasons, The formula is written as:

$$-2\ln(g(x \mid \hat{\theta})) + 2\hat{tr}[I(\theta_0)\Sigma]$$

In the our case, we have $\hat{tr}[I(\theta_0)\Sigma] = K$, and then we get

$AIC = -2\ln(g(x \mid \hat{\theta})) + 2K$ and the proof is complete.

7.15 AIC_c: A second order improvement

AIC might perform poorly if there is too many parameters in relation to the size of the sample. Sugiura (1978) derived a second order variant of AIC that is called AIC_c.

AIC_c=$AIC + ((2k(k+1))/(n-k-1))$ where n is the sample size [6].

AIC_c is used in the case where the sample size is small relative to the number of parameters and usually when $\frac{n}{k} < 40$.

If n is large with respect to k, then the term $((2k(k+1))/(n-k-1))$ will be very small and close to zero ($lim(((2k(k+1))/(n-k-1))) = 0$) and then AIC_c becomes the same as AIC. AIC_c merely has an additional bias term correction to take under consideration the sample size when it is small.

It is important to mention that one must use AIC or AIC_c consistently in the analysis rather than mixing the two criterions.

Definition 7.9. AIC differences are defined as $\Delta_i = AIC_i - AIC_{min}$, where AIC_{min} is the smallest value of the AIC values for all the set of candidate models.

These differences estimate the relative expected K-L differences between f and g_i. Δ_i values are easy to interpret and give a quick comparison and ranking of candidate models. Here the model estimated to be best has $\Delta_i = \Delta_{min} = 0$.

It is important to note that it is not the absolute size of the AIC value, it is the relative values and hence the Δ_i values are important.

The larger Δ_i is, the less possible it is that the fitted model g_i is the K-L best model, given the data.

Definition 7.10. Given the data and the set of R models, Akaike's Weight for a model i, is defined to be w_i=$\frac{\exp(-1/2\Delta_i)}{\Sigma_{r=1}^{R}\exp(-1/2\Delta_r)}$.

It is important to note that the w_i depends on the entire set of models; therefore if a model is added or dropped during the analysis, then the w_i must be recomputed for all the models in the newly defined set. A given w_i is considered as the weight of evidence in favor of model i being the best model in the R candidate models. As we can see Akaike weight provides an effective way to scale and interpret the Δ_i. Therefore, given that there are only R models, it is convenient to normalize the relative likelihoods to sum to 1. For the best model in the set of candidate models, $\Delta_{min} = 0$, hence for that model $\exp(-1/2\Delta_{min}) = 1$. It is clear that the bigger the Δ_i the smaller the w_i

Definition 7.11. The ratio of evidence of a model g is the Akaike weight of that model deviled by the Akaike weight of the best model.

The ratio of evidence of a model gives an evidence of a kind of weak or strong support for the best model versus any other model in the set of candidate models. Such ratios represent the evidence about fitted models as to which is better in a K-L in information sense. In particular there is often interest in the ratio $\frac{w_i}{w_j}$ where model i is the estimated best model and j indexes the rest of the models in the set. These ratios are not affected by any other model and therefore it does not depend on the full set of R models, just on models i, j. These evidence ratios are invariant to all other models besides i and j.

7.16 Confidence Set of Models

As we stated earlier, data analysis involves the proper tradeoff between bias and variance or in other words tradeoff between under fitting and over fitting. AIC gives a very efficient way to rank the models, as in this approach, given a set

of candidate models , one choose the best model where the information loss is minimized. It is perfectly reasonable that several models would serve nearly equally well in approximating the information in a set of data. Researcher must admit that there are sometimes competing models and the data don't support selecting only one model. When more than one model has substantial support, some form of multi model inference should be considered and in that case the confidence intervals are important.

Definition 7.12. (Confidence set for the $K - L$ best model)
For a 95% confidence set on the actual K-L best model, one of the approaches is based on Akaike weights, interpreted as approximate probabilities of each model being the actual best model, given the data. In this approach we sum the Akaike weights from largest to smallest until the sum is just ≥ 0.95, then the corresponding subset of models is a type of confidence set on the K-L best models.

Another approach to developing a confidence set of models is an approach that based on the idea of a Δ_i being a random variable with a sampling distribution. in this approach we look at the corresponding values of Δ and interpret the following empirical support (for a 95% confidence set):

- $0 \leq \Delta_i \leq 2$, Substantial support.
- $4 \leq \Delta_i \leq 7$, Considerable support.
- $\Delta_i > 10$, essentially no support.

7.17 Relative Importance of Variables

The issue of selecting a good model is important but on the other hand in the same model to decide which variables should be included in a certain model and to compare the importance of the variables in the same model is also essential. The principal component analysis (PCA) method that is usually used in this situation. There is another approach using the information theory approach and namely the Akaike weight is used in that matter.

For estimating the relative importance of predictor variables x_j can be made by summing the Akaike weights across all the models in the set where the variables j occurs say $w_+(j)$. The larger the $w_+(j)$ the more important variable j is, relative to the other variables. Using the $w_+(j)$, all the variables can be ranked in their importance. We can extent this idea to find the importance of subsets of variables and that can be done by summing the Akaike weights of all models that include the particular variables and then find $w_+(i, j, k, ...)$ where $i, j, k, ...$ are the variables we are interested in. Using $w_+(j)$, all variables can be ranked in their importance.

Example 5. Consider the hypothetical example of three regressors, x_1, x_2, x_3, and a search for the best of the eight possible models of simple linear type: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$. The following table shows the appearance of each regressor in the eight models along with hypothetical Akaike weights ω_i.(a 1 denotes that x_i is in the model and a 0 means it is excluded.)

x_1	x_2	x_3	w_i
0	0	0	0.00
1	0	0	0.10
0	1	0	0.01
0	0	1	0.05
1	1	0	0.04
1	0	1	0.50
0	1	1	0.15
1	1	1	.15

The best model is the model that has Akaike weight of 0.5. The sum of the weights for variable x_1 is 0.79. This is an evidence of of the importance of this variable. Note also that variable x_2 was not included in the best model but that does not mean it is not of zero importance. in fact the sum of the weights of the variable x_2 is .35. The sum of the weights for the variable x_3 is 0.85. This concludes that the most important variable here is x_3 and the least important one is x_1.

7.18 Model Averaging

When several models compete for the first place, a further investigation is needed. In other words if it is the case that no single model is clearly superior to the other models in the set , then it is risky to only have the decision made on the AIC values. So we consider model-based inference for prediction. Assume we have a set of R models each having the parameter θ as the predicted value of interest. To conduct model averaging, the estimate of the parameter for each model is weighted by the Akaike weights as follows:

Model-averaged estimate $= \hat{\bar{\theta}} = \Sigma_{r=1}^{R} \omega_i \hat{\theta}_i$ where $\hat{\bar{\theta}}$ denotes a model averaged estimate of θ.

When interpreting AIC values, one should keep in mind the following remarks:

• AIC, AIC_c are not tests, they are just criterions to rank the set of candidate models.

• AIC, AIC_c cant be used to compare models of different Data set.

• Order is not important in computing AIC, AIC_c values.

• There is a theoretical basis to information-theoretic approaches to model selection criteria, while the use of null hypothesis testing for model selection is not.

• AIC ranking of the models is more stable than the R^2 ranking.