ESCUELA NAVAL DE SUBOFICIALES A.R.C. "BARRANQUILLA"

SEXTANTE

Un modelo de regresión lineal aplicando lógica difusa

Alex Gutiérrez & Wilfrido Ferreira

Departamento de Ciencias Básicas, Universidad Autónoma del Caribe, Barranquilla, Colombia.
alex.gutierrez@uac.edu.co, wilfrido.ferreira@uac.edu.co

Recibido: Abril 25, 2019.
Recibido en su versión corregida: Septiembre 10, 2020.
Aceptación: Octubre 15, 2020.

Cómo citar: Gutiérrez, A. & Ferreira, W. (2020). Un modelo de regresión lineal aplicando lógica difusa. Revista Sextante, 23, pp. 48 - 54, 2020.

Resumen
En este artículo se hace una aplicación de la lógica difusa a modelo de regresión lineal con parámetros difusos, utilizando los criterios que sugiere Bo Yuan y Klir [2], comparando los resultados obtenidos con el modelo de regresión lineal probabilístico.

Palabras claves: *Conjuntos difusos; Lógica difusa; Números difusos; Regresión lineal.*

A linear regression model applying fuzzy logic

Abstract
This article makes an application of Fuzzy logic to a linear regression model with fuzzy parameters using the criteria suggested by Bo Yuan and Klir [2], comparing the results obtained with the probabilistic linear regression model.

Keywords: *Fuzzy logic; Fuzzy numbers; Fuzzy sets; Linear regression.*

1. Introducción

La lógica difusa o borrosa nació en 1965 cuando el Dr. Lofti Zadeh publicó un artículo titulado "Conjuntos Difusos" en la revista Information and Control" [1]. La teoría de los conjuntos borrosos es un acercamiento entre la precisión de las matemáticas clásicas y la imprecisión del mundo real, el cual se le ha intentado ajustar a modelos matemáticos que no permiten lo borroso o lo impreciso, y como consecuencias nos lleva a resultados indeseables. Este trabajo se hace con la finalidad de divulgar el modelo de regresión lineal, usando la teoría de conjuntos borrosos. El problema consiste en encontrar los parámetros $a_1, a_2, \dots, a_n$ tal que nos permita ajustar de la mejor manera la función lineal $Y = a_1 X_1 + a_2 X_2 + \cdots + a_n X_n$ donde los valores de entrada son datos ordinarios y los de salida son números difusos (triangulares y simétricos) y $a_1, a_2, \dots, a_n$ son valores reales, utilizando los criterios que sugiere Bo Yuan y Klir [2].

2. Conceptos básicos

Sea *X* un conjunto y *A* un subconjunto de *X*. Para denotar que un elemento $x \in X$ pertenece a *A*, se puede utilizar el concepto de función característica que se relaciona en la Ecuación 1:

$$\mu_A: X \to \{1,1\}$$
$$\mu_A = \begin{cases} 1 & x \in A \\ 0 & x \notin A \end{cases} \quad (1)$$

Supongamos que la función característica puede tomar cualquier valor en el intervalo [0, 1]. Así un elemento podrá no pertenecer a *A*, pertenecer "un poco", pertenecer "bastante" o ser de *A*; según que *A* fuese cero, próximo a uno o igual a uno.

Definición: Sea X un conjunto, llamado referencial, y x ∈ X. Un subconjunto difuso A de X es un conjunto de pares (x, $\mu_A(x)$) para todo x ∈ X donde $\mu_A(x)$ se le llama función de membresía y representa el grado de pertenencia de x a A. [10, 11, 12, 13]

Ejemplo: un corredor de bienes raíces desea clasificar las casas que ofrece a sus clientes. Tomando el criterio de comodidad, consideremos el número de cuartos con los que cuentan las casas. Sea $X = \{1,2,3,4,5,6,7,8,9,10\}$ el conjunto de casas disponibles y sea x el número de cuartos de una casa.

Entonces podemos definir el conjunto difuso de acuerdo con el criterio de comodidad como A= Casas cómodas para una familia de 4 integrantes (ver Ecuación 2).

$$A = \{(1, 0.2), (2, 0.5), (3, 0.8), (4, 1), (5, 0.7), (6, 0.3)\} \quad (2)$$

Definición: definamos al número difuso A como un subconjunto difuso sobre la recta real, que cumple las tres propiedades denotadas en las Ecuaciones 3, 4 y 5:

1. Normalidad
$$\sup_{x \in X} \mu_A(x) = 1 \quad (3)$$
2. Nivel ∝ : para todo ∝∈ [0,1] el conjunto,
$$A^{\propto} = \{x \in X: \mu_A(x) \geq \propto\} \quad (4)$$

Sea compacto.

3. Convexidad: para todo ∝ y β ∈ [0,1] con,
$$\propto > \beta \text{ se tiene que } A^{\propto} \subset A^{\beta}. \quad (5)$$

Observación: notemos que las condiciones relacionadas en las Ecuaciones 4 y 5 implican que todos los niveles ∝ son intervalos cerrados sobre la recta real, a los que se les dota de una altura o grado de pertenencia sobre el conjunto difuso.

Los números difusos pueden ser considerados también como una extensión del concepto de intervalo de confianza. En lugar de considerar el intervalo de confianza a un único nivel de confianza, se consideran varios dentro del intervalo [0,1], dándole al 1 el máximo de posibilidad y al 0 el mínimo. El nivel de confianza no dará una hipótesis restrictiva. Así el nivel de posibilidad α ∈ [0,1], genera un intervalo de confianza: A(α)=[α^1, α^2], el cual es una función monótona decreciente de α (ver Ecuación 6).

$$Si\ \alpha' > \alpha \quad entonces \quad A(\alpha') \subset A(\alpha) \quad (6)$$

Para ejemplificar un número difuso y sus elementos veamos la Figura 1. En donde Soporte $S(A) = A^0 = (0,5)$, la moda $A^1 = [2,3]$ y el nivel $-0.6\ A^{0.6} = [1,4]$

Números difusos triangulares: existen varios tipos de números difusos [7,8,9], pero en este

trabajo mencionaremos los números difusos triangulares. Definimos un número triangular difuso como aquel subconjunto de $\mathbb{R}$ con función de membresía μ_A descrita por la Ecuación 7 y esbozada en la Figura 2. Donde $[a_1, a_2]$ es el soporte del intervalo difuso A y $(a_c, 1)$ es el punto máximo.

$$\mu_A(x) = \begin{cases} \dfrac{x - a_1}{a_c - a_1} & para\ a_1 \le x \le a_c \\ \dfrac{x - a_2}{a_c - a_2} & para\ a_c \le x \le a_2 \\ 0 & en\ otro\ caso \end{cases} \quad (7)$$

Para algunas aplicaciones el punto $a_c \in (a_1, a_2)$ se localiza a la mitad del intervalo del soporte del número difuso A, es decir $a_c = (a_1 + a_2)/2$. Sustituyendo este valor en la Ecuación 7 se obtiene la Ecuación 8:

$$\mu_A(x) = \begin{cases} 2\dfrac{x - a_1}{a_c - a_1} & para\ a_1 \le x \le a_c \\ 2\dfrac{x - a_2}{a_c - a_2} & para\ a_c \le x \le a_2 \\ 0 & en\ otro\ caso \end{cases} \quad (8)$$

Entonces decimos que la Ecuación 8 representa un número difuso triangular simétrico con centro en $a_c = (a_1 + a_2)/2$. Este tipo de número se ocupa frecuentemente en aplicaciones de controladores difusos, teoría de las decisiones, negocios, [10] etc. Usualmente denotamos a un número triangular como $A = [a_1, a_2]$ y si además el simétrico lo denotamos como $A = (a_c - a_2)$, donde a_c tiene la membresía total y a_2 representa la incertidumbre o imprecisión para a_c.

El análisis de regresión es un área dentro de la estadística que nos ayuda a determinar y evaluar la relación entre una variable determinada (variable dependiente) y una o más variables independientes (o explicativas). Para ello es necesario estimar los parámetros de dicho modelo, el cual toma la siguiente forma lineal (ver Ecuación 9).

$$y = a_0 + a_1 x_1 + a_2 x_2 + \cdots + a_n x_n \quad (9)$$

Donde y es la variable dependiente, $x_1, x_2, \ldots, x_n$ son las variables independientes y $a_0, a_1, a_2, \ldots, a_n$ son los parámetros. Los problemas de análisis de regresión que se formulan bajo estos términos se les conocen como regresiones lineales

Tomemos como ejemplo un problema de regresión lineal con una sola variable como se muestra en la Ecuación 10.

$$y = a_0 + a_1 x_1 \quad (10)$$

$A = \{(x_i f(x_i))\}_{i=1}^{m}$ Un conjunto de m observaciones. Nuestro objetivo es encontrar a_0 y a_1, para los cuales el error total de los puntos estimados con los puntos observados sea máximo.

Basándonos en el método de mínimos cuadrados, el error se expresa como la Ecuación 11.

$$error\ Total = \sum_{i=1}^{m} [a_0 + a_1 x - f(x_i)]^2 \quad (11)$$

Derivando con respecto a a_0 y a a_1, igualando a cero y resolviendo el sistema de ecuaciones se obtiene que los valores óptimos para a_0 y a_1 están dados por la Ecuación 12.

$$a_0 = \frac{[\sum_{i=1}^{m} f(x_i)][\sum_{i=1}^{m} f(x^2{}_i)] - [\sum_{i=1}^{m} x_i][\sum_{i=1}^{m} f(x_i)x_i]}{m\sum_{i=1}^{m} x_i^2 - [\sum_{i=1}^{m} x_i]^2}$$

$$a_1 = \frac{m\sum_{i=1}^{m} f(x_i)x_i - [\sum_{i=1}^{m} f(x_i)][\sum_{i=1}^{m} (x_i)]}{m\sum_{i=1}^{m} x_i^2 - [\sum_{i=1}^{m} x_i]^2} \quad (12)$$

Una de las ventajas que nos motiva a hacer análisis difuso en lugar de análisis lineal es que, aunque la relación difusa sea menos precisa, ésta resulta estar más pegada a la realidad, y además que en algunas aplicaciones la naturaleza de los datos es tal que éstos están en términos difusos.

Desarrollaremos el ejemplo de la regresión lineal que involucra datos ordinarios y se estiman parámetros difusos basándonos en el libro Fuzzy Set and Fuzzy Logic, George J. Klir [1].

3. Regresión lineal con parámetros difusos

Este modelo de regresión difusa se expresa como sigue la Ecuación 13.

$$y = c_1 x_1 + c_2 x_2 + \cdots + c_n x_n \quad (13)$$

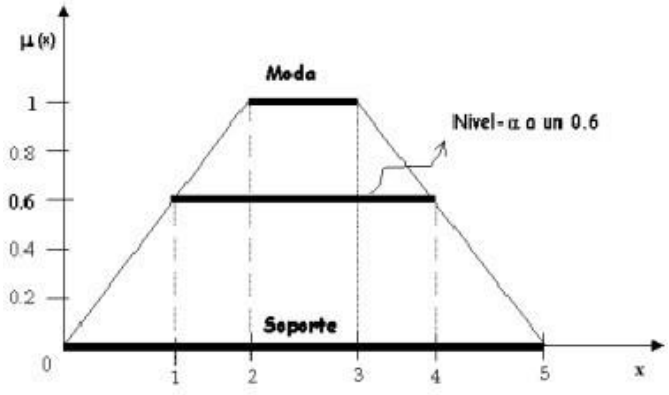

Figura 1. Número difuso y sus elementos.
Fuente: Los autores.

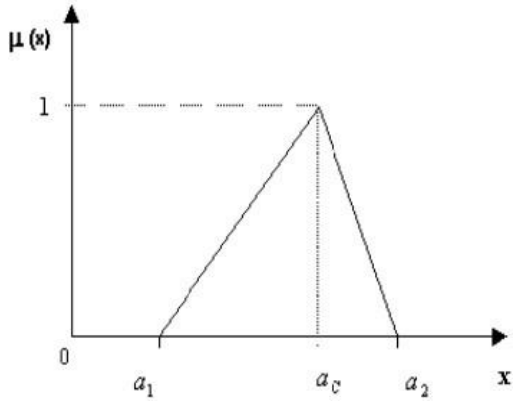

Figura 2. Número difuso triangular
Fuente: Los autores.

Donde $c_1 + c_2 + \cdots + c_n$ son números difusos y $x_1 + x_2 + \cdots + x_n$ son valores reales de la variable de entrada. Para cada $x_1, x_2, \cdots, x_n$ de los valores de las variables de entrada, el valor de la variable de salida Y definida por la Ecuación 13 es también un número difuso.

Sea el conjunto de datos $\{(a_1, b_1), (a_2, b_2), \ldots, (a_n, b_n)\}$. Al igual que en la regresión lineal el problema consiste en encontrar los parámetros difusos $c_1, c_2, \cdots, c_n$ para los cuales la Ecuación 13 exprese el mejor ajuste para los datos de entrada, acorde a algunos criterios de bondad de ajustes.

Asumiremos que los parámetros de la ecuación son números difusos triangulares simétricos, definidos como se muestra en la Ecuación 14. Donde c_i es el punto para el cual $C_i(c_i) = 1$ y $s_i > 0$ es el incremento de C_i (la mitad de la amplitud del soporte de C_i), ver Figura 3.

$$C_i(c) = \begin{cases} 1 - \dfrac{|c - c_i|}{s_i} & c_i - s_i \le c \le c_i + s_i \\ 0 & en\ otro\ caso \end{cases} \quad (14)$$

Por otro lado, dado que C_i se encuentra alrededor de c_i denotado como $C_i = (c_i, s_i)$ para $i = 1,2,3, \ldots, n$, entonces resulta fácil probar que Y en la Ecuación 7 es también un número difuso triangular dado por la Ecuación 15, donde T denota la transpuesta como se relaciona en la Ecuación 16.

$$Y(y) = \begin{cases} 1 - \dfrac{|y - X^T c|}{S^T |X|} & si\ X \neq 0 \\ 0 & si\ X = 0, y \neq 0 \\ 1 & si\ X = 0, y = 0 \end{cases} \quad (15)$$

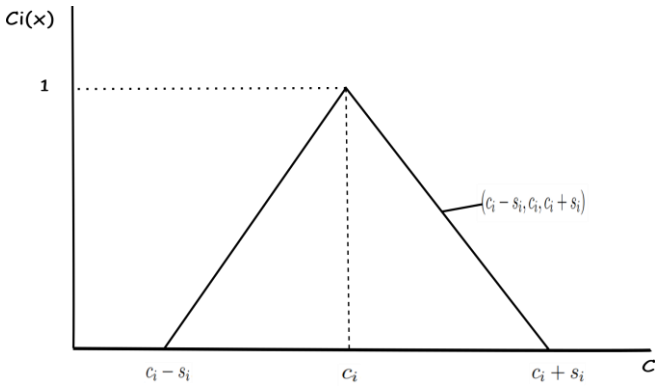

Figura 3. Número difuso triangular simétrico.[5]
Fuente: Los autores.

$$X = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}, c = \begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ c_n \end{pmatrix}, s = \begin{pmatrix} s_1 \\ s_2 \\ \vdots \\ s_n \end{pmatrix}, |X| = \begin{pmatrix} |x_1| \\ |x_2| \\ \vdots \\ |x_n| \end{pmatrix} \quad (16)$$

El problema consiste en encontrar los parámetros difusos $C_1, C_2, \ldots, C_n$ éste puede ser resuelto encontrando los vectores c y s para los cuales $Y(y)$ dado en la Ecuación 14, se ajusta de mejor forma a las observaciones. Para lograr el mejor ajuste se consideran los dos criterios siguientes:

1. Consideremos que Y_j representa a un número difuso definido como en la Ecuación 14 y que $X = a_j$, entonces, para cada observación (a_j, b_j) y con $h \in [0,1]$ (confianza para la variable Y_j) el valor $Y_j(b_j) \ge h$.

2. Especificando el valor $h \in [0,1]$ y puesto que la imprecisión de cada parámetro difuso C_i dada por la Ecuación 15 puede ser expresada en términos de su incremento s_i, es necesario minimizar la ambigüedad o la imprecisión de los parámetros difusos C_i.

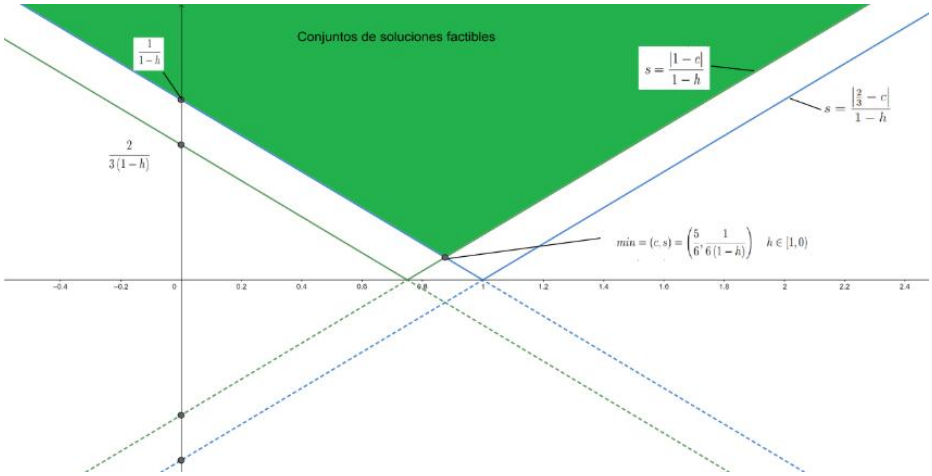

Figura 4. Región factible dadas las rectas [5].
Fuente: Los autores.

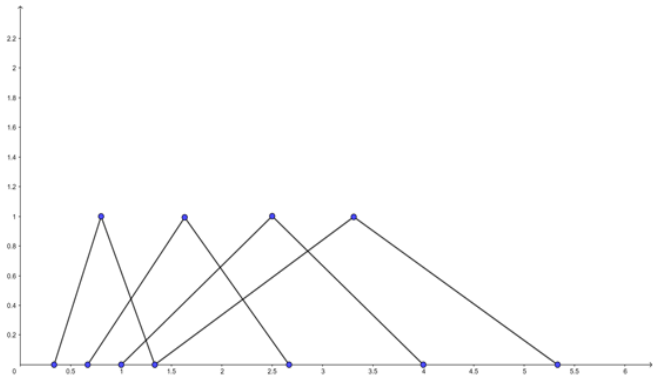

Figura 5. Conjuntos difusos $Y_j = Ca_j$ [6].
Fuente: Los autores.

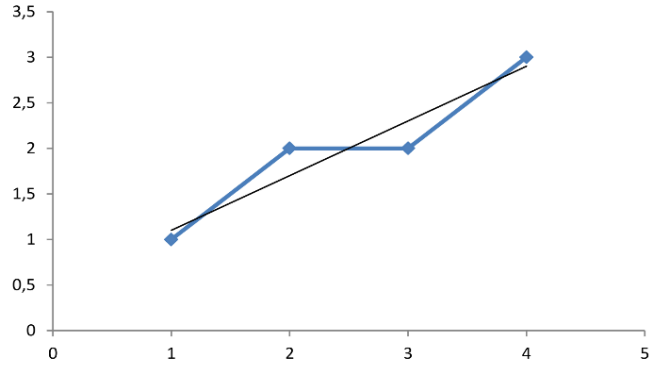

Figura 6. Modelo lineal
Fuente: Los autores.

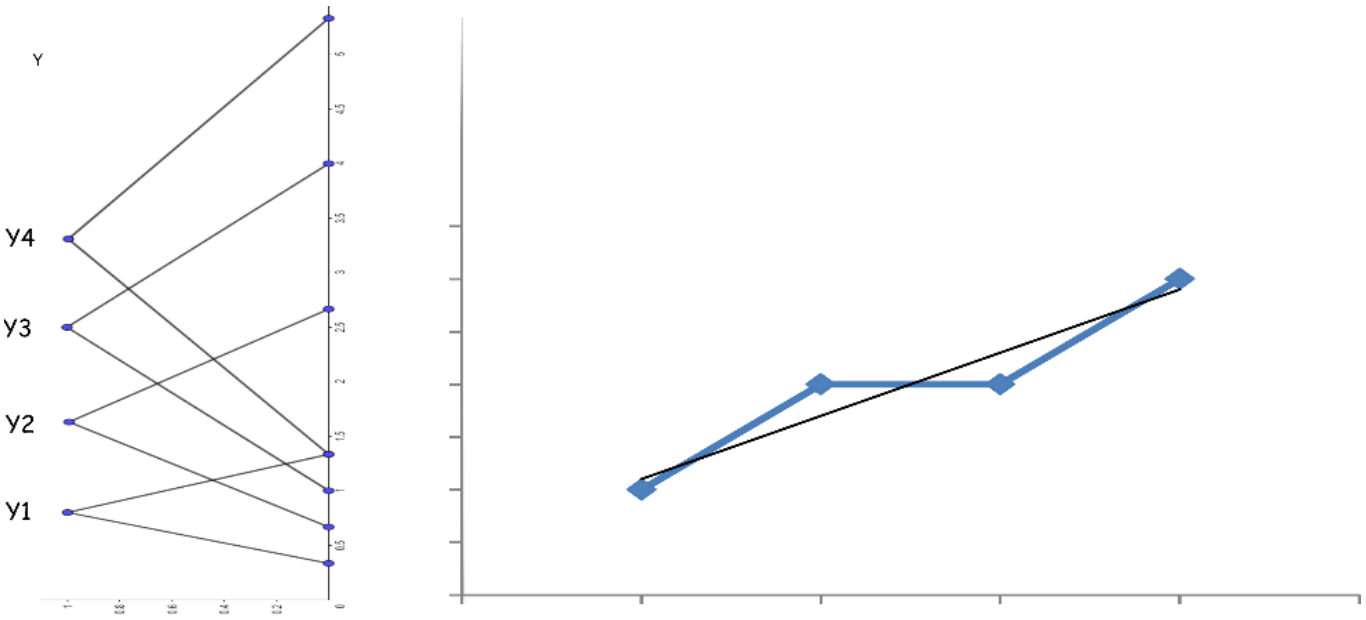

Figura 7. Comparando los dos modelos [5].
Fuente: Los autores.

Con los criterios 1 y 2, la Ecuación 7 se puede formular como la Ecuación 17.

$$\begin{aligned} &\min \sum_{i=1}^{n} s_i \\ &(1\text{-}h)S^T|a_j|\text{-}\left|b_j - a_j^T\, c\right| \geq 0 \qquad j=1,2,3,\ldots,n \\ &s_i \geq 0 \qquad\qquad\qquad\quad i=1,2,3,\ldots,n \end{aligned} \tag{17}$$

Se trasforma en un problema clásico de programación lineal. Notemos que las ecuaciones dadas en la Ecuación 17 se obtuvieron de un despeje de la Ecuación 14 sólo que se especifica el valor de $h \in [0,1]$.

Ejemplo: Dadas las siguientes observaciones:

1	2	3	4
1	2	2	3

Asumimos que la regresión lineal difusa para estos datos se ajusta al modelo de la Ecuación 18. Donde *C= (c,s)* es un parámetro difuso triangular simétrico. Entonces el problema de programación lineal toma la forma de la Ecuación 19. Que se reduce a la Ecuación 20.

$$Y = C_x \tag{18}$$

$$\begin{aligned} &min\ s \\ &(1-h)s - |1-c| \geq 0 \\ &2(1-h)s - |2-2c| \geq 0 \\ &3(1-h)s - |2-3c| \geq 0 \\ &4(1-h)s - |3-4c| \geq 0 \\ &s \geq 0 \\ &h \in [0,1], \quad número\ fijo \end{aligned} \tag{19}$$

$$\begin{aligned} &min\ s \\ &s \geq \frac{1}{(1-h)} max\ \left(|1-c|, \left|\frac{2}{3}-c\right|, \left|\frac{3}{4}-c\right|\right) \\ &h \in [0,1], \quad número\ fijo \end{aligned} \tag{20}$$

En la Figura 4 se grafican las rectas para $h = 0$. La recta de s=|3/4-c|/(1-h) no es gráfica porque sale de la región de soluciones.

Resolviendo el problema con ayuda de la Figura 4, y realizando cálculos necesarios, encontramos que los valores óptimos para c y s son c=5/6 y s=1/6(1-h).

Por ejemplo: si consideramos a h=2/3 entonces C =(5⁄6,1⁄2). Los conjuntos difusos para los valoresa1=1, a2=2, a3=3 y a4=4 son: Y1=(5⁄6,1⁄2) a1=(5⁄6,1⁄2), Y2=(5/6,1/2) a2 =(5⁄3,1), Y3=(5⁄6,1⁄2) a3=(5⁄2,3⁄2), Y4=(5⁄6,1⁄2) a4=(10⁄3,2) que se muestran en la Figura 5. Se ajustan los datos con un modelo lineal probabilístico como se enseña en la Figura 6. comparándose los dos modelos; el ajuste de la recta para los datos dados y la difusividad para cada observación tal cual como se muestra en la Figura 7.

4. Conclusiones

Se puede apreciar que la regresión lineal difusa, aplicando el criterio que sugiere Bo Yuan y Klir [1] no es tan precisa como la regresión lineal probabilística, pero una ventaja al aplicar la regresión difusa es que se ajusta de una mejor manera a nuestro lenguaje impreciso.

La lógica difusa abre muchas puertas para poder seguir estudiando una cantidad de fenómenos físicos, químicos, estadísticos, etc. Por esta razón, la estimación de coeficientes difusos de mínimos cuadrados en un modelo de regresión difuso es un buen trabajo de estudio y abre muchas puertas para continuar con la investigación con estos conjuntos borrosos aplicados a la estadística.

5. Referencias

[1] J. Klir, George.Yuan Bo, V. M. (1995) Fuzzy Set and Fuzzy Logic: Theory and Aplications". Ed. Prentice Hall, USA

[2] Zadeh, L. A. (1965) Information and Control: Fuzzy Subsets". 8. pp.338-353

[3] Devore, Jay L. (1998); "Probabilidad y Estadista para Ingeniería y Ciencias", Ed. International Thomson Editores, México.

[4] Kaufmann, Arnold-M. Gupta(1985); " Introduction to Fuzzy Arithmetic: Theory and Applications", Ed. Van Nostrand Rinhold Company,USA.

[5] Donoso, Sergio, (2006). "Analisis de Regresion Difusa: Nuevos Enfoques y Aplicaciones", Tesis de Doctoral, Universidad de Granada, Facultad de Ingeniería.

[6] Arias Martínez, Carlos Humberto (1988); "Conjuntos Borrosos, Teoría Básica y Aplicaciones al Análisis de Decisiones", Tesis de maestria en IDO, Facultad de Ingeniería, UNAM

[7] Buckley, J. J. (2006). Fuzzy prediction in linear regression. Fuzzy Probability and Statistics, 177-179.

[8] Wu, H. C. (2008). Fuzzy linear regression model based on fuzzy scalar product. Soft Computing, 12(5), 469-477.

[9] Karim, S. A. A., & Kamsani, N. F. (2020). Water Quality Index Prediction Using Multiple Linear Fuzzy Regression Model: Case Study in Perak River, Malaysia. Springer Nature.

[10] Xu, J. W., & Xu, R. N. (2016). Statistical Diagnostic of Center Fuzzy Linear Regression Model Based on Fuzzy Decentring Degree. In Fuzzy Systems & Operations Research and Management (pp. 233-241). Springer, Cham.

[11] Nong, X. (2011, December). A new fuzzy linear regression model for least square estimate. In International Conference on Information and Business Intelligence (pp. 709-715). Springer, Berlin, Heidelberg.

[12] Buckley, J. J., & Jowers, L. J. (2007). Fuzzy Linear Regression I. In Monte Carlo Methods in Fuzzy Optimization (pp. 117-125). Springer, Berlin, Heidelberg.

[13] Arnold, B. F., & Gerke, O. (2003). Testing fuzzy linear hypotheses in linear regression models. Metrika, 57(1), 81-95.