DEVELOPMENT OF A NEW HYBRID MODEL AND ITS APPLICATION IN BIOSTATISTICS

FARAZ AHMED FAROOQI

UNIVERSITI SAINS MALAYSIA

2024

DEVELOPMENT OF A NEW HYBRID MODEL AND ITS APPLICATION IN BIOSTATISTICS

by

FARAZ AHMED FAROOQI

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

September 2024

ACKNOWLEDGEMENT

First and foremost, endless gratitude and thanks to **ALLAH** Almighty for giving me the strength and patience to complete this doctoral research. I would like to express my endless gratitude to my supervisor, Associate Professor Ts. Dr. Wan Muhamad Amir W Ahmad for his continuous support, patience, kindness, and encouragement and for his valuable supervision and guidance in achieving this research. I have benefited enormously from his knowledge, experience, excellence, and expertise in biostatistics. I really extend my heartfelt gratitude to him for his steadfast support and for inspiring me to pursue excellence in every facet of this work. I would like to extend my gratitude to the College of Dentistry and Universiti Sains Malaysia (USM) for giving me this opportunity to study at such a wonderful and prestigious college and university. I would like to express my sincere appreciation to Prof. Jehan Al-Humaid, Dean of the College of Dentistry, Imam Abdulrahman Bin Faisal University, Saudi Arabia, as well as Dr. Muhanad Alhareky for their unwavering support and guidance from the very beginning.

I dedicate this thesis to my family members, particularly my parents and wife, for their unwavering support, prayers, and encouragement throughout this research. Additionally, I extend my dedication to each of my siblings, whom I deeply appreciate for consistently offering me love and prayers. Last but not least my sincere thanks to my friends and colleagues, especially Dr. Soban Qadir Khan who inspired, supported, and encouraged me from time to time.

TABLE OF CONTENTS

ACKNOWLEDGEMENT ..... **ii**

TABLE OF CONTENTS ..... **iii**

LIST OF TABLES ..... **viii**

LIST OF FIGURES ..... **ix**

LIST OF ABBREVIATIONS ..... **x**

LIST OF APPENDICES ..... **xi**

ABSTRAK ..... **xii**

ABSTRACT ..... **xiii**

CHAPTER 1 INTRODUCTION **1**

1.1 Preview of the Chapter 1

1.2 Background of the Study 1

1.3 Problem Statement 5

1.4 The Rationale of the Study 7

1.5 Conceptual Framework of the Study 8

1.6 Research Objectives 9

1.6.1 General Objectives 9

1.6.2 Specific Objectives 9

1.7 Scope of the Study 9

1.8 Contribution and Significance of the Study 10

1.9 Limitations of the Study 11

1.10 Thesis Organization 12

CHAPTER 2 LITREATURE REVIEW **13**

2.1 Overview of the Chapter 13

2.2 Introduction to Multiple Linear Regression with Qualitative Predictors 13

2.2.1 The History of Linear Regression 14

2.2.2 Qualitative Predictors in Multiple Linear Regression (MLR) 15
2.2.3 Advantages of Qualitative Predictors in MLR 17
2.2.4 Practice Toward QPV and Regression Approach 18
2.2.5 Role of Qualitative Predictor Variables 19
2.2.6 Linear Model to the Health Sciences Research 19
2.2.7 The Accessibility of Linear Regression 20
2.3 The Fuzzy Regression 21
2.3.1 Introduction of Fuzzy Sets and Membership Functions 22
2.3.2 Fuzzy Logic Models 22
2.3.3 Classification of Fuzzy Sets 23
2.3.4 The Accessibility of Fuzzy Regression 24
2.3.5 The Fuzzy Modelling in Medical 25
2.4 Artificial Neural Network (ANN) 25
2.4.1 Multilayer Feedforward Neural Network (MLFNN) 27
2.4.2 Single-Layer Feed-Forward Neural Network 28
2.4.3 Multilayer Feedforward Neural Network in Health Science 28
2.4.4 The R's Neuralnet Package 29
2.4.5 The Accessibility to Multilayer Feedforward Neural Network 29
2.5 Literature Review 30
2.5.1 Application of Qualitative Predictor in Biostatistics 30
2.5.2 Application of Linear Regression in Biostatistics 31
2.5.3 Application of Fuzzy in Biostatistics 33
2.5.4 Application of Multilayer Feedforward Neural Network in Biostatistics 35
2.5.5 The Hybrid Technique for a Better Result 36
2.6 Reviews of Methodology and Limitations 38
2.7 Concluding Remarks 39

CHAPTER 3 METHODOLOGY 40

3.1 Overview of Chapter 40

3.2 Study Period and Location 41

3.3 Study Design 41

3.4 Research Tool and Data 41

3.5 Advantages of Using R-Software 42

3.6 Methodology Development 43

3.6.1 The Linear Regression Model 43

3.6.1(a) The Predictor Variables 46

3.6.1(b) The Qualitative Predictor Variables (QPV) 47

3.6.1(c) The Data Arrangement 49

3.6.2 Method for Improving Computational Efficiency 51

3.6.2(a) Case Resampling (Bootstrap) 51

3.6.2(b) The R Syntax for Bootstrap 52

3.6.3 Fitting and Testing Multiple Linear Regression Model 53

3.6.4 Fuzzy Linear Regression 54

3.6.4(a) Fitting Fuzzy Linear Regression (FLR) on Proposed Model 57

3.6.4(b) The R Syntax for the Fuzzy Linear Regression 57

3.6.5 Multilayer Perceptron 58

3.6.5(a) Multilayer Feedforward Neural Network (MLFFNN) 60

3.6.5(b) Fitting Multilayer Feedforward Neural Network 60

3.6.5(c) R Syntax for the Multilayer Feedforward Neural Network 61

3.7 The Case Study 62

3.7.1 Case Study I 63

3.7.2 Case Study II 69

3.7.3 Case Study III 74

3.8 Ethical Approval 80
3.9 Conceptual Framework for Proposed Methodology 81
3.10 Summary 82
CHAPTER 4 RESULTS 83
4.1 Chapter Overview 83
4.2 Case Study I: Effectiveness of New Treatment Type 83
4.2.1 Arrangement of Qualitative Predictor Variables 84
4.2.2 Diagnostic Test for the Validation of the Variable Selection 85
4.2.3 Regression Model with Qualitative Predictors Variable 86
4.2.4 Fuzzy Regression Modelling 88
4.2.5 Performance of Derived Models 90
4.2.6 Summary Case Study I 92
4.3 Case Study II: Dental Caries Among the School Children 93
4.3.1 Arrangement of Qualitative Predictor Variables 94
4.3.2 Diagnostic Test for the Validation of the Variable Selection 95
4.3.3 Regression Model with Qualitative Predictors Variable 96
4.3.4 Fuzzy Regression Modelling 97
4.3.5 Performance of Derived Models 99
4.3.6 Summary Case Study II 101
4.4 Case Study III: Body Mass Index (BMI) 102
4.4.1 Arrangement of Qualitative Predictor Variables 103
4.4.2 Diagnostic Test for the Validation of the Variable Selection 103
4.4.3 Regression Model with Qualitative Predictors Variable 105
4.4.4 Fuzzy Regression Modelling 106
4.4.5 Performance of Derived Models 108
4.4.6 Summary Case Study III 110
4.5 Summary 111

CHAPTER 5 DISCUSSION 113

5.1 Overview of the Chapter 113

5.2 The Qualitative Variable for Regression Modelling 113

5.3 Implementation of Developed Methodology 114

5.4 Conclusions From the Results Obtained Using the Hybrid Methodology ... 117

5.5 Conclusion 118

CHAPTER 6 CONCLUSION AND RECOMMENDATIONS 120

6.1 Overview of Chapter 120

6.2 Summary and Conclusion 120

6.3 Recommendations and Future Directions 121

6.4 Potential Future Work 122

REFERENCES 124

APPENDICES

LIST OF PUBLICATION

LIST OF TABLES

Page

Table 3.1: General Data distribution with *k*-independent variables.....44

Table 3.2: General distribution of data for equation 3.5.....50

Table 3.3: Effectiveness of treatment among the different types of treatment**.**.....64

Table 3.4: Transformed data for regression analysis. 64

Table 3.5: dmft score along with school-going children's age and brushing habits..69

Table 3.6: Transformed data for the regression modeling. 70

Table 3.7: BMI score with the waist circumference and blood pressure 75

Table 3.8: Transformed Data for Regression Modelling. 76

Table 4.1: Transformed data for regression analysis. 85

Table 4.2: Estimated parameters from multiple linear regression 87

Table 4.3: Estimated parameters of fuzzy regression 88

Table 4.4: Comparison of actual and predicted values via MLR and FLR.....91

Table 4.5: Transformed data for regression analysis. 95

Table 4.6: Estimated parameters from multiple linear regression 97

Table 4.7: Estimated Parameters of Fuzzy Regression.....98

Table 4.8: Comparison of actual and predicted values via MLR and FLR.....100

Table 4.9: Transformed data for regression analysis. 103

Table 4.10: Estimated parameters from multiple linear regression 105

Table 4.12: Estimated Parameters of Fuzzy Regression.....106

Table 4.13: Comparison of actual and predicted values via MLR and FLR.....109

LIST OF FIGURES

Page

Figure 1.1: Conceptual Framework of the Study 8

Figure 2.1: Architectural illustration of NN with input, hidden, and output layers ... 27

Figure 3.1: Fuzzy coefficient 56

Figure 3.2: Fuzzy parameters 56

Figure 3.3: Illustration of MLP 59

Figure 3.4: The Conceptual Framework of the Proposed Methodology 81

Figure 4.1: Architectural structure of MLFFNN for case study I 86

Figure 4.2: Architectural structure of MLFFNN for case study II 96

Figure 4.3: Architectural structure of MLFFNN for case study III 104

LIST OF ABBREVIATIONS

ANN	Artificial Neural Networking
ANOVA	Analysis of Variance
BMI	Body Mass Index
BP	Blood Pressure
dmft	Decayed, Missing, Filled Teeth
FLR	Fuzzy Linear Regression
GPA	Grade Point Average
GNU	Gnu's Not Unix
HDL	High-Density Lipoprotein
HRV	Heart Rate Variability
HR	Hear Rate
MLE	Maximum Likelihood Estimator
MLFFNN	Multilayer Feedforward Neural Network
MLR	Multiple Linear Regression
MLP	Multilayer Perceptron
MSE	Mean Square Error
MTT	More Than Twice Daily
NN	Neural Network
PLRLS	Possibilistic Linear Regression with Least Squares
QPV	Qualitative Predictive Variables
RMSE	Root Mean Square Error

LIST OF APPENDICES

Appendix A Ethical approval letter granted by Human Research Ethics Committee of the Universiti Sains Malaysia

PEMBANGUNAN SUATU MODEL HIBRID BAHARU SERTA APLIKASINYA DALAM BIOSTATISTIK

ABSTRAK

Regresi linear, merupakan suatu alatan asas dalam analisis statistik yang membolehkan penyiasatan hubungan di antara pembolehubah. Walaupun digunakan secara meluas, analisis regresi tradisional menghadapi cabaran apabila dikendalikan dengan Pembolehubah Peramal Kualitatif (QPV), Rangkaian Neural Hadapan Suapan Berbilang Lapisan (MLFFNN), dan Regresi Linear Kabur. Terdapat suatu jurang yang signifikan dalam pemahaman bagaimana untuk mengintegrasikan regresi linear berganda dengan pendekatan lain untuk meningkatkan tahap ketepatan dan kebolehramalan model. Hal ini menekankan kepada keperluan untuk pembangunan model hibrid. Penggabungan regresi linear berganda (MLR) dengan teknik canggih seperti regresi kabur dan rangkaian neural, dapat mengatasi kelemahan MLR dalam menangani data yang kompleks dan meningkatkan ketepatan serta generalisasi model. Pendekatan hibrid ini penting untuk mengatasi cabaran dalam biostatistik dan meningkatkan prestasi peramalan. Kajian ini menggunakan metodologi menyeluruh yang mengintegrasikan beberapa teknik, seperti penukaran QPV, butstrap, MLFFNN, dan regresi kabur. Kegunaan metodologi yang dibangunkan ini ditunjukkan dengan menggunakan tiga set data sekunder. Kesemua hasil yang diperolehi menunjukkan statistik yang signifikan, dengan ketepatan tinggi yang diperolehi menerusi nilai R^2. Selain itu, nilai ralat min kuasa yang kecil mengesahkan suatu hubungan rapat di antara nilai ramalan dengan nilai sebenar. Semua kes menunjukkan keunggulan kaedah ini, memberikan penyelidik alat yang tepat untuk membuat inferens dan ramalan biostatistik. Untuk penyelidikan masa depan, pendekatan yang dicadangkan ini akan disesuaikan untuk aplikasi dalam pelbagai bidang.

DEVELOPMENT OF A NEW HYBRID MODEL AND ITS APPLICATION IN BIOSTATISTICS

ABSTRACT

Linear regression, a fundamental tool in statistical analysis, enables the exploration of relationships between variables. Despite its widespread use, traditional regression analysis encounters challenges when handling qualitative predictive variables (QPV), Multilayer Layer Feedforward Neural Network (MLFFNN), and Fuzzy Linear Regression. There is a significant gap in understanding how to integrate multiple linear regression with other approaches to enhance model accuracy and predictability. This highlights the need for the development of hybrid models. Integrating Multiple Linear Regression (MLR) with advanced techniques, such as fuzzy regression and neural networks, addresses MLR's limitations in handling complex data and improves model accuracy and generalizability. This hybrid approach is crucial for overcoming challenges in biostatistics and enhancing predictive performance. This study utilizes a comprehensive methodology that integrates several techniques, such as transforming QPV, bootstrapping, MLFFNN, and employing fuzzy regression. The utility of the developed methodology is demonstrated using three secondary datasets. All obtained results demonstrate statistical significance, with high accuracy reflected in the R^2 values. Additionally, small mean squared errors confirm a close alignment between predicted and actual values. All cases show the method's superiority, offering researchers precise tools for biostatistical inferences and forecasts. Future work will adapt this approach for other regression types and explore its application across various domains.

CHAPTER 1

INTRODUCTION

1.1 Preview of the Chapter

This is the introductory chapter of the study, providing a concise overview of the entire study. This chapter provides an overview of the study's background and introduces several statistical techniques, including regression analysis, qualitative predictors, bootstrapping, fuzzy linear regression, neural network, and their applications. Following this, the chapter explores the identification of the research problem and the rationale behind its investigation, highlighting gaps in existing literature that necessitate further exploration. The study's conceptual framework is a graphic that shows the complete concept of the research. Subsequently, the chapter outlines the research objectives, both general and specific. The next section is the scope of the study, followed by the study's significance and its contribution to the literature. Furthermore, the chapter discusses the significance of the study and its potential contributions to the existing literature, while also acknowledging its limitations. Finally, an outline of the entire thesis is provided.

1.2 Background of the Study

The hybrid methodology combines two or more approaches to enhance the accuracy and predictability of the models. Hybrid methodologies promote better results than standalone methods. This research aims to develop a hybrid method that integrates linear regression qualitative predictors, bootstrapping technique, fuzzy linear regression, and neural networking. The succeeding sections of the study describe the different approaches combined in this research.

Regression analysis is a statistical tool used to estimate relationships between dependent and independent variables. It may be used to determine the strength of a relationship between variables and to model the future relationship between them (Yip *et al*., 2017). Linear Regression (LR) is the most common type of regression analysis, which estimates the strength of one dependent variable and an independent variable. The independent variables can be nominal (qualitative), ordinal (Likert scale), or scale (height, weight, age, BMI [Body Mass Index], and the dependent variable, which is also known as the regressed or response variable, should be scaled or continuous (quantitative) in nature. Regression Analysis is a technique almost applicable in all disciplines/fields, such as health sciences, physical and chemical sciences, financials, marketing, medical science, engineering, and social sciences (Klees, 2016). Multiple linear regression (MLR) is an extended form of simple linear regression (SLR) involving multiple explanatory variables. The primary purpose of a regression equation is twofold: first, it serves to predict the value of the dependent variable when values for the independent variables are provided; second, it quantifies the relationship between the independent and dependent variables. Researchers from the medical field typically use regression analysis effectively for making a diagnosis or assessing prognosis in medical sciences (Sox *et al*., 2014).

In certain scenarios, the dependent variable may not only be influenced by quantitative variables but also by qualitative variables. These qualitative variables, also known as categorical variables (like gender, intelligence quotient (IQ) level, and marital status), and these variables cannot be used directly in the regression analysis. In some situations, qualitative predictor variables seem essential, and researchers are interested in including those variables in the model. However, the procedure of incorporating these variables into the regression model is not straightforward, so it is

not common among researchers to include qualitative predictive variables in the regression analysis (Loewen *et al*., 2014). However, these variables can be defined within the scope of regression analysis by introducing the method of dummy variables (Schroeder *et al*., 2016). Dummy variables typically take on two possible values, commonly 0 and 1, to indicate the absence or presence of a particular characteristic or attribute. The numerical values of indicator variables are not meant to represent a quantitative ordering of the categories but rather to denote category or class membership. The advantage of dummification is its direct interpretability in regression analysis. In general, if there is a qualitative predictor with *k* categories, then the number of dummy variables required to encode *k*-categories of the Qualitative Predictor Variable (QPV) is *k*-1.

Like any other quantitative variable, the dummy variable can be employed in regression analysis, although its meaning differs from that of the quantitative variable (Das, 2019). Furthermore, these transformations of categorical variables sometimes cause fuzziness in the data, or sometimes the variable itself has a fuzzy nature rather than being crisp. Such fuzziness can lead to parameter uncertainty or vagueness (Hernandez, 2021). If the variables are ambiguous, the risk of losing essential information increases. Thus, a promising technique has been developed known as fuzzy linear regression (FLR), which is superior to linear regression in such situations (Tanak *et al*., 1984). FLR extends linear regression by incorporating fuzzy logic principles, allowing for the modeling of uncertainty and imprecision in the data (Mendel, 2024). By representing variables and relationships in terms of fuzzy sets, FLR can accommodate ambiguous data more flexibly than traditional linear regression methods. This makes FLR particularly well-suited for scenarios where the variables

are not clearly defined or where there is uncertainty in the data, ultimately enhancing the model's accuracy and robustness.

The regular approach to statistical interference usually depends on the perfect model and ideal assumptions. Often these assumptions do not fulfil a small sample size, like standard errors are based on asymptomatic theory. A very promising computerize alternative method was introduced by Efron in 1979, called the Bootstrapping approach (LaFontaine, 2021). Bootstrapping is a statistical procedure that involves creating a sample distribution for statistics by repeatedly resampling the original sample. The advantage of using bootstrapping is that it does not create a new sample; it uses the original data values to substitute the population and draws simulated samples within the sample. The use of bootstrapping can improve the efficiency and precision of statistical inference without the need for additional data collection.

For the predictions to be accurate from the derived model, they must be carefully validated for their reliability. However, a multilayer layer feedforward neural network (MLFFNN) can be used as a further process of validation of the regression models and to improve its predictability and accuracy. MLFFNN is used to improve the model’s precision and move predicted values closer to reality. MLFFNN also helps to pick the best regression model among the different models derived from the same datasets. After training, the MLFFNN computes the MSE for each regression model. The model with the smallest MSE is considered the best, as it has the least amount of error in its predictions. This approach allows the MLFFNN to effectively compare and rank the regression models based on their predictive performance. Although neural networks are complex, they are also flexible at the same time and can be used for classification and regression validation (Zhang *et al*., 2018). Generally, the neural network is used to develop an efficient model. It can also be defined as reading the

input data, producing the predictive model, measuring the error (Mean Square Error, MSE.net) in the model, and repeatedly implementing necessary corrections to the model until the model with the least error is found.

Hence, the primary goal of the research is to develop a hybrid methodology that enables researchers to achieve more precise predictive models. This technique will utilize bootstrapping to improve the models' predictability, linear regression with qualitative predictors, fuzzy linear regression, and construction of a multilayer feedforward neural network for validation purposes. This hybrid approach is a more robust estimation of model parameters and an improvement in short-term prediction accuracy. It is expected that the proposed hybrid approach will benefit the researchers for more accurate prediction and validation.

1.3 Problem Statement

Regression analysis is widely used in the medical sector since it reveals a functional association between two or more related variables. It allows the identification and characterization of relationships between various variables (Tolles *et al*., 2016). Sometimes researchers may misuse qualitative predictive variables (QPV) in regression analysis, leading to difficulties in selecting the appropriate method for handling them in regression modeling. This can result in inaccuracies and challenges in effectively incorporating qualitative predictors into the regression framework. These qualitative variables cannot be incorporated into the regression model directly without further transformation, which is the limitation of these predictors. One of the significant challenges is the improper encoding of qualitative variables (such as categorical data) into a format suitable for regression analysis. Standard techniques like one-hot encoding or label encoding might not always capture

the true relationship between the qualitative variables and the dependent variable. This can lead to loss of information, multicollinearity, and biased model estimates. To help the researcher transform qualitative predictors into linear regression is one of the objectives of this research.

Another drawback of linear regression is that the underlying relationship is assumed to be precise, as it gives a precise value of response for a set of values of explanatory variables (quantitative and qualitative) (Van-Kuijk et al., 2019). The idea that a linear relationship may completely capture the relationship between variables is typically overly restrictive. This assumption can lead to overfitting, in which the model becomes excessively complicated by fitting to uncertainty in the training data, capturing patterns that do not generalize to new data. Conversely, it can lead to underfitting, in which the model is too straightforward and fails to capture the underlying patterns and relationships, resulting in poor prediction performance. These issues arise because linear regression models are limited by their linearity, which may fail to accurately represent the underlying, sometimes complex, nature of data connections. Subsequently, another issue is how to deal with data ambiguity and what could be the best strategy to counter it.

On the other hand, there is a lack of understanding about integrating MLR with other approaches to improve model accuracy and predictability. By exploring alternatives to traditional linear regression or augmenting it with other methods, hybrid model will contribute to the development of models that can better handle complex relationships and ambiguous data. This includes incorporating techniques like fuzzy regression, or machine learning approaches that offer greater flexibility and robustness. As a result, hybrid models are in demand to generate more advantageous outcomes in accordance with the needs of present and future researchers. The creation

of a hybrid model that combines the strengths of linear regression with other advanced techniques (e.g., bootstrapping, FLR, neural networks) will be a key contribution. This model will aim to improve predictive accuracy, handle both qualitative and quantitative variables more effectively, and offer better generalizability across different datasets. In biostatistics, Multiple Linear Regression (MLR) encounters many challenges. Multicollinearity between predictors can destabilize the model and make interpretation difficult. Nonlinear interactions are common in biological data (for example, when categorical characteristics are added). MLR struggles with qualitative predictors, especially when there are many categories. MLR is based on various assumptions (e.g., linearity, independence, normalcy of errors), which are frequently violated in real-world biostatistical data.

1.4 The Rationale of the Study

In some situations, standalone models do not provide accurate or precise results as expected. Furthermore, if the dataset obtained is not crisp or the underlying relationship is not as exact as predicted, regression performed alone may produce inaccurate findings. Furthermore, there is no integrated process from the qualitative predictor's transformation to the final model development, testing, and validation.

A hybrid model combines many approaches to achieve more precise results. This proposed hybrid modeling and prediction method offers several advantages over standalone methods. The hybrid technique enables a more robust estimate of model parameters in model fitting, accuracy, and necessary parameter estimation. Hence this is the remedy that would address the issue.

1.5 Conceptual Framework of the Study

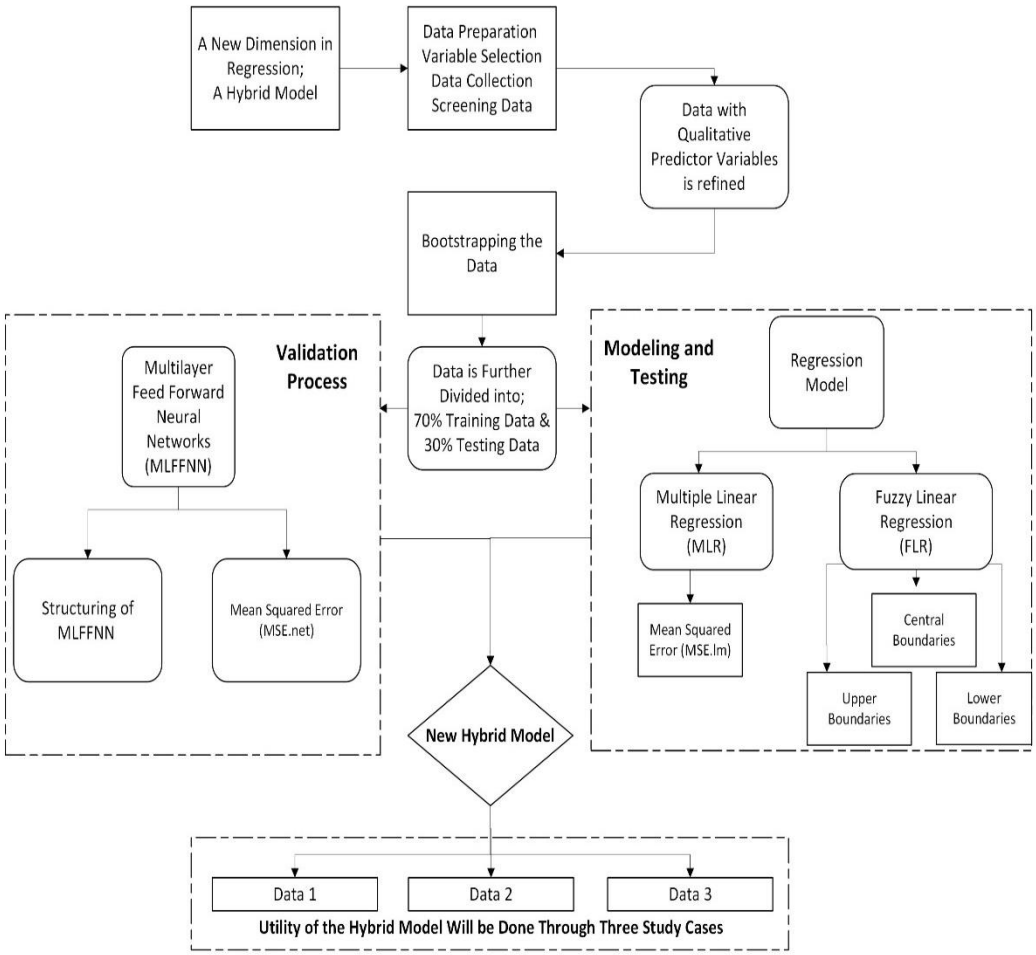

Figure 1.1: Conceptual Framework of the Study

1.6 Research Objectives

1.6.1 General Objectives

To develop a methodology with a *k*-category of qualitative predictor variables using a regression approach, validate using a multilayer feedforward neural network (MLFFNN), and apply it to biostatistics.

1.6.2 Specific Objectives

a) To develop a template for transforming *k*-category qualitative predictors in a multiple linear regression model.

b) To develop and validate a hybrid bootstrap methodology, multiple linear regression with *k*-categories of qualitative predictors, multilayer feedforward neural network, and a fuzzy linear regression model.

c) To make statistical inference of the obtained results from the hybrid bootstrap methodology.

1.7 Scope of the Study

Nowadays, hybrid methods have become more popular as a method for prediction models. In regression analysis, it is assumed that the cause-and-effect relationship between the variables remains unchanged which sometimes misleads the results (Ghassami *et al.*, 2017). Qualitative predictors influence the intercept of the regression line. The focus of the current study is to develop a hybrid methodology (computational statistics), and this approach will begin with the appropriate method of adding qualitative predictors in multiple linear regression, this transformation of variables is known as dummification. This dummification has the benefit of being interpretable in regression models directly. Further steps involve applying bootstrap

techniques to enhance the accuracy of estimators and fit the fuzzy linear regression. The final step is constructing a neural network that evaluates the model through the input and output variables method (MLFFNN).

As the preceding section states, the study aims to develop, test, and validate a hybrid method. Therefore, the model will be tested through three different secondary datasets belonging to the biostatistics field. Extracted datasets based on dental caries measurement among school children, the effect of different types of treatment on depression, and prediction of BMI through hypersensitivity.

1.8 Contribution and Significance of the Study

There are many methods to calculate estimators and develop models to predict specific variables (quantitative and qualitative). A critical situation is how to deal with qualitative predictors while fitting a regression model, which is currently not well discussed. Researchers and academicians are eager for a technique that would allow them to employ qualitative predictors in linear regression. Existing available methodologies are abandoned to provide a comprehensive framework for modeling with validation for datasets from *k*-category qualitative predictors. Thus, this study aims to develop and validate methodologies for transforming qualitative predictors in a way that preserves their fundamental properties and relationships with other variables. This will help in better capturing the nuances of categorical data, leading to more accurate and interpretable models. Integrating Multiple Linear Regression (MLR) with advanced approaches like bootstrapping, FLR, MLFFNN improves the model's precision and addresses its shortcomings. This integration allows the model to handle uncertainty more effectively while also improving forecast robustness.

The novelty of the proposed hybrid model lies not merely in the integration of existing techniques, but in the strategic and innovative combination of these methods to address specific challenges in regression analysis that have not been fully explored. By systematically transforming qualitative predictive variables, handling data ambiguity, and enhancing prediction accuracy through a customized hybrid approach, this research goes beyond the application of individual methods. It contributes a consistent and robust framework tailored to the complex needs of modern medical data analysis, offering a novel solution that improves both the reliability and interpretability of regression models in ways that isolated techniques cannot achieve on their own.

1.9 Limitations of the Study

The study's objective is to propose a hybrid methodology that can be used for precise predictability for both types of variables, quantitative and qualitative, and testing and validation. Hence, one of the limitations of the hybrid model is that it will be limited to few statistical methods: bootstrapping, MLR with qualitative predictors, fitting of fuzzy linear regression, and construction of MLFFNN. As described earlier, the novel approach will be evaluated for accuracy. Therefore, the datasets used to test the methodology will be secondary. These datasets will be extracted from previously published articles and belong to one single domain, health science. In other words, the dataset will not belong to various fields, which is another limitation of the study (secondary datasets). Another limitation of this study is that the datasets included have only three variables. Hence, datasets with a single dependent variable and two independent variables (quantitative and *k*-category qualitative) will be considered for methodology testing. Finally, the limitation is the concerns of the usage of the

software. All methods performed, tested, or validated in this research will be based only on the R-Software.

1.10 Thesis Organization

This research is categorized into the following chapters. Chapter 1 is the introduction, background of the study, significance, rationale for the study, and scope of the study. This chapter will also cover the contribution and limitations of the study and present the study's conceptual framework. Chapter 2 comprises a detailed literature review of the application of qualitative predictors, the importance of linear regression in biostatistics, the application of fuzzy regression, the need for combining different statistical methods, and the history of these methods. Chapter 3 is the methodology that introduces the methods used, study design, time of study, the data, methodology building, software, and programming details. A flowchart of the findings is followed by case-by-case methodology development. Chapter 4 is based on the results derived from the methodology built in the previous chapter. This chapter will discuss the results of each case in depth and will summarize them at the end. Chapter 5 is the discussion, where the objectives of the study will be discussed one by one with the conclusion (goodness of methodology and study design) drawn from them. Chapter 6 is the final chapter of the current research, summarizing the work done. It includes recommendation, future direction, and potential future work.

CHAPTER 2

LITERATURE REVIEW

2.1 Overview of the Chapter

This is the second chapter of the study, which provides an overview of the study's literature review. This chapter addresses the historical methods of statistical analysis in the field of biological science or biometry, followed by a discussion and some justification of its research benefits. The history and application of QPV in regression will also be presented. Fuzzy linear regression's origins and the development of MLFFNN's architecture will also be covered in this second chapter. The chapter will also cover the applications of linear regression with QPV, FLR, and MLFFNN in the field of biostatistics. A hybrid model consisting of primary statistical methods for getting better results will be the conclusive part of the chapter. Furthermore, a review of the proposed hybrid methodology and its limitations will be presented in the concluding remarks.

2.2 Introduction to Multiple Linear Regression with Qualitative Predictors

MLR is a statistical approach that is used in place of ordinary linear regression in situations in which there are a large number of independent variables to take into consideration. Through multiple regression analysis, researchers may simultaneously examine the impact of several independent factors on a single dependent variable. Sometimes, in multiple linear regression, qualitative factors in addition to scale variables impact the result (Das, 2019). The way of using those qualitative variables in regression analysis is to "quantify" them by considering artificial variables having

values of 0 or 1. The process through which qualitative factors are converted into quantitative ones is known as dummy variable approach (Das, 2019).

2.2.1 The History of Linear Regression

The least squares approach is claimed to be the first kind of "regression", developed by Legendre (1805) and subsequently Gauss (Gauss, 1809; Legendre, 1805). Later, in the 19th century, Charles Darwin's cousin, Francis Galton, invented the term "regression" to describe biological phenomena in which "the height of tall progenitors" regresses down towards a normal average (Krashniak & Lame, 2021). He termed this inheritance as "regression towards the mean" and found a method of predicting the values of one variable using another variable (quantitative). A few years later, three statisticians, Francis Edgeworth, Karl Pearson, and George Yule, developed precise mathematical formulae for regression analysis. Then they continued to expand it to a broader statistical context. Since then, regression research has continued in a variety of domains, including physics, medicine, finance, and health care.

Regression is a statistical method that allows us to examine the relationship between one dependent variable (also called regressor or response variable) and one or more independent variables (also known as an explanatory variable). The regression analysis has various kinds depending on the number of variables and the nature of the variables. Simple linear regression is used to build a model function stating the linear relationship between two variables (response and predictor) and can be defined by the following formula:

$$y = \beta_0 + \beta_1 x + \varepsilon \qquad (2.1)$$

where,

y = is the dependent variable

x = is the independent variable

β_0 = is the intercept

β_1= is the slope of x and is also called the regression coefficient

ε is the error term, which is also called statistical error.

If the mean value ε is 0 and the variance is σ^2 then the expected response at any value of the predictor variable will be:

$$E(y/x) = \mu_{x/y} = E(\beta_0 + \beta_1 x + \varepsilon) = \beta_0 + \beta_1 x$$

and variance will be:

$$V(y/x) = Var(a + bx + \varepsilon) = \sigma^2$$

Equation (2.1) presents the regression line for two continuous variables (x, y) then the height of the regression line at any value of "*x"* is the expected value of "*y"* and "β_1" can be interpreted as the change in the mean of *Y* distribution produced by a unit change in "*x"*. Slope "β_0" will be the mean of the distribution of "*y"* when "x" is 0 and if "x" is not 0 then β_0 will not have any interpretation with "y". A casual change in the predictive variable can be explained by the explanatory variable.

2.2.2 Qualitative Predictors in Multiple Linear Regression (MLR)

In the early 1900s, qualitative and categorical data analysis were initially applied, and Karl Pearson and George Udny were the pioneers in this field (Agresti, 2012). Qualitative variables are the type of variables that are not numerical and explain the data that fits into the categories. Qualitative variables are used to classify data that

share the same attribute, which might be nominal (e.g., gender, IQ level, hair colour) or ordinal (e.g.: age groups, BMI categories) (Agresti, 2012). Similar to quantitative variables, there are several statistical methods to analyze qualitative variables (Bernard *et al*., 2016).

Qualitative variables like any other variables can be very useful predictors in multiple linear regression analysis. Because of their categorical nature, they cannot be incorporated into the regression directly and need some transformation (Panchan, 2019). This transformation of variables into a numerical representation is called dummification. Dummy variables usually take only two values "0" and "1" to transform categorical variables into quantitative variables. If qualitative variables have α categories, α-1 dummy variables will be needed to introduce into the model. For example, there are 2 independent variables (x_1, x_2) one of them is continuous but x_2 has 3 sub-categories (low, moderate, and high). Therefore, they will require 2 (α-1) additional dummy variables.

Given below is the multiple regression equation with two independent and one dependent variable.

$$y_i = \beta_o + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$$

where x_{i2} has 3 subcategories so the dummy variable will be x_2 (moderate), x_3 (high) added to the above equation and let us define dummy variables as:

$$x_2 = \begin{cases} 1 & \text{if moderate caries} \\ 0 & \text{Otherwise} \end{cases} \qquad x_3 = \begin{cases} 1 & \text{if high caries} \\ 0 & \text{Otherwise} \end{cases}$$

Hence, the model equation will be as:

$$y = \beta_o + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon_i$$

To understand the above equation and co-efficient, let's consider the above equation for low caries when (x_2 and $x_3 = 0$)

$$y = \beta_o + \beta_1 x_1 + \beta_2(0) + \beta_3(0)$$

$$y = \beta_o + \beta_1 x_1 \qquad \text{For Low Caries status}$$

For moderate caries, consider $x_2 = 1$, and $x_3 = 0$ in the above equation. Hence the model will be as follows:

$$y = \beta_o + \beta_1 x_1 + \beta_2(1) + \beta_3(0)$$

$$y = (\beta_o + \beta_2) + \beta_1 x_1 \qquad \text{For Moderate Caries status}$$

For high caries, consider $x_2 = 0$, and $x_3 = 1$in the above equation. Hence the model for high caries will be as follows:

$$y = \beta_o + \beta_1 x_1 + \beta_2(0) + \beta_3(1)$$

$$y = (\beta_o + \beta_3) + \beta_1 x_1 \qquad \text{For High Caries status}$$

The category that is not included in the model is devoted to the base category because of the interpretation of regression coefficients based on the based category. The numerical values assigned to the categorical values are not meant to be in quantitative order of category but only used to identify the category.

2.2.3 Advantages of Qualitative Predictors in MLR

In multivariate linear regression analysis, the dependent variable is often impacted not only by elements that can be easily quantified on a well-defined scale but also by qualitative variables. In some situations, the quantitative variables are not enough to describe the dependent variable. Therefore, researchers should incorporate qualitative predictors or categorical variables into multiple linear regression along with

quantitative variables (Venkataramana *et al*., 2016). Qualitative variables may be defined as gender, socioeconomic background, ethnicity, geographic region, marital status, BMI categories, blood groups, or cholesterol levels. Especially in biological sciences, demographic characteristics play a vital role in investigating factors associated with outcome variables. Adding qualitative predictors in the data enriches the quantitative results. The use of both qualitative and quantitative data can improve an evaluation by ensuring that the limits of one type of data are balanced by the strengths of the other (Brannen, 2017). Numerous researchers have integrated qualitative predictors as an essential variable within prediction models (Ahmad *et al*., 2016; Nawai *et al*., 2018).

2.2.4 Practice Toward QPV and Regression Approach

There is no difficulty in training a linear regression model if all the predictor variables are numerical. The problem arises when one or more predictor variables (independent variables) are categorical. Qualitative variables can be brought within the regression scope using a method known as dummy variables (Cottrell, 2015). A common approach of the dummy variable is previously described in earlier sections of the research to avoid the difficulty. Prediction models with qualitative predictors are very commonly used in social and biological sciences.

Recently, Ahmad *et al*. (2022), used a mixed dataset, waist circumference (numeric), high-density lipoprotein (HDL) (numeric), and hypertension (qualitative) with 3 categories (normal, borderline, and high blood pressure) to predict the model which provides a relationship between these measures and triglycerides. The authors used a dummy variable method to incorporate qualitative predictors into MLR and developed a significant model that shows the significant effect of HDL, waist, and

hypertension status on triglycerides (Ahmad *et al*., 2022). Similarly, in another study mixed dataset was used in which diabetes mellitus (dependent) was predicted using BMI, height, weight, blood pressure (quantitative variable), and cholesterol level (qualitative variable) in the MLR model (Nawi *et al*., 2018).

2.2.5 Role of Qualitative Predictor Variables

Medical research often requires investigating the relationship between the response variable and one or more predictor variables. Changing a continuous predictive variable into categorical variables by characterizing it makes researchers' goal easier to analyze the relationship between outcome variables and explanatory variables for diagnostics or treatment decisions (Bolotin, 2007). Qualitative predictors play a vital role in prediction models, especially in health sciences-related research. Researchers may want to investigate if there is any evidence of a relationship between a qualitative (grouping) predictor (such as treatment group or patient gender) and a quantitative result (e.g., blood pressure, BMI). Qualitative predictors made linear regression more flexible (Schober & Vetter, 2021).

2.2.6 Linear Model to the Health Sciences Research

Regression analysis is frequently used to illustrate associations between variables that are thought to be biologically connected (Shapiro, 2014). There are numerous types of regression analysis along with qualitative predictors playing their role in prediction and associations in almost every field of science. Regression models have become standard tools in medical research (Vech, 2012). Several studies have used multiple linear regression models to predict illness prevalence and health-related physical fitness (Shah *et al*., 2020).

Several regression analysis techniques are applicable in health sciences, depending on the application and nature of the variables (Room, 2021). Simple linear regression, multiple linear regression, logistic regression, and exponential regression are commonly used for forecasting in health science research (Ahmad *et al*., 2014; Zabor *et al*., 2022; Ahmad *et al*., 2018). According to Yergens *et al*., (2014) in the overview of statistical methods reported in medical research, regression was mentioned 80 times (79.8%) more than any statistical test and the most popular type mentioned was logistics regression (67%) (Yergens *et al*., 2014).

2.2.7 The Accessibility of Linear Regression

Linear regression analysis is used to predict the value of one variable based on the value of another variable employing a simple mathematical function. The linear regression equation measures the two variables' straight-line relationship. The linear regression model includes assumptions, parameter estimation methods, and usage. The most common assumptions of LR required to build a precise linear model are:

i) The relationship between the variables should be linear.

ii) Observations are independent of each other.

iii) The error terms of the model are normally distributed.

Regression models' accuracy also depends on the size of the samples. Thus, if the pre-conditions are satisfied probability of getting accurate models is high (Maturo, 2016). Linear regression is a very adaptable approach that may be used to answer a wide range of research questions and study objectives. Researchers may want to investigate the proof of an association between a continuous dependent variable (like age, income, heart rate) and one or more qualitative independent variables (like BMI levels, blood pressure level, cholesterol levels, gender, treatment types, or disease

stages) (Schober & Vetter, 2021). These qualitative predictors go over a transformation process called dummification to be incorporated into linear regression. Linear regression not only provides a test of association but also quantifies the strength and direction of the association.

2.3 The Fuzzy Regression

The term fuzzy logic was coined by scientist Lotfi Zadeh in a proposal of fuzzy set theory in 1960 (Freitas *et al.*, 2017). Later, as an extension of the fuzzy set, he (Zadeh) introduced possibility theory in 1978, in which he described that possibility theory is a mathematical way of dealing with uncertain types of data (Zadeh, 1978). Therefore, to investigate the relationship between these variables a very promising technique of fuzzy regression has been developed (Hesamian & Akbari, 2020). In contrast to conventional linear regression, in which the parameters are considered to be random variables with probability distribution functions, the coefficients in fuzzy regression are subject to the possibility theory (Pérez *et al*., 2015).

Fuzzy regression is used in evaluating the functional relationship between the dependent and independent variables in a fuzzy environment. Tanaka *et al.,* (1982), developed a fuzzy regression model with fuzzy responses, fuzzy parameters, and non-crisp response data (Tanaka *et al.,* 1982). Their approach, later, was handled by many authors. Since possibility theory was introduced by Zadeh, it is observed that possibility distribution can represent a certain type of impression. In general, fuzzy datasets are not statistical but possibilistic. The possibility model is a novel interpretation of fuzzy equations that deals with linear regression analysis using probabilistic linear systems (Chen & Nien, 2020). In the literature, several types of fuzzy regression models are introduced, and various methods for estimating the

models' fuzzy parameters are provided. In general, there are two approaches in the analysis of fuzzy regression models: the possibilistic approach and the fuzzy least squares model (Pérez *et al*., 2015).

2.3.1 Introduction of Fuzzy Sets and Membership Functions

The collection of objects or numbers is usually called classical sets or crisp sets, in which the boundaries are defined in exact locations. Whereas, if there exists uncertainty about the location of boundaries or ambiguousness is present in the sets, these sets are defined as fuzzy sets (Afful-Dadzie *et al*., 2017). Fuzzy sets were first introduced by La Zadeh in 1965 as an extension of the classical notion of sets. In other words, fuzzy set theory is an extension of classical set theory where a number has a degree of membership. A membership function provides a measure of the degree of similarity of an element to a fuzzy set (Bandemer & Näther, 2012). The membership function for a fuzzy set A on the universe of discourse X is defined as μ_A: X which means the member function maps every element of the universe of discourse X to the interval of 0,1. It can be written as $\mu_A : X \rightarrow [0,1]$. The value 0 means that X is not a member of the fuzzy set whereas 1 means that X is a full member of the fuzzy set. These values characterize the fuzzy members. There are different shapes of membership function, Triangular function, Trapezoidal function, Gaussian function, etc. (Mikkili & Panda, 2013). Fuzziness in the data is best described by the membership function. The fuzzy membership function is used to transform the fuzzy inference system's crisp input.

2.3.2 Fuzzy Logic Models

Fuzzy logic was derived from fuzzy set theory in 1965 by Iranian mathematician La Zadeh. Classical logic simply allows either true or false conclusions

whereas, true and false sometimes are not enough when describing human reasoning. Fuzzy logic took the whole interval of 0 means false and 1 means true to describe human reasoning (Dewidar *et al*., 2019). Most natural languages are fuzzy because they involve imprecise and vague terms (Hooda & Raich, 2017). Fuzzy propositions are linguistic statements that reflect subjective thoughts and can be interpreted slightly differently by different people (Hooda & Raich, 2017). The fuzzy logic model is a logical mathematical approach that focuses on the IF-THEN rule structure that allows the human cognitive process to be mathematically replicated (Dewidar *et al*., 2019). Fuzzy logic has also shown to be an effective decision-making tool for expert systems and pattern categorization systems. Some medical expert systems have previously incorporated fuzzy set theory. Fuzzy logic has been proven a useful prediction method for uncertain and vague information. Fuzzy logic is especially appealing because of its ability to solve issues in the absence of precise mathematical models (Sivarao, 2009). This idea has shown to be an effective method for dealing with linguistically stated objectives. The linguistic terms like extremely low, very low, very high, medium, and extremely high are considered fuzzy sets. Kovac et al. (2013) used a fuzzy logic model and a traditional regression model to predict the surface roughness in the machine. The cutting speed was measured as highest, high, low, and lowest as independent variables, and study results revealed that the fuzzy logic model can predict the surface roughness in the machine more effectively than the traditional regression model (Kovac *et al*., 2013).

2.3.3 Classification of Fuzzy Sets

If the set of observations has some heterogeneity or the set consists of several homogeneous subsets (categories), the linear regression may not be applied because the heterogeneity of the set is a violation of one of LR's key assumptions. There is an

alternative method to solve the problem; first, a fuzzy classification of the set (observations) is determined using the data matrix. Then FLR is tested for each fuzzy class of objects (Lowen, 2012). Researchers have used crisp datasets to fit fuzzy regression models and have obtained better results than ordinary regression.

Fuzzy sets find frequent application within medical sciences, exemplified by scenarios where the treatment of elderly patients experiencing substantial back pain necessitates the prolonged application of acupuncture to specific anatomical positions. In this medical context, descriptors such as “severe,” “elderly,” and “certain points” exhibit inherent vagueness and fuzziness, reflecting the imprecision often encountered in clinical terminology. FLR is the most suitable statistical method for assessing relationships between definite output and uncertain sets (fuzzy sets) of explanatory variables (Gürsel, 2016). The fuzzy set offers several qualities that make it suited for formalizing the uncertain information that is typically used in medical diagnosis and therapy (Iakovidis, 2010).

2.3.4 The Accessibility of Fuzzy Regression

Sometimes, researchers encounter non-precise data, small sample sizes, or fuzziness in data. In these situations, using linear regression for parameter estimation will not provide precise results. Hence fuzzy linear regression is superior to classical linear regression in these situations. Fuzzy regression is one of the most reliable approaches to handle complex problems including uncertainty and ambiguity in data (Campos, 2020). A fuzzy regression method based on minimizing fuzziness was used for model development for forecasting the agriculture system. It has been found that the average widths for fuzzy linear regression models are much lower compared to linear regression models for all values of fitness criterion (Kumar & Srinivas, *et al*.,