MODELING *K*-FACTORS ANALYSIS IN DESIGN OF EXPERIMENT (DOE) TOWARDS REGRESSION APPROACH USING MULTI-LAYER FEED-FORWARD NEURAL NETWORK (MLFF): ITS' APPLICATION IN BIOSTATISTICS

SOBAN QADIR

UNIVERSITI SAINS MALAYSIA

2022

MODELING *K*-FACTORS ANALYSIS IN DESIGN OF EXPERIMENT (DOE) TOWARDS REGRESSION APPROACH USING MULTI-LAYER FEED-FORWARD NEURAL NETWORK (MLFF): ITS' APPLICATION IN BIOSTATISTICS

by

SOBAN QADIR

Thesis submitted in fulfilment of the requirements
for the degree of
Doctor of Philosophy

November 2022

ACKNOWLEDGMENT

I am grateful to Allah, the Almighty and the Most Beneficent, for His Mercy in completing this doctoral research. I am very thankful to my supervisor, Associate Professor Ts. Dr. Wan Muhamad Amir W Ahmad for his continual support, encouragement, and endurance during this research work, His knowledge, experience, and expertise in statistics benefited me. His supreme virtue has given me essential capabilities of wisdom, good health, and patience to go through the whole doctoral research study journey. I want to extend my sincere thanks to the College of Dentistry and Universiti Sains Malaysia (USM) for giving me this opportunity to study in such a wonderful and prestigious college and university. I would also like to thank Dr. Fahad Al-Harbi (former) and Dr. Jehan Al-Humaid, Dean of College of Dentistry, Imam Abdulrahman Bin Faisal University, Saudi Arabia, and Dr. Faisal Al-Onaizan for their continuous support from the day first.

Completing this work was not possible without the prayers of my parents, my wife, and my family and their unrestrained support. My sincere thanks to my friends and colleagues, especially Dr. Imran Alam Moheet, who inspired, supported, and encouraged me from time to time.

TABLE OF CONTENTS

LIST OF TABLES

Page

LIST OF FIGURES

Page

LIST OF ABBREVIATIONS

ANN	Artificial Neural Networking
ANOVA	Analysis of Variance
BMI	Body Mass Index
COVID	Coronavirus
DOE	Design of Experiment
DW	Distilled Water
FLR	Fuzzy Linear Regression
GDP	Gross Domestic Product
HRQOL	Health Related Quality of Life
KAP	Knowledge, Attitude and Practice
mg	Milligram
ml	Millilitre
MLE	Maximum Likelihood Estimator
MLFF	Multilayer Feedforward
MLR	Multiple Linear Regression
MSE	Mean Square Error
PMMA	Polymethyl Methacrylate
PLRLS	Possibilistic linear regression with Least Square
QoL	Quality of Life
USM	Universiti Sains Malaysia
ZrO_2	Zirconium Oxide

PEMODELAN ANALISIS K-FAKTOR DALAM REKA BENTUK EKSPERIMEN (DOE) KE ARAH PENDEKATAN REGRESI MENGGUNAKAN RANGKAIAN NEURAL SUAPAN KEHADAPAN BERBILANG LAPISAN (MLFF): APLIKASINYA DALAM BIOSTATISTIK

ABSTRAK

Reka Bentuk Eksperimen (DOE) adalah merupakan salah suatu metodologi statistik terkenal dan seringkali digunakan secara meluas. Keputusan daripada DOE ini, boleh memberikan suatu hasil yang amat bernilai terutamanya apabila penyelidik mengkaji sifat hubungan di antara pembolehubah. Sejumlah besar kajian yang telah dijalankan pada masa kini, adalah berkisarkan kepada penghasilan keputusan yang lebih tepat. Oleh hal yang demikian, kajian yang melibatkan pembinaan metodologi penyelidikan secara saintifik semakin diberi perhatian utama dari semasa ke semasa. Kajian yang dijalankan ini, bertujuan untuk membangunkan suatu kaedah terbaik untuk menganalisis data, terutamanya melibatkan gabungan DOE, butstrap dan regresi linear serta Rangkaian Neural Suapan Kehadapan Berbilang Lapisan (MLFF) yang menerusi bahasa pengaturcaraan R. Tesis memberi penekanan terhadap penghasilan suatu model regresi yang tepat dan sah yang melibatkan beberapa gabungan kaedah utama. Berdasarkan keputusan yang diperoleh, dapat dirumuskan bahawa metodologi yang dibangunkan ini menunjukkan suatu keputusan yang memberangsangkan bagi teknik pemodelan regresi. Sebagai kesimpulannya, kaedah yang dibina ini dapat digunakan secara berkesan terutamanya apabila melaksanakan pemodelan regresi terhadap reka bentuk eksperimen.

MODELING *K*-FACTORS ANALYSIS IN DESIGN OF EXPERIMENT (DOE) TOWARDS REGRESSION APPROACH USING MULTI-LAYER FEED-FORWARD NEURAL NETWORK (MLFF): ITS' APPLICATION IN BIOSTATISTICS

ABSTRACT

Design of Experiments (DOE) is one of the well-known and widely used statistical methodologies. The results of this DOE provide a very valuable result especially when a researcher studying the relationship between variables. A large number of studies that have been carried out today are hoping for a more accurate result. Indeed, the number of studies involving the development of scientific research methodology is increasing over time. This study aims to develop the best method for data analysis, especially involving a combination of DOE, bootstrap, and linear regression as well as a multi-layer feed-forward neural network (MLFF) through the R programming language. The thesis emphasizes the development of an accurate and valid regression model that involves several combinations of key methods. Based on the results obtained, it can be concluded that this developed methodology shows results encouraging for modeling techniques. In conclusion, this method can be used effectively, especially when performing regression modeling on experimental design.

CHAPTER 1

INTRODUCTION

1.1 Preview of the Chapter

This chapter summarises the entirety of the thesis. In this chapter, the background of the study and an introduction to the design of the experiment (DOE), and an analysis of variance (ANOVA) are briefly provided. The following section addresses the problem statement followed by the rationale of the study. The conceptual framework depicts the entire investigation in the form of a flowchart. The objective section includes the general and specific goals of the current study. Considering the study objectives, this chapter focuses on the scope of the study, as well as its significance and contribution. The last section of this chapter will show the limitations of the study, followed by the thesis organization outlining the complete workflow of the current analysis.

1.2 Background of the Study

The experimental design includes planning, creating, and analyzing experiments to obtain accurate and effective results (Durakovic, 2017). Successful experimental design requires careful planning, adequate design, and the selection of relevant statistical tests for analysis. Randomization, replication, and blocking are the three essential principles of experimental design (Lúcio &

Sari, 2017; Webster & Lark, 2018). The process by which the trials of an experiment are performed is referred to as randomization. As a result, randomization aids in reducing bias. Replication is the repetition of an entire experimental treatment, whereas blocking is the arrangement of similar experimental units in a group (block). As a result, replication and blocking help to improve experiment precision (Lúcio & Sari, 2017). The DOE is a statistical tool employed in various research sectors (Montgomery, 2017: Chien *et al*., 2014; Antony, 2014). Therefore, it has a wide range of applications in both science and non-science topics. Although repeated observations or replication produce variability in the gathered data, it is one of the most important aspects of the experimental design (Casler, 2015). Therefore, the *k*-factors analysis of variance (ANOVA) was developed by Fisher in 1925 to analyze the variation caused by different factors (Cortina *et al*., 2017).

Analysis of variance (ANOVA) is used to investigate the relationship between the variance within classes and the overall variance. The dependent variable must be continuous and the independent variables nominal. Therefore, ANOVA is used to determine the effect of the independent variables (also known as factors) on the dependent variable. This is called the main effect when one independent variable is studied for its effect on the dependent variable. In contrast, the interaction effect occurs when two or more independent variables are examined for their combined effect on the dependent variable (Brauer & Curtin, 2018).

However, ANOVA can only be used to study the connection between dependent and independent variables; it cannot make predictions. Researchers cannot use DOE study designs to make predictions. However, the most effective technique for using DOE for prediction is to transform it into a linear form. Furthermore, certain variables are fuzzy rather than crisp. Thus, fuzzy regression is superior to linear regression in dealing with imprecise relationships between variables (Bas *et al*., 2021).

Thus, the current research aims to apply the DOE study design to linear and fuzzy regression by transforming it into a linear form. In addition, bootstrapping was used to improve the predictability and accuracy of the regression model. In addition, the model was validated using a multilayer feedforward (MLFF) neural network. The proposed approach is expected to help the researchers, particularly in terms of factor validation and prediction (Ahmad *et al*., 2020).

1.3 Problem Statement

Applying DOE study design in biological and healthcare research is prevalent (Gu *et al*., 2020; Memon *et al*., 2019; Hamid *et al*., 2019). However, owing to the limitations of DOE, as discussed in the previous section, the potential associations among the variables should be investigated. There is a scarcity of literature in the biological sciences studies to provide a method of using DOE study designs for regression modeling in conjunction with the model

validation process. Another issue is dealing with data ambiguity that arises as a result of data transformation. However, the demand for data modeling to forecast the future increases. Therefore, this study focuses on developing high-accuracy models with high reliability to maximize the relevance of the findings.

Another problem is that linear regression has some limitations, one of which is that it can only be used to regress crisp data and if the relationship between the variables is not crisp then the accuracy of the model is questioned (van Kuijk *et al*., 2019; Chen *et al*., 2021). Therefore, if the relationship between the variables is ambiguous or imprecise, linear regression may not be the best option because critical information may be lost (Chen *et al*., 2021). Consequently, another issue is how to deal with data ambiguity and the best strategy to incorporate it into the model to increase model accuracy and predictability.

On the other hand, DOE can provide valuable results for studying the associations between variables. However, it cannot be used to make predictions, and combining it with other methods can improve its precision and yield results that meet contemporary research demands.

1.4 Rationale of the Study

In the design of experiments, there is no mechanism for validating fixed factors owing to the limitations of the DOE. Furthermore, the linearity

requirement of linear regression limits its use to crisp datasets. In addition, no robust methodology exists that includes a data transformation process, a method for boosting parameter estimation, a method for dealing with ambiguous relationships between variables, and a method for model validation. Hence, it is necessary to devise remedies to address this issue. This issue can be solved by refining the current method or developing a new one. Rather than establishing a new approach, this study aims to improve the present methodology while still addressing this problem.

1.5 Conceptual Framework of the Study

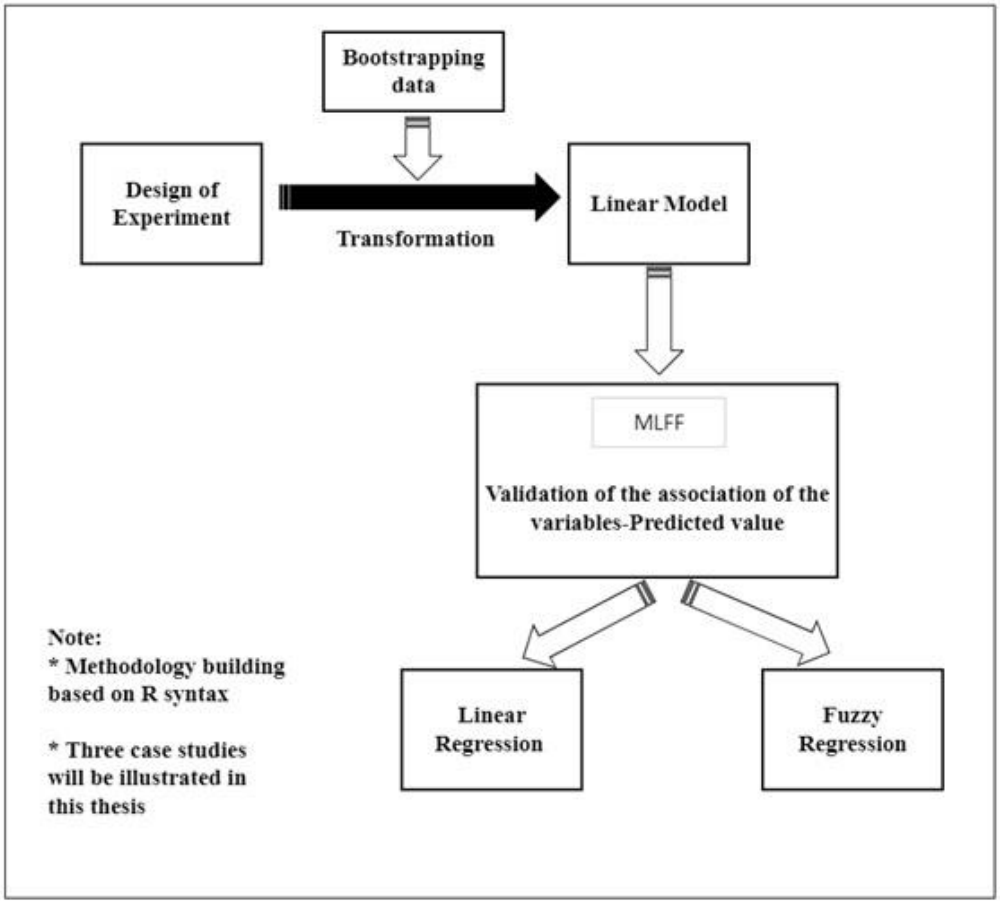

Figure 1.1 Conceptual Framework of the Study

Figure 1.1 depicts the steps involved in using DOE study designs for regression modeling and model validation. The first step is to convert the DOE study design dataset into a linear form and bootstrap the transformed data. Validation of the developed regression model via MLFF neural network using dependent and independent variables. To model the transformed data, the developed methodology employs linear and fuzzy regression models. The syntax is written in the R programming language, and three different retrospective datasets are used to demonstrate the developed methodology.

1.6 Research Objectives

1.6.1 General Objectives:

To model *k*-factors analysis in the design of experiment (DOE) towards a regression approach using a multilayer feed-forward neural network (MLFF) and its application in biostatistics.

1.6.2 Specific Objectives:

a) To transform the *k*-factors design of the experiment (DOE) into a linear form.
b) To develop a hybrid regression method from the DOE *k*-factors, considering bootstrapping, Linear, and fuzzy regression to improve parameter estimation.

c) To validate the regression model obtained in (b) through the *input* and *output* variables using a multilayer feedforward neural network approach.

1.7 Scope of the Study

The DOE technique is used to determine cause-and-effect linkages. The statistical methods available under DOE cannot be utilized to estimate or predict the future value of any parameter. Regression analysis is one of the most effective and powerful methods in this regard because it can be used for both estimation and prediction. Factors were defined as independent variables in the case of DOE when utilized in the regression analysis. Therefore, the preliminary step is to harmonize the DOE and regression approaches when the estimation has to be drawn from the DOE (Ahmad *et al*., 2020).

In the health and biological sciences, there is no robust methodology that allows the use of DOE study designs for regression modeling and model validation. Therefore, the focus of this study was to develop a methodology that will allow researchers to use DOE study designs related to health sciences and biological sciences to make predictions. Hence, the process included transforming the DOE research design into a linear form and employing data bootstrapping to improve the accuracy of the results. The next step is to fit the linear and fuzzy regression models. Utilizing the technique, the input and output variables applied to the MLFF are evaluated. As previously stated, this thesis aims to develop a methodology. Therefore, retrospective data are used to test

this methodology. In addition, the datasets used in this work are connected to biostatistics and are used for various medical studies that include the influence of multiple medications on pain relief, effect of hormonal treatment on plasma calcium, and the effect of denture base materials on tooth surface roughness.

1.8 Significance of the Study

There is an increasing need to make future research predictions. Researchers and academics are looking forward to developing a methodology that can utilize DOE study designs for future predictions. This research will play a significant role in developing a method of using DOE study designs for prediction through regression modeling. Therefore, the contribution of this work to the scientific literature is to provide an opportunity to predict through *k*-factors DOE study designs used for health-related studies. Similarly, the existing methodology lacks a comprehensive method of using datasets from *k*-factors DOE study designs for modeling with validation. Thus, this work provides a method for linearizing the datasets belonging to DOE and a method of dealing with the vague relationship between the dependent and independent variable(s). In addition, the MLFF neural network was used to validate the derived model. Therefore, the significance of this work is that it enables researchers and academicians to use biostatistics-related DOE studies to make inferences and predictions by developing a robust methodology. Accordingly, the developed methodology ensures high predictability and accuracy in the derived model.

1.9 Limitations of the Study

This project aims to develop an integrated methodology for DOE study design that can be used for regression analysis. Therefore, retrospective study data will be used to test the developed method using three different datasets extracted from books and published articles. In addition, only specific statistical tools are employed in the methodology (e.g. bootstrap, multiple linear regression (MLR), fuzzy regression, and MLFF). Second, all three datasets share a common characteristic: they are all related to medical or health sciences. In other words, datasets are not diverse across various scientific and nonscientific fields. Another limitation of this study is that its datasets have one, two, or three factors. Hence, datasets with a single dependent variable and *k*-independent variables (factors) will be considered for methodology testing.

1.10 Thesis Organization

This thesis is divided into six sections. The first chapter serves as an introduction, focusing on the definition of DOE and the various types of statistical analyses associated with it. This chapter discusses the objectives, scope, importance, and limitations of the current study.

The history of DOE, linear and fuzzy regression, as well as their history in biostatistics, are discussed in the second chapter. This chapter also explores the conflicts and gaps among these strategies. The third chapter, Methodology,

contains detailed information on the data transformation procedure and the formulation of statistical programming used to test the method. The results obtained from the proposed statistical programming are presented in Chapter four. This chapter also includes a discussion of the findings obtained using various statistical techniques. Chapter five discusses the findings with regard to each study objective. Chapter 6 is the last chapter of this thesis. It summarises the entire thesis and provides recommendations for future work.

CHAPTER 2

LITERATURE REVIEW

2.1 Overview of the Chapter

The significance of statistical analysis in biological sciences research is discussed in Sections 2.2 and 2.3. Section 2.3 also briefly describes the history and accessibility of the DOE, linear regression, fuzzy regression, and MLFF neural networking. The application of statistical methods presented in various scientific publications is further discussed in Section 2.4. Section 2.5 provides an overview of these statistical tools (DOE, linear regression, fuzzy regression, and MLFF) and their limitations. Hence, owing to the limitations of the statistical methods, the need for a combined technique is highlighted in Section 2.6. The last section provides concluding remarks.

2.2 Approach to Statistical Methods in the Field of Biological Science

In biological sciences, statistical methods are primarily applied to make inferences and draw conclusions about the data (Meeker *et al*., 2021; Mead *et al*., 2017). Testing the statistical significance between variables in any scientific experiment is commonly used to test the proposed hypothesis, leading to conclusions. The most frequently used statistical methods in biological studies to investigate the relationships between variables are the chi-square test,

Student's t-test, *k*-factors ANOVA, and post-hoc tests (Mead *et al*., 2017). Further, statistical methods enable researchers to use their data to forecast trends. Numerous regression analysis techniques are available, depending on the application and nature of the variables (Room, 2021). Simple linear regression, multiple linear regression, logistic regression, exponential regression, and other statistical methods are commonly used for forecasting in biological sciences research (Nawi *et al*., 2018; Ahmad *et al*., 2017; Ahmad *et al*., 2018).

2.2.1 Advantages of Statistical Analysis in Research

The use of forecasting methods as statistical analysis tools in research is becoming increasingly important (Moons *et al*., 2019). The scientific community and authorities in health care and educational systems are more interested in learning about future patterns or behaviours to adjust and prepare policies and plans accordingly. Therefore, the use of statistical forecasting methods is on the rise. There has been an increase in regression analysis in various scientific experiments, whether related to education, clinical, or nonclinical research (Khan *et al*., 2019; Nawi *et al*., 2018; Amir *et al*., 2016).

The developed regression models must be validated and accurate (Yuan *et al*., 2017). The results of these studies, which predict future trends, have the potential to be used for future planning in terms of policies, changes in current study designs, and approaches to improve study outcomes. Therefore, neural networks can be used to validate regression models (Mohamed *et al*., 2011).

Researchers have used this statistical technique to validate their regression models and demonstrate the accuracy of the forecasting capabilities of the models (Ahmad *et al*., 2018; Mohamed *et al*., 2011).

2.3 Practice Towards Design of Experiment and Regression Approach Conceptual Framework

The Experiments which were used to develop DOE and regression were related to biological sciences (Attah *et al*., 2020; Krashniak & Lamm; 2021). Since then, different types of ANOVAs have been used in numerous biological and public health-related studies to study the relationship between different variables and regression approaches for modeling and prediction (Wang *et al*., 2021; Sonneville *et al*., 2021; Madadizadeh, 2020). A brief history of these statistical analysis methods and their accessibility is discussed in the following sections.

2.3.1 History of Design of Experiment

'Experientia' and 'experimentum' were the terminologies used by Bacon (Schwarz, 2012) which are defined as 'the unforced observation which might call the experience' and 'the contrived experience which we might call an experiment, respectively. Bacon's 'contrived experiment' was formulated by Ronald Aylmer Fisher 260 years later (Cortina *et al*., 2017).

Fisher single-handedly developed the field of design of experiments and published his books on experimental design titled 'Statistical Methods for Research Workers' (Politis *et al*., 2017) and 'Design of Experiments' (Runyan, 2019). Randomization, replication, and blocking are fundamental DOE concepts, and Fisher integrated them into a unified approach called 'Design of experiments'. To investigate the variation due to the replication of the observations included in any experiment, R. A. Fisher (1925) coined the terms 'variance' and 'analysis of variance' (Cortina *et al*., 2017). Analysis of variance (ANOVA) is a statistical method used to determine variation introduced by one or more elements or stages of an experimental process (Chambers *et al*., 2017). ANOVA is currently the most frequently used statistical approach for determining the significance of treatments(Lix & Keselman, 2018). Scheffe (1959) provided a well-known definition of ANOVA.

'The analysis of variance is a statistical technique for analyzing measurements depending on several kinds of simultaneous effects, deciding which effect is important and estimating the effect. The measurements or observations may be in an experimental science, such as genetic or nonexperimental one, such as astronomy' (Attah *et al*., 2020).

2.3.2 Accessibility of Design of Experiment

This methodology primarily examines the variation in the mean of each set of individual characteristics such as educational attainment, occupation, and

treatment group. When each group has a different treatment effect, ANOVA can investigate the differences between the means of k population groups (k =1,2,...,n) (Kim, 2017). Consequently, ANOVA tests the hypothesis that $\mu_1 = \mu_2 = \ldots = \mu_n$ (Brereton, 2019). Therefore, ANOVA can be used to determine whether distinct treatment effects cause significant variation in means. A one-way ANOVA is used to examine the effect of a single factor behind mean variation. However, many experiments have examined the effects of two or more factors causing mean variation. The most effective method involves considering two or more factors to determine the ones with the greatest and most significant effects (Kim, 2017; Durakovic, 2017).

The model effects can be unknown constants or random variables. If it is an unknown constant, then the effect is known as a fixed effect. Otherwise, it is called a random effect. Hence, among the statistical tests, ANOVA is commonly used in almost every kind of research, such as medical, engineering, and social sciences, to check the statistical significance between variables or factors (Hlongwa & Rispel, 2021; Gharacheh *et al*., 2020; Cho *et al*., 2020). ANOVA is useful for inferring some effects, regardless of the magnitude of the other effects.

2.3.3 History of Linear Regression

Legendre (1805) and Gauss (1809) devised the least square method, which is also known as the earliest version of regression (Farebrother, 2018). The term

'regression' was coined by Francis Galton in a research paper in which he investigated the tendency of having tall children with tall parents or vice versa and discovered that a child's average height tended to move or 'regress' towards the average height of a child in a population as a whole (Krashniak & Lamm; 2021). Karl Pearson corroborated Galton's universal regression law, discovering that regressing tall and short sons towards the average height in the entire population yielded the same results. (Kellicott; 2020). The modern definition of regression is different from that described by Galton. Specifically, 'regression analysis is the study of dependency of one variable (dependent variable) towards one or more other variables (explanatory variables), to estimate and predict (population) mean or average' (Mahbobi & Tiemann, 2016).

A linear regression model is defined as a model that has a linear relationship between the independent variable(s) and dependent variable (Wang *et al*., 2018). A linear regression function is defined as a response or dependent derived from '*n*' independent variables. Hence, the linear regression model can be expressed as

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_n x_{in} + \varepsilon_i \qquad (2.1)$$

where y_i is the *i*th observation of the response variable, x_{in} is the *i*th value of the *n*th independent variable, and ε_i is a random error term.

2.3.4 Accessibility of Linear Regression

Linear regression is used to make predictions; however, its accuracy depends on the size and quality of the data. The performance of the linear regression model improves as the size of the dataset increases. The linear regression model includes assumptions, parameter estimation methods, and usage. Generally, (i) it is generally assumed that the error terms are normally distributed. (ii) The other assumption for linear regression is based on the classical set theory. A classical set has crisp (exact) boundaries, which means that there is no uncertainty about where the set's boundaries are. Therefore, probability theory concepts and techniques are employed. In regression analysis, the dependent variable is not only influenced by scale variables (e.g., blood pressure, blood sugar, temperature, and body mass index.); it can also be affected by qualitative variables (e.g., gender, nationality, race, etc.). Hence, dummy variables are used in regression analysis, which helps to transform qualitative variables into quantitative variables (Islam *et al*., 2021; Antunes *et al*., 2021).

2.3.5 History of Fuzzy Regression

In 1978, Zadeh connected the theory of possibility to fuzzy sets. Zadeh (1978) demonstrated that a fuzzy variable is associated with a possible distribution in the same manner that a random variable is associated with a probability distribution (Klir & Yuan, 2020). Hence, to investigate the vague relationship between the dependent and independent variable(s), a new

regression approach known as fuzzy regression was developed (Hesamian & Akbari, 2020). Fuzzy regression is based on the fuzzy set theory and possibility theory (Cattaneo, 2017; Feng *et al*., 2018). Tanaka *et al*. first proposed that unfitted errors can be interpreted as model structure fuzziness (Eren, 2018).

Wiener (1894–1964) and Shannon's (1916–2001) remarkable work on the statistical theory of communication has gained worldwide acceptance. The theory states that information is intrinsically statistical and, thus, must be dealt with using statistical methods (Xiong & Proctor, 2018). Thus, the primary concern is the meaning of the information rather than its measurement. Therefore, information rather than probabilistic analysis is possible (Pasman & Rogers, 2020). Gaines and Kohout (1975) first coined the term 'possibilistic' for information analysis in their paper on possible automata (Teodorescu *et al*., 2017). Hence, possibility theory is used to analyze possible information (Boukhari & Omri, 2021).

Since then, several fuzzy regression methods have been developed that use various criteria to maximize fuzzy models, including the development of fuzzy regression with the lowest fuzziness criterion, fuzzy least-square regression, and interval regression (Škrjanc *et al*., 2019; López *et al*., 2019).

2.3.6 Accessibility to Fuzzy Regression

When the data are fuzzier, fuzzy linear regression outperforms statistical linear regression, and the aptness of statistical regression suffers (Sharma & Singh, 2018; Yaseen *et al*., 2018). Fuzzy regression involves assumptions, methods for estimating parameters, and applications (Chen & Nien, 2020). Fuzzy set theory underpins fuzzy regression. Consequently, a membership function is used to specify the degree to which an object belongs to a set. A fuzzy linear regression model with 'n' independent variables can be written as

$$Y = a_0 + a_1x_1 + a_2x_2 + \cdots + a_nx_n \tag{2.2}$$

where $a_0, a_1, a_2, \ldots, a_n$ are unknown fuzzy parameters with components α (denotes the mean or center value) and c (indicates dispersion or spread). The dependent variable is denoted from 'Y' whereas '$x_1, x_2, \ldots, x_n$' are showing denoting the independent variable. Therefore, Equation (2.2) can be written as

$$Y = (\alpha_0, c_0) + (\alpha_1, c_1)x_1 + \cdots + (\alpha_n, c_n)x_n \tag{2.3}$$

2.3.7 Multilayer Feedforward Neural Network

The basic building blocks of artificial neural networks (ANN) are elements known as 'neurons' (Pazikadin, 2020). The biological brain and nervous system inspired the name, structure, and design of the ANN (Moayedi *et al*., 2020). There are many different forms of neural networks, but their underlying principles are the same (Hasson *et al*., 2020). The most primitive ANN type is neurons with vector $\tilde{A} = (a_1, a_2, \ldots, a_n) \in R^n$ as the input and with some internal

weights $W=(w_0,w_1,\ldots,w_n) \in R^{n+1}$. Figure 1 depicts the general structure of a neural network, with the input or independent variables depicted in the first or input layer. Figure 1 also shows the input 'x_n', weight 'v_{ji}', and bias 'v_{j0}' values needed to calculate the value for each node in the hidden layer. The values calculated for the first hidden layer nodes served as the input layer to generate the next layer. (Mijwel *et al*., 2019).

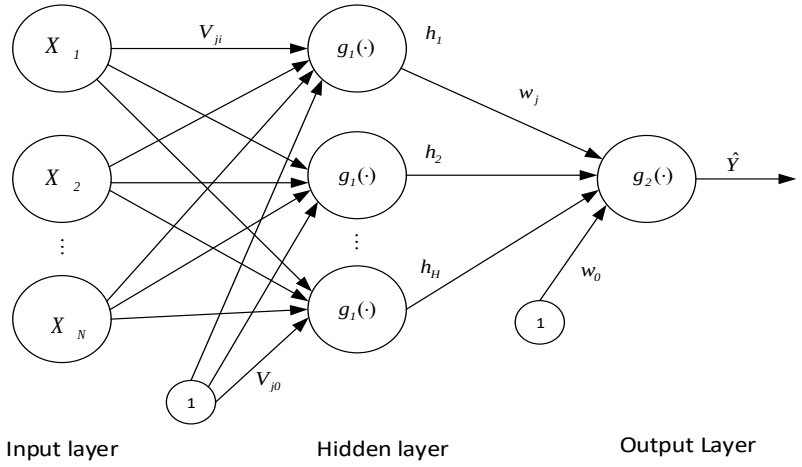

Figure 2.1 Multilayer feedforward neural network model architecture with N input nodes, H hidden nodes, and one output node.
(*Sources: Mohamed et al., 2011*)

2.3.8 Accessibility to Multilayer Feedforward Neural Network

Depending on the connection between the neurons, neural networking architecture falls into two groups, namely, "feed-forward neural network" and "recurrent neural network." The network is called a "feed-forward neural network" if the synaptic connection between the neurons solely conveys information in the forward direction. There is no "feedback" from the neuron

output to the input. However, if synaptic connections transfer information from output neurons to input neurons (feedback), the network is described as a "recurrent neural network" (Li *et al*., 2018). "Single-layer or multiple-layer" feed-forward neural networks are available. If the network includes input and output layers, it is referred to as a single-layer feed-forward neural network; however, if the network has one or more hidden layers in addition to the input and output levels, it is referred to as a MLFF (Faris *et al*., 2019).

2.4 Literature Review

2.4.1 Application of Design of Experiment in Biostatistics

The *k*-factors DOE is widely used in biostatistics. There are a variety of biological domains in which DOE is used to conduct research and is embedded with several types of ANOVA to make the analysis easier. This statistical analysis includes clinical, nonclinical, and survey investigations (Gu *et al*., 2020; Song *et al*., 2021; Memon *et al*., 2019; Hamid *et al*., 2019). ANOVA is commonly used in survey-based research that evaluates knowledge, attitude, and practice (KAP); quality of life (QoL) or health-related quality of life (HRQOL); laboratory-based studies; and other types of analysis (Azlan *et al*., 2020; Saqlain *et al*., 2020; Martino *et al*., 2020; Alzayyat *et al*., 2021).

Community-based KAP studies examine people's KAP regarding disease prevention, disease treatment, drug usage, and attitudes of healthcare

practitioners toward patient management, among other variables. Studies have been conducted in China, Iran, Ethiopia, and Nigeria to assess the general public's knowledge, attitude, and practices regarding coronavirus protection (COVID-19) (Iheanacho *et al*., 2021; Erfani *et al*., 2020; Zhong *et al*., 2020; Yoseph *et al*., 2021). Furthermore, KAP studies on chronic disease management, such as diabetes management during Ramadan fasting, healthcare practitioners' understanding of patient management, and healthcare workers', KAP toward disease, were conducted (Al-Hariri *et al*., 2019; Shah *et al*., 2015; Huynh *et al*., 2020). Regarding analysis, ANOVA is used in all of these types of study designs. In most KAP research, KAP levels are separated into levels that function as 'factors', and ANOVA is used to evaluate it with any scale variable.

Regarding the analysis, health-related quality of life (HRQOL) studies are similar to the KAP studies. In HRQOL research, levels (or factors) are usually created, which are then evaluated with other variables, and ANOVA is used when the primary variable is factored in and analyzed with scale variables (Saxena *et al.,* 2020 2019; Campbell *et al*., 2019; Pope, 2021). On the other hand, laboratory-based studies frequently use experimental designs that make using ANOVA simple. The impact of numerous elements on a study subject or material is commonly investigated in these investigations. For example, a study was conducted to determine the flexural strength of denture-based materials mixed with various concentrations of zirconium oxide nanoparticles (nano-ZrO2) and placed in different solutions for various durations. As a result, *k*-

factors ANOVA was used to determine the effect of independent variables (solutions and time durations) on flexural strength (Gad *et al*., 2021).

Similarly, other researchers have examined how various circumstances affect the different qualities of the studied material (AlBin-Ameer *et al*., 2020; Alzayyat *et al*., 2021). Therefore, DOE has many health and biological science research applications. The application of DOE in biostatistics has been demonstrated in various study domains, as shown in the examples.

Sipiyaruk *et al*. (2021) evaluated the trend using statistical analysis methods in dental research. They found that the use of *k*-factors ANOVA was among the top three statistical analysis tools used for analysis in dental research (Sipiyaruk *et al*., 2021). Similarly, another study evaluated the statistical analysis techniques used in American Physical Therapy Association journals between 2011 and 2012. It was reported that ANOVA was among the top three statistical tests used to study the association between group(s) (Tilson *et al*., 2016). A systematic review and bibliographic study evaluated the global trends in clinical biomarker studies. The other parameters used as trends included the topic of the study, software used for data analysis, and statistical tests used to analyze the data. Reportedly, *k*-factors ANOVA is one of the most commonly used statistical analysis tools (Jovicic, 2021).

2.4.2 Application of Linear Regression in Biostatistics

There are various cases/experiments in biostatistics in which the relationship between variables must be investigated to forecast future values and determine which factor(s) produce an outcome. Accordingly, regression analysis is utilized to realize such goals. The early experiments that were performed and used to drive the regression technique were historically tied to the biological sciences and, hence, were examined using biostatistics. Galton (1875) conducted a sweat pea experiment exploring the relationship between the weights of daughter and mother seeds. The graph depicts a straight line with a positive slope. As a result, he concluded that the straight regression line was correct. Following Galton's invention of regression, there has been continuous research and advancement in the regression methodology. Various regression models have been devised to assist in the regression of the datasets. Numerous approaches have been developed for estimating linear regression parameters. The linear regression models used for regression parameter estimations were least squares and maximum likelihood estimation (Ross, 2020).

Multiple linear regression was employed in studies in which the data were linearly distributed. Amir *et al*. employed triglyceride level as a dependent variable for MLR models and regressed it with blood parameters, weight, hip circumference, and lipid-lowering medication (Amir *et al*., 2016). Similarly, in another study, diabetes mellitus was determined in patients (dependent variable) using BMI, cholesterol level, height, weight, and systolic blood pressure as