DOKUZ EYLÜL UNIVERSITY
GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

FUZZY LINEAR REGRESSION

by
Bekir ÇETİNTAV

August, 2012
İZMİR

FUZZY LINEAR REGRESSION

A Thesis Submitted to the
Graduate School of Natural and Applied Sciences of Dokuz Eylül University
In Partial Fulfillment of the Requirements for
the Degree of Master of Science in Statistics

by
Bekir ÇETİNTAV

August, 2012
İZMİR

M.Sc. THESIS EXAMINATION RESULT FORM

We have read the thesis entitled **"FUZZY LINEAR REGRESSION"** completed by **BEKİR ÇETİNTAV** under supervision of **ASSIST. PROF. Dr. A.FIRAT ÖZDEMİR** and we certify that in our opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

Assist. Prof. Dr. A.Fırat ÖZDEMİR

Supervisor

Prof. Dr. Efendi NASIBOĞLU

(Jury Member)

Assist. Prof. Dr. Ilhan KARAKILIÇ

(Jury Member)

Prof. Dr. Mustafa SABUNCU

Director

Graduate School of Natural and Applied Sciences

ACKNOWLEDGMENTS

I am grateful to Prof. Dr. Serdar KURT, my first supervisor, and Assist. Prof. Dr. A.Fırat ÖZDEMİR, my second supervisor, for their encouragement and insight throughout my research and for guiding me through the entire thesis process from start to finish. Has it not been for their faith in me, I would not have been able to finish the thesis.

I'm also grateful for the insights and efforts put forth by the examining committee; Prof. Dr. Efendi NASIBOĞLU and Assist. Prof. Ilhan KARAKILIÇ .

I wish to give a heartfelt thanks to my father, Savaş ÇETİNTAV, and my mother, Bahriye ÇETİNTAV. They offered their endless support and constant prayers. This academic journey would not have been possible without their love, patience, and sacrifices along the way.

Bekir ÇETİNTAV

FUZZY LINEAR REGRESSION

ABSTRACT

Linear Programming (LP) methods are commonly used to construct Fuzzy Linear Regression (FLR) models. Probabilistic Fuzzy Linear Regression (PFLR) (Tanaka, 1989) and Unrestricted Fuzzy Linear Regression (UFLR) (Lee and Chang, 1994) are two of the mostly applied models that employ LP methods. In this study, commonly used models which employ LP methods are given. Also a new modified fuzzy linear regression model which use LP methods is proposed. Proposed model divides total vagueness into two parts as explained and unexplained. It tries to minimize only the explained vagueness not the unexplained one. Four numerical applications with four different data sets were performed in which PFLR, UFLR and proposed model were compared in terms of mean squared error (MSE) and total fuzziness and it is concluded the proposed model is acceptable.

Keywords : fuzzy linear regression (FLR), linear programming (LP) methods for fuzzy regression.

BULANIK DOĞRUSAL REGRESYON

ÖZ

Bu Doğrusal Programlama (LP) yöntemleri Bulanık Doğrusal Regresyon (FLR) modellerinin kurulmasında sıkça kullanılmaktadır. Olasılıksal Bulanık Doğrusal Regresyon (PFLR) (Tanaka, 1989) ve Sınırlanmamış Bulanık Doğrusal Regresyon (UFLR) (Lee and Chang, 1994) modelleri en sık uygulanan modellerden ikisidir. Bu çalışmada doğrusal programlamayla çalışan modellerin en sık kullanılanlarına yer verilmiştir. Ayrıca modifiye edilmiş ve doğrusal programlamayla çalışan yeni bir bulanık doğrusal regresyon modeli önerilmiştir. Önerilen model bualnıklığı açıklanan ve açıklanamayan olmak üzere ikiye bölmektedir ve sadece açıklanan bulanıklığı minimize etmeye çalışmaktadır. Hata kareler ortalaması ve toplam bulanıklık açısından PFLR, UFLR ve yeni önerilen modeli karşılaştırmak için dört ayrı veri setiiyle dört ayrı uygulama yapılmıştır ve sonuçlar kabul edilebilir bulunmuştur.

Anahtar sözcükler : bulanık doğrusal regresyon, bulanık regresyonda doğrusal programla ile çalışan yöntemler.

CONTENTS

Page

M.Sc. THESIS EXAMINATION RESULT FORM iii

ACKNOWLEDGMENTS ii

ABSTRACT iv

ÖZ v

CHAPTER ONE - INTRODUCTION 1

CHAPTER TWO - BASIC PRINCIPLES OF REGRESSION ANALYSIS AND FUZZY SET THEORY 3

2.1 Fuzzy Set Theory 3

2.1.1 Fuzzy sets and Classical-Crisp sets 3

2.1.2 Membership function 4

2.1.3 Operations with fuzzy numbers 6

2.1.3.1 Arithmetic operations 7

2.1.3.2 Fuzzy-set operations 7

2.1.4 Fuzzification and defuzzification 8

2.2 Linear Regression Analysis 9

2.2.1 Linear regression models 9

2.2.2 Parameter estimation 10

2.2.3 Analyzing linear regression model 11

CHAPTER THREE - FUZZY LINEAR REGRESSION 13

3.1 Introduction to Fuzzy Linear Regression 13

3.2 LP Methods for FLR 14

3.2.1 Tanaka’s First Model 15

3.2.2 Tanaka’s Second Model (Possibilistic Fuzzy Linear Regression-PFLR).. 16

3.2.3 Peters’ Model 18

3.2.4 Lee and Chang’s Model (Unrestricted in Sign Fuzzy Linear Regression-UFLR) 19

3.2.3 Chen’s Model... 20

3.1.5.1 Detection of Abnormal Data 20

3.1.5.2 Determination the Expected Number of Outliers................................ 21

3.1.5.1 Modifying the Outlier and Redrawing the Model 21

CHAPTER FOUR - PROPOSED NEW MODEL 24

4.1 Fair Fuzzy Linear Regression (FFLR) 24

4.2 Some remarks on FFLR Model 25

CHAPTER FIVE - APPLICATION 27

5.1 Example 1 27

5.2 Example 2 30

5.2 Example 3 33

5.2 Example 4 38

CHAPTER SIX - CONCLUSION 40

REFERENCES 43

APPENDIX 1 - GENERALIZED HYPERBOLIC DISTRIBUTIONS WITH R……………………………………………………………………………..46

CHAPTER ONE
INTRODUCTION

An indispensable part of human nature is to comprehend the objects and events; shortly we can call units. He needs to measure some characteristics of those units by hand, eye or improved tools and the values are obtained by that measuring are called data. Usually numbers of these characteristics are more than one. So a group of characteristics which are related with each other occurs and that relation is not easy to measure. Therefore special methods are improved in science of Statistics. The characteristics are called variables and one of the methods of estimating the relation between variables is called Regression Analysis (RA).

RA is a commonly used methodology for analyzing relationships and correlations between a response variable, also called dependent variable, and one or more explanatory variables, independent variables. For example, the relation between a student's amount of study (hour) and exam point could be analyzed by RA. Beyond the correlation analysis, RA also has prediction capability. For previous example; a researcher could predict a student's specific exam point if he knows how many hours the student studied for that exam and past information for these two variables.

RA has many kinds of sub-models for different cases. For instance, there are several models like linear, quadratic, cubic, polynomial, etc. for different shapes of relation; piecewise, spline, etc. for conflicting trends of relation or robust, logistic, ridge, etc. for different cases of data. Based on this it could be said that RA is a powerful and "temporizing" methodology for data analysis.

Statistics is science of data and all statistical methods, including RA, operate with data. In other words the statistical methods make inferences based on existing data. So the form of data is very important. As well as the existing data have classical-precise form, they may have uncertain-fuzzy form. For previous example; a student may study for a specific exam exactly four hour or he may study "hard" or "too much". First one is crisp data and it could be used by classical methods but second one is uncertain and a different concept is needed for it. That concept is called Fuzzy Set Theory.

The Fuzzy Set Theory was first presented in 1965 by Lotfi Zadeh to deal with approximate reasoning and imprecise-uncertain knowledge. It has been developed for different types of analysis methods, such as fuzzy numbers, fuzzy relations and fuzzy inference systems. This theory has been adapted to topics of various sciences and combined with various methods. One of these sciences is Statistics and one of the methods is RA.

In this study, the main aim is to study the combination of RA methods and Fuzzy Set Theory. This combination is called Fuzzy Regression (FR). In FR, regression analysis is implemented for fuzzy spaces. Basic principles of RA and Fuzzy Set Theory were given in chapter 2. Fuzzy Linear Regression (FLR) methods and Linear Programming (LP) approach was given in chapter 3. A new approach for FLR based on LP method is introduced with several numerical applications in chapter 4. And Chapter 5 includes conclusions.

CHAPTER TWO
BASIC PRINCIPLES OF REGRESSION ANALYSIS AND FUZZY SET THEORY

2.1 Fuzzy Set Theory

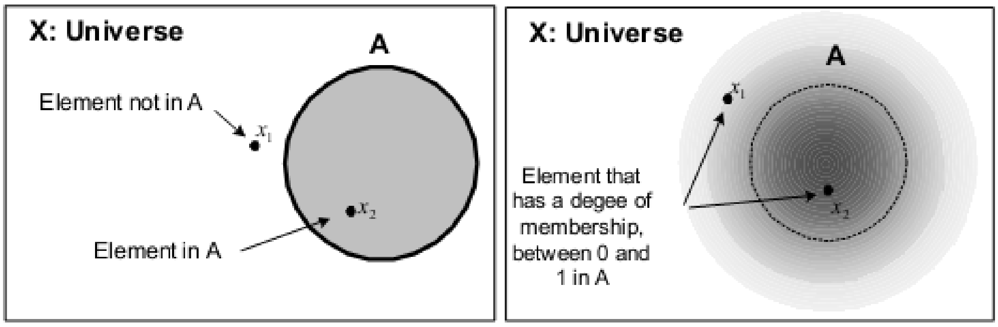
The concept of fuzzy sets and approximate reasoning was first introduced by Professor Lofti Zadeh at the University of California in 1965. Since this introduction, it has been used in many areas of different kinds of science, especially in engineering. This theory is a branch of a set theory that is useful for the representation of imprecise knowledge of the type that is prevalent in human concept formation and reasoning because fuzzy theory can represent a type of uncertainty due to vagueness or fuzziness (Yager 1986, Yen and Langari 1998).

The main difference of fuzzy sets is their uncertain or vague boundaries. Because, the opposite of classical-crisp sets, fuzzy sets have more flexible sense of membership. In this chapter, fuzzy sets are described and compared with classical sets. The concept of fuzzy membership function is defined and a group of frequently-used membership functions are given. Some fuzzy operations, fuzzification and defuzzification methods are given, too.

2.1.1 Fuzzy sets and Classical-Crisp sets

A classic set can be defined as a collection of objects in a given domain. That means an object should either belong to the set or not belong to the set. There is a sharp boundary between members of the set and those are not in the set. So the concept of classical-crisp sets can be defined as “0 or 1” or “black or white” sets. An object should either completely belongs to the set or does not belong to the set at all.

However the events, concepts or memberships are not always so sharp like “black or white”. There are sometimes grey zones. For example; two cases of sets are dealt. The first one is the sets of married and unmarried people and the second one is sets of happy and unhappy couples. Memberships of the first one are crisp because a person is either married or not. What about the second one? If a question of “are you happy?” is asked to couples, there will be kinds of answers as “very happy”, “too happy”, “happy”, “so so”, “unhappy”, “I regret to be married” and etc. for the concept of classical set, “very happy”, “too happy”, “happy”, “so so” (and also similar answers) will be equal members of the set of happy. Starting from this point, the question should be asked: is the happiness of couples whose answers are “too happy” equal with couples whose answers are “so so”?


X: Universe
A
Element not in A
x1
x2
Element in A
X: Universe
A
x1
Element that has a degee of membership, between 0 and 1 in A
x2

(a) classic set

(b) a fuzzy set

Figure 2.1 Presentation of Classical set and Fuzzy set (Dongmin Lee, 2006)

A fuzzy set is as a set with un-sharp and vague boundaries. It generalizes the notion of membership from a “0 or 1” or “black or white” binary categorization in classical set theory into one that allows partial membership. So it includes the grey zone. Fuzzy set theory can overcome the limitations of the classical set theory by allowing membership in a set to be a matter of degree. The degree of membership in a fuzzy set is represented by a number between 0 and 1; 0 means entirely not in the set, 1 means completely in the set and there are infinite number of membership degrees between 0 and 1.

Prima facie, fuzzy set concept is very similar to probability concept. Even though they have some similarities, they are different tools with different logics. Probability measures “likelihood of occurrence.” This probability is related to the following question, “How often or frequently does it happen?” While a fuzzy set measure “the degree of certainty” and is related to following question, “How sure are you that it happens?”(Dongmin Lee, 2006).

2.1.2 Membership function

Membership function is the major element of the fuzzy set theory because it allows the fuzzy approach to evaluate uncertain and ambiguous cases. The main role of the membership function is to represent a human perception, which is usually individual and subjective, as a member of a fuzzy set.

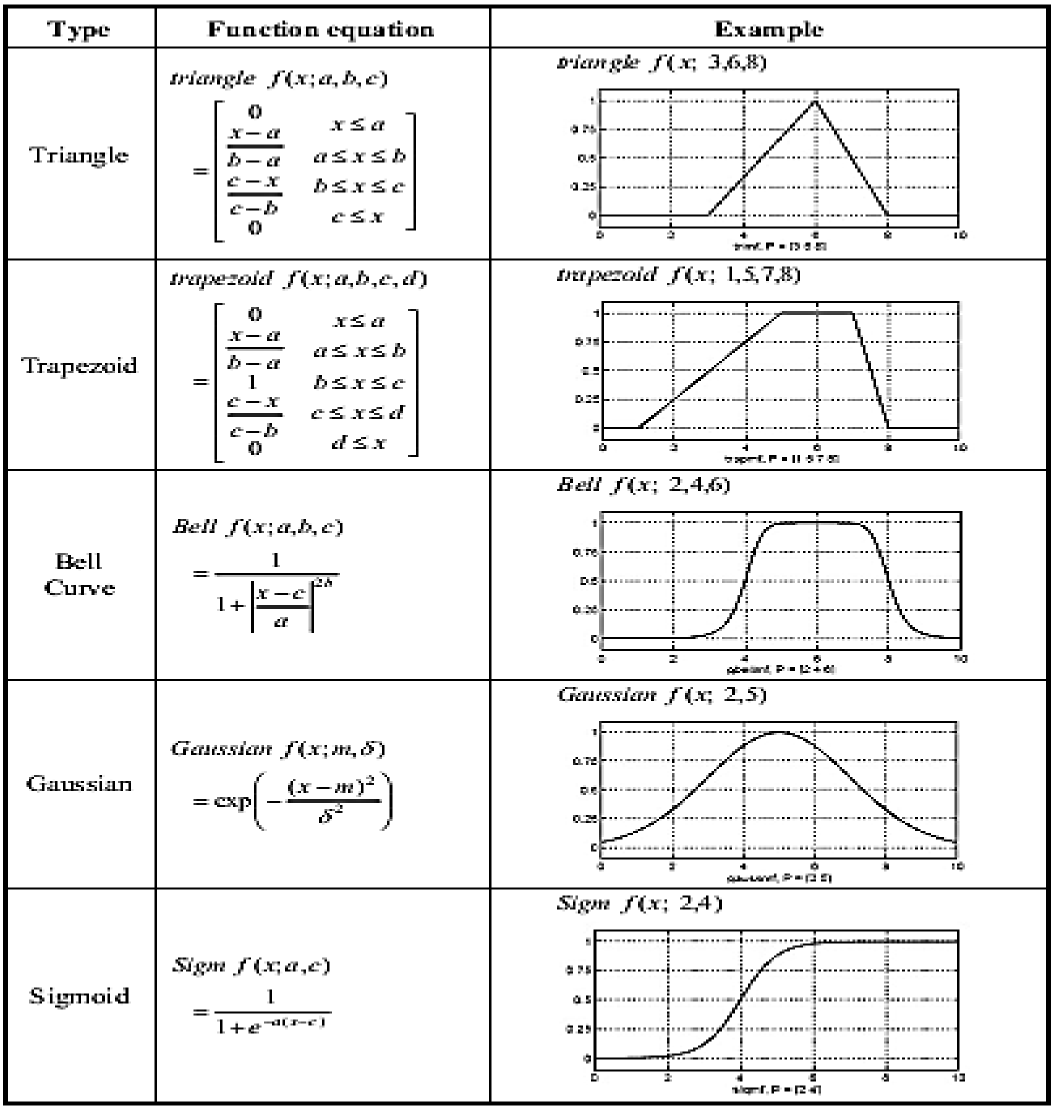
Fuzzy set theory allows membership in a set to be a matter of degree. That means an element could be in a specific fuzzy set with a membership degree between 0 and 1. “0” means that element doesn’t belong to that set and “1” means it is completely in the set. So the element has a membership degree which is depending on the degree of belonging to the set.

$$\mu_A(y) : x \rightarrow [0,1] \tag{2.1}$$

The characteristics vary from set to set so naturally the membership functions vary from set to set. Therefore the prior task in an analysis is to determine the optimal membership function. Also, one of the most difficult tasks for applying fuzzy sets is to correctly measure it.

There are numerous types of fuzzy membership functions including triangular, trapezoidal, bell-shaped, S-shaped, Gaussian, sigmoid and etc. the most commonly used in practice are triangular, trapezoidal, bell curves, Gaussian, and sigmoid functions. This is because they are easy to use for arithmetic and fuzzy operations.

Figure below describes the formulas and parameters of each membership function. It also shows examples of each. As can be seen below there are parameters which control the exact shape of the membership function and also the function values. There are 2 parameters for Gaussian and sigmoid functions, three for the triangular and bell-shape curve functions and four for the trapezoidal function. So these parameters have a major role to determine appropriate membership function.



Type	Function equation	Example
Triangle	$triangle\ f(x;a,b,c)$ $= \begin{bmatrix} 0 & x \le a \\ \frac{x-a}{b-a} & a \le x \le b \\ \frac{c-x}{c-b} & b \le x \le c \\ 0 & c \le x \end{bmatrix}$	$triangle\ f(x;\ 3,6,8)$
Trapezoid	$trapezoid\ f(x;a,b,c,d)$ $= \begin{bmatrix} 0 & x \le a \\ \frac{x-a}{b-a} & a \le x \le b \\ 1 & b \le x \le c \\ \frac{c-x}{c-b} & c \le x \le d \\ 0 & d \le x \end{bmatrix}$	$trapezoid\ f(x;\ 1,5,7,8)$
Bell Curve	$Bell\ f(x;a,b,c)$ $= \frac{1}{1+\left\lvert \frac{x-c}{a} \right\rvert^{2b}}$	$Bell\ f(x;\ 2,4,6)$
Gaussian	$Gaussian\ f(x;m,\delta)$ $= \exp\left(-\frac{(x-m)^2}{\delta^2}\right)$	$Gaussian\ f(x;\ 2,5)$
Sigmoid	$Sigm\ f(x;a,c)$ $= \frac{1}{1+e^{-a(x-c)}}$	$Sigm\ f(x;\ 2,4)$

Figure 2.2 Describing the formulas and parameters of each membership function (Dongmin Lee, 2006)

2.1.3 Operations with fuzzy numbers

There are two basic sets of operations with fuzzy numbers, arithmetic and fuzzy-set operations. Arithmetic operations are inverse, addition, subtraction, multiplication, division and etc. Fuzzy-set operations are union, intersection, complement, Cartesian product and etc. In this part, triangular membership functions are used for example because easy to use and commonly-used for FLR, the main subject.

2.1.3.1 Arithmetic operations

Suppose there are two triangular fuzzy numbers, $Y_i = (l_i, y_i, r_i)$ and $X_j = (l_j, x_j, r_j)$

Inverse : $Y_i^{-1} = (y_i^{-2} l_i,\ \ y_i^{-1}, y_i^{-2} r_i)$ (2.2)

Addition : $Y_i + X_j = (l_i + l_j, y_i + x_j, r_i + r_j)$ (2.3)

Subtraction : $Y_i - X_j = (l_i + l_j, y_i - x_j, r_i + r_j)$ (2.4)

Multiplication with constant : $a * Y_i = (a * l_i, a * y_i, a * r_i)$ for $a > 0$ and $a \in R$

$a * Y_i = (-a * l_i, a * y_i, -a * r_i)$ for $a < 0$ and $a \in R$ (2.5)

2.1.3.2 Fuzzy-set operations

Like the other notions, all set operations are redefined for (or extended to) the case of fuzzy concept. These operations are different from classical sets for basic concept. Also another major difference is there could be various kinds of operators for one specific operation. For instance, there are standard (Zadeh's), probabilistic, bounded, drastic, Yager's, Hamacher's and etc. products (or intersections) for fuzzy intersection operation. Although it makes a bit confusion selecting appropriate one, it provides practicality and flexibility in data analysis.

To give an idea, Zadeh's standard forms are given below.

Union (OR): The membership function $\mu_{A \cup B}(y)$ of the union $A \cup B$ is point-wise defined for all $x \in U$ (universal set) by

$$\mu_{A \cup B}(y) = \max\{\mu_A(y),\ \mu_B(y)\} \qquad (2.6)$$

Intersection (AND): The membership function $\mu_{A \cap B}(y)$ of the intersection $A \cap B$ is point-wise defined for all $x \in U$ (universal set) by

$$\mu_{A \cap B}(y) = \min\{\mu_A(y),\ \mu_B(y)\} \qquad (2.7)$$

Complement: The membership function $\mu_{\bar{A}}(y)$ of the complement $\bar{A}$ is point-wise defined for all $x \in U$(universal set) by

$$\mu_{\bar{A}}(y) = 1 - \mu_A(y) \qquad (2.8)$$

2.1.4 Fuzzification and defuzzification

Fuzzification is the process of transforming crisp input values into fuzzy values by using membership functions which are appropriate for the data sets. As stated before, it is very important determining the optimal membership function for the data set.

Defuzzification, as its name implies, reverse process of fuzzification. It is the process of transforming fuzzy values into crisp values by special methods as center of area (COA), bisector, middle of maximum (MOM) and etc. There are also many kinds of defuzzification methods. To give an idea, five methods which are supported in the MATLAB Fuzzy Logic Toolbox™ are given.

Center of area (COA): it is also known Centroid method. Centroid defuzzification returns the center of area under the curve. If the area is thought of as a plate of equal density, the centroid is the point along the x axis about which this shape would balance.

$$y_{COA}^{*} = \frac{\int \mu_A(y)\, y\, dy}{\int \mu_A(y)\, dy} \tag{2.9}$$

Bisector: The bisector is the vertical line that will divide the region into two sub-regions of equal area. It is sometimes, but not always coincident with the centroid line.

Middle, Smallest, and Largest of Maximum (MOM, SOM, and LOM): MOM, SOM, and LOM stand for Middle, Smallest, and Largest of Maximum, respectively. These three methods key off the maximum value assumed by the aggregate membership function. In this example, because there is a plateau at the maximum value, they are distinct. If the aggregate membership function has a unique maximum, then MOM, SOM, and LOM all take on the same value.

2.2 Linear Regression Analysis

Regression Analysis is a commonly-used statistical methodology for analyzing relationships and correlations between two or more variables so that one variable can be predicted from the other or others. That relation is not a functional relation, a statistical relation.

Functional relation, it is also called deterministic relation, is a perfect relation where all observations fall directly on the line of functional relationship. A statistical relation, unlike functional relation, is not a perfect one. The observations do not fall directly on the line (or curve) of relationship. (It is also called probabilistic relation.)

A regression model is a formal means of expressing the two essential ingredients of a statistical relation:

1. A tendency of the response variable Y to vary with the predictor variable(s) X in a systematic fashion.
2. A scattering of points around the curve of a statistical relationship.

These two characteristics are embodied in a regression model by postulating that:

1. There is a probability distribution of Y for each level of X.
2. The means of these probability distributions vary in some systematic fashion with X.(J. Neter and at al. Applied linear regression models, Third edition,Irwin,1996)

2.2.1 Linear regression models

Practical applications of regression analysis utilize models that have one response (independent) and one or more predictor (dependent) variables. It is called simple regression model when there is only one predictor is related with response variable. When there are more than one predictor variables, it is called multiple regression model. And if these relations between response and predictors are linear, also it is called first-order, it becomes a linear model. So the General Linear Regression Model is,

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_i X_{ij} + \varepsilon_i \tag{2.10}$$

Where

Y_i is the vector of dependent variable,

X_{ij} are vectors of independent variables,

β_i determines the contribution of the independent variable X_{ij},

ε_j is the error term (random error).

Also there are some assumptions about random error ε_j.(McClave)

1. The mean is equal to 0
2. The variance is equal to σ^2
3. The probability distribution is a normal distribution
4. Random errors are independent (in a probabilistic sense).

2.2.2 Parameter estimation

There are several methods for parameter estimation in linear regression models. These are also methods of Statistical Inference. The most common ones are Least Square Estimation (LSE) and Maximum Likelihood Estimation (MLE).A general and brief information is given below.

Least Square Estimation (LSE): The method of least squares is about estimating parameters by minimizing the squared discrepancies between observed data, on the one hand, and their expected values on the other. So it tries to minimize the following function. (For linear regression models and *j*= number of ind. variables)

$$Q = \sum_{i=1}^{n}(Y_i - \left(\beta_0 + \beta_i X_{ij}\right))^2 \tag{2.11}$$

For a brief illustration of parameter estimation, consider a straight-line (Simple Linear) regression model,

$$Y_i = \beta_0 + \beta_1 X_1 + \varepsilon_i. \tag{2.12}$$

For this model the least squares estimates of the parameters would be computed by minimizing following equation,

$$Q = \sum(Y_i - (\beta_0 + \beta_1 X_{i1}))^2. \tag{2.13}$$

Doing this by

1. Taking partial derivatives of Q with respect to β_0 and β_1,
2. Setting each partial derivative equal to zero
3. Solving the resulting model of two equations with two unknowns.

It yields the following estimators for the parameters:

$$\beta_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} \tag{2.14}$$

$$\beta_0 = \bar{y} - \beta_1 \bar{x}. \tag{2.15}$$

LSE, which unlike MLE requires no or minimal distributional assumptions, is useful for obtaining a descriptive measure for the purpose of summarizing observed data, but it has no basis for testing hypotheses or constructing confidence intervals. (I. J. Myung, 2003)

Maximum Likelihood Estimation (MLE): MLE is a standard approach to parameter estimation and inference in statistics. it makes the known likelihood distribution a maximum and has many optimal properties in estimation: sufficiency (complete information about the parameter of interest contained in its MLE estimator); consistency (true parameter value that generated the data recovered asymptotically, i.e. for data of sufficiently large samples); efficiency (lowest-possible variance of parameter estimates achieved asymptotically); and etc.

The joint density of n independent and identically distributed (i.i.d.) observations from this process is the product of the individual densities.

$$f(y_1, \dots \dots, y_n \backslash \theta) = \prod_{i=1}^{n} f(y_i \backslash \theta) = L(\theta \backslash y_i) \tag{2.16}$$

This joint density is the likelihood function, defined as a function of the unknown parameter vector, $\boldsymbol{\theta}$, where y is used to indicate the collection of sample data. In this classical estimation framework, the parameters are assumed to be fixed constants that we hope to learn about from the data.

Many of the inference methods in statistics are developed based on MLE. For example, MLE is a prerequisite for the chi-square test, the G-square test, Bayesian methods, inference with missing data, modeling of random effects, and many model selection criteria such as the Akaike information criterion and the Bayesian information criteria. (I. J. Myung, 2003)

2.2.3 Analyzing linear regression model

McClave suggests a stepwise procedure to analyze a linear regression model.

Step1: Hypothesize the deterministic component of the model. This component relates the mean $E(y)$ to the independent variables $x_1, x_2, \dots \dots, x_n$. Involved here is the choice of the independent variables to be included in the model.

Step2: Use the sample data to estimate the unknown parameters $\beta_0, \beta_1, \dots \dots, \beta_k$ in the model.

Step3: Specify the probability distribution of the random-error term, and estimate the standard deviation σ of this distribution.

Step4: Check that the assumptions about error term are satisfied, and make modifications to the model if necessary.

Step5: Statistically evaluate the usefulness of the model.

Step6: When you are satisfied that the model is useful, use it for prediction, estimation and other purposes.

CHAPTER THREE
FUZZY LINEAR REGRESSION

3.1 Introduction to Fuzzy Linear Regression

As stated before, Regression analysis is a commonly used methodology for analyzing relationships and correlations between a response variable, also called dependent variable, and one or more explanatory variables, independent variables. For example; a researcher could predict a student's a specific exam point if he knows how many hours the student studied for that exam and past information for these two variables. The Classical Linear Regression (CLR) model can be stated as follows:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_i X_{ij} + \varepsilon_i$$

$$for\ i = 1, \ldots, n(\#\ of\ obs.) \quad j = 1, \ldots, N\ (\#\ of\ ind.var.) \tag{3.1}$$

In CLR model; the deviation between the observed value and estimated value of dependent variable Y_i is generally regarded as *error* and that *error* is normally distributed with zero mean. That error is a kind of uncertainty and it may be result from kinds of things, such as wrong selected or inadequate independent variables, lack of fit and etc. There are also β_i parameters which present the magnitude of the independent variables' effect on the dependent variable. Several methods have been constructed for estimation of parameters. (The Least Square (LS) method is frequently used one)

After improvements of fuzzy set theory, it has been successfully demonstrated in many applications, such as: reliability, quality control, econometrics, engineering applications, etc. The common point of these different areas is that there are data with vagueness (or fuzzy data). So there are need some special tools for applications of these data. Because the original vagueness is not taken into account in the analysis when the fuzzy data is analyzed through nonfuzzy techniques and it makes the model inaccurate. Therefore Fuzzy Regression (FR) models have been constructed to restore regression analysis for fuzzy space. Although it makes the model imprecise, FR models could be used for analyzing the crisp data. Because the crisp data is also a kind of fuzzy data (Even though it is degenerated). For example some FR models could be used when some properties of CLR are not maintained.

Different from the main idea of CLR, the deviation between the observed value and estimated value of dependent variable Y_i can be defined as "fuzziness" and it depends on the fuzziness of the system structure in FLR. So FLR model is roughly like as follows:

$$Y_i^* = A_i X_{ij}$$

$$for\ i = 1, \dots, n(\#\ of\ obs.) \quad j = 1, \dots, N\ (\#\ of\ ind.var.) \tag{3.2}$$

Recent years, many kinds of fuzzy regression models have been constructed to restore regression analysis. These models can be roughly categorized into three groups, linear programming (LP) methods, multi-objective (MO) techniques (Nasrabadi M.M., Nasrabadi E., 2004; Nasrabadi M.M. at al., 2005; Özelkan E.C., Duckstein L., 2000) and least square (LS) methods (D'Urso P., Gastaldi T., 2000; C. Kao, Chyu, C.L., 2002; Coppi R. at al., 2006; Chen L.H. and Hsueh C.C., 2009).

3.2 LP Methods for FLR

The LP methods are the first approaches for FLR. Therefore they are the most famous ones. As can be understood from the name, the LP models are used to estimate the parameters in these methods and the main purpose is to minimize the fuzziness of the estimated regression model. Therefore they are also called The Minimum Uncertainty methods.

The LP methods are commonly used for fuzzy linear regression (FLR) because they are simple and easy to apply. Also it needs nearly no assumption. But it doesn't mean these methods are appropriate for all kinds of data sets. They also have some weaknesses; (i) they are extremely sensitive to outliers (W.L. Hung, M.S. Yang, 2006) ; (ii) when there is an outlier they don't allow all observations for estimation and (iii) estimated fuzziness per unit increases as number of observations increase

(D.T. Redden, W.H. Woodall, 1994) . The multi-objective (MO) techniques are proposed to solve some of these weaknesses (J. Lu, R. Wang, 2009) , but these techniques are not as simple as LP methods. Also they are not as good as the other methods (especially LS methods) for predictability.

3.2.1 Tanaka's First Model

Tanaka at al. proposed the first FLR model in 1982. According to that article; the deviation between the observed value and estimated value of dependent variable Y_i can be defined as "fuzziness" and it depends on the fuzziness of the system structure (H. Tanaka at al., 1982). That is also the main idea of the LP methods. The fuzzy model is;

$$Y_i^* = A_i X_{ij}$$

$$for\ i = 1, \dots, n(\#\ of\ obs.) \quad j = 1, \dots, N\ (\#\ of\ ind.var.) \tag{3.3}$$

The model consists of fuzzy parameters such as $A_i = (\alpha_i, c_i)$ and dependent variable $Y_i^* = (y_i, e_i)$. They both have triangular membership functions. α_i is the center and c_i is the fuzziness of the fuzzy parameter A_i and (as Figure 1.2 shows) observed Y_i has center "y_i" and fuzziness "e_i" .Also estimated Y_i^* is similar. Membership function of Y_i is as follows.

$$\mu_{Y_i}(y) = 1 - \frac{|y_i - y|}{e_i} \tag{3.4}$$

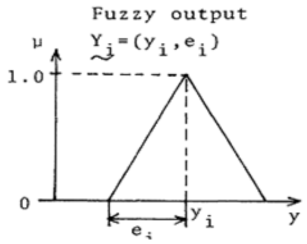


Figure 3.1 Membership function of Y_i

So the model can be presented as;

$$(y_i^*, e_i^*) = (\alpha_0, c_0)x_0 + (\alpha_1, c_1)x_1 + \cdots + (\alpha_n, c_n)x_n \tag{3.5}$$

Tanaka (1982) proposed a linear programming model to obtain the estimations of parameters. Basic ideas of this model;

1. It should minimize the total fuzziness of the parameters. (Sum of c_i).
2. The (membership function of) estimated Y_i^* should include the (membership function of) observed Y_i(see also figure 2.2).
3. There should be a threshold value ***H***, which presents the degree of fitting value of estimated Y_i^* to observed Y_i(see also figure 2.2).
4. The fuzziness of a parameter should be nonnegative.

The properties of ***H;***

- The threshold value "H" is defined between "0" and "1" but it could not be "1".
- It is a lower limit of fitting and generally researcher decides that value. So it makes the model flexible.
- Higher "*H*" values approximate the center of estimated Y_i^* to the center of observed Y_i, but it increases the vagueness (fuzziness) of estimated Y_i^* .
- The value of H is also interested with the researchers trust on the data. (Y.S. Chen, 2001)

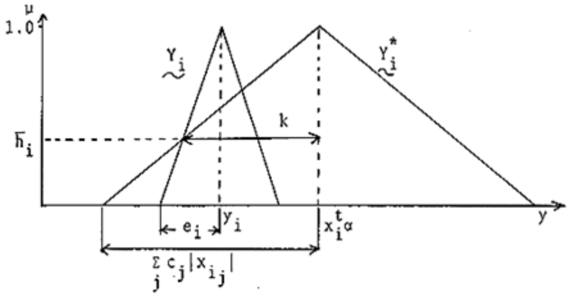


Figure 3.2 Presentation of $h_i(is\ an\ estimator\ for\ H)$

The linear programing model:

$$\min z = c_0 + c_1 + c_2 + \cdots + c_n$$

Subject to

$$\alpha^t x_i + (1 - H)c^t|x_i| \leq y_i + (1 - H)e_i$$

$$\alpha^t x_i - (1 - H)c^t|x_i| \geq y_i - (1 - H)e_i$$

$$c_i \geq 0\ for\ i = 1,2,\ldots,N \tag{3.6}$$

$(n\ is\ the\ \#\ of\ ind. variables\ and\ N\ is\ the\ \#\ of\ obsevations.\quad ^t$ means transpose.$)$

The first two constraints are “density constraints” which make the estimated Y_i^* to include observed Y_i in the model. So they should be generate for all data (total number of data is “N”).

The last one is “constraint of sign” that makes the fuzziness parameters c_i nonnegative.

3.2.2 Tanaka’s Second Model (Possibilistic Fuzzy Linear Regression-PFLR)

Tanaka modified his first model in 1987 and 1989(H. Tanaka, 1987 and Tanaka at al., 1989). The total “fuzziness” of the parameters (sum of c_i) was minimized in the first model. On the contrary, the second model try to minimize the total fuzziness of the model .That model is called Possibilistic Fuzzy Linear Regression (PFLR).

$$\min z = \sum_{i=1}^{N} (c_0|x_0| + c_1|x_{1i}| + c_2|x_{2i}| + \cdots + c_n|x_{ni}|)$$

Subject to

$$\alpha^t x_i + (1-H)c^t|x_i| \geq y_i + (1-H)e_i$$
$$\alpha^t x_i - (1-H)c^t|x_i| \leq y_i - (1-H)e_i$$
$$c_i \geq 0 \; for \; i = 1,2,\ldots,N$$
$$n \; is \; the \; \# \; of \; ind.variables \; and \; N \; is \; the \; \# \; of \; obsevations \qquad (3.7)$$

Tanaka at al. modified only the objective function by multiplying “fuzziness” of the parameters (c_i) to absolute value of independent variable(s) (x_i). All other parts are the same with his first model.

That modification reduced the fuzziness of the model significantly and brought it to the level required to be. But Tanaka’s basic ideas (approach) did not change. (Figure 3.3 is illustrative for Tanaka’s basic ideas).

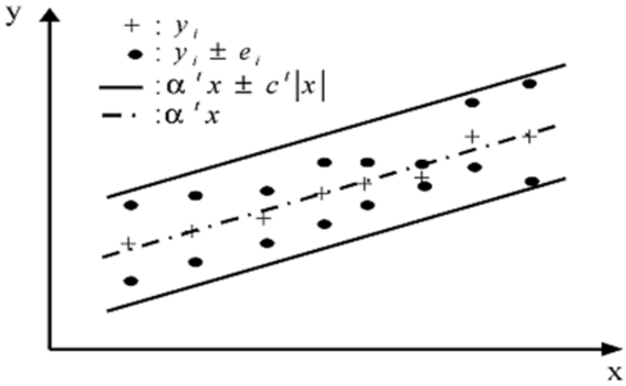


Figure 3.3 Presentation of Tanaka's approaches

As can be seen, Tanaka's models try to include the observed data and that causes two problems: it increases the fuzziness if there is an outlier in the data and model could not catch the trends (shrinking or expanding)

3.2.3 Peters' Model

In order to treat outlier problem, Peters modified Tanaka's second approach (PFLR) for non-fuzzy input (observed) data (G.Peters, 1994). $(Y_i = (y_i, e_i)\ and\ e_i = 0\ for\ all\ i)$. He introduced new variable and constants. λ is the variable presents the membership degree of the solution in a set of good solutions. So the model tries to maximize it.

$$max\ \lambda$$

Subject to

$$(1-\lambda)p_0 - \sum_{i=1}^{N}(c_0|x_0| + c_1|x_{1i}| + c_2|x_{2i}| + \cdots + c_n|x_{ni}|)\ \geq -d_0\ \ (objective\ function)$$

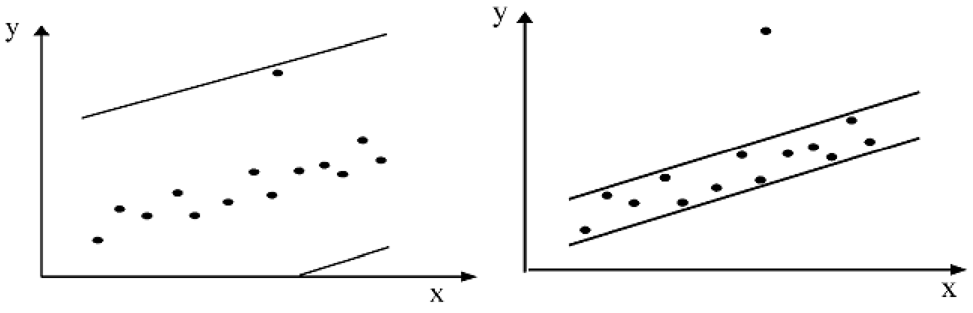
$$(1-\lambda)p_i + \alpha^t x_i + (1-H)c^t|x_i| \geq y_i\ \ (upper\ limits)$$

$$-(1-\lambda)p_i + \alpha^t x_i - (1-H)c^t|x_i| \leq y_i\ (lower\ limits)$$

$$-\lambda \geq -1\ \ and\ c_i, \lambda \geq 0\ for\ i = 1,2,\dots,N$$

$$n\ is\ the\ \#\ of\ ind.variables\ and\ N\ is\ the\ \#\ of\ obsevations \qquad (3.8)$$

The constant d_0 is the ideal value of (old) objective function. So it should be zero generally because minimum fuzziness is desired in FLR. (However it could not be zero in practice). The other constants p_0 and p_i are width constants of the model fuzziness. Peters treated the bounds of the interval, which include the data in Tanaka's model, as fuzzy with those three constants. Because generally d_0 is supposed zero, the bounds of the interval change with different values of p_0 and p_i in the model. (Figure 2.4)



(a) **(b)**

Figure 3.4 Peter's model, **(a)** for higher $\boldsymbol{p_0}$ and smaller $\boldsymbol{p_i}$,**(b)** For smaller $\boldsymbol{p_0}$ and higher $\boldsymbol{p_i}$

As can be seen; the point is determining the (suitable) values of p_0 and p_i. But that is not easy and it should be done in a context-dependent way (Y.S. Chen, 2001). After that, there would be occurred two or more results, for one or more outliers, with different p_0 and p_i values.

3.2.4 Lee and Chang's Model (Unrestricted in Sign Fuzzy Linear Regression-UFLR)

Another problem in PFLR is conflicting trends. In the cases where shrinking or expanding trends in the observations exist, PFLR frequently misinterprets the model. In order to avoid that problem Lee and Chang suggested canceling the *constraint of sign* in the PFLR model and called new model Unrestricted in Sign Fuzzy Linear Regression (UFLR)(E.S. Lee, P.T. Chang, 1994).

$$\min z = \sum_{i=1}^{N}(c_0|x_0| + c_1|x_{1i}| + c_2|x_{2i}| + \cdots + c_n|x_{ni}|)$$

Subject to

$$\alpha^t x_i + (1-H)c^t|x_i| \geq y_i + (1-H)e_i$$

$$\alpha^t x_i - (1-H)c^t|x_i| \leq y_i - (1-H)e_i$$
$$for\ i = 1,2,\dots,N$$
$$n\ is\ the\ \#\ of\ ind.variables\ and\ N\ is\ the\ \#\ of\ obsevations \quad (3.9)$$

UFLR model is very similar with PFLR. Only difference is there is no constraint for fuzziness of parameters (c_i), those could be negative in this model. That means some independent variables could affect the fuzziness of the model negatively. In other word some independent variables decrease the total fuzziness of the model. With that change model could capture the different trends. (Figure 3.5)

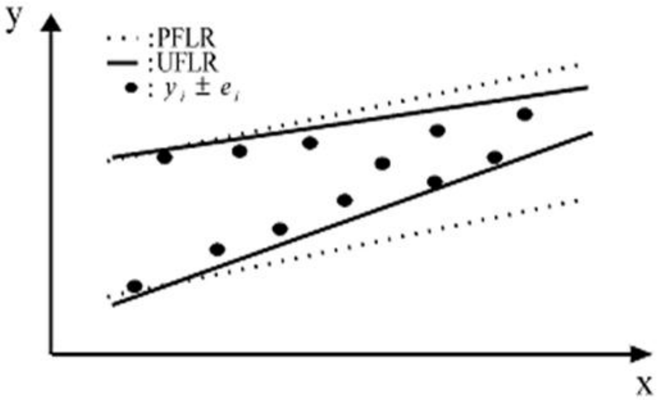


Figure 3.5 UFLR model (comparison with PFLR)

UFLR model works well in the data sets which have trend; however there is confusion about “negative fuzziness” and also outliers create problems like in the PFLR model.

3.2.5 Chen’s Model

In order to treat outlier problem in UFLR model, Chen suggested a three-step procedure (Y.S. Chen, 2001) . First step is detection of abnormal data. Second step is determination the expected number of outliers. And the model is redrawn after modification of outlier data. (Figure 3.6)

2.1.5.1. Detection of Abnormal Data

According to Chen, that is an abnormal data or potentially outlier if the difference between total fuzziness of estimated value ($c^t|x_i|$) and fuzziness of observed data (e_i) is greater than a constant of “k”. He suggested adding that to UFLR model as constraints (one constraint for each data). So new LP model will find an *infeasible solution* if there is an outlier.

New LP model:

$$\min z = \sum_{i=1}^{N} (c_0|x_0| + c_1|x_{1i}| + c_2|x_{2i}| + \cdots + c_n|x_{ni}|)$$

Subject to

$$\alpha^t x_i + (1-H)c^t|x_i| \geq y_i + (1-H)e_i$$
$$\alpha^t x_i - (1-H)c^t|x_i| \leq y_i - (1-H)e_i$$
$$c^t|x_i| - e_i \leq k \quad (new\ constraints)$$
$$for\ i = 1,2,\ldots,N$$
$$n\ is\ the\ \#\ of\ ind.variables\ and\ N\ is\ the\ \#\ of\ obsevations \qquad (3.10)$$

The point in that model is determining “k”. For small values of k, model would be a strict model and treat normal data as abnormal. For large values of k model would be liberal and could not detect the potentially outliers. Chen suggested some kind of ways to determine “k”. These are; “$Max\{e_i\}_{i=1}^N - Max\{e_i\}_{j\neq i=1}^N$” , “$Max\{e_i\}_{i=1}^N$” , “$Min\{e_i\}_{i=1}^N$” , “$(Max\{e_i\}_{i=1}^N + Min\{e_i\}_{i=1}^N)/2$” , “$Max\{e_i\}_{i=1}^N - Min\{e_i\}_{i=1}^N$” , “$\bar{e}$” and “$3s_e$”.

3.2.5.2. Determination the Expected Number of Outliers

According to Chen, confidence interval concept is useful for determination the expected number of outliers. If there are N data, $(1-\alpha) * N$ data are normal data and $m = (\alpha) * N$ data are abnormal data. (If m is not integer, it should be rounded to the upper integer). “$(1-\alpha)$” is the confidence level of data set and $\alpha = 0{,}05$ or less.

That step correspond to have an idea for how many times the model should run and to control that “k” is suitable for data set or not.

3.2.5.3. Modifying the Outlier and Redrawing the Model

If decision maker is not satisfied and thinks there is more than one outlier, the model should be redrawn. So it should be modified by eliminating the effect of first outlier for detection new outlier or abnormal data.

The fuzziness of a normal data (e_i) should be either $e_{i-1} \leq e_i \leq e_{i+1}$ or $e_{i-1} \geq e_i \geq e_{i+1}$ for ascending or descending data set. So there is a “λ_i” cut for estimated value Y_i^* such as:

$$e_i' = (1-\lambda_i)e_i = \frac{e_{i-r}+\cdots+e_{i-1}+e_{i+1}+\cdots+e_{i+r}}{2r} \qquad (3.11)$$

$$(\lambda_i) = 1 - \frac{e_{i-r}+\cdots+e_{i-1}+e_{i+1}+\cdots+e_{i+r}}{2re_i}. \qquad (3.12)$$

“λ_i” also gives the influence of the abnormal data on the data set and the value of “r” is important for reliability of λ_i. Larger values of r increases the reliability of λ_i. After obtaining the value of λ_i, the bound constraints (upper and lower limits) of abnormal data should be modified as:

$$\alpha^t x_i + (1-H)c^t|x_i| \geq y_i + (1-\lambda_i)e_i$$
$$\alpha^t x_i - (1-H)c^t|x_i| \leq y_i - (1-\lambda_i)e_i$$
$$(i\ is\ the\ abnormal\ value) \qquad (3.13)$$

Those three steps are repeated m times (m is the expected number of outliers). After the decision maker is satisfied with that all outliers have been detected, the modified model is obtained (figure 2.6)

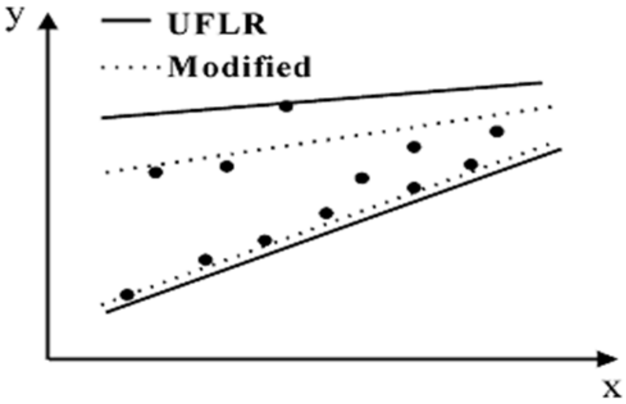


Figure 3.6 Chen’s model (comparison with UFLR)

Chen’s procedure is useful for fuzzy input-fuzzy output cases because it needs a non-zero fuzziness value ($e_i \neq 0$) for estimated Y_i^*. For non-fuzzy input-fuzzy output cases ($e_i = 0$), Peters’ model should be used.

CHAPTER FOUR
PROPOSED NEW MODEL

Since it is simple and easy to apply, the most widely used approach while constructing FLR models is linear programming. However, there are some points that should be discussed in detail: Redden and Woodall has stated that (i) they are extremely sensitive to outliers; (ii) when there is an outlier, they don't allow all observations for estimation (iii) as the number of observations increase, estimated fuzziness per unit also increases (D.T. Redden, W.H. Woodall, 1994). Peter's and Chen's models which were given in chapter 3 have tried to solve this problem.

Although prediction and estimation are the two main goals in regression analysis, these two models are not satisfactory enough in this respect. (M. Modarres at al, 2005). And this makes them a little bit inadequate.

In FLR based on LP methods, the deviation between the observed value and estimated value of dependent variable Y_i can be defined as "vagueness" and it depends on the fuzziness of the system structure. In other words, vagueness results from the system parameters included in the model. The main goal of LP methods (for FLR) is to minimize that vagueness. However, there might be several problems in a linear regression model like model specification, variable selection or lack of fit. Vagueness caused by problems given above and some other similar problems may be defined as "unexplained vagueness". In literature, FLR models based on LP methods ignore this unexplained part and focus on vagueness resulted from the parameters in the model. But it is not "fair".

3.1 Fair Fuzzy Linear Regression (FFLR)

Proposed model FFLR divides total vagueness into two parts as explained and unexplained. Explained vagueness (or fuzziness) is caused by independent variables included in the model. And unexplained vagueness (or fuzziness) is caused by problems mentioned above.

A new parameter F is added to the model to represent the unexplained vagueness part. It has a triangular membership function with center “0” and fuzziness “f”, F= (0, f). (Also it can be seen in figure 4.1)

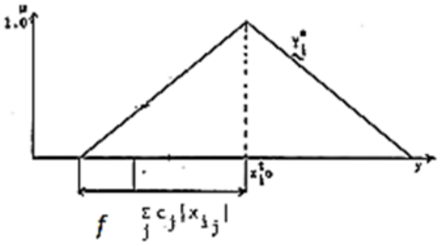


Figure 4.1 Membership function with two kind of vagueness

Then the regression function becomes as follows.

$$Y_i^* = A_i X_{ij} + F$$

$$(y_i^*, e_i^*) = (\alpha_0, c_0)x_0 + \cdots + (\alpha_n, c_n)x_n + (0, f) \tag{4.1}$$

FFLR and other LP based models in the literature are very similar in estimating the model parameters. The only difference is that boundary constraints are modified and there is an additional constraint for the new parameter F. The LP model is as follows;

$$\min z = \sum_{i=1}^{N} (c_0|x_0| + c_1|x_{1i}| + c_2|x_{2i}| + \cdots + c_n|x_{ni}|) \tag{4.2}$$

Subject to

$$\alpha^t x_i + (1 - H)c^t|x_i| + (\mathbf{1} - \boldsymbol{H})\boldsymbol{f} \geq y_i + (1 - H)e_i$$

$$\alpha^t x_i - (1 - H)c^t|x_i| - (\mathbf{1} - \boldsymbol{H})\boldsymbol{f} \leq y_i - (1 - H)e_i \tag{4.3}$$

$$N\, f \leq \sum c^t|x_i| \tag{4.4}$$

$$c_i \geq 0 \tag{4.5}$$

$$and\, f \geq 0 \quad for\; i = 1,2, \dots, N$$

$$n\; is\; the\; \#\; of\; ind. variables\; and\; N\; is\; the\; \#\; of\; obsevations \tag{4.6}$$

3.2 Some remarks on FFLR Model

FFLR can be introduced as a modified version of PFLR and UFLR. Although the objective function (4.2) of the FFLR is the same with these two model’s objective functions, it tries to minimize only the explained vagueness not the unexplained one represented by F in equation (4.1).

In general, a model can minimize the vagueness caused by it's independent parameters (explained vagueness). Therefore the proposed FFLR model aims to minimize explained vagueness but optimize the remaining unexplained vagueness part. That is why the objective function does not include *f*, the vagueness of the unexplained part.

Except for the new parameter f, the boundary constraints (4.3) of FFLR are similar to PFLR and UFLR. The main idea doesn't change. All models aim to get the estimated Y_i^* to include observed Y_i.

There is a new constraint (4.4) which makes the model meaningful by limiting *F* in that the unexplained vagueness part could not be greater than the explained vagueness part.

The constraint of sign (4.5) is optional. It could be used if the vagueness of the parameters are considered nonnegative as in the PFLR model, or it may be cancelled to catch the trend (if it exists), as in Lee and Chang's UFLR model.

CHAPTER FIVE
APPLICATION

In this section, three simulated data sets and one real-world data set are used to illustrate how the proposed model (FFLR) performs. In first data set, all parameters are positive. There is a negative parameter in the second one. In the third one, there is a simulated data set, too. But the independent variables came from populations which have non-symmetric G-H distributions. That gives how FFLR model performs for non-symmetric data sets.The last data set is from Tanaka's article to see how it works for real-world data. There are different kinds of independent variables and we have no idea which kind of distribution they have.

R is used to generate data for all simulated data sets. Linear programming parts are done with WinQSB. Also Minitab 14 is used for some illustrations.

The results of FFLR model are compared with Tanaka's PFLR and Lee and Chang's UFLR models. Chen's and Peter's models are ignored because they are modified versions of UFLR and PFLR for treating outlier problem and this study is not interested in outlier problem.

Three criterias have been used to compare these models. Two of them are for comparing predictability, total sum of square (***SS Total***) and mean square error (***MS Error***), which are the most-famous tools for determining predictability. The last one is for total vagueness of the model (***Total Vagueness***) which shows the sum of the estimated values' fuzziness.

For simplicity, the observations are assumed to be symmetric triangular fuzzy numbers and are denoted by $Y_i = (y_i, e_i)$. Also estimated fuzzy parameters are same as, $A_i = (\alpha_i, c_i)$. The threshold value is H=0, 5 for all models.

Example 1: The data set is in Table 5.1, is obtained by a simulation study. The distributions of independent variables are $X_1, X_2 \sim N(2,1)$ and $X_3 \sim N(4,1)$. The dependent variable is calculated from following equation, $Y \; = 2 * X_1 + 3 * X_2 + 2 * X_3 + e$ where $e \sim N(0{,}1)$. The fuzziness of the dependent variable is $e_i \sim N(4{,}1)$.

Table 5.1 Data set-1

	Y	e_i	X_1	X_2	X_3
1	19,3	3,41	1,39	3,12	3,17
2	20,2	3,57	3	1,73	3,91
3	19,2	4,27	2,37	1,14	4,93
4	15,7	5,39	0,43	2,44	3,62
5	20,8	3,82	2,32	2,61	4,06
6	17,7	4,06	0,68	2,13	4,35
7	24,9	3,29	2,28	3,26	4,46
8	20,9	4,32	5,05	0,51	4,91
9	32,2	4,13	2,31	5,12	5,42
10	19,9	5,3	3,08	1,93	4,07
11	18,5	4,22	0,8	2,28	4,16
12	22,6	5,15	3,53	2,48	3,64
13	24,1	4,22	1,83	3,77	4,62
14	20,4	4,29	1,17	2,72	5,05
15	23,9	4,18	2,46	2,71	4,48
16	20,3	4,87	2,99	2,4	3,36
17	18,1	2,26	1,59	1,72	4,53
18	30,1	2,31	3,81	3,37	5,61
19	22,1	3,04	1,91	2,92	4,23
20	18,5	3,72	1,51	1,72	4,68

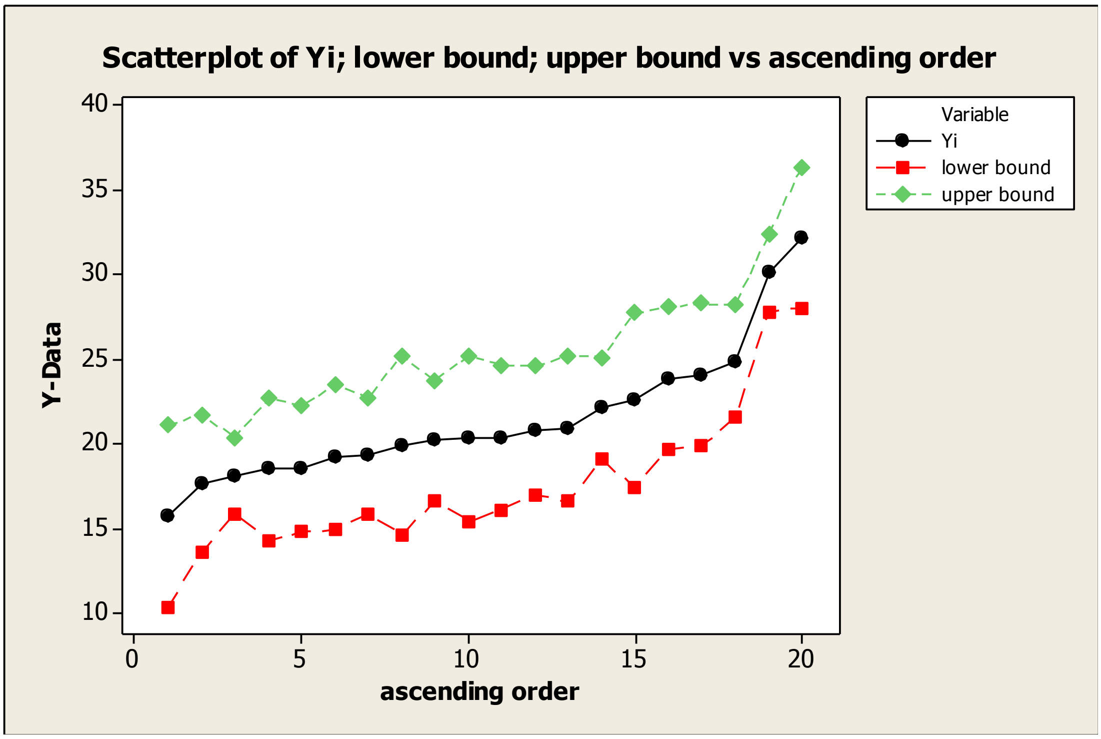


Figure 5.1 Data set-1

There are two cases for comparisons of predictability and fuzziness, FFLR-PFLR and FFLR-UFLR. Because the fuzzy part of the parameters (c_i) are must be nonnegative in PFLR model, but they are unrestricted in UFLR. Proposed FFLR model modifies PFLR and UFLR models on their own conditions. The estimated parameters and final results are given.

Table 5.2 Estimated parameters of data set-1, for PFLR and FFLR

		PFLR			FFLR			
n		A_1	A_2	A_3	A_1	A_2	A_3	F
10	a_i	1,8861	2,8745	2,1824	1,9263	3,0446	2,0635	
	c_i	0	0,1931	1,3588	0,0594	0	0,6614	2,9735
13	a_i	1,9849	3,0931	2,0793	1,9102	3,1089	2,1105	
	c_i	0	0	1,6009	0	0	0,7663	3,2602
16	a_i	1,9849	3,0931	2,0793	1,9329	3,1826	2,0621	
	c_i	0	0	1,6009	0	0	0,7632	3,2785
18	a_i	1,87	3,2388	2,0406	1,8779	3,2843	2,0264	
	c_i	0	0,9907	1,0218	0	0,5181	0,46	3,3101
20	a_i	1,8755	3,2364	2,0399	1,8788	3,2737	2,0316	
	c_i	0	0,9901	1,022	0	0,526	0,456	3,3064

Table 5.3 Comparisons of PFLR-FFLR for data set-1

n	PFLR			FFLR		
	SS Total	MS error	Total Vagueness	SS Total	MS error	Total Vagueness
10	7,341	0,734	62,921	6,399	0,640	59,467
13	6,476	0,498	88,552	5,905	0,454	84,770
16	8,835	0,552	109,181	8,303	0,519	104,506
18	9,347	0,519	125,071	8,919	0,496	119,163
20	9,740	0,487	138,759	0,107	0,005	132,258

The results are compared for different number of data size. n=10 means that the first ten observations in the data set are used and the order is same as in the table 5.1. (Note: as shown in the table 5.1, there is no descending or ascending order in the set)

Table 5.4 Estimated parameters of data set-1, for UFLR and FFLR

		UFLR				FFLR			
n		A_1	A_2	A_3		A_1	A_2	A_3	F
10	a_i	2,0678	3,4569	1,7683		2,0982	3,5921	1,6735	
	c_i	-0,3695	-0,9808	2,1949		-0,2853	-1,0969	1,4509	2,9396
13	a_i	2,0353	3,7465	1,677		1,9378	3,5024	1,8692	
	c_i	-0,3189	-0,0268	1,746		-0,1899	0,0154	0,8462	3,2139
16	a_i	1,9609	3,5091	1,8459		1,9533	3,5149	1,8597	
	c_i	-0,1705	0,4464	1,4091		-0,1555	0,0465	0,8147	3,2429
18	a_i	1,9609	3,5091	1,8459		1,9528	3,5127	1,8615	
	c_i	-0,1705	0,4464	1,4091		-0,1527	0,0487	0,8043	3,2753
20	a_i	1,9609	3,5091	1,8459		1,9527	3,5122	1,8619	
	c_i	-0,1705	0,4464	1,4091		-0,1521	0,0492	0,8018	3,2831

Table 5.5 Comparisons of UFLR-FFLR for data set-1

n		**UFLR**			**FFLR**	
	SS Total	**MS error**	**Total Vagueness**	**SS Total**	**MS error**	**Total Vagueness**
10	6,470	0,647	62,158	7,584	0,758	58,781
13	11,555	0,889	86,437	7,938	0,611	83,569
16	10,425	0,652	108,026	10,189	0,637	103,776
18	10,935	0,607	123,678	10,548	0,586	117,910
20	11,697	0,585	137,725	11,165	0,558	131,323

Proposed FFLR model gives better results from PFLR for both predictability and fuzziness. Also FFLR model gives better results from UFLR for sample sizes 13,16,18,20. Only for sample size 10, UFLR model is better than FFLR for predictability.

Example 2: The data set, is in Table 5.6, is obtained by a simulation study , too. The distributions of independent variables are $X_1, X_2, X_3 \sim N(2,1)$ .The dependent variable is calculated from following equation, $Y = 2 * X_1 - 3 * X_2 + 4 * X_3 + e$ where $e \sim N(0,1)$. The fuzziness of the independent variable is $e_i \sim N(4,1)$. Different from data set-1, there is an independent variable which has negative effect on the dependent variable.

Table 5.6 Data set-2

	Y	e_i	X_1	X_2	X_3
1	0,87	0,41	1,39	2,12	1,17
2	11,64	3,57	3,00	0,73	1,91
3	16,18	4,27	2,37	0,14	2,93
4	2,34	5,39	0,43	1,44	1,62
5	7,29	3,82	2,32	1,61	2,06
6	7,62	4,06	0,68	1,13	2,35
7	8,24	3,29	2,28	2,26	2,46
8	17,67	4,32	5,05	1,49	2,91
9	6,31	4,13	2,31	4,12	3,42
10	10,47	5,30	3,08	0,93	2,07
11	7,13	4,22	0,80	1,28	2,16
12	8,98	5,15	3,53	1,48	1,64
13	4,71	4,22	1,83	2,77	2,62
14	8,22	4,29	1,17	1,72	3,05
15	10,59	4,18	2,46	1,71	2,48
16	6,61	4,87	2,99	1,40	1,36
17	10,86	2,26	1,59	0,72	2,53
18	15,04	2,31	3,81	2,37	3,61
19	7,03	3,04	1,91	1,92	2,23
20	11,53	3,72	1,51	0,72	2,68

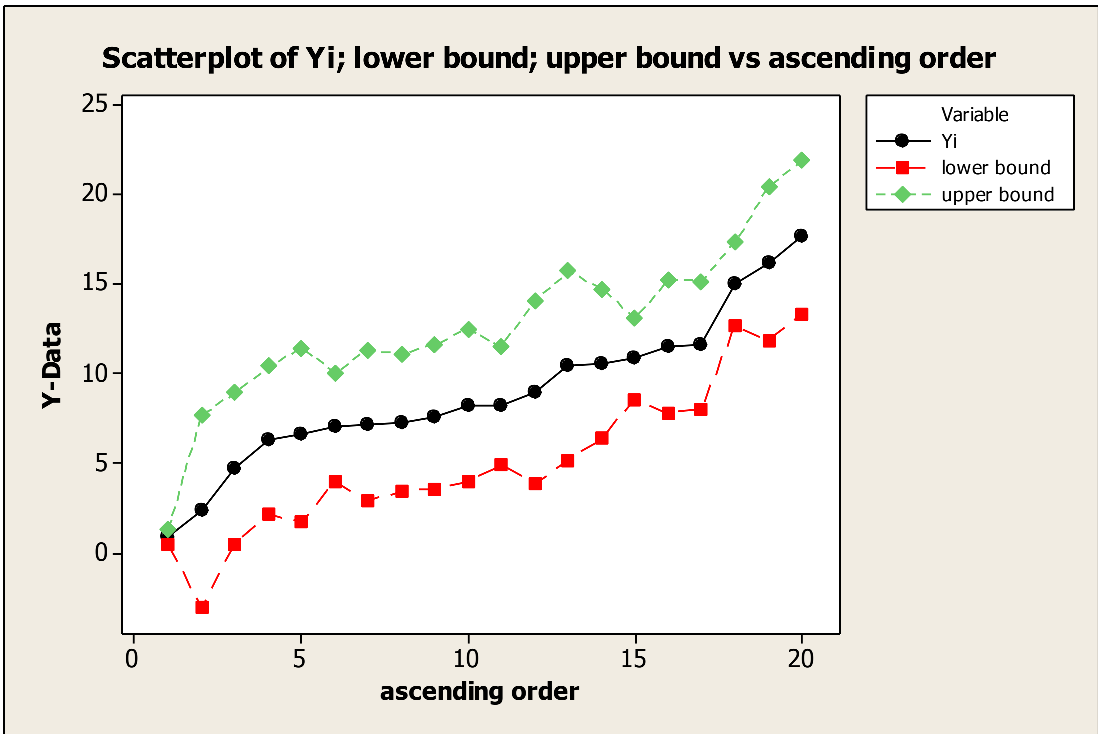


Figure 5.2 Data set-2

As in Example 1; there are two cases for comparisons of predictability and fuzziness. The estimated parameters and final results are given.

Table 5.7 Estimated parameters of data set-2, for PFLR and FFLR

		PFLR				FFLR			
n		A_1	A_2	A_3		A_1	A_2	A_3	F
10	a_i	1,5832	-4,1745	4,7349		2,0089	-3,0762	3,6456	
	c_i	0	1,0296	2,4119		0,0408	0	1,3501	3,1853
13	a_i	2,1097	-3,2129	3,7404		1,9713	-3,1377	3,8462	
	c_i	0	0,0244	3,3055		0	0	1,6888	3,0948
16	a_i	2,0374	-3,414	3,9383		1,9844	-3,0473	3,7887	
	c_i	0,1446	0,4266	2,9096		0	0	1,5235	3,4479
18	a_i	1,9651	-2,9981	3,7186		1,9855	-3,0345	3,7814	
	c_i	0	1,2585	2,4702		0	0	1,4926	3,5118
20	a_i	1,9651	-2,9981	3,7186		1,9856	-3,0331	3,7806	
	c_i	0	1,2585	2,4702		0	0	1,4892	3,519

Table 5.8 Comparisons of PFLR-FFLR for data set-2

n	PFLR			FFLR		
	SS Total	MS error	Total Vagueness	SS Total	MS error	Total Vagueness
10	29,627	2,963	71,675	12,267	1,227	63,705
13	11,341	0,872	97,442	9,560	0,735	89,748
16	15,415	0,963	121,750	11,428	0,714	110,332
18	15,525	0,862	141,638	12,468	0,693	126,424
20	16,644	0,832	157,089	13,196	0,660	140,760

The results are compared for different number of data size. n=10 means that the first ten observations in the data set are used and the order is same as in the table 5.6. (Note: as shown in the table 5.6, there is no descending or ascending order in the set)

Table 5.9 Estimated parameters of data set-2, for UFLR and FFLR

		UFLR				FFLR			
n		A_1	A_2	A_3		A_1	A_2	A_3	F
10	a_i	2,107	-2,718	3,3012		2,3181	-2,2048	2,789	
	c_i	-1,0476	-1,8835	5,2794		-0,5865	-1,6812	3,0925	3,0533
13	a_i	2,1584	-3,0776	3,6072		2,1255	-3,0548	3,6574	
	c_i	-0,0973	-0,2462	3,5718		-0,1447	-1,0715	2,5773	3,02
16	a_i	2,0374	-3,414	3,9383		2,018	-3,2589	3,9032	
	c_i	0,1446	0,4266	2,9096		-0,0046	-0,6131	1,9599	3,4162
18	a_i	1,9431	-2,8715	3,6518		2,0418	-3,2318	3,8643	
	c_i	-0,044	1,5116	2,3365		0,0403	-0,5923	1,8602	3,5006
20	a_i	1,9431	-2,8715	3,6518		1,9609	-2,709	3,5846	
	c_i	-0,044	1,5116	2,3365		-0,122	0,4485	1,2975	3,5133

Table 5.10 Comparisons of UFLR-FFLR for data set-2

n	UFLR			FFLR		
	SS Total	MS error	Total Vagueness	SS Total	MS error	Total Vagueness
10	13,088	1,309	66,818	16,644	1,664	61,066
13	11,052	0,850	96,603	9,542	0,734	87,583
16	15,415	0,963	121,750	11,981	0,749	109,320
18	16,052	0,892	141,614	12,799	0,711	126,021
20	17,294	0,865	156,926	15,043	0,752	140,535

The results show that proposed FFLR model gives better results from PFLR for both predictability and fuzziness. Also FFLR model gives better results from UFLR for sample sizes 13,16,18,20. Only for sample size 10, UFLR model is better than FFLR for predictability.

Example 3: The data set, is in Table 5.11, is obtained by a simulation study, too. The distributions of independent variables X_1, X_2, X_3 are generated from Generalized Hyperbolic distribution.The dependent variable is calculated from following equation, $Y = 2 * X_1 - 3 * X_2 + 4 * X_3 + e$ there e is coming from Generalized Hyperbolic distribution. The fuzziness of the independent variable is e_i.and it is coming from Generalized Hyperbolic distribution, too. (Descriptive statistics and graphs of generated populations are given below. Details about generation are given in Appendix 1.)

Different from data set-1 and set-2, independent variables are coming from a non-symmetric distribution.

Table 5.11 Data set-3

	Y	e_i	X_1	X_2	X_3
1	23,02	3,70	8,00	1,45	3,41
2	2,79	5,63	1,57	4,13	3,02
3	22,26	4,30	0,75	3,17	7,15
4	0,71	3,38	2,82	3,83	1,32
5	-3,54	11,88	6,96	8,92	1,75
6	18,71	7,61	3,24	1,72	4,13
7	5,68	9,12	0,96	2,37	1,93
8	17,73	4,63	3,77	3,37	4,92
9	19,17	5,04	7,48	3,92	4,06
10	2,34	3,90	1,44	1,89	1,54
11	8,05	3,93	4,39	3,24	2,35
12	16,96	2,80	6,57	3,42	3,87
13	7,75	4,07	1,53	1,67	2,10
14	33,54	3,30	2,17	4,16	8,99
15	37,76	3,53	13,68	2,17	3,78
16	8,56	5,30	2,73	2,28	2,41
17	-12,55	2,59	0,66	9,33	2,79
18	-7,23	3,79	1,84	4,21	-0,09
19	7,96	8,77	3,30	3,40	2,45
20	5,43	3,25	1,55	0,51	0,69

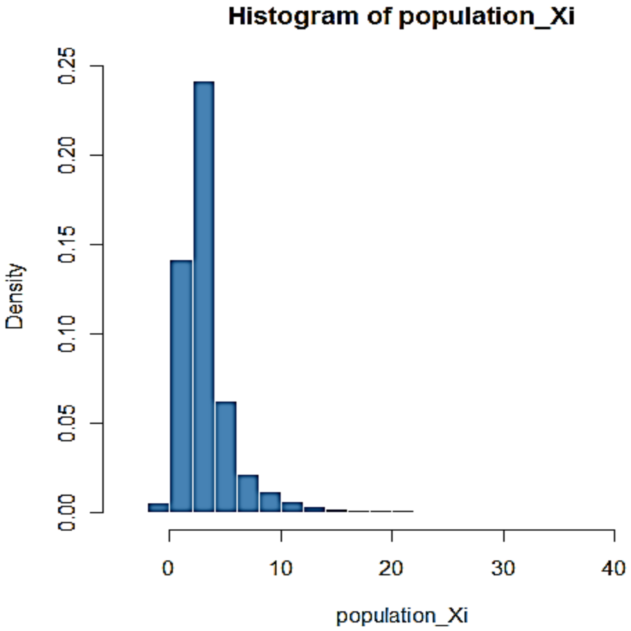


Figure 5.3 Population of independent variables

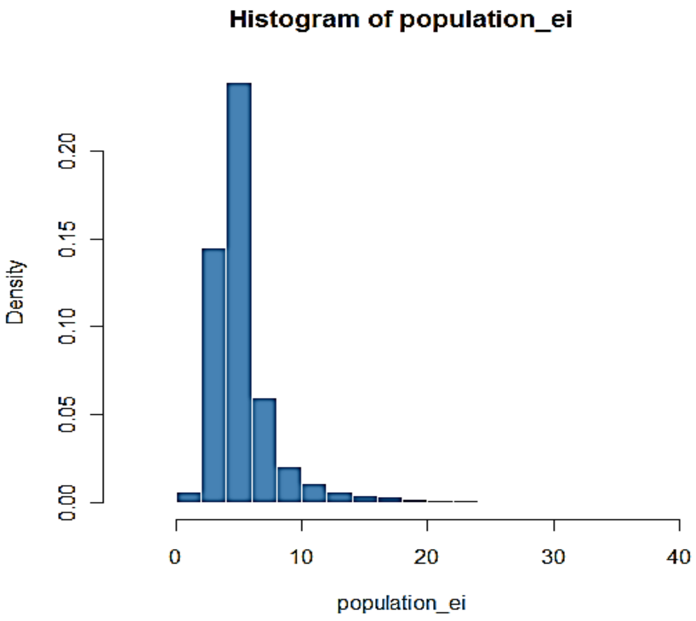


Figure 5.4 Population of fuzziness (e_i)

Table 5.13 Descriptive statistics of generated populations

Summary (population_Xi)

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-3,817	1,875	2,565	3,307	3,770	44,970

Summary (population_ei)

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-0.2743	3,859	4,550	5,303	5,711	38,670

Summary (population_e)

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-4,953	-0,126	0,117	0,398	1,815	33,630

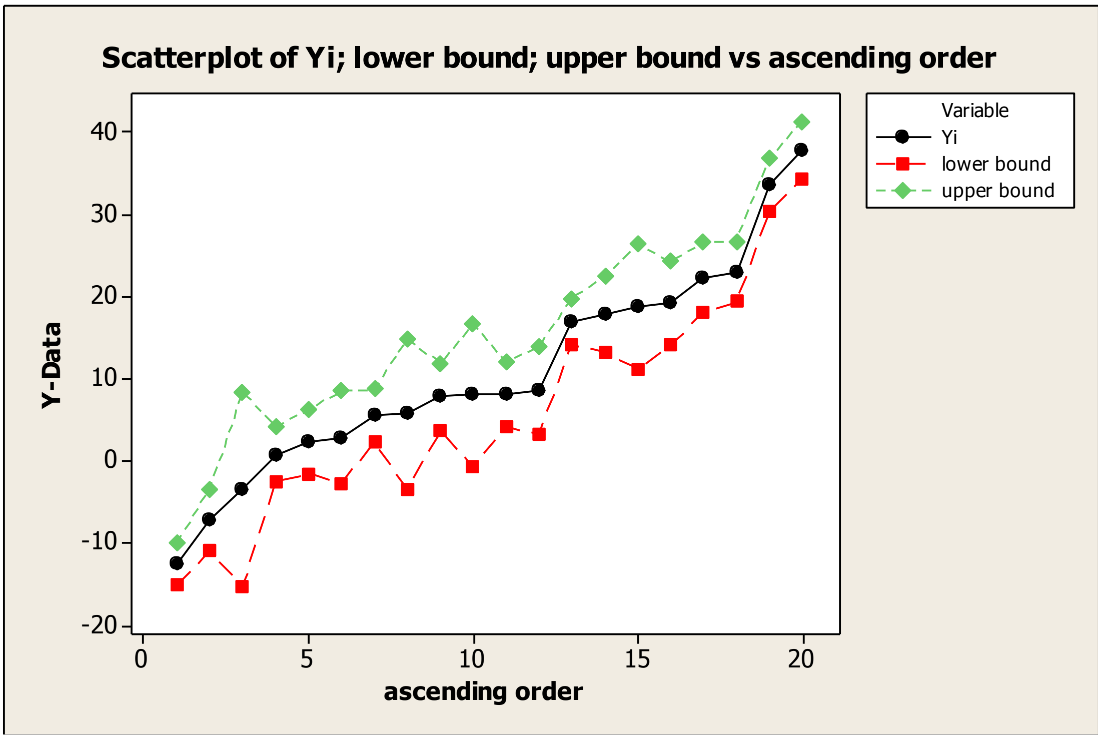


Figure 5.5 Data set-3

There are two cases for comparisons of predictability and fuzziness. The estimated parameters and final results are given.

Table 5.13Estimated parameters of data set-3 for PFLR and FFLR

		PFLR				FFLR			
n		A_1	A_2	A_3		A_1	A_2	A_3	F
10	a_i	1,1521	-2,087	4,5165		1,5235	-2,173	4,4624	
	c_i	0	4,5311	0		0	1,8195	0	6,3247
13	a_i	1,1521	-2,087	4,5165		1,5127	-2,159	4,4581	
	c_i	0	4,5311	0		0	1,8663	0	6,1859
16	a_i	1,6875	-2,133	4,1903		1,6547	-2,109	4,3238	
	c_i	0	4,6369	0,1047		0	1,8949	0	6,1228
18	a_i	1,6875	-2,133	4,1903		1,6472	-2,152	4,3643	
	c_i	0	4,6369	0,1047		0	1,7811	0	6,4555
20	a_i	2,2243	-2,803	4,6277		2,2243	-2,803	4,6277	
	c_i	0	1,4195	4,2942		0	1,4195	4,2942	3,519

Table 5.14 Comparisons of PFLR-FFLR for data set-3

n	PFLR			FFLR		
	SS Total	MS error	Total Vagueness	SS Total	MS error	Total Vagueness
10	53,11	5,31	148,93	41,67	4,17	116,73
13	72,37	5,57	187,69	48,72	3,75	151,54
16	79,93	5,00	234,83	99,88	6,24	185,48
18	110,99	6,17	289,24	131,78	7,32	218,45
20	205,24	10,26	366,72	136,84	6,84	252,79

The results are compared for different number of data size. n=10 means that the first ten observations in the data set are used and the order is same as in the table 5.6. (Note: as shown in the table 5.11, there is no descending or ascending order in the set)

Table 5.15 Estimated parameters of data set-3, for UFLR and FFLR

		UFLR				FFLR			
n		A_1	A_2	A_3		A_1	A_2	A_3	F
10	a_i	1,0719	-2,938	5,4843		1,2292	-2,372	4,8747	
	c_i	-1,235	3,0255	2,6996		-1,087	1,4018	1,4139	5,548
13	a_i	1,0719	-2,938	5,4843		1,2235	-2,393	4,8967	
	c_i	-1,235	3,0255	2,6993		-1,093	1,4605	1,4603	5,3427
16	a_i	1,7142	-2,005	4,0197		1,7746	-1,791	3,883	
	c_i	0,0539	4,893	-0,237		-0,004	2,4875	-0,6099	5,8615
18	a_i	1,8178	-2,497	4,4294		1,6198	-2,635	4,9689	
	c_i	-1,047	1,6187	4,6201		-0,407	0,6031	1,7322	6,3076
20	a_i	2,2833	-2,7	4,4465		1,617	-2,635	4,977	
	c_i	-0,117	1,2134	4,6567		-0,411	0,613	1,7865	6,1548

Table 5.16 Comparisons of UFLR-FFLR for data set-3

n	UFLR			FFLR		
	SS Total	MS error	Total Vagueness	SS Total	MS error	Total Vagueness
10	135,52	13,55	141,08	63,76	6,38	102,15
13	141,46	10,88	172,57	75,52	5,81	129,81
16	88,41	5,53	232,45	129,85	8,12	177,50
18	82,23	4,57	301,76	151,91	8,44	219,13
20	205,31	10,27	366,35	157,16	7,86	246,19

The results show that proposed FFLR model gives better results from PFLR and UFLR for for sample sizes 10, 13, 20. FFLR model doesn't give better results from PFLR and UFLR for sample sizes 13 and 16.

Example 4:

Data set-4 is from Tanaka's article "Linear regression analysis with fuzzy model, IEEE Transactions on Systems, Man and Cybernetics 12 (1982)", which is the first article of FLR. There are 5 independent variables (X_1 represents the constant), which are rank of material, first floor space (m^2), second floor space (m^2), number of rooms, number of Japanese-style rooms. Independent variable Y is fuzzy prices of the houses.

Table 5.17 Data set-4

	Y	e_i	X_1	X_2	X_3	X_4	X_5	X_6
1	6060	550	1	1	38,09	36,43	5	1
2	7100	50	1	1	62,1	26,5	6	1
3	8080	400	1	1	63,76	44,71	7	1
4	8260	150	1	1	74,52	38,09	8	1
5	8650	750	1	1	75,38	41,4	7	2
6	8520	450	1	2	52,99	26,49	4	2
7	9170	700	1	2	62,93	26,49	5	2
8	10310	200	1	2	72,04	33,12	6	3
9	10920	600	1	2	76,12	43,06	7	2
10	12030	100	1	2	90,26	42,64	7	2
11	13940	350	1	3	85,7	31,33	6	3
12	14200	250	1	3	95,27	27,64	6	3
13	16010	300	1	3	105,98	27,64	6	3
14	16320	500	1	3	79,25	66,81	6	3
15	16990	650	1	3	120,5	32,25	6	3

There are two cases for comparisons of predictability and fuzziness. The estimated parameters and final results are given.

Table 5.18Estimated parameters of data set-4, for PFLR and FFLR

PFLR

	A_1	A_2	A_3	A_4	A_5	A_6
a_i	-188,656	2280,207	105,035	82,198	-519,045	-553,430
c_i	313,649	480,099	0,000	0,000	0,000	0,000

FFLR

	A_1	A_2	A_3	A_4	A_5	A_6	F
a_i	-350,294	2199,389	105,035	82,198	-519,045	-391,793	
c_i	0,000	318,461	0,000	0,000	0,000	0,000	636,923

Table 5.19Estimated parameters of data set-4, for UFLR and FFLR

UFLR

	A_1	A_2	A_3	A_4	A_5	A_6
a_i	-374,525	2190,861	107,954	87,803	-631,368	-259,694
c_i	2494,279	120,372	18,490	5,389	-350,401	-514,377

FFLR

	A_1	A_2	A_3	A_4	A_5	A_6	F
a_i	-374,525	2190,861	107,954	87,803	-631,368	-259,694	
c_i	1940,292	120,372	18,490	5,389	-350,401	-514,377	553,988

As shown below, proposed FFLR model gives better results from PFLR for both predictability and fuzziness. However it gives same results with UFLR.

Table 5.20 Comparisons for data set-4

	SS Total	MS error	Total Vagueness		SS Total	MS error	Total Vagueness
PFLR	2099050	139936,7	19107,68	**UFLR**	1692787	112852,5	16617,56
FFLR	1754053	116936,9	19107,68	**FFLR**	1692787	112852,5	16617,57

CHAPTER SIX

CONCLUSION

In this study, a new LP model is developed for FLR. This new model modifies previous LP models by dividing total vagueness into two parts as explained and unexplained and aims to minimize only explained vagueness. So the estimations of parameters (centers of parameters) and unexplained vagueness are optimized.

Four numerical applications with four different data sets were performed and PFLR, UFLR and proposed model were compared in terms of mean squared error (MSE) and total fuzziness.

The results from first two examples indicate that the proposed method usually has better performance than the previous studies and improves predictability of LP methods for normal data. (Because data are in these examples have normal distribution.)

The third example is to show the new model's performance with asymmetric data. The results indicate that the proposed method usually has better performance than the previous studies and improves predictability of LP methods. However it gives worse results for some size of n. The asymmetric type of data may cause it in two ways. i) There may be abnormal values of $\boldsymbol{Y}$ which might be treated as an outlier. ii) there may be abnormal values for $\boldsymbol{e_i}$ (fuzziness of $\boldsymbol{Y}$**)** so it may disrupt te structure of the LP model. For both two cases, it could be said that new model is more sensitive than previous ones for asymmetric data.

The last one is a real data set from Tanaka's article "Linear regression analysis with fuzzy model, IEEE Transactions on Systems, Man and Cybernetics 12 (1982)", which is the first article of FLR. There are different kinds of independent variables and we have no idea which kind of distribution they have. The new model gives better results from PFLR for both predictability and fuzziness. However it gives same results with UFLR.

Consequently, the results from numerical examples indicate that the proposed method FFLR generally has better performance than its counterparts in literature and improves predictability of LP methods. So it can be said that FFLR would be a remarkable alternative to existing LP models.

REFERENCES

Chen Y.S. (2001). Outliers detection and confidence interval modification in fuzzy regression, *Fuzzy Sets and Systems, 119* , 259-272.

Chen L.H. and Hsueh C.C. (2009) Fuzzy Regression Models Using the Least-Squares Method Based on t he Concept of Distance*, IEEE Transactions On Fuzzy Systems, 17*, 6.

Coppi R., D'Urso P. (2006) Giordani P., Santoro A., Least squares estimation of a linear regression model with LR fuzzy response, *Computational Statistics and Data Analysis, 51* ,267 – 286.

D'Urso P., Gastaldi T. (2000). A least-squares approach to fuzzy linear regression analysis. *Computational Statistics and Data Analysis, 34*, 427-440.

Fuzzy Logic Toolbox™, MATLAB, MathWorks.

Hung W.L., Yang M.S. (2006) An omission approach for detecting outliers in fuzzy regression models, *Fuzzy Sets and Systems, 157* (23), 3109 – 3122.

Kao C., Chyu C.L. (2002) A fuzzy linear regression model with better explanatory power, *Fuzzy Sets and Systems*, *126* 401 – 409.

Lee D. (2006). A PhD Thesis in Civil Engineering, the Pennsylvania State University.

Lee E.S., Chang P.T. (1994) Fuzzy linear regression analysis with spread unconstrained in sign, *Comp. Math. Appl., 28*(4), 61-70.

Lu J., Wang R. (2009) An enhanced fuzzy linear regression model with more flexible spreads, *Fuzzy Sets and Systems, 160*, 2505 – 2523.

McClave J.T. & Sincich T. (2009). *Statistics* (11th edition). Pearson Education Inc.,

Myung I.J. (2003). Tutorial on maximum likelihood estimation, *Journal of Mathematical Psychology, 47*, 90–103

Nasrabadi M.M., Nasrabadi E. (2004). A mathematical-programming approach to fuzzy linear regression analysis, *Applied Mathematics and Computation, 155,* (3) 873 – 881.

Nasrabadi M.M., Nasrabadi E., Nasrabady A.R. (2005). Fuzzy linear regression analysis: a multi-objective programming approach, *Applied Mathematics and Computation, 163*, 245 – 251.

Neter J., & Kunter M.H., & Nachtsheim C. J., & Wasserman W.(1996) *Applied linear regression models*, (6th edition), Irwin.

Özelkan E.C., Duckstein L. (2000). Multi-objective f uzzy regression: a general framework, *Computers & Operations Research, 27,* 635 – 652.

Peters G. (1994). Fuzzy linear regression with fuzzy intervals, *Fuzzy Sets and Systems, 63,* 45-55.

Redden D.T., Woodall W.H. (1994). Properties of certain fuzzy linear regression methods, *Fuzzy Sets and Systems, 64,* 361 – 375.

Tanaka H. (1987). Fuzzy data analysis by possibilistic linear models. *Fuzzy Sets and Systems*, *24*, 363-375.

Tanaka H., Hayashi I., Watada J. (1989). Possibilistic linear regression analysis for fuzzy data. *European J. Oper. Res., 40,* 389-396.

Tanaka H., Uejima S., Asai K. (1982) Linear regression analysis with fuzzy model, *IEEE Transactions on Systems, Man and Cybernetics*, *12*, 903–907.

Yager R. R. (1986) *An Introduction to Fuzzy Theory, Applications of Fuzzy Set Theory in Human Factors*, Elsevier Science, Amsterdam.

Yen J. and Langari R. (1998) *Fuzzy Logic: Intelligence, Control, and Information*, Prentice Hall, New Jersey.

Zadeh L. A.(1965). Fuzzy Sets, *Information and Control*, *8*, 338–353.

APPENDIX 1

GENERALIZED HYPERBOLIC DISTRIBUTIONS WITH R

Description

Calculates moments of the generalized hyperbbolic distribution function.

Usage

dgh(x, alpha = 1, beta = 0, delta = 1, mu = 0, lambda = -1/2, log = FALSE)

pgh(q, alpha = 1, beta = 0, delta = 1, mu = 0, lambda = -1/2)

qgh(p, alpha = 1, beta = 0, delta = 1, mu = 0, lambda = -1/2)

rgh(n, alpha = 1, beta = 0, delta = 1, mu = 0, lambda = -1/2)

Arguments

alpha, beta, delta, mu, lambda

Numeric values. Alpha is the first shape parameter; beta is the second shape parameter in the range (0, alpha); delta is the scale parameter, must be zero or positive; mu is the location parameter, by default 0; and lambda defines the sublclass, by default -1/2. These are the meanings of the parameters in the first parameterization pm=1 which is the default parameterization. In the second parameterization, pm=2 alpha and beta take the meaning of the shape parameters (usually named) zeta and rho. In the third parameterization, pm=3 alpha and beta take the meaning of the shape parameters (usually named) xi and chi. In the fourth parameterization, pm=4 alpha and beta take the meaning of the shape parameters (usually named) a.bar and b.bar. log a logical flag by default FALSE.

n number of observations.

p a numeric vector of probabilities.

x, q a numeric vector of quantiles.

... arguments to be passed to the function integrate.

Details

The generator rgh is based on the GH algorithm given by Scott (2004).