MODELING CROSS SECTIONAL DATA OF THE SALE PRICE OF RESIDENTIAL PROPERTIES IN AMES USING FUZZY REGRESSION ANALYSIS

MOTURI CELESTINA BOSIBORI

A Research Project Submitted in Partial Fulfillment for the degree of Masters of Science in Statistics of Kirinyaga University

NOVEMBER, 2023

DECLARATION

"This research project is my original work and has not been presented elsewhere for a degree award in any other University."

Signature .... Date ....

Moturi Celestina Bosibori- PA200/S/12176/20

"This research project has been submitted for examination with our approval as University supervisors".

Signature .... Date ....

Dr. Victor Muthama Musau, PhD

Lecturer

Kirinyaga University

Signature .... Date...........................

Dr. Muriungi Robert Gitunga, PhD

Lecturer

Meru University of Science and Technology

DEDICATION

This project is dedicated to my lovely parents Mr and Mrs Moturi, my daughter Joy Keilah Kerubo Ondieki and my husband Dennis Ondieki.

ACKNOWLEDGEMENTS

I give thanks to God for giving me this chance to undertake this study, also express my sincere gratitude to my supervisors, Dr. Victor Muthama Musau and Dr. Robert Muriungi Gitunga for the guidance they accorded me.

ABSTRACT

Fuzzy Regression Analysis (FRA), also known as non-statistical regression analysis, is an approach used to establish an ambiguous connection between input and output variables. FRA serves as an alternative method to classical Regression Analysis CRA. The models that are used to model cross sectional data are statistical regression models which are based on linearity, normality and homoscedasticity assumptions. However, these assumptions may not hold true. Thus, fuzzy regression analysis gives a solution to challenges that may arise when using statistical regression models. The main objective of the study was modeling cross sectional data of the sale price of the residential properties in Ames using fuzzy regression analysis and the specific objectives were performing diagnostic tests on fuzzy regression assumptions, fitting the model of fuzzy linear regression and evaluating the fuzzy model of linear regression. Secondary data accessed from Ames assessors office was used. Data visualization indicated price fluctuations of the residential properties which was uncertain. Diagnostic tests of normality, linearity, multicollinearity and homoscedasticity were performed to ascertain the application of fuzzy regression analysis. After verifying the assumptions, fuzzy regression analysis was applicable to model this application. Three fuzzy regression methods: possibilistic linear regression methods with least squares, possibilistic linear regression and fuzzy least absolute residuals were employed to fit the fuzzy linear regression model (FLRM). Fitting FLRM involved conversion of real value observations of the response variable into fuzzy numbers. The fitted models using fuzzy regression methods were assessed based on total fit error and goodness of fit measure. According to the study's findings, method based on fuzzy least squares gave a better compatibility of the fuzzy linear regression model than possibilistic methods. Also, methods based on possibilistic regression indicated the range in which the price of the residential properties can vary to the smallest and largest value using predictor variables present. Therefore, to model cross sectional data of the price of residential properties which may change at a given time, models based on fuzzy least squares methods were preferred compared to possibilistic linear regression methods. The study recommended that diagnostic tests to be performed on any given data set to determine the model to be used when fitting the data, fuzzy linear regression models to be used to fit a given data that assumes classical regression assumptions and fuzzy least squares methods be used in modeling cross sectional data of the sale price of the residential properties for optimal results when making decisions on the range at which the sale price may range.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

CRA Clasical Regression Analysis.

FLAR Fuzzy Least Absolute Residual.

FLR Fuzzy Linear Regression.

FLRM Fuzzy Linear Regression Model.

FLS Fuzzy Least Squares.

FRA Fuzzy Regression Analysis.

GOF Goodness of Fit.

MAE Mean Absolute Error.

MLR Multiple Linear Regression.

MLRM Multiple Linear Regression Model.

MOFLR Multi Objective Fuzzy Linear Regression.

nsTFN non symmetric Triangular Fuzzy Numbers.

OLS Ordinary Least Squares.

PLR Possibilistic Linear regression.

PLRLS Possibilistic Linear Regression with Least Squares.

RMSE Root Mean Square Error.

sTFN Symmetric Triangular Fuzzy Number.

TEF Total Error Fit.

TFN Triangular Fuzzy Number.

VIF Variance Inflation Factor.

CHAPTER ONE

INTRODUCTION

1.1 Background of the study

Regression analysis is a statistical technique that gives a correlation between a response variable and an explanatory variable (Montgomery *et al*., 2021). Its main focus is to determine the effect of an explanatory variable and prediction of an effect. The basic types of regression analysis pointed out by Gogtay *et al*. (2017) include simple and multiple linear regression analyses which are chosen depending on the type of variables one is dealing with. Statistical regression assumptions which include linearity, normality, homoscedasticity and non multicollinearity have to be justified for one to perform the analysis. In addition, the data in use should be precise and sufficient enough to support statistical regression analysis. If the assumptions do not hold true as Arkes (2019) claims, it will be difficult for a researcher to proceed with the analysis. This will result in the change of a data set and discarding certain observations which may cause a loss of information for decision makers.

Chukhrova and Johannssen (2019) pointed out that, despite statistical regression analysis being the most reliable approach in determining the effect of independent variables to the dependent variables, problems may arise. Fuzzy regression analysis was introduced to address the shortcomings of statistical regression analysis. These shortcomings include difficulty in verifying distribution assumptions, insufficient data to support statistical regression analysis and involvement of human judgment when collecting data. Also, the casual relationship between a response and an explanatory variable that was fuzzy in nature was investigated using fuzzy regression analysis.

A study by Khademi *et al*. (2017) modeled concrete's compressive strength with 173 different designs using an artificial neural network, multiple linear regression model and adaptive neuro inference system. The study concluded that multiple linear regression analysis was unreliable because of the non linearity which was observed in the parameters that were used. On the other hand, the other two methods presented more accurate results. Selçuk Öğüt (2006) modeled car ownership in Turkey. According to the study, the interrelationship between the predictor variables proved a challenge for a classical regression analysis to be performed which resulted in the use of fuzzy regression analysis.

Literature revealed that, for one to perform a regression analysis, regression assumptions may hold true. Hence, it is in this context an approach that is used to model cross sectional data was established. In particular, this study aimed to model the cross sectional data for the sale price of residential properties in Ames and the model fit for the data was determined. Determining the model fit for the cross sectional data with uncertainties will put any researcher in the best position to offer some alternatives to the problem.

From the given discussion, it was evident that modeling cross sectional data with uncertainties could result to unreliable results. It was for this reason a fuzzy regression analysis method was used to model cross sectional data for the sale price of residential properties which is uncertain to attain accurate and reliable results.

1.2 Statement of the problem

Regression analysis has been used as a statistical technique to investigate the quantitative correlation between output variable and input variables. This technique is applied if the statistical regression assumptions hold true. Linearity, normality, homoscedasticity and absence of multi colinearity are some of the assumptions that have to be true for classical regression analysis to be applied (Al-Kandari *et al* 2020).

To establish consistency and robustness in the models, Pandelara *et al* (2022) points out that traditional statistical models have a criteria and assumptions about the underlying data such as precision or normality. However, fuzzy regression gives an extension of classical regression that is based on possibility theory rather than probability theory and is employed when available data is restricted or variables interact in a specific way.

Further Chukhrova and Johannssen (2019) pointed out that, despite statistical regression analysis being the most reliable approach in determining the effect of independent variables to the dependent variables, problems may arise. Fuzzy regression analysis was introduced to address the shortcomings of statistical regression analysis. These shortcomings include difficulty in verifying distribution assumptions, insufficient data to support statistical regression analysis and involvement of human judgment when collecting data. Also, the casual relationship between a response and an explanatory variable that was fuzzy in nature was investigated using fuzzy regression analysis.

Due to the use of cross sectional data in this study, uncertainity between the sale price of the residential properties in Ames and the lot area, garage area and total basement square feet was visualized which classical regression analysis has not been addressed from Chukhrova and Johannssen (2019) and Pandelara *et al* (2022) studies. Also classical regression assumption such as normality, linearity and homoscedasticity were not upheld. Therefore, this study modeled cross sectional data of the sale price of residential properties in Ames using fuzzy regression analysis.

1.3 Significance of the study

The study determined a fuzzy regression technique for modeling cross sectional data that is uncertain and where non-normality, non-linearity and heteroscedasticity exist. Fuzzy regression technique is of great importance to any researcher interested in modeling cross sectional data where statistical regression assumptions do not hold true. Furthermore, the results enable for policy formulation based on the data's fuzziness description where stake holders can determine the value of residential property within a given range.

1.4 Research Objectives

1.4.1 General Objective

The general objective of this study was to model cross sectional data of the sale price of residential properties in Ames using fuzzy regression analysis.

1.4.2 Specific Objectives

The specific objectives of the study were to:

(i) Perform diagnostic tests on fuzzy regression assumptions

(ii) Fit the fuzzy linear regression model.

(iii) Evaluate the fuzzy linear regression model.

1.5 Research Questions

(i) Cross sectional data assumes the properties of fuzzy regression.

(ii) Fuzzy linear models can fit cross sectional data.

(iii) Fuzzy linear regression model are efficient in fitting cross sectional data.

1.6 Scope of the study

The study was limited to data acquisition and analysis, derivation, fitting and evaluation of fuzzy linear regression models using fuzzy regression analysis.

CHAPTER TWO

LITERATURE REVIEW

2.1 Introduction

The chapter highlights streams of thoughts that have contributed to fuzzy linear regression. To begin with, linear regresion has been discussed as it forms the basis of fuzzy regression analysis.

2.2 Linear Regression

In real life situations, outcomes are as a result of a reason. Rheumatology study done by Lunt (2015) the outcome is referred to as a response variable denoted by Y and the reason is referred to as explanatory variables denoted by $X_1, \ldots, X_k$. k indicates that there are more variables. During the computation of these models errors do occur which are assumed to follow a normal distribution. The model is referred to as a linear regression model and is represented as $\hat{Y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \cdots + \beta_n x_n$ where β_0 is the intercept, $\beta_1, \ldots, \beta_n$ are unknown regression parameters and $x_1, \ldots, x_n$ are the explanatory variables. Hypothesis testing and confidence intervals are used to make inferences on the type of data being used. Before fitting the linear regression model using any given data, checking the assumptions of regression is useful. Linear regression analysis therefore is considered to be an effective method of modeling various relationships when the validity of assumptions is observed (Schneider *et al.*,2010).

According to Marchionni *et al.* (2014) multiple linear regression approach was used to establish the cost functions, based on the hydraulic parameters and known physical characteristics of assets of sewer systems. Data from a construction group under the administration of the Aguas de portugal was analyzed following a four step procedure where the functions of costs were evaluated based on additive multiple linear regression. Uni-variate and polynomial regression were considered as special cases in the multiple linear regression analysis. The model was developed by the R software. Some variables used in this study indicated a reasonably good correlation and others indicated a weak correlation, an increase in the amount of the available data was found to be an option to improve the observed weak correlation. A comparison of the functions of cost with other studies showed that the generated cost functions were highly robust and reliable compared to the functions utilized in the study. The study recommended the use of more data to improve the cost function in the analysis

2.3 Fuzzy Regression

Fuzzy regression analysis was used to model ownership of cars in Turkey (Selçuk Öğüt, 2006). Ownership of cars was not common in Turkey before 1970, thereafter ownership of cars became common in the country creating a database to form a reliable model. The study used the Pearson correlation coefficient to check the inter correlation between the predictor variables. According to the study, there was an inter correlation between the explanatory variables making classical regression analysis unfit for analyzing the data. The fuzzy regression technique was used to model the data despite the interrelationship observed between the variables. The possibilistic linear regression method was employed to model the data where a three step procedure was used to model these relationships. The efficiency of these models was assessed using the total error of squares, as the variables were increased in the models, the total error of squares diminished. The use of fuzzy linear regression indicated broad ranges in estimation which is not helpful in application. For the ranges between the lower and upper bound to decrease, and observed values to be situated within range, the study recommended increasing the quantity of model variables.

Chen *et al*. (2006) modeled the degree of comfort in AIR CONDITIONED rooms. Comfort was considered to be subjective hence making it difficult for statistical approaches to handle the subjective terms effectively. Data was collected through experiments conducted in three AIR CONDITIONED setups. Five explanatory variables were considered to evaluate the quality of comfort. Possible correlation between the factors was examined at the start and at the end of the experiment using principal component analysis which was also considered in the removal of the existing outliers. Subjective individual feeling was observed to dominate in the air conditioned rooms. The influence of the subjective independent variable led to the use of fuzzy approaches. Two fuzzy regression models were utilized to analyze the thermal comfortability due to space limitations and limited data. The study suggested an extensive further research study using increased data sets and other fuzzy methods to assess the various pros and cons of the method.

The relationship between the targets and ecological parameters of a regulatory system was considered to be uncertain. Kropat *et al*. (2016) introduced the theory of fuzzy target environment systems where related fuzzy regression models were discussed using a possibilistic regression model. The crisp predictor-fuzzy response factors and fuzzy coefficients were used with a representation of symmetric triangular fuzzy numbers. Fuzzy regression analysis centered on symmetric triangular fuzzy coefficients was found to be rigid in representing the variance between datasets. This resulted in the adoption of a regression analysis of target ecological data centered on asymmetric triangular fuzzy number where the possibilistic model was obtained. It was proposed that techniques from fuzzy least squares regression centered on a reduction of the total square error of the output should be considered and solved.

The fuzzy logic approach has been used in real estate valuation to predict the performance of the real estate (Sarip *et al*., 2016). To estimate the property value, the study used fuzzy least squares models and adaptive neuro fuzzy inference system. The performance evaluation of the models was done by testing the effectiveness and performance using the fuzzy approaches by comparing their prediction results using the MAE evaluation criterion. The study established that higher ranking performances of the proposed fuzzy least squares regression based approach is significant in real estate valuation. Further, the study suggested other models to be applied to other datasets involving different areas to check if the results are different from the conclusion made in the study.

Customer satisfaction with a new product design has also been modeled by fuzzy regression analysis. According to Nazari-Shirkouhi and Keramati (2017), the best fuzzy regression model for product design and policy making was determined. Data was collected from a study conducted in a freezer refrigeration industry using four aspects of determining product features. A two stage algorithm determined the required model. The index of confidence, was the first stage followed by error based measures as the outputs for data envelopment were applied. Six models were considered to be efficient in modeling the correlation between the level of customer satisfaction and the four aspects used as input variables. A change in the order of the two-stage algorithm was applied to check if there was any effect on the models. It was observed that the first step gave the required results than the second step. It was suggested that the proposed methodology for product design is better since it took into account multiple computation.

Multiple linear regression analysis and fuzzy logic models have been utilized for the functional service forecasting of the cultural heritage. The fuzzy inference system model under fuzzy logic models was used to determine the classification with regard to built heritage life and functional service. Hence, giving preference to preventive and maintenance protective measures to corresponding categories of building thus maximizing cost incurred in maintenance actions. On the other hand, MLR model has been used to arrange the variables affecting approximations of heritage buildings. The study highlighted 17 susceptibility and risk factors affecting building service life. Based on the samples performance, it was found that this method was complex and required specific software to compute the functionality index. A multiple linear regression analysis described a simple model to forecast the serviceability of the building whereby variables with a maximum contribution on degradation phenomena were identified. One hundred churches with similar constructive attributes and contrasting serviceability levels were examined. From this research, it was reported that decision makers can develop the most suitable approach for the heritage building taking into account efficient maintenance strategies (Prieto *et al.*, 2017).

Further, FLRM and MLRM have been applied to predict mulberry leaf yield. The research investigated the effects of the factors that determine the production of the leaf yield mainly the quantity of leaves for each plant and the percentage moisture content by estimating the parameters of both models. The parameters of the MLRM were computed using the least square method and assumptions of linear regression were taken care of throughout the study. Consequently, the parameters of FLRM were approximated by decreasing the cumulative uncertainty of the model data combination with respect to the limitation that every data point lies in the range of the estimated value of the output variable. To show the relationship between leaf yield and its attributed characteristics, the models were fitted by a statistical software. From their analysis, the average width under MLRM was higher than under FLRM. Therefore, FLRM which had the least width was considered to be the most efficient (Bhavyashree *et al.*, 2017) .

Fuzzy linear regression analysis study by Zhou *et al.* (2018) formulated a relation of house price with some factors to provide a framework for governmental agencies to facilitate decision making processes while controlling the prices of houses in the real estate market in China. The factors influencing house price affordability were considered to support decision making process which included policy factors and non policy factors. A survey conducted by interview questionnaires on two hundred employees showed that 53 employees provided responses that were contrary to the actual situation. A statistical analysis was done on the remaining 147 questionnaires where two observations detected as outliers were rejected. The remaining 145 observations were used to formulate FLRM. House price affordability was analyzed using classical distance measures of the fuzzy linear regression model (Diamond, 1988).

Modeling of FLRM and comparison of derived coefficients was done using non-symmetric and symmetric triangular fuzzy numbers to obtain a fitting degree of the approximated and real values of house prices affordability. The fitting difference of the approximated ranged between 0 and 0.8 for each set of data indicating that this model had a good fitting object since the degree of fit ranged from 0 to $+\infty$. The fitting difference of FLRM using non-symmetric coefficients was minimal compared to that with symmetric fuzzy coefficients. Therefore, the method used to construct FLRM using non-symmetric fuzzy coefficients provided the most appropriate fitting. It was concluded that the selected policy variables had a maximum effect on house prices affordability with real estate tax reporting the greatest impact. The non-policy variables demonstrated that customers possess personal housing expectations and preferences (Zhou *et al.*,2018). The study recommended that for optimal strategy, micro and macro parameters be considered for further research.

A study by Sorkheh *et al.* (2018) compared FLRM and MLRM lentil yield management. FLR method was used to estimate the yield by comparing inference capabilities with the multiple linear regression approach to investigate the efficient model. Fuzzy linear equations and multiple linear equations were obtained using a statistical software. Close data were organized as fuzzy numbers and the yield of the lentil genotypes was modeled by the FLRM andMLRM. The close data organized as fuzzy numbers were defuzzified using the center of area method to obtain the actual value representing it. In comparison of the FLRM to MLRM, FLRM was the best model in terms of the quantity of involved variables leading to the easiness of calculation with regard to it's RMSE which is the minimum error value. Based on the study, MLRM was found to be the most preferred model in modeling data while fuzzy linear regression analysis was found to be useful method especially when dealing with large data.

Further, Pandit *et al.* (2021) studied the efficiency of the statistical regression models and fuzzy regression model using sweet corn yield data affected by the cumulative dry matter of weed and cumulative density of weed. Data collected from these variables was utilized to evaluate the efficiency of these approaches. The simple linear regression model assessed each independent variable with respect to the FLRM while the MLRM assessed both independent variables with respect to the fuzzy linear regression model. The mean width of the forecast interval was used to calculate the model efficiency. During the assessment, the cumulative dry matter of weed and cumulative density of weed showed a negative result on the yield of sweet corn with respect to the simple linear regression model. MLRM taking into consideration both cumulative dry matter of the weed and cumulative density of the weed as regressors showed sign alteration in case of total density of weed even though there were significant at one percent level, this alteration of sign was considered to be multi collinearity. In fuzzy linear regression model indistinctness attributed with the center was less when the input variables were included in the model instead of only one of these variables when they were incorporated within the model, the average widths of the simple linear regression model and MLRM were large compared to fuzzy linear regression model hence indicating maximum correlative efficiency of the fuzzy linear regression method.

CHAPTER THREE

METHODOLOGY

3.1 Introduction

The chapter outlined diagnostic tests for verifying fuzzy linear regression assumptions, methods for fitting FLRM and evaluation techniques of FLRM. The model of multiple linear regression was also discussed as it created the basis for using fuzzy regression models.

3.2 Descriptive Statistics

Secondary data was used in this study. The data set contained assessed values for individual residential properties sold in Ames. The value of the residential properties was measured as price which was determined by the size of lot area, total basement square feet and garage area of the residential properties. To achieve the objectives of the study 600 observations were used from the collected data. Three predictor variables were used which included lot area indicating the lot size in square feet (x_1), total basement area in square feet (x_2) and size of garage area in square feet (x_3). On the other hand, the response variable was the price of a residential property sold in Ames in US dollars (y). Since Kenya is a developing country, the data set used from Ames will help the country in the sector of housing when determining the prices of houses which are constructed under the affordable housing program. The analysis of the data was carried out by the use of the R statistical software.

3.3 Linear Regression

According to Kumari and Yadav (2018), linear regression is an analytical step for determining the response variable using an explanatory variable. An association between two variables is measured in which a response variable is predicted grounded on at least one explanatory variable.

3.3.1 Multiple linear regression model

The use of multiple explanatory variables and one response variable resulted in the use of a model of Multiple Linear Regression (MLR). The MLR model with k predictor variables is represented in Equation 3.1;

$$y = \beta_0 + \beta_1 x_1 + \ldots \beta_k x_k + \varepsilon \tag{3.1}$$

where y denotes the response variable, $x_1, \ldots, x_k$ represent k explanatory variables and $\beta_0, \ldots, \beta_k$ denote the unknown regression coefficients of the model to be estimated. β_0 is the value of y when the value of $x_1, \ldots, x_k$ is equal to zero and $\beta_1, \ldots, \beta_k$ represent measures of the expected changes in y per unit change in $x_1, \ldots, x_k$ respectively.

3.3.2 Assumptions of a linear regression model

Assumptions of linear regression analysis have to be satisfied before performing linear regression analysis (Uyanık and Güler, 2013).

According to Ernst and Albers (2017) assumptions of the model on the basis of ordinary least squares (OLS) method include:

i. Independence of observations to ensure that the explanatory variables aren't too highly correlated;

ii. When all variables are held constant, the relationship between each explanatory variable and response variable is thought to linear.

iii. Distribution of the residuals should have a normal distribution that is, the residuals being distributed around zero; and

iv. The distribution of an error term should have constant variance for all observations.

3.3.3 Diagnostic Testing

To evaluate whether the assumptions of OLS are upheld, diagnostic tests were conducted. Hickey *et al* 2019 pointed out the diagnostic tests to be done to check whether the data in use is unusual or helpful and to check if any modifications are to be done. The main diagnostic tests performed in this study were:

(a) Test for normality

Normal q-q plot with the Shapiro wilk test were utilized to verify if the response variable follows a normal distribution. q-q plot gave a graphical visualization of normality. If the residuals observed from the q-q plot were not deviating severely from the straight line then the residuals had a normal distribution.
Since the graphic couldn't be enough to conclude that normality exists, a quantitative result was computed. Shapiro wilk test which is an hypothesis test was used with a null hypothesis that the data set has been generated from a normal distribution. If the p-value exceeded 0.05 observed from the Shapiro Wilk test, then the response variable follows a normal distribution under the following hypothesis:

H_0 : The variable follows a normal distribution

H_1 : The variable does not follow a normal distribution

(b) Test for linearity

Scatter plots were used to test linearity between the response variable and the explanatory variable. In multiple linear regression analysis, a scatter plot for each explanatory was plotted to check if linearity exists. If a pattern shows up in the plots no linear observation will be observed.

(c) Test for heteroscedasticity

Heteroscedasticity inflates the standard error raising the likelihood of making a type two error and failing to rule out an inaccurate assumption about a coefficient. Breusch-pagan test for heteroscedasticity in linear regression was applied to verify this assumption of no homoscedasticity. The variance of the errors from a regression was tested whether it is dependent on the values of the independent variables. If the $p-$value was below 0.05 then there was an indication of heteroscedasticity, that is the variance across entities is not constant. Using the following hypothesis:

H_o : No heteroscedasticity

H_1 : Heteroscedasticity present

(d) Test for multicollinearity

The degree of association present among the input factors in a model of regression was tested using the variance Inflation Factor (VIF). VIF quantifies the severity of multicollinearity in an ordinary leasr squares regression analysis. It provides an index that measures how much the variance of an estimated regression coefficient is increased because of collinearity. If the value of VIF is less than 1 it could indicate that there is no association in between the independent variables and a value greater 5 could indicate a severe correlation.

However, in some situations these assumptions may fail to be satisfied, as reported by Ayinde *et al* (2012) making it difficult to fit the linear regression model. Alternative method was considered to ensure that the inferences to be made were significant on the data in use for example the use of non statistical regression analysis methods in which fuzzy regression is considered in this case.

3.4 Fuzzy Linear Regression

Fuzzy regression gives a different view of statistical regression. It was applied to estimate the practical correlation in a fuzzy setting between the dependent and independent variables. The response and the explanatory variables are required to follow a normal distribution in statistical regression analysis. However, there are cases where these variables may not follow this distribution, regression assumptions may not hold true and imprecision between variables could exist. Estimating the regression coefficients and making subsequent prediction becomes a challenge to the classical regression analysis, hence the use of fuzzy linear regression analysis (Tanaka *et al.*,1989).

3.4.1 Fuzzy Numbers

In fuzzy domain crisp numbers are represented as fuzzy (uncertain) numbers. A fuzzy real or actual number $\tilde{A}$ is defined as uncertain set of actual numbers R. Every actual value number $x \in R$ belongs to the uncertain set $\tilde{A}$ with 0 to 1 membership degree according to a membership function $\mu_{\tilde{A}}(x) : x \rightarrow [0, 1]$. A membership degree of 0 implies that the value of x in the real number is excluded in the fuzzy number $\tilde{A}$ while a membership degree of 1 denotes that x's value in the real number is incorporated in $\tilde{A}$. A group of all observations in real value observations such that their membership function is greater than zero is referred to as the support for fuzzy number $\tilde{A}$. Fuzzy numbers are described in different functions of membership. Fuzzy numbers with triangular membership functions were used in this study.

A triangular fuzzy number (TFN) $\tilde{A}$ is characterized by a membership function which is triangular $\mu_{\tilde{A}}(x)$ that dictates a joint set of attainable values for a fuzzy number $\tilde{A}$, a set of x with $\mu_{\tilde{A}}(x) > 0$ where $\mu_{\tilde{A}}(x) = 1$ to the greatest extent of x.

$$\mu_{\tilde{A}}(x) = \begin{cases} \frac{a-x}{\underline{\alpha}} & \text{if} \quad a - \underline{\alpha} < x < a \\ 1 & \text{if} \quad x = a \\ \frac{x-a}{\overline{\alpha}} & \text{if} \quad a < x < a + \overline{\alpha} \\ 0 & \text{otherwise} \end{cases} \tag{3.2}$$

A fuzzy number $\tilde{A}$ which is triangular is defined by its mean value a, left width $\underline{\alpha}$ and right width $\overline{\alpha}$ Škrabánek and Martínková (2021) as shown in Figure 3.1

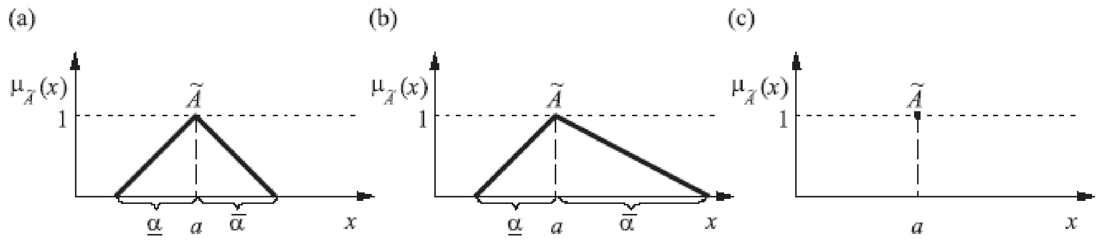

Figure 3.1: Triangular Fuzzy Number

Figure 3.1 (a) illustrates a triangular fuzzy number in symmetric form whose left and right widths are similar. A triangular fuzzy number whose left and right spreads are unequal illustrates an asymmetric triangular fuzzy number as evident in Figure 3.1 (b). Figure 3.1 (c) expresses a special triangular fuzzy number whose widths are equal to zero.

3.4.2 Fuzzification

The process of fuzzification involves changing actual numbers to fuzzy value observations. The methods used in fuzzy linear regression model requires the variables either the response or the explanatory variable to be fuzzy. This implies that the values of these variables are not regular real numbers but rather to a connected set of potential values with each value having a weight between 0 and 1.

A fuzzy number is represented by three points for example a set of three numbers (m_1, m_2, m_3), where m_1, m_2 and m_3 represent the left spread, the central value that is the most probable value and the right spread respectively.

Spreads represent the closeness to the real value observation that is the maximal deviation from the average score or observation which is the central value or the mode or peak point within the fuzzy number.

The fuzzy methods used in this study necessitated the real value observation of the response variable be modified into a fuzzy number, that is the response variable (Y) observed to take this form $\tilde{Y} = (y_i, \underline{v_i}, \overline{v_i})$.

Where y_i was the possible value of the response variable, $\underline{v_i}$ and $\overline{v_i}$ were the left and right spreads.

Methods of fuzzification

The basic methods that can deduce triangular fuzzy numbers from actual value observations include;

(i) The mean and zero methods calculate the sTFN, the mean method determines the possible TFN value as the average of x given y and standard deviations given as left and right spreads and the zero method inserts zeros to both spreads. These methods compute sTFN

(ii) The median and error method can enumerate either sTFN or nsTFN based on the data present. The median method provides the central values as a median and left and right spreads are determined as distance of the first and third quartile from the median. The error method utilizes a user characterized numeric value of vector for the spreads.

(iii) Simulation method through various statistical distribution methods can also be applied to generate fuzzy numbers if the discussed methods fail to give the results required.

3.4.3 Fuzzy Linear Regression Model

The FLRM has the following structure:

$$\tilde{Y}_j = \tilde{A}_0 + \tilde{A}_1 x_{1j} + \ldots \tilde{A}_i x_{ij} + \cdots + \tilde{A}_n x_{nj} \quad \text{with} \quad j = 1, \ldots, m, i = 1, \ldots, n \tag{3.3}$$

where; n is the quantity of independent variables x_{ij}, m is the quantity of data, and $\tilde{Y}_j$ is the fuzzy predicted value of the output variable considering the j^{th} data.

Because uncertainty is incorporated into the model via fuzzy numbers, there is no error term in the fuzzy linear regression model. Different fuzzy linear models are used depending on the observations that variables used take, sensitivity to outliers and the number of explanatory variables. For example,

i. When the explanatory variables are actual value observations and the response variable takes a non-symmetric observations. The Fuzzy least squares (FLS) model and Fuzzy least absolute error model (FLAR) are used, however FLS model works well for simple fuzzy linear regression whereas FLAR supports multiple linear regression.

ii. When explanatory variables take actual value observations and response variable assumes symmetric fuzzy observations the possibilistic linear regression model is used, it also support multiple regression.

iii. Fuzzy linear regression model with multiple objectives abbreviated as MOFLR is used when both observations of the response variable and the explanatory variable are symmetric fuzzy observations.

iv. The least squares possibilistic linear regression model is used when the independent and dependent variables are actual value observations but the estimated response variable is non-symmetric observation.

Table 3.1: Characteristics of fuzzy linear regression procedure

Method	m	x,$\tilde{X}$	y,$\tilde{Y}$	$\hat{\tilde{Y}}$	sensistivity to outliers
FLS	1	R	nsTFN	nsTFN	medium
FLAR	∞	R	nsTFN	nsTFN	medium
PLR	∞	R	sTFN	sTFN	very high
OPLR	∞	R	sTFN	sTFN	low
MOFLR	∞	sTFN	sTFN	sTFN	medium
PLRLS	∞	R	R	nsTFN	very high

The model should be applied if the following assumptions hold true;

3.4.4 Assumptions of fuzzy linear regression model

Assumptions of the model of fuzzy linear regression were on the basis of problems that arise from statistical regression analysis.

(i) The residual variation is unequal throughout a range of measured values.

(ii) Correlation of the same variables between two successive time intervals due to incorrect specification of the model.

(iii) Distribution of the error term is due to fuzziness unlike randomness which is observed in the error term in statistical linear regression model.

(iv) Availability of inadequate data when the number of observations is insufficient.

3.4.5 Estimation of the fuzzy regression model parameters

Two approaches were used to estimate parameters in fuzzy regression model namely, the minimum fuzziness criterion and the least squares criterion.

Minimum fuzziness criterion

The minimum fuzziness criterion was used when the total spread of the fuzzy numbers was required to be small. Coefficients of the FLRM in case they were fuzzy triangular numbers, the function Y_j will also be a fuzzy number which is triangular with the following mean $(Y_{a,j})$ and spread $Y_{c,j}$.

$$\text{centre} : Y_{a,j} = a_o + a_1 x_{1j} + \cdots + a_i x_{ij} + \cdots + a_n x_{nj}$$

$$\text{Width} : Y_{c,j} = c_0 + c_1|x_{1j}| + \cdots + c_i|x_{ij}| + \cdots + c_n|x_{nj}| \tag{3.4}$$

The degree $h(0 \leq h < 1)$ is chosen so as the available data will be included in the inferred fuzzy number Y_j.

When there is sufficient data $(m > 10)$ (where m is the number of observation) $h = 0$ can be set which maximizes the regression ambiguity. The chosen value for h must be greater than 0 or equal to 0 but in any case less than 1 since $h = 1$ will turn the FLRM into classical linear regression model. Since the total fuzzy number spread $Y_j, j = 1, 2, \ldots, m$ was required to take the least value as possible it was minimized as in Equation 3.5

$$J = \min\{mc_o + \sum_{j=1}^{m} \sum_{i=1}^{n} c_i|x_{ij}|\} \tag{3.5}$$

Subject to

$$y_j \geq \sum_{i=0}^{n} a_i x_{ij} - (1-h) \sum_{i=0}^{n} c_i|x_{ij}| \tag{3.6}$$

$$y_j \leq \sum_{i=0}^{n} a_i x_{ij} + (1-h) \sum_{i=0}^{n} c_i|x_{ij}| \tag{3.7}$$

$c_i \geq 0$ where $i = 0, 1, \ldots, n$ and $j = 0, 1, \ldots, m$

If $h \neq 0$ is singled out the uncertainty of the generated model will be maximum compared to $h = 0$.

To derive the results concerning center values and spreads of the fuzzy triangular number that represent the coefficients of fuzzy linear regression, computer software was used. To completely exploit fuzzy linear regression, coefficients of fuzzy regression model were data driven.

Least squares criterion

A study by Diamond (1988) proposed a metric fuzzy number distance method for developing fuzzy linear regression models. To execute the approximation concepts employed in statistical regression analysis, the rules employed in statistical regression analysis objective functions have to be replaced by fuzzy distance measures. A fuzzy version of least squares on the basis of distance is defined as

$$\text{minimize} S(\tilde{A}) = \sum_{i=1}^{n} d(\tilde{Y}_i^*, Y_i)^2 = ((y_i - y_{iL}) - (\hat{y}_i - \hat{y_{iL}}))^2 + (y_i - \hat{y}_i)^2 + ((y_i + y_{iR}) - (\hat{y}_i + \hat{y_{iR}}))^2 \tag{3.8}$$

in which $\tilde{Y}_i^*, i = 1, 2, ..., n$ is the fuzzy result.
Y_i is the observed fuzzy response variable.
$Y_i = (y_{iL}, y_i, y_{iR})$
Assuming that Y_i was composed of a non-symmetric triangular fuzzy numbers then $Y_i = (y_{iL}, y_i, y_{iR})$. The coefficients of regression are asymmetric triangular fuzzy numbers with it's equivalent least square given as in Equation 3.9

$$S_{asy}^{D}(\tilde{A}_0, \tilde{A}_1, \ldots, \tilde{A}_m) = \sum_{i=1}^{m} \left\{ \left(\sum_{k=0}^{m} a_{kL} x_{ik} - y_{iL} \right)^2 + \left(\sum_{k=0}^{m} a_k x_{ik} - y_i \right)^2 + \left(\sum_{k=0}^{m} a_{kR} x_{ik} - y_{iR} \right)^2 \right\} \tag{3.9}$$

To calculate the solution of $\tilde{A}_k = (a_{kL}, a_k, a_{kR})$ the subsequent set of equations were generated for $k = 0, 1, \ldots, m$ by computing the partial derivatives a_{kL}, a_k and a_{kR} to be equal to 0.

$$\begin{aligned} \sum_{i=1}^{n}\left(x_{ik}\sum_{k=0}^{n}a_{kL}x_{ik}\right) &= \sum_{i=1}^{n}y_{iL}x_{ik} \\ \sum_{i=1}^{n}\left(x_{ik}\sum_{k=o}^{m}a_{k}x_{ik}\right) &= \sum_{i=1}^{n}y_{i}x_{ik} \\ \sum_{i=1}^{n}\left(x_{ik}\sum_{k=0}^{m}a_{kR}x_{ik}\right) &= \sum_{i=1}^{n}y_{iR}x_{ik} \end{aligned} \tag{3.10}$$

Solving these equations gives values of a_{kL}, a_k and a_{kR} where the fuzzy parameters are estimated and hence FLRM inferred. If the regression coefficients are taken to be sTFN that is $\tilde{A}_j = (a_j, c_j)$ the distance expression was given as:

$$S_{sy}^{D}(\tilde{A}_0, \tilde{A}_1, \ldots, \tilde{A}_m) = \sum_{i=1}^{m}\left\{\left[\sum_{j=0}^{m}(a_j - c_j)x_{ij} - y_{iL}\right]^2 + \left(\sum_{j=0}^{m}a_j x_{ij-y_i}\right)^2 + \left[\sum_{j=0}^{m}(a_j + c_j)x_{ij} - y_{iR}\right]^2\right\} \tag{3.11}$$

Similarly, the partial derivatives of a_k and c_k $k = 0, 1, \ldots, m$ are equated to 0. Whereby the subsequent set of equations are determined;

$$\begin{cases} 2\sum_{i=1}^{n}(x_{ik}\sum_{j=0}^{m}c_j x_{ij}) = \sum_{i=1}^{n}[x_{ik}(Y_{iR} - Y_{iL})] \\ 3\sum_{i=1}^{n}(x_{ik}\sum_{j=0}^{m}a_j x_{ij}) = \sum_{i=1}^{n}[x_{ik(y_{iR}+y_{iL+y_i})}] \end{cases} \tag{3.12}$$

for $k = 0, 1, \ldots, m$

Also figuring out these equations, parameters were estimated which were symmetric triangular fuzzy numbers.

3.4.6 Evaluation of the Model

Fuzzy regression model was assessed on the basis of squared distances and deviations of membership functions from the observed $\tilde{Y}$ and predicted $\hat{\tilde{Y}}$ fuzzy output observations $\tilde{Y}$. The squared distances were expressed by the measure of goodness of fit (G) between the observed $\tilde{Y}$

and predicted $\hat{\tilde{Y}}$ fuzzy response value. The measure G was obtained out of the squared distance between fuzzy numbers which are triangular as given in Equation 3.8 and it was defined as in Equation 3.13

$$G = \frac{1}{n}\sum_{i=1}^{n}([(y_i - y_{iL}) - (\hat{y}_i - \hat{y_{iL}})]^2 + (y_i - \hat{y}_i)^2 + [(y_i + y_{iR}) - (\hat{y}_i + \hat{y_{iR}})]^2) \quad (3.13)$$

Low G values could imply that the model sharply corresponds to the observations.

Model predictions and output membership functions were assessed by the total error of fit. It was computed as shown in Equation 3.14

$$\sum E = \sum_{i=1}^{n} E_i \quad (3.14)$$

where E_i was a contrast of membership functions between the i^{th} observation $\tilde{Y}_i$ and the i^{th} model prediction $\hat{\tilde{Y}}_i$ with respect to $\tilde{Y}_i$'s membership function.

CHAPTER FOUR

RESULTS AND FINDINGS

4.1 Introduction

This chapter presents results and discussions of data analysis using fuzzy linear regression methods. The uncertain change of the sale price in the residential properties resulted in the conversion of the real value observations of the sale price (the response variable) to uncertain number was outlined. Fuzzy linear regression model was fitted, evaluated and discussed using three fuzzy regression methods.

4.2 Diagnostic test of fuzzy linear regression assumptions

The study summarized the data set as presented in Table 4.1

Table 4.1: Summary of data used

	price	lot area	Total basement.sf	garage area
Min.	12789	1476	160.0	0.0
1st Qu.	128425	7433	795.8	336.0
Median	158000	9074	992.0	475.5
Mean	168188	8899	1034.3	457.5
3rd Qu.	197525	10625	1258.0	566.8
Max.	337500	31770	2110.0	932.0

Before fitting the fuzzy linear regression model, normality, linearity, heteroscedasticity and multi collinearity tests were performed to check if the fuzzy linear regression assumptions hold true.

4.2.1 Test for normality

Observations from the q-q plot as shown in Figure 4.1 indicated that, not all the values of price fall along the reference line denoted by the blue line.

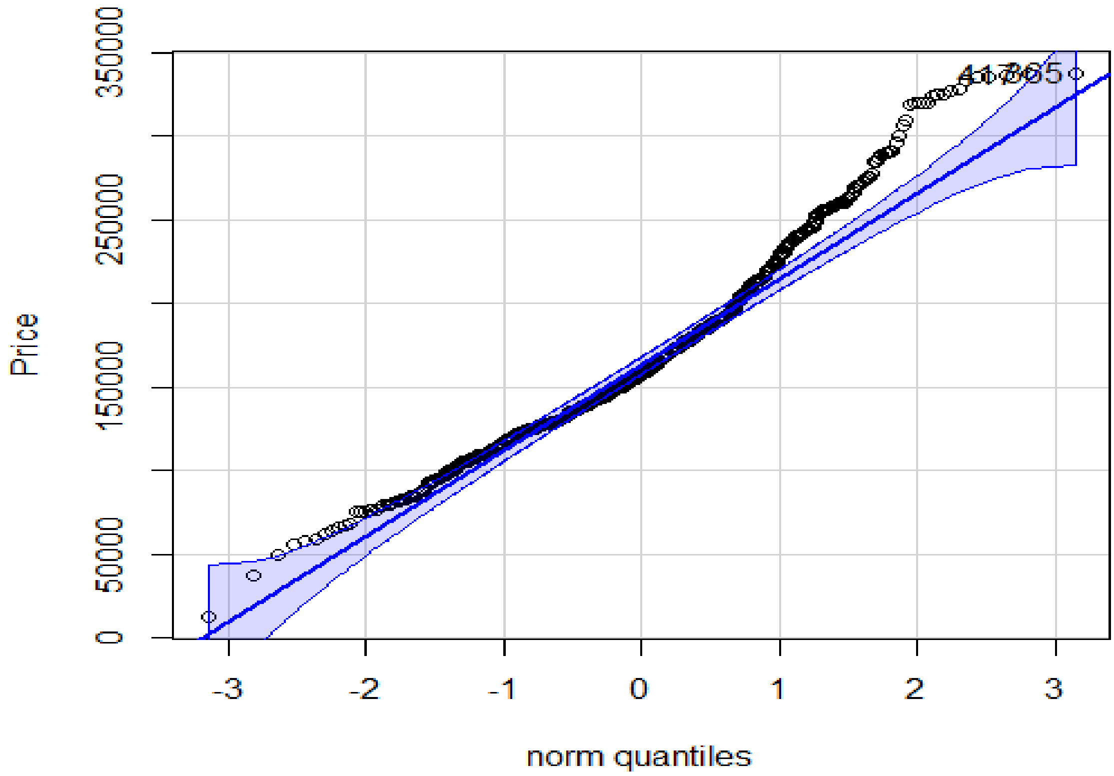

Figure 4.1: Q-Q Plot of price

Results from Shapiro wilk test using the $p-$ value > 0.05 level of significance, shows that the $p-$ value was $6.837\text{e}-11$ which is less than 0.05. Therefore the null hypothesis was rejected. From the two analyses it was assumed that non-normality exists.

4.2.2 Test for linearity

A scatter plot of each independent variable and dependent variable was plotted to check for linear correlation between the predictor and response variable used in the analysis as depicted in Figure 4.2 to Figure 4.4.

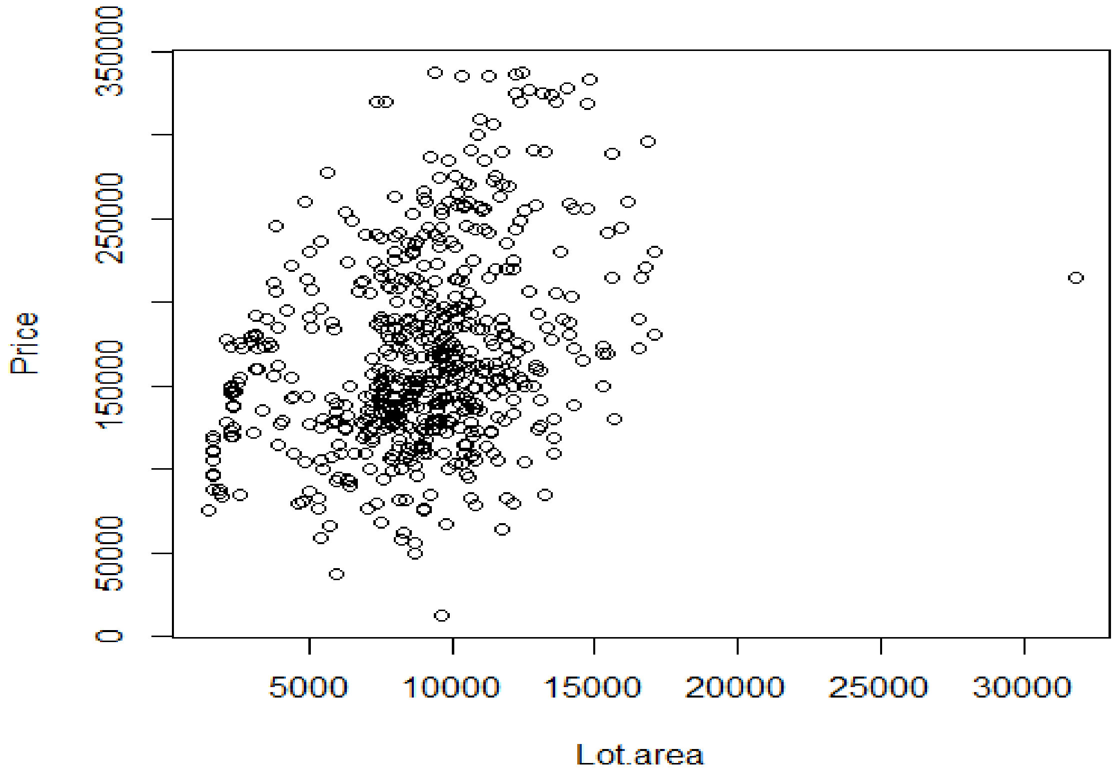

Figure 4.2: Scatter plot of Lot Area

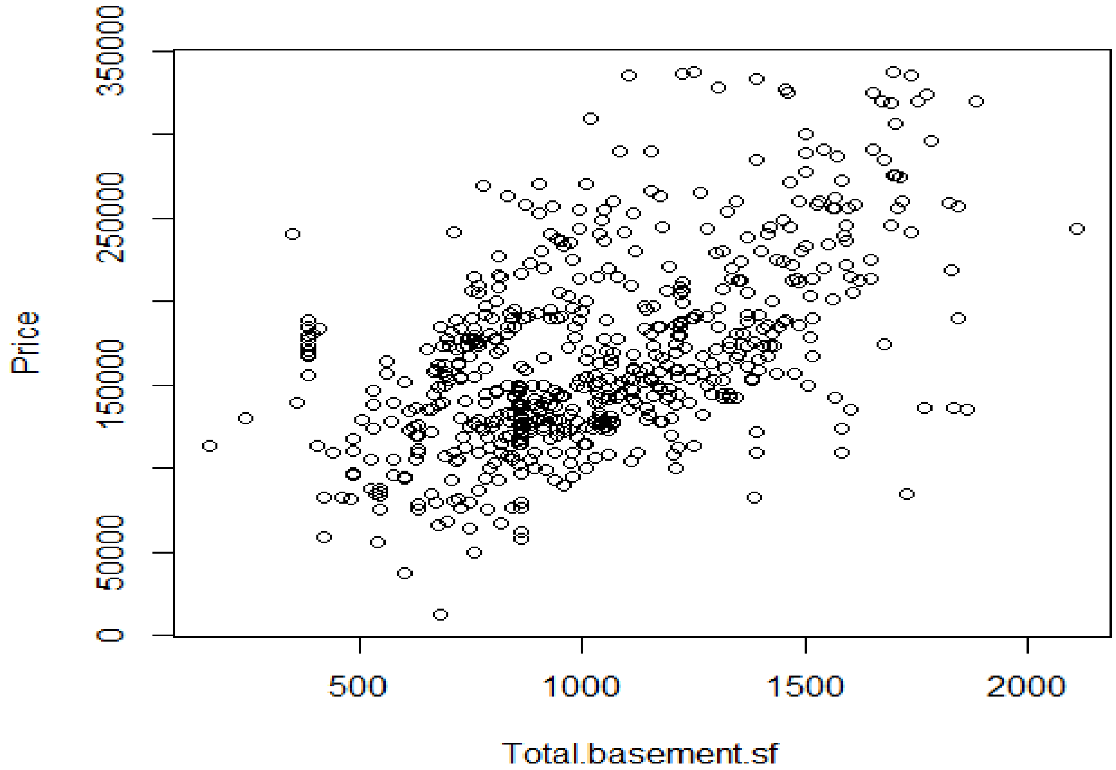

Figure 4.3: Scatter plot of Total basement Square Feet

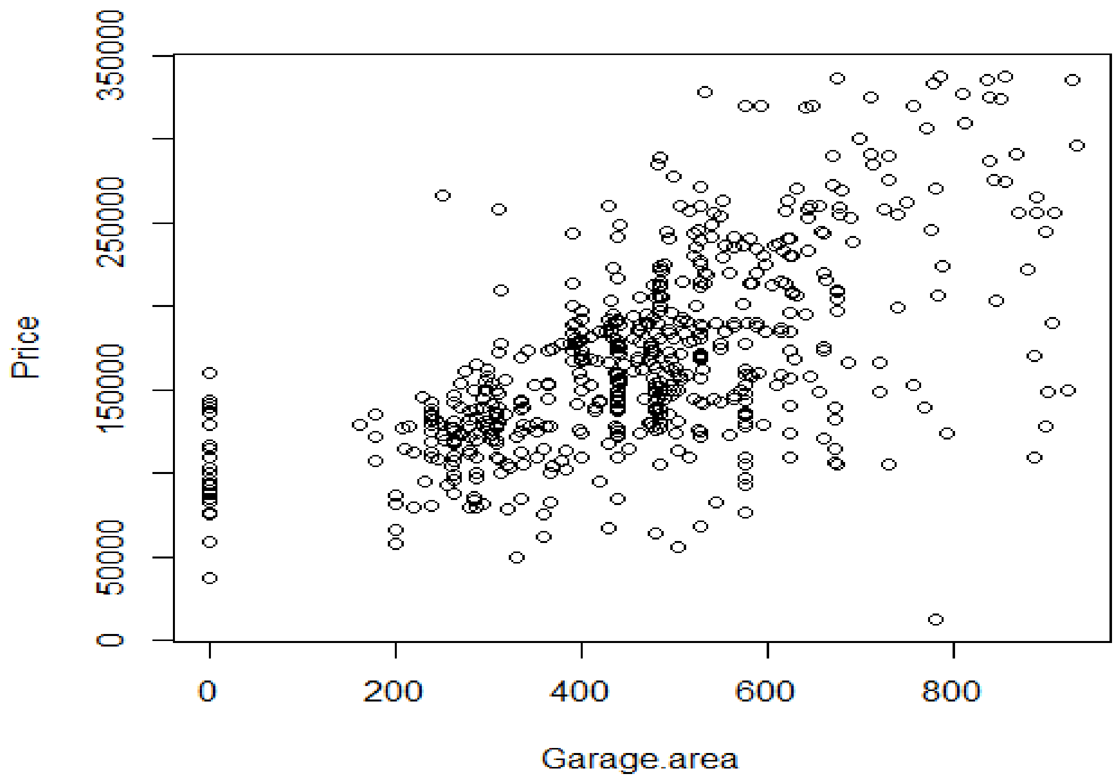

Figure 4.4: Scatter plot of Garage area

From the scatter plots, it was deduced that the data points in Figure 4.2, Figure 4.3 and Figure 4.4 indicate a non-linear relationship between lot area versus price, total basement floor square feet versus price and garage area versus price.

4.2.3 Test for homoscedasticity

Breusch-pagan test results indicated a $p-$value$= 1.64e-14$ which is less than 0.05 significance level. Thus the null hypothesis was rejected showing that heteroscedasticity exists among the variables.

4.2.4 Test for multi collinearity

Scores computed by variance inflation factor as shown in Table 4.2 were less than 5 which was the reference value. This shows that there was no multi-collinearity among the variables.

Table 4.2: VIF values

lot area	total basement area	garage area
1.114201	1.245163	1.230160

4.3 Fitting fuzzy linear regression model

Data visualization indicated that, there was an uncertain change in the sale price of the residential properties when there is a change in the area size of lot area, total basement floor squar e feet and garage area respectively. Non-linearity, heteroscedasticity and non-normality among the variables was also observed.

Applying classical regression analysis where there is uncertainty, non-normality, heteroscedasticity and non linearity among variables becomes a challenge which can be addressed by fuzzy regression analysis.

Fitting the fuzzy linear regression model involved the conversion of real numbers to fuzzy numbers and using the PLRLS, PLR and FLAR methods to fit the fuzzy linear regression model. The unknown parameters of the model were estimated and the fitted models were expressed using the central location of the fuzzy regression model, the maximum and minimum limit of the model support interval and visualised on their respective plots.

4.3.1 Fuzzification

Potential fluctuations (uncertainty) of the prices motivated the fuzzification of the sale price of the residential properties to fuzzy numbers. A statistical software R was used to fuzzify price values where uniform distribution function was used to impute the spreads of the price. The error and simulation methods were used for fuzzification. The error method was used to fuzzify the real numbers of price to fuzzy numbers. This resulted to the values of price in terms of symmetric triangular fuzzy numbers. An illustration of the section of symmetric triangular fuzzy numbers is presented in Table 4.3.

Table 4.3: Symmetric Fuzzy number.

price	left spread	right spread
215000	9976.275	9976.275
105000	10533.471	10533.471
172000	10284.042	10284.042
244000	10478.234	10478.234
189900	10553.245	10553.245
195500	10925.274	10925.274
213500	9730.295	9730.295

Unequal spreads of the triangular fuzzy number were simulated to generate the non symmetric triangular fuzzy number. Left spreads were simulated within a range of $10000 - 12000$ and the right spreads within a range of $5000 - 8000$ respectively using the uniform distribution. Table 4.4 shows the subsection of the simulated results.

Table 4.4: Non symmetric Fuzzy number.

Price	left spread	right spread
215000	10373.98	6438.885
105000	10136.96	6727.261
172000	10317.42	6625.119
244000	11683.12	6904.427
189900	10746.00	6259.994
195500	11173.55	6705.355
213500	11936.85	5617.864

4.3.2 Fitting fuzzy linear regression model using PLRLS

Using Equation 3.3 the regression coefficients of the model were estimated using the R software. Focusing on each explanatory variable coefficients of the spreads and centers of the model were given in Table 4.5.

Table 4.5: Central and Spreads coefficients of FLR model using PLRLS method

Variable	Spread		center	
	coefficient	value	coefficient	value
Lot Area	a_{0L}	38631.09	a_0	116408.5
	a_{0R}	142008.1	a_1	5.818
	a_{1L}	12.55		
	a_{1R}	2.5145		
Total basement SF	a_{0L}	105884.36	a_0	72110.26
	a_{0R}	124525.63	a_1	92.89
	a_{1L}	24.21		
	a_{1R}	32.44		
Garage Area	a_{0L}	43156.903	a_0	81056.90
	a_{0R}	130290.07	a_1	190.46
	a_{1L}	222.656		
	a_{1R}	28.39		

where (a_0, a_1) are the central values, (a_{0L}, a_{1L}) are left spreads and (a_{0R}, a_{1R}) are the right spreads coefficients of the fuzzy linear regression model.

Lot area

Central location of the fuzzy regression model:

Price $= 116408.5 + 5.8188 *$ Lot area

Minimum limit of the model support interval:

Price $= 77777.43 - 6.7304 *$ Lot area

Maximum limit of the model support interval:

Price $= 258416.7 + 8.3333 *$ Lot area

Total basement area

Central location of the fuzzy regression model:

Price $= 72110.26 + 92.8909 *$ Total basement sf

Minimum limit of the model support interval:

Price $= 33774.1 + 68.6772 *$ Total basement sf

Maximum limit of the model support interval:

Price $= 196635.9 + 125.3298 *$ Total basement sf

Garage area

Central location of the fuzzy regression model:

Price $= 81056.9 + 190.4624 *$ Garage area

Minimum limit of the model support interval:

Price $= 37900 - 32.1936 *$ Garage area

Maximum limit of the model support interval:

Price $= 211347 + 218.8612 *$ Garage area

Spreads of the fuzzy coefficients were required to be as small as possible. From Table 4.5 spreads are observed to have a wide range. The total error of fit was used as a performance indicator. In this case, it was calculated as $9.440e + 13$ for the lot area, $8.670e + 13$ for the total basement area and $8.648e + 13$ for the garage area.

Fitting the model using three variables from Equation (3.3) the estimated parameters are given as shown in Table 4.6

Table 4.6: Parameters of FLR model using PLRLS method.

	center	left spread	right spread
(Intercept)	27420.012701	10018.590675	$2.516520e-14$
Lot area	2.120028	4.344706	8.104145
Total basement area	56.132987	0.000000	40.62421
Garage area	139.558384	166.723898	0.000000

Central location of the fuzzy regression model:

Price $= 27420.01 + 2.12 *$ lot area $+ 56.133 *$ total basement area $+ 139.5584 *$ garage area

Minimum limit of the model support interval:

Price $= 17401.42 - 2.2247 *$ lot area $+ 56.133 *$ total basement.sf $- 27.1655 *$ garage area

Maximum limit of the model support interval:

Price $= 27420.01 + 10.2241 *$ lot area $+ 96.7572 *$ total basement area $+ 139.5584 *$ garage area

It was observed that, when fitting the model using multi variables the spreads of the fuzzy numbers are minimized as compared to when fitting a single predictor variable to the model. Also, the total error of fit calculated in this model was $7.1725e + 13$ which was minimum compared to the error of fit calculated when using each single variable.

However, the total error of fit was observed to be too complex when using this method leading to the use of another method which was the PLR method.

4.3.3 Fitting the fuzzy linear regression model using the PLR method

To reduce the total error of fit, the fuzzy linear regression model was fitted using a new method. The regression coefficients of the model were calculated and given as shown in Table 4.7

Table 4.7: Central and Spreads coefficients of FLR model using PLR method

Variable	Spread		center	
	coefficient	value	coefficient	value
Lot Area	α_0	101303.9	a_0	169811.10
	α_1	7.412	a_1	0.6144
Total basement SF	α_0	124835.77	a_0	80718.02
	α_2	28.77	a_2	98.07
Garage Area	α_0	96685.02	a_0	124915.36
	α_3	125.83	a_3	93.31

Lot area

Central location of the fuzzy regression model:

Price $= 169811 + 0.6144 *$ lot.area

Minimum limit of the model support interval:

Price $= 68507.04 - 6.7978 *$ lot.area

Maximum limit of the model support interval:

Price $= 271114.9 + 8.0266 *$ lot.area

Total basement area

Central location of the fuzzy regression model:

Price $= 80718.02 + 98.0736 *$ total basement area

Minimum limit of the model support interval:

Price $= 44117.75 + 69.3004 *$ total basement area

Maximum limit of the model support interval:

Price $= 205553.8 + 126.8468 *$ total basement area

Garage area

Central location of the fuzzy regression model:

Price $= 124915.4 + 93.3129 *$ garage area

Minimum limit of the model support interval:

Price $= 28230.34 + -32.5159 *$ garage area

Maximum limit of the model support interval:

Price $= 221600.4 + 219.1417 *$ garage area

Spreads of the variables were observed to be minimum in this method as compared to the previous method. The total error of fit for the variables using this method was calculated as 9582.3 for the lot area, 8804.4 for the total basement area and 8769.33 for the garage area. The total error of fit using this method was observed to be minimal as compared to the previous method. Model coefficients for the multi variables were also calculated and presented in Table 4.8.

Table 4.8: Coefficients of the PLR model.

	center	left spread	right spread
(Intercept)	57443.50	42808.531	42808.531
Lot area	0.4562	6.198	6.198
Total basement area	79.36	0.000	0.000
Garage area	43.06	56.04	56.04

Central location of the fuzzy regression model:

Price $= 57443.5 + 0.4562 *$ lot area $+ 79.3641 *$ total basement area $+ 43.0582 *$ garage area

Minimum limit of the model support interval:

Price $= 14634.97 - 5.7423 *$ lot area $+ 79.3641 *$ total basement area $- 12.985 *$ garage area

Maximum limit of the model support interval:

Price $= 100252 + 6.6547 *$ lot area $+ 79.3641 *$ total basement area $+ 99.1014 *$ garage area

Minimal spreads are also observed and the total error of fit is 6964.94 which is minimal compared to the previous method. This indicated that a model fitted using the PLR method is better compared to the model fitted using the PLRLS method.

4.3.4 Fitting fuzzy linear regression method using the FLAR method

The FLAR method was also used to model the cross sectional data. The reason for this approach was because the method is a statistic based method unlike the other methods which are possibilistic. The regression coefficients were asymmetric triangular fuzzy numbers which were estimated based on Equation 3.9. The regression coefficients are expressed in Table 4.9.

Table 4.9: Central and Spreads coefficients of FLR model using FLAR method

Variable	Spread		center	
	coefficient	value	coefficient	value
Lot Area	a_{0L}	10013.66	a_0	113701.0
	a_{0R}	10013.66	a_1	4.7715
	a_{1L}	0.00		
	a_{1R}	0.00		
Total basement SF	a_{0L}	10001.02	a_0	61451.87
	a_{0R}	10001.02	a_2	93.58
	a_{2L}	0.01120		
	a_{2R}	0.01120		
Garage Area	a_{0L}	10015.04	a_0	76500
	a_{0R}	10015.04	a_3	197.397
	a_{3L}	0.00		
	a_{3R}	0.00		

Lot area

Central location of the fuzzy regression model:

Price $= 113701 + 4.7715 *$ lot area

Minimum limit of the model support interval:

Price $= 103687.3 + 4.7715 *$ lot area

Maximum limit of the model support interval:

Price $= 123714.6 + 4.7715 *$ lot area

Total basement area

Central location of the fuzzy regression model:

Price $= 61451.87 + 93.5829 *$ total basement area

Minimum limit of the model support interval:

Price $= 51450.85 + 93.5717 *$ total basement area

Maximum limit of the model support interval:

Price $= 71452.9 + 93.5941 *$ total basement area

Garage area

Central location of the fuzzy regression model:

Price $= 76500 + 197.397 *$ garage area

Minimum limit of the model support interval:

Price $= 66484.96 + 197.397 *$ garage area

Maximum limit of the model support interval:

Price $= 86515.04 + 197.397 *$ garage area

Using this method, it was observed that spreads were equivalent to zero indicating fuzziness in the model was minimal compared to the other models. Total error of fit for the models using single variables was, 1067.72 for the lot area, 1069.97 for the total basement area and 1029.54 for the garage area which was minimal compared to the previous two models.

Using the three variables for modeling the estimated parameters was also given as in Table 4.10.

Table 4.10: Coefficients when using FLAR method.

	center	left spread	right spread
(Intercept)	34521.06	10001.02	10001.02
Lot area	1.3686	0.000	0.000
Total basement area	47.61	0.0112	0.0112
Garage area	153.763	0.000	0.000

Central location of the fuzzy regression model:

Price $= 34521.06 + 1.3686 *$ lot area $+47.6059 *$ total basement area $+153.7627 *$ garage area

Minimum limit of the model support interval:

Price $= 24520.04 + 1.3686 *$ lot area $+ 47.5947 *$ total basement area $+ 153.7627 *$ garage area

Maximum limit of the model support interval:

Price $= 44522.08 + 1.3686 *$ lot area $+ 47.6171 *$ total basement.sf $+ 153.7627 *$ garage area

The total error of fit was calculated as 1016.92 indicating that the model fitted using the FLAR method is better than the PLR and PLRLS.

At the next stage, the FLAR method was also fitted using fuzzy numbers which were obtained through the simulation method. This approach was useful because it was used to check if the FLAR method was the best when fitting fuzzy linear regression model. Table 4.11 presents the model coefficients.

Table 4.11: Asymmetric coefficients using FLAR method.

	center	left spread	right spread
(Intercept)	34521.04	10967.92	6413.87
Lot area	1.369	0.00	0.000
Total basement area	47.61	0.00	0.1776
Garage area	153.76	0.00	0.000

The total error of fit of this model was determined as 1040.23 which is minimum compared to the models fitted using PLR and PLRLS methods. From the results of the study, different fuzzy regression methods used to model the sale price of the residential properties gives a different total error of fit and different range of the spreads.

Models using the PLRLS method were visualized on individual plots while models of the PLR and FLAR method were visualized on the same plots using the coefficients for the models central location and both support limits as presented in Figure 4.5, Figure 4.6, Figure 4.7, Figure 4.8, Figure 4.9, and Figure 4.10 where the total basement area variable was used for illustration.

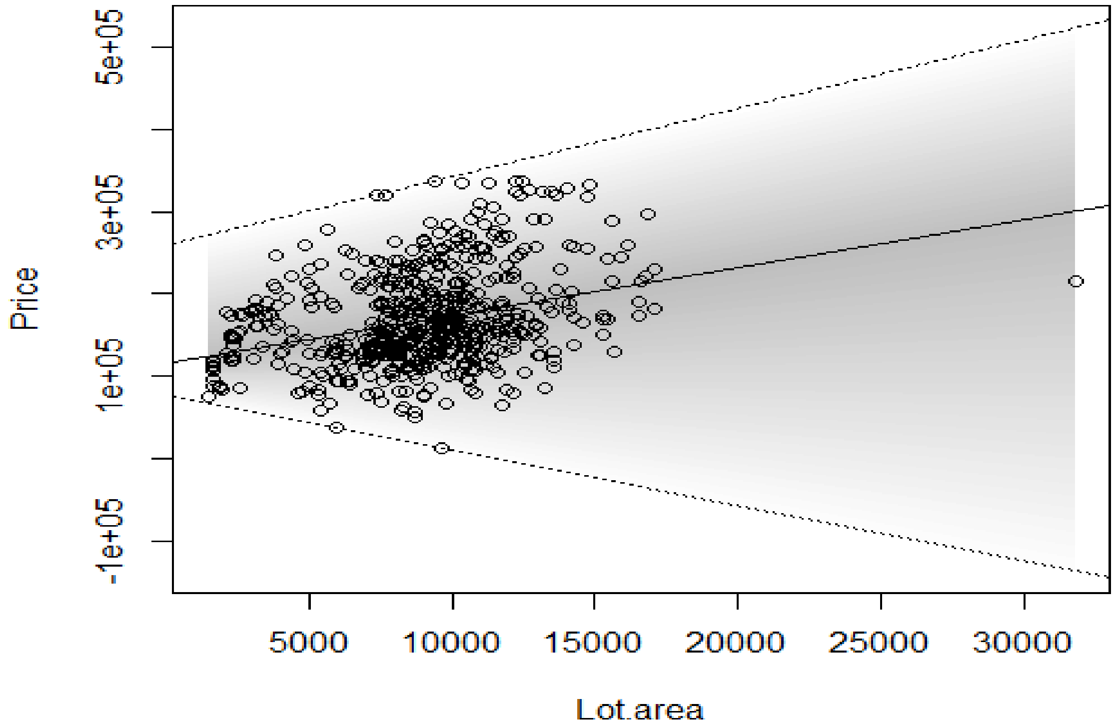

Figure 4.5: FLR model of price versus lot area using the PLRLS method

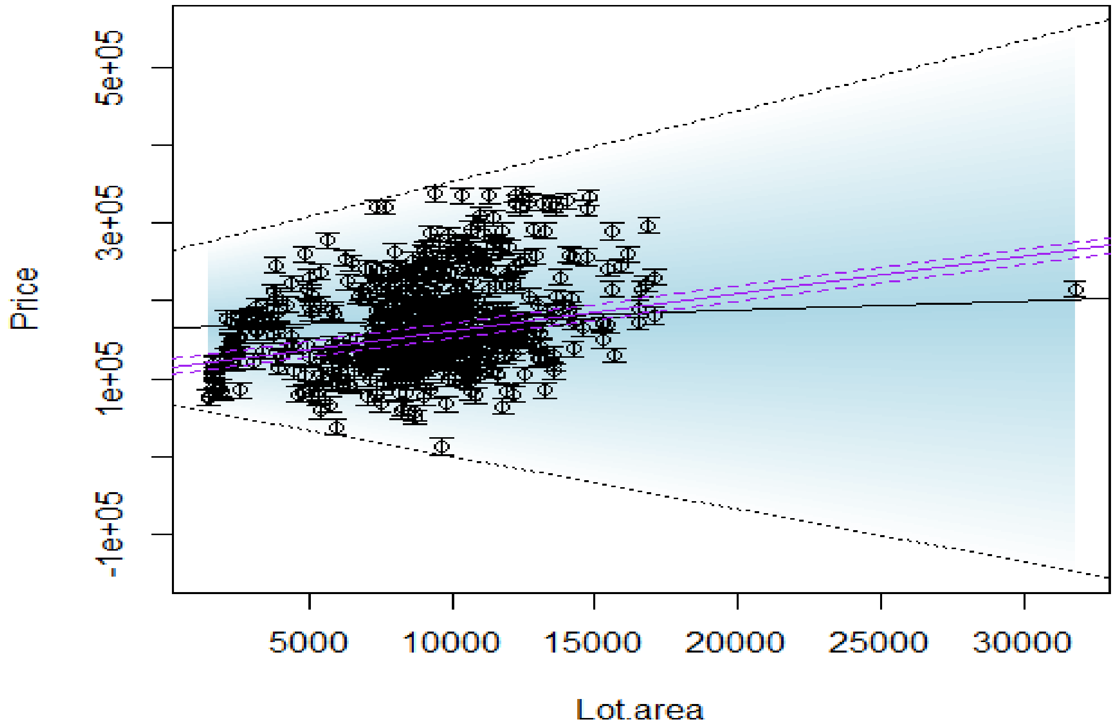

Figure 4.6: FLR model of price versus lot area using the PLR and FLAR methods

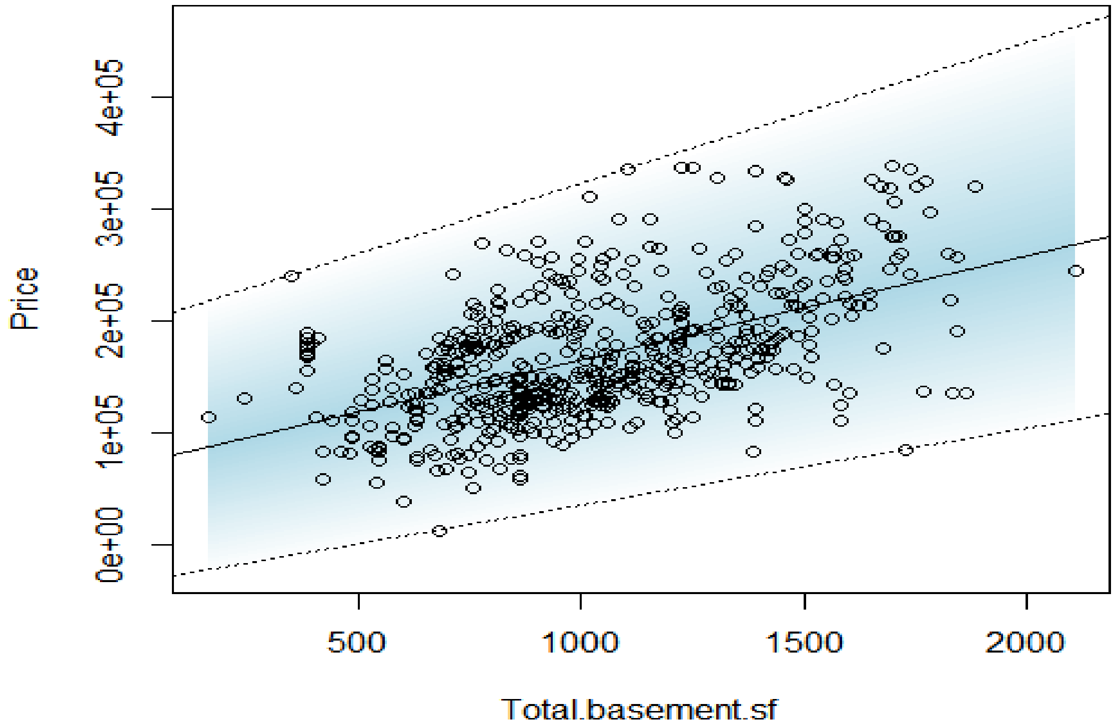

Figure 4.7: FLR model of price versus total basement area using the PLRLS method

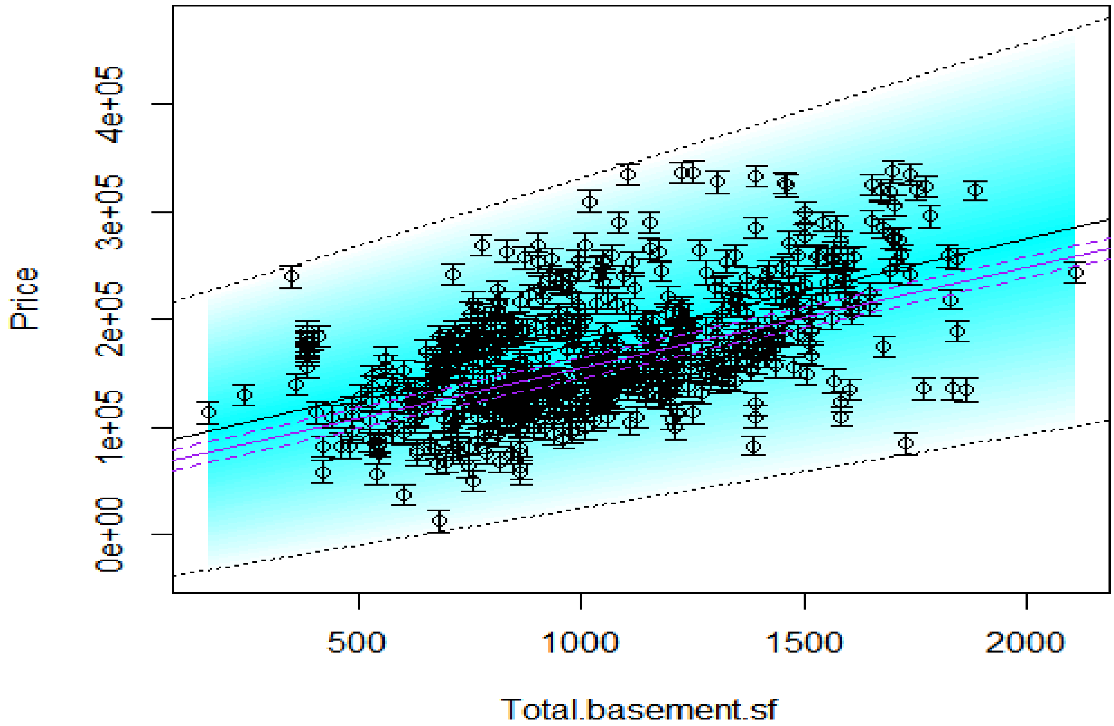

Figure 4.8: FLR model of price versus total basement area using the PLR and FLAR methods

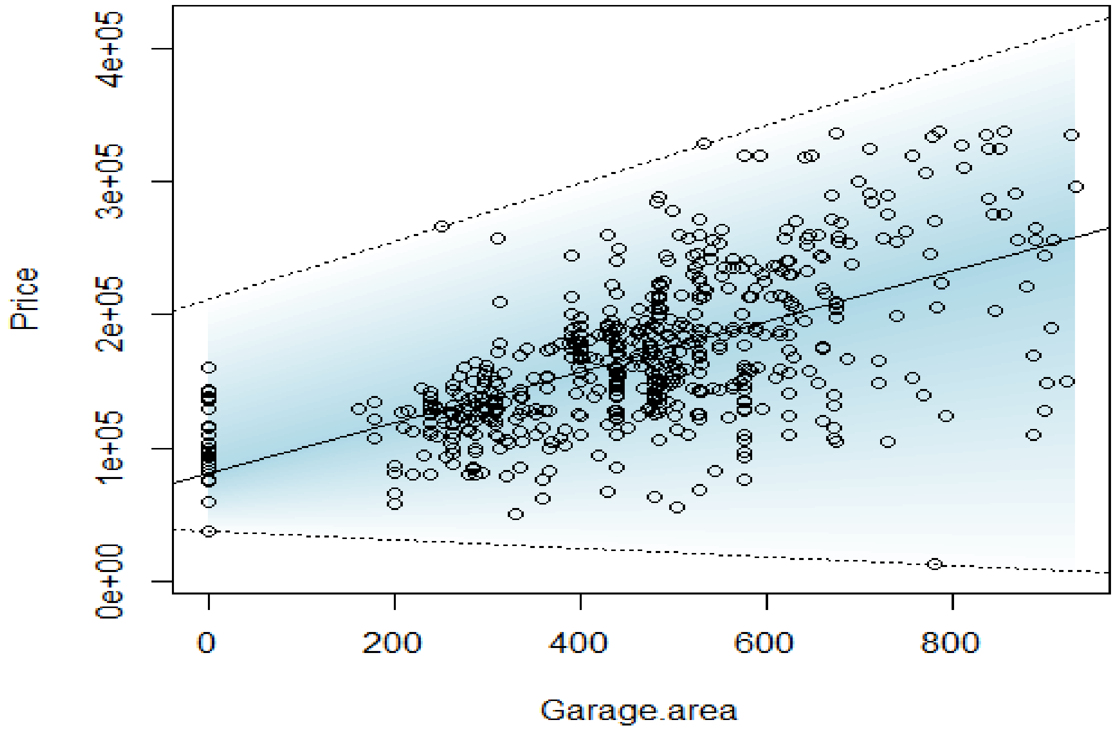

Figure 4.9: FLR model of price versus garage area using the PLRLS method

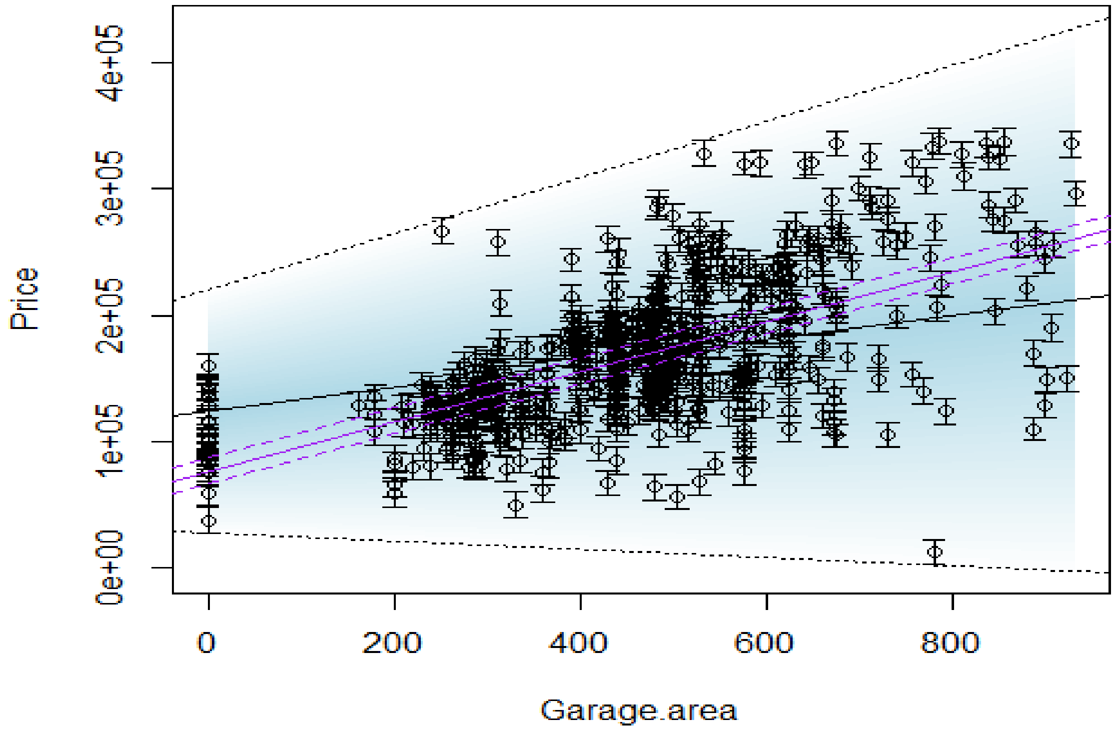

Figure 4.10: FLR model of price versus garage area using the PLR and FLAR methodS

The plotted results showed that the fuzzy linear regression model based on possibillistic methods includes all the observations of the sale price which is a result of the wider spreads as indicated by the light blue shaded region. The purple dotted lines define the fuzzy linear regression model based on the FLAR methods indicating small spreads compared to other models.

4.4 Evaluating the fuzzy linear regression models

Goodness of fit measure was used to evaluate how fit the models were when using the PLRLS, PLR and FLAR methods.

Based on the non-zero spreads in some variables and the number of observations that were used in the analysis, goodness of fit (GOF) measure was suitable to evaluate these models.

Table 4.12 and Table 4.13 indicated the GOF and TEF values which were used for evaluation.

Table 4.12: Evaluation summary of the FLR model using one variable.

		Real numbers	symmetric numbers		Asymmetric numbers
		PLRLS	PLR	FLAR	FLAR
L.A	GOF	60204601702	60476872415	9353241672	9350184563
	TEF	9.440e+13	9582.3	1067.72	1087.63
T.B	GOF	49311080120	49580636932	7290781522	7288068902
	TEF	8.670e+13	8804.4	1069.97	1089.26
G.A	GOF	49489585440	49864900581	6206571194	6208347958
	TEF	8.648e+13	8769.33	1029.54	1050.95

Table 4.13: Evaluation summary of the FLR model using fuzzy regression methods.

	Real numbers	Symmetric numbers		Asymmetric numbers
	PLRLS	PLR	FLAR	FLAR
GOF	36118147000	33104827755	5086862808	5086385724
TEF	7.172498e+13	6964.94	1016.92	1040.23

Comparing the goodness of fit measure of the models using the alternative fuzzy regression methods shows that the model fitted using the FLAR method had a lower measure of goodness of fit (G) thus fitting the data better.

Evaluating these models indicated that both the total error of fit (TEF) and GOF are in agreement that the fitted model using the FLAR method fits the data better.

CHAPTER FIVE

SUMMARY, CONCLUSION AND RECOMMENDATIONS

5.1 Introduction

This chapter presents the summary, conclusion and the recommendations based on the results of the study.

5.2 Summary

Modeling sale price of residential properties using fuzzy regression analysis focused on determining the performance of the fuzzy regression methods in data analysis. The study used secondary data comprising 600 observations of the residential properties. Quantitative research method was used to determine the variables predicting the price of residential properties. Diagnostic tests were done to test if the variables satisfied fuzzy regression assumptions and the R software was used to analyze the data. Based on the study findings, fuzzy least squares methods fit fuzzy linear regression models better than the possibilistic linear regression methods.

5.3 Conclusion

The fuzzy regression methods that have been used indicate that the possibilistic linear regression (PLR) method gives a clear explanation of all the values that are used in the study in determining the price of individual residential properties. The fuzzy least absolute residual (FLAR) method provided a statistical view that can be used in determining the fuzziness in the given data. Also, total error fit (TEF) and goodness of fit measure of the FLAR method are minimal than the PLR and the (PLRLS) methods. This indicates that models based on FLAR methods are efficient.

5.4 Recommendations

The following are the recommendations for this research:

(i) Diagnostic tests to be performed on any given data set to determine the model to be used when fitting the data.

(ii) Fuzzy linear regression models to be used to fit a given data that assumes classical regression assumptions.

(iii) Fuzzy linear regression models are efficient when fitting cross sectional data.

For further research, the study recommends use of fuzzy linear regression analysis to fit longitudinal data.

References

Al-Kandari, M., Adjenughwure, K., and Papadopoulos, K. (2020). A fuzzy-statistical tolerance interval from residuals of crisp linear regression models. *Mathematics*, 8(9):1422.

Arkes, J. (2019). *Regression analysis: A practical introduction*, volume 49 of *984*. Taylor and Francis.

Ayinde, K., Apata, E. O., and Alaba, O. O. (2012). Estimators of linear regression model and prediction under some assumptions violation. *Open Journal of Statistics*, 2(5):534–546.

Bhavyashree, S., Mishra, M., and Girisha, G. (2017). Fuzzy regression and multiple linear regression models for predicting mulberry leaf yield: A comparative study. *International Journal of Agricutural and Statistical Sciences*, 13(1):149–152.

Chen, K., Rys, M., and Lee, E. S. (2006). Modeling of thermal comfort in air conditioned rooms by fuzzy regression analysis. *Mathematical and computer modelling*, 43(7-8):809–819.

Chukhrova, N. and Johannssen, A. (2019). Fuzzy regression analysis: systematic review and bibliography. *Applied Soft Computing*, 84:1–29.

Diamond, P. (1988). Fuzzy least squares. *Information Sciences*, 46(3):141–157.

Ernst, A. F. and Albers, C. J. (2017). Regression assumptions in clinical psychology research practice—a systematic review of common misconceptions. *PeerJ*, 5:1–16.

Gogtay, N., Deshpande, S., and Thatte, U. (2017). Principles of regression analysis. *Journal of the Association of Physicians of India*, 65(48):48–52.

Hickey, G. L., Kontopantelis, E., Takkenberg, J. J., and Beyersdorf, F. (2019). Statistical primer: checking model assumptions with regression diagnostics. *Interactive cardiovascular and thoracic surgery*, 28(1):1–8.

Khademi, F., Akbari, M., Jamal, S. M., and Nikoo, M. (2017). Multiple linear regression, artificial neural network, and fuzzy logic prediction of 28 days compressive strength of concrete. *Frontiers of Structural and Civil Engineering*, 11:90–99.

Kropat, E., Özmen, A., Weber, G.-W., Meyer-Nieberg, S., and Defterli, O. (2016). Fuzzy prediction strategies for gene-environment networks–fuzzy regression analysis for two-modal regulatory systems. *RAIRO-Operations Research-Recherche Opérationnelle*, 50(2):413–435.

Kumari, K. and Yadav (2018). Linear regression analysis study. *Journal of the practice of Cardiovascular Sciences*, 4(1):33.

Lunt, M. (2015). Introduction to statistical modelling: linear regression. *Rheumatology*, 54(7):1137–1140.

Marchionni, V., Lopes, N., Mamouros, L., and Covas, D. (2014). Modelling sewer systems costs with multiple linear regression. *Water resources management*, 28:4415–4431.

Montgomery, D. C., Peck, E. A., and Vining, G. G. (2021). *Introduction to linear regression analysis*. John Wiley & Sons.

Nazari-Shirkouhi, S. and Keramati, A. (2017). Modeling customer satisfaction with new product design using a flexible fuzzy regression-data envelopment analysis algorithm. *Applied Mathematical Modelling*, 50:755–771.

Pandelara, D., Kristjanpoller, W., Michell, K., and Minutolo, M. C. (2022). A fuzzy regression causality approach to analyze relationship between electrical consumption and gdp. *Energy*, 239:122459.

Pandit, P., Dey, P., and Krishnamurthy, K. (2021). Comparative assessment of multiple linear regression and fuzzy linear regression models. *SN Computer Science*, 2(2):1–8.

Prieto, A. J., Silva, A., de Brito, J., Macías-Bernal, J. M., and Alejandre, F. J. (2017). Multiple linear regression and fuzzy logic models applied to the functional service life prediction of cultural heritage. *Journal of Cultural Heritage*, 27:20–35.

Sarip, A. G., Hafez, M. B., and Daud, M. N. (2016). Application of fuzzy regression model for real estate price prediction. *Malaysian Journal of Computer Science*, 29(1):15–27.

Schneider, A., Hommel, G., and Blettner, M. (2010). Linear regression analysis: part 14 of a series on evaluation of scientific publications. *Deutsches Ärzteblatt International*, 107(44):776–782.

Selçuk Öğüt, K. (2006). Modeling car ownership in turkey using fuzzy regression. *Transportation planning and technology*, 29(3):233–248.

Škrabánek, P. and Martínková, N. (2021). Algorithm 1017: fuzzyreg: An r package for fitting fuzzy regression models. *ACM Transactions on Mathematical Software (TOMS)*, 47(3):1–18.

Sorkheh, K., Kazemifard, A., and Rajabpoor, S. (2018). A comparative study of fuzzy linear regression and multiple linear regression in agricultural studies: a case study of lentil yield management. *Turkish Journal of Agriculture and Forestry*, 42(6):402–411.

Tanaka, H., Hayashi, I., and Watada, J. (1989). Possibilistic linear regression analysis for fuzzy data. *European Journal of Operational Research*, 40(3):389–396.

Uyanık, G. K. and Güler, N. (2013). A study on multiple linear regression analysis. *Procedia-Social and Behavioral Sciences*, 106:234–240.

Zhou, J., Zhang, H., Gu, Y., and Pantelous, A. A. (2018). Affordable levels of house prices using fuzzy linear regression analysis: the case of shanghai. *Soft Computing*, 22(16):5407–5418.

APPENDIX

STUDY CODES

```
D1<-as.data.frame(D)

D1

library("fuzzyreg")

v<-fuzzylm(Price Lot.area+Total.basement.sf+Garage.area,data=D1,method="PLRLS")

v

GOF(v)

TEF(v)

summary(v)

symmetric fuzzy numbers

F1<-fuzzify(x=dat1price,method="err",err=rnorm(600,mean(10000),sd=500))

F1

F1a<-F1[,-4]

F1a

head(F1a,7)

F2<-cbind(F1a,Lot.area,Total.basement.sf,Garage.area)

F2

head(F2,7)

F3<-fuzzylm(Price Lot.area+Total.basement.sf+Garage.area,data=F2,method="plr",fuzzy.left.y="Al")

F3

GOF(F3)

TEF(F3)

summary(F3)

F3a<-fuzzylm(Price Lot.area+Total.basement.sf+Garage.area,data=F2,method="flar",fuzzy.left.y="Al",fuzzy.
= "Ar")
```

F3a

GOF(F3a)

TEF(F3a)

summary(F3a)

plot((F3a))

non-symmetric fuzzy numbers

Pricel<-runif(600,min=10000,max=12000) simulating the left spread

Pricer<-runif(600,min=5000,max=8000) simulating the right spread

Dtt<-cbind(Price,Pricel,Pricer) binding the central value and the spreads

Dtt (non symmetric)

Dttt<-as.data.frame(Dtt)

Dttt

head(Dttt,7)

F4<-cbind(Dttt,Lot.area,Total.basement.sf,Garage.area)

F4

head(F4,7)

F5<-fuzzylm(Price Lot.area+Total.basement.sf+Garage.area,data=F4,method="flar",fuzzy.left.y="Pricel",fuzz

= "Pricer")

F5

Fitting by the use of each independent variable in FLR

using fuzzified spreads by the error method

Lot Area

A1<-fuzzylm(Price Lot.area,data=D1,method="PLRLS")

A1

plot(A1,res=30,"lightblue",main="")

TEF(A1)

GOF(A1)

summary(A1)

```
Flar
A4<-cbind(Dttt,Lot.area)
head(A4,5)
A5<-fuzzylm(Price Lot.area,data=A4,method="flar",fuzzy.left.y="Pricel",fuzzy.right.y = "Pricer")
A5
plot(A5,res=30,col="orange",main="")
summary(A5)
plr
A6<-cbind(F1a,Lot.area)
head(A6,6)
A7<-fuzzylm(Price Lot.area,data=A6,method="plr",fuzzy.left.y="Al")
A7
plot(A7,res=30,col="lightblue",main="")
GOF(A7)
TEF(A7)
summary(A7)
a7<-fuzzylm(Price Lot.area,data=A6,method="flar",fuzzy.left.y="Al",fuzzy.right.y ="Ar" )
a7
abline(coef(a7)[, 1], col = "red")
abline(coef(a7)[, 2], col = "red", lty = 2)
abline(coef(a7)[, 3], col ="red", lty = 2)
GOF(a7)
TEF(a7)
summary(a7)
```

REPUBLIC OF KENYA

NATIONAL COMMISSION FOR SCIENCE,TECHNOLOGY & INNOVATION

Ref No: **888443**

Date of Issue: **10/October/2023**

RESEARCH LICENSE

This is to Certify that Miss.. celestina bosibori moturi of Kirinyaga University, has been licensed to conduct research as per the provision of the Science, Technology and Innovation Act, 2013 (Rev.2014) in Kirinyaga on the topic: modeling cross sectional data of the sale price of residential properties in AMES using fuzzy regression analysis for the period ending : 10/October/2024.

License No: **NACOSTI/P/23/30211**

888443

Applicant Identification Number

Director General
NATIONAL COMMISSION FOR SCIENCE,TECHNOLOGY & INNOVATION

Verification QR Code

NOTE: This is a computer generated License. To verify the authenticity of this document, Scan the QR Code using QR scanner application.

See overleaf for conditions

THE SCIENCE, TECHNOLOGY AND INNOVATION ACT, 2013 (Rev. 2014)

Legal Notice No. 108: The Science, Technology and Innovation (Research Licensing) Regulations, 2014

The National Commission for Science, Technology and Innovation, hereafter referred to as the Commission, was the established under the Science, Technology and Innovation Act 2013 (Revised 2014) herein after referred to as the Act. The objective of the Commission shall be to regulate and assure quality in the science, technology and innovation sector and advise the Government in matters related thereto.

CONDITIONS OF THE RESEARCH LICENSE

1. The License is granted subject to provisions of the Constitution of Kenya, the Science, Technology and Innovation Act, and other relevant laws, policies and regulations. Accordingly, the licensee shall adhere to such procedures, standards, code of ethics and guidelines as may be prescribed by regulations made under the Act, or prescribed by provisions of International treaties of which Kenya is a signatory to
2. The research and its related activities as well as outcomes shall be beneficial to the country and shall not in any way;
    - i. Endanger national security
    - ii. Adversely affect the lives of Kenyans
    - iii. Be in contravention of Kenya's international obligations including Biological Weapons Convention (BWC), Comprehensive Nuclear-Test-Ban Treaty Organization (CTBTO), Chemical, Biological, Radiological and Nuclear (CBRN).
    - iv. Result in exploitation of intellectual property rights of communities in Kenya
    - v. Adversely affect the environment
    - vi. Adversely affect the rights of communities
    - vii. Endanger public safety and national cohesion
    - viii. Plagiarize someone else's work
3. The License is valid for the proposed research, location and specified period.
4. The license any rights thereunder are non-transferable
5. The Commission reserves the right to cancel the research at any time during the research period if in the opinion of the Commission the research is not implemented in conformity with the provisions of the Act or any other written law.
6. The Licensee shall inform the relevant County Director of Education, County Commissioner and County Governor before commencement of the research.
7. Excavation, filming, movement, and collection of specimens are subject to further necessary clearance from relevant Government Agencies.
8. The License does not give authority to transfer research materials.
9. The Commission may monitor and evaluate the licensed research project for the purpose of assessing and evaluating compliance with the conditions of the License.
10. The Licensee shall submit one hard copy, and upload a soft copy of their final report (thesis) onto a platform designated by the Commission within one year of completion of the research.
11. The Commission reserves the right to modify the conditions of the License including cancellation without prior notice.
12. Research, findings and information regarding research systems shall be stored or disseminated, utilized or applied in such a manner as may be prescribed by the Commission from time to time.
13. The Licensee shall disclose to the Commission, the relevant Institutional Scientific and Ethical Review Committee, and the relevant national agencies any inventions and discoveries that are of National strategic importance.
14. The Commission shall have powers to acquire from any person the right in, or to, any scientific innovation, invention or patent of strategic importance to the country.
15. Relevant Institutional Scientific and Ethical Review Committee shall monitor and evaluate the research periodically, and make a report of its findings to the Commission for necessary action.

National Commission for Science, Technology and
Innovation(NACOSTI),
Off Waiyaki Way, Upper Kabete,
P. O. Box 30623 - 00100 Nairobi, KENYA
Telephone: 020 4007000, 0713788787, 0735404245
E-mail: dg@nacosti.go.ke
Website: www.nacosti.go.ke