İNSAN&İNSAN, Sayı/Issue 6, Güz/Fall 2015, 27-36
ISSN: 2148-7537, www.insanveinsan.org

Bağımsız Değişkenin Pareto Dağılımına Sahip Olması Durumunda Üyelik Fonksiyonunun Dayalı Parametre Tahmini

TÜRKAN ERBAY DALKILIÇ*
tedalkilic@gmail.com

TUĞBA KARAN**
tugba.karan@tuik.gov.tr

Özet: *Basit ve çoklu doğrusal regresyon çözümlemesinde klasik metotlardan yararlanılıyorsa, modele ilişkin parametrelerin tahmini bazı varsayımlara dayanır. Doğrusal regresyon modellerinin bilinmeyen parametrelerinin tahmininde verilerin normal dağılım dışında bir dağılıma sahip olması durumu tahmin sürecinde klasik çözümlemelerin dışına çıkılmasını gerektirir. Böyle durumlarda bulanık mantığa dayalı çözümleme yöntemleri alternatif yöntemler olarak kendini göstermektedir. Bu çalışmada, bağımsız değişkenlerden herhangi birinin Pareto dağılımına sahip olması ve veri setinde aykırı gözlemlerin mevcut olması durumunda, kurulacak çoklu doğrusal regresyon modelinin bilinmeyen parametrelerini tahmin etmek için bir algoritma önerilerek, algoritmadan elde edilen tahminler literatürde yer alan mevcut yöntemlerden elde edilen tahminler ile karşılaştırılmıştır.*

Anahtar kelimeler: *Parametre Tahmini, Bulanık Üyelik Fonksiyonu, Pareto Dağılımı.*

Giriş

Regresyon çözümlemesi yöntemlerinin her biri için veri analizi oldukça önemlidir. Çünkü veri setinde yer alan değişkenlerin hangi dağılımdan geldikleri regresyon modelindeki parametre tahminleri için belirleyici rol oynamaktadır (Zadeh, 1965). Ayrıca veri setinde aykırı gözlemlerin varlığı da kullanılacak yöntemin önemini artırmaktadır. Çünkü aykırı gözlemler yapılacak parametre tahminlerini ana veri setinden uzaklaştıracak önemde olabilir. Basit ve çoklu doğrusal regresyon modellerinin bu varsayım bozulmalarından etkilenme düzeyini en aza indirgemek üzere literatürde Huber, Hampel, Andrews ve Tukey gibi güçlü (robast), model tahmin yöntemleri mevcuttur.

Pareto analizi, değişik sayıdaki önemli nedenleri daha az önemde olan nedenlerden ayırmak için kullanılan bir yöntemdir. Willfredo Pareto isimli İtalyan ekonomist Pareto Prensibi de denilen 80/20 kuralını ortaya çıkarmış ve bu kural Pareto analizi yönteminin temelini oluşturmuştur. Willfredo Pareto ilk olarak İtalya topraklarının % 80' ine nüfusun %20'sinin sahip olduğunu fark ederek bu prensibi kurmuştur.

Bu çalışmada, bağımsız değişkenlerden en az birinin Pareto dağılımına sahip olması durumunda, kurulacak çoklu doğrusal regresyon modelinin bilinmeyen paramet-

* Doç. Dr., Karadeniz Teknik Üniversitesi, Fen Fakültesi İstatistik ve Bilgisayar Bilimleri Bölümü.
** İstatistikçi, TÜİK İstanbul Bölge Müdürlüğü.

relerini tahmin etmede normal dağılıma uymama sorunlarından mümkün olduğu kadar az etkilenen bir tahmin yöntemi önerilmiştir. Bağımsız değişkenlerden Pareto dağılımına sahip olan değişkenin modele katkısı, Pareto dağılımına uygun elde edilen üyelik fonksiyonu ile derecelendirilmiştir. Bu anlamda kullanılan yöntem güçlü (robast) yöntemler sınıfında değerlendirilebilir.

Zadeh, (1965) Bulanık yaklaşım konusundaki ilk ciddi adımı 1965 yılında yayınlanan "Bulanık Kümeler" adlı makale ile atmıştır. Zadeh bulanık mantığın matematik ve bilgisayar bilimleri alanlarındaki uyarlanabilirliği üzerinde durmuştur.

Civanlar ve Trussel (1986) "İstatistiksel veriler kullanılarak üyelik fonksiyonlarının oluşturulması" başlıklı çalışmalarında üyelik fonksiyonunun belirlenmesinin, bulanık küme teorisinin pratik uygulamalarında önemli olduğu belirtilmiş ve elemanları bilinen bir olasılık yoğunluk fonksiyonu ile belirleyici niteliklere sahip, bulanık kümelere dair üyelik fonksiyonunun belirlenmesi için bir yöntem sunmuşlardır. Çalışmada üyelik fonksiyonunun sağlaması gereken koşullar da verilmiştir.

Dombi (1990) üyelik fonksiyonları üzerine yaptığı çalışmada kullanılan farklı üyelik fonksiyonlarını tanımlamış, üyelik fonksiyonlarının kurulması için gerekli özellikler ve üyelik fonksiyonlarının matematiksel formları ile ilgili bilgi vermiştir.

Erbay Dalkılıç (2005) bağımsız değişkenlerin üstel dağılımdan gelmesi durumunda switching regresyonda bulanık sinir ağları yaklaşımı ile parametre tahmini yapmış ve mevcut tahmin yöntemlerinden elde edilen sonuçlar ile karşılaştırmıştır.

Çalışmanın uygulama kısmında önerilen yöntemin geçerliliğinin irdelenebilmesi için sayısal örneklere yer verilmiş ve ele alınan veri setleri için elde edilen sonuçlar literatürde var olan klasik yöntemlerden elde edilen sonuçlar ile karşılaştırılmıştır.

1. Üyelik Fonksiyonu

Ele alınan problemlerin yapılarına göre üyelik fonksiyonları farklılıklar gösterebilmektedir. Bununla birlikte üyelik fonksiyonlarının ortak birtakım özellikleri bulunmaktadır; üyelik fonksiyonları sürekli fonksiyonlardır ve bir [a,b] aralığını $\mu(x)$ fonksiyonu yardımı ile [0,1] aralığına dönüştürürler. Üyelik fonksiyonunun doğrusal biçimde ya da doğrusallaşabilen bir yapıda olması büyük önem taşımaktadır. Optimal üyelik fonksiyonunun bulunabilmesi için;

$$1.\ E\{\mu(x) \mid x \text{ bir olasılık yoğunluk fonksiyonuna göre dağılsın}\} \geq 0$$

$$2.\ 0 \leq \mu(x) \leq 1$$

$$3.\ \int \mu^2(x)\, d(x) \text{ en küçüklenmelidir.}$$

biçiminde verilen koşulların sağlanması gerekmektedir. Bu koşullar altında optimal üyelik fonksiyonu;

$$\mu(x) = \begin{cases} \lambda p(x) & \text{eğer } \lambda p(x) < 1 \\ 1 & \text{eğer } \lambda p(x) \geq 1 \end{cases} \qquad (1)$$

biçimindedir. Burada,

$p(x)$: olasılık yoğunluk fonksiyonu
λ : sabit

tir (Civanlar ve Trussel, 1986). Verilen üyelik fonksiyonunda $p(x)$, ilgilenilen dağılıma ilişkin olasılık yoğunluk fonksiyonu olduğundan formu belirlidir. Ancak λ sabiti,

$$P: \min_{\mu} f(\mu) = \frac{1}{2} \int_{-\infty}^{+\infty} \mu^2 (x)\, d(x)$$

$$G(\mu) = c - E\{\mu\} = c - \int_{-\infty}^{+\infty} \mu(x)p(x)\, d(x) \leq 0 \qquad (2)$$

$$\mu \in \Omega = \{\mu(x) \mid 0 \leq \mu(x) \leq 1\}$$

ile tanımlanan problemin çözümü ile elde edilebilir. P ile verilen problem, optimal üyelik fonksiyonu için tanımlanan koşullardan oluşturulmaktadır. λ sabitinin elde edilmesi için, Eşitlik (2)' de P ile ifade edilen problem, Langrange yöntemi ile çözümlenebilir.

$$L(\mu, \lambda) = \frac{1}{2} \int_{-\infty}^{+\infty} \mu^2 (x)d(x) + \lambda \left\{ c - \int_{-\infty}^{+\infty} \mu(x)p(x)d(x) \right\} \quad (3)$$

biçimindedir. Burada, $\lambda \geq 0$ langrange çarpanı ve $c < 1$ olacak biçimde bir sabittir.

Eşitlik (1)' de verilen üyelik fonksiyonunun değerleri Eşitlik (3)' te yerine konularak λ sabitinin değerini verebilecek fonksiyona ulaşılmaya çalışılır. Bunun için $(\lambda p(x)) \leq 1$ ve $(\lambda p(x)) > 1$ biçimindeki iki durum için çözümleme yapılır:

i) $(\lambda p(x)) \leq 1 \Longrightarrow \mu(x) = \lambda p(x)$

olur. Bu durumda,

$$L = \frac{1}{2} \int_{-\infty}^{+\infty} \lambda^2 p^2(x)d(x) - \lambda \int_{-\infty}^{+\infty} \lambda p^2(x)d(x) + \lambda c$$

$$L = -\frac{1}{2} \int_{-\infty}^{+\infty} \lambda^2 p^2(x)d(x) + \lambda c$$

elde edilir. Bu fonksiyonun λ sabitine göre türevi alındığında,

$$\frac{\partial L}{\partial \lambda} = -\lambda \int_{-\infty}^{+\infty} p^2(x)d(x) + c \qquad (4)$$

eşitliğine ulaşılır.

ii) $(\lambda p(x)) > 1 \Longrightarrow \mu(x) = 1$

olur. Bu durumda,

$$L = \frac{1}{2} \int_{-\infty}^{+\infty} d(x) - \int_{-\infty}^{+\infty} \lambda p(x) d(x) + \lambda c$$

elde edilir. Bu fonksiyonun λ sabitine göre türevi alındığında;

$$\frac{\partial L}{\partial \lambda} = 0 - \int_{-\infty}^{+\infty} p(x) d(x) + c \qquad (5)$$

eşitliğine ulaşılır. İkinci aşamada elde edilen (5) eşitliği λ'dan bağımsız olduğu için, birinci aşamada elde edilen (4) eşitliği kullanarak λ parametresi,

$$\frac{\partial L}{\partial \lambda} = -\lambda \int_{-\infty}^{+\infty} p^2(x) d(x) + c \qquad (6)$$

eşitliğinden,

$$\lambda = \frac{c}{\int_{-\infty}^{+\infty} p^2(x) d(x)} \qquad (7)$$

biçiminde elde edilir (Klir, Yuan, 1995; Baykal, Beyan, 2004; Chen, Wang, 1999).

2. Pareto Dağılımı için Optimal Üyelik Fonksiyonunun Belirlenmesi

Bu çalışmada veri setinde yer alan değişkenlerden herhangi birinin Pareto dağılımına sahip olması ve aykırı değerlerin varlığı durumunda doğrusal regresyon modelinin parametrelerinin tahmini ile ilgilenildiğinden, Pareto dağılımına ilişkin üyelik fonksiyonu Civanlar ve Trussel (1986) tarafından önerilen bu yöntem kullanılarak aşağıdaki gibi elde edilmiştir:

Eğer X Pareto dağılımına sahip rastgele bir değişken ise olasılık yoğunluk fonksiyonu,

$$f(x;k,x_m) = k \frac{x_m^k}{x^{k+1}} \quad x \geq x_m \qquad (8)$$

biçimindedir. Pareto dağılımı için optimal üyelik fonksiyonun elde edilmesi için öncelikle dağılıma uygun λ parametresi belirlenmelidir. Pareto dağılımına ilişkin eşitlik ile verilen olasılık yoğunluk fonksiyonu eşitliğinde yerine konulduğunda λ sabiti,

$$\lambda = \frac{c}{\int_{x_m}^{+\infty} p^2(x)d(x)} = \frac{c}{\int_{x_m}^{+\infty} k\left[\frac{x_m{}^k}{x^{k+1}}\right]^2 d(x)}$$

$$\lambda = \frac{c}{k^2\, x_m{}^{2k} \int_{x_m}^{+\infty} \frac{1}{x^{2(k+1)}}\, d(x)}$$

$$\lambda = c\,\frac{2(k+1)x_m}{k^2} \tag{9}$$

biçiminde elde edilir. Üyelik fonksiyonu, (9) eşitliği ile elde edilen λ parametresini kullanarak,

$$\mu(x)=\lambda p(x)= c\,\frac{2(k+1)x_m}{k^2}\, k\,\frac{x_m{}^k}{x^{k+1}}$$

$$\mu(x) = \frac{2(k+1)}{k}\left(\frac{x_m}{x}\right)^{k+1} \tag{10}$$

biçiminde elde edilir.

3. Uygulama

Önerilen algoritmanın geçerliliğinin sınanması için bu bölümde ele alınan veri seti 149 gözlemden oluşmakta ve verilerin bir kısmı Tablo 1'de yer almaktadır. Veri setinde bir bağımlı ve üç bağımsız değişken bulunmaktadır. Bağımlı değişken Y, bağımsız değişkenler X_1, X_2, ve X_3 ile ifade edilmiştir. Tablo 1 ayrıca önerilen algoritmadan ve EKK yönteminden elde edilen tahminleri ve tahminlere ilişkin hataları da içermektedir (Gamgam, Altunkaynak, 2012; Köseoğlu, Yamak, 2004).

H_0 : $O_i = e_i$ (gözlenen frekans beklenen frekanslara uygundur.)

H_1: $O_i \neq e_i$ (gözlenen frekans beklenen frekanslara uygun değildir.)

Bağımsız değişkenlerden X_3'e ilişkin Kolmogrov-Smirrnov (K-S) uyum iyiliği testine göre *($k = 2{,}8$, $x_m = 5{,}2$)* parametreleri ile Pareto dağılımına sahip olduğu hesaplanmıştır.

$$K\text{ - }S_{\text{Hesaplanan Değer}} = (0{,}0472) < K\text{ - }S_{\text{Tablo Değeri}} = (0{,}1335)$$

olduğundan X_3 rasgele değişkeninin dağılımının Pareto dağılımından farkı yoktur biçiminde kurulan yokluk hipotezi, H_0 kabul edilir.

Tablo 1. Veri Setine İlişkin Tahminler ve Hatalar

Gözlem No	X_1	X_2	X_3	Y	$\hat{Y}_{(ÖA)i}$	$e_{(ÖA)i}$	$\hat{Y}_{(EKK)i}$	$e_{(EKK)i}$
1	69.5800	1.6133	6.0000	31.7880	39.6349	-7.8469	45.6409	-13.8529
2	76.6400		6.2000	27.9600	39.4771	-11.5171	45.0669	-17.1069
3	81.9200	1.5900	7.5000	43.0870	39.3869	3.7001	43.8577	-0.7707
4	80.9200		7.6000	41.1750	39.4268	1.7482	43.7737	-2.5987
5	85.2200	2.1600	6.6000	44.9820	39.8287	5.1533	45.2810	-0.2990
6	88.6000		7.8000	43.0630	40.1280	2.9350	43.6270	-0.5640
7	89.8800	2.1900	6.2000	42.1820	41.1277	1.0543	46.3927	-4.2107
8	86.8000		8.5000	40.9500	40.6731	0.2769	44.0117	-3.0617
9	91.8200	2.1467	7.5000	45.2670	40.2464	5.0206	44.0345	1.2325
10	94.7200		8.7000	41.5210	41.0775	0.4435	43.2670	-1.7460
11	93.2400	2.3867	6.2000	40.5280	40.6164	-0.0884	46.0086	-5.4806
12	95.1800		5.4000	40.8660	43.1497	-2.2837	48.7825	-7.9165
13	98.4400	2.4200	7.5000	39.2390	40.3473	-1.1083	43.8335	-4.5945
14	97.3600		5.8000	32.9150	40.2679	-7.3529	45.8507	-12.9357
15	101.3400	3.1467	7.9000	39.5440	39.5991	-0.0551	42.6576	-3.1136
16	100.5000		6.0000	26.9140	40.1354	-13.2214	45.3526	-18.4386
17	97.6800	2.3767	8.6000	38.4840	41.1938	-2.7098	43.3464	-4.8624
18	97.4200		17.9000	44.1730	40.8017	3.3713	27.1834	16.9896
19	105.4000	3.0500	6.6000	29.4420	42.2717	-12.8297	45.8328	-16.3908
.	.		.	.	.	.	.	.
.	.	2.2733	.	.	.	.	.	.
.	.		.	.	.	.	.	.
142	188.1800	2.9967	5.2000	41.5530	40.9170	0.6360	48.8308	-7.2778
143	194.5600		11.6000	52.6630	40.8762	11.7868	35.1598	17.5032
144	194.1800	2.3633	9.3000	45.3860	40.9183	4.4677	38.8686	6.5174
145	195.6800		14.7000	42.6410	38.0316	4.6094	27.8065	14.8345
146	196.0200	1.9100	13.3000	44.7110	40.6469	4.0641	32.6420	12.0690
147	199.9400		6.2000	41.6830	41.8662	-0.1832	44.8849	-3.2019
148	200.6000	2.1567	9.9000	41.1220	40.2164	0.9056	37.3797	3.7423
149	201.1400		11.3000	24.3580	38.1484	-13.7904	33.5026	-9.1446
		1.8733						
		3.0433						
		2.9000						
		2.6833						
		.						
		.						
		.						
		4.0633						
		2.7867						
		2.6733						
		1.7800						
		2.9967						
		3.1067						
		2.5267						
		1.7967						
	HATA				$\varepsilon_{ÖA}$ = 119.0887		ε_{EKK} = 201.1761	

Verilere ilişkin çoklu doğrusal regresyon modelini elde etmeden önce Pareto dağılımı gösterdiği belirlenen X_3 değişkenine ilişkin üyelik dereceleri,

$$\mu(X_{3i}) = \frac{2(k+1)}{k}\left(\frac{X_{3m}}{X_{3i}}\right)^{k+1}$$

ile belirlenir, burada $k = 2{,}8$ ve $X_{3m} = 5{,}2$ dir. Ayrıca, $[0,1]$ aralığında değerler alabilen c sabiti 0.4 olarak belirlenmiştir. Bu sabit, üyelik fonksiyonunun olasılık yoğunluk fonksiyonuna uyumunun kontrolü ile elde edilmiştir. Önerilen algoritma için çoklu doğrusal regresyon modeline ilişkin parametre değerleri, X matrisinde X_3 rasgele değişkeni üyelik fonksiyonundan elde edilen üyelik dereceleri ile ağırlıklandırılmış olmak üzere;

$B=(X^T X)^{-1} X^T Y$

eşitliğinden ve MATLAB da yazılan ve önerilen algoritmanın işleyişini sağlayan program ile;

$$\hat{Y}_{ÖA} = 37.6244 + 0.01785X_1 + 0.9888X_2 + 100.9175X_3 \quad (11)$$

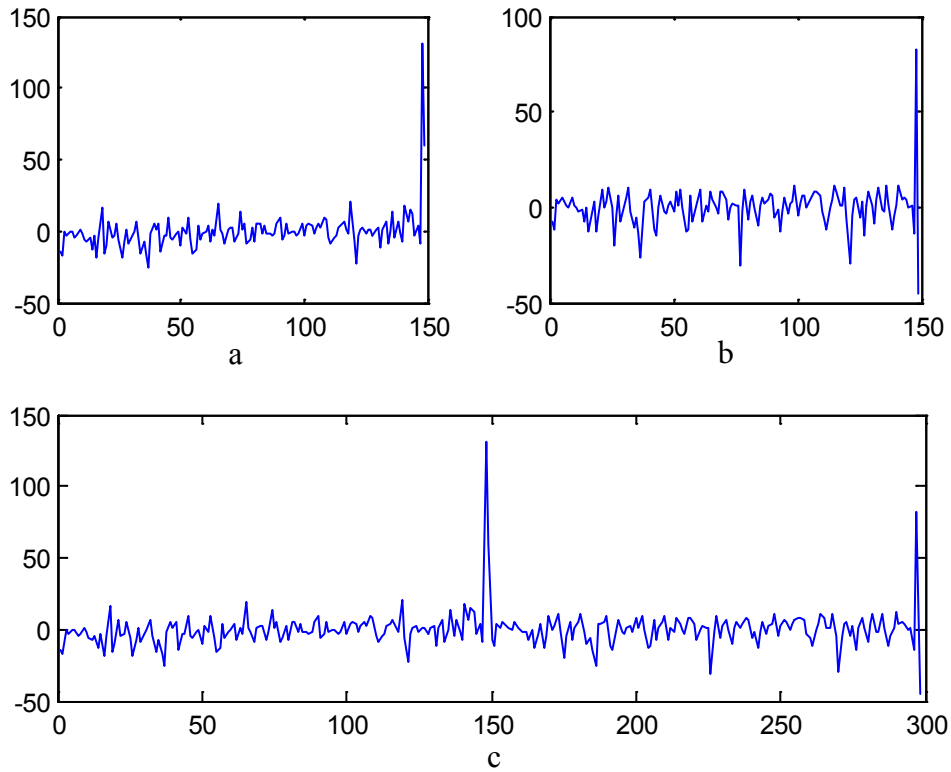
biçiminde elde edilmiştir. En Küçük Kareler yöntemine ilişkin model ise;

$$\hat{Y}_{EKK} = 54.4619 - 0.0263x_1 + 2.0167x_2 - 1.7076x_3 \quad (12)$$

biçimindedir.

Önerilen algoritmadan elde edilen modelin tahmin değerlerine ilişkin hata $\varepsilon_{ÖA} = 119.0887$, En Küçük Kareler yönteminden elde edilen modelin tahmin değerlerine ilişkin hata $\varepsilon_{EKK} = 201.1761$ olarak elde edilmiştir. Şekil 1-a En küçük Kareler modelinden elde edilen hata miktarlarını, Şekil 1-b önerilen algoritmadan elde edilen tahminlerin hata miktarlarını göstermektedir. Tüm modellerden elde edilen hatalar grafikleri karşılaştırmalı olarak Şekil 1-c'de yer almaktadır.

Önerilen algoritma ile elde edilen tahminlerin En Küçük Kareler yönteminden elde edilen modele ilişkin tahminlerden daha küçük hatalara sahip olduğu gözlenmiştir.



Şekil 1. Tablo 1'de yer alan veri setine ilişkin hata grafikleri

Sonuç ve Öneriler

Çoklu doğrusal regresyon analizinde kurulacak modele ilişkin bağımsız değişkenlerden herhangi biri Pareto dağılıma sahip olduğunda klasik yöntemlerin dışına çıkılması gerektiği belirtilmişti. Bu çalışmada çoklu doğrusal regresyon modeline ilişkin parametre tahmini sürecinde, Pareto dağılımı gösteren bağımsız değişkenin ağırlıklandırılmasında kullanılmak üzere, Pareto dağılımına uygun, dağılımın olasılık yoğunluk fonksiyonuna dayalı üyelik fonksiyonu elde edilmiştir. Çoklu doğrusal regresyon modelinin bilinmeyen parametreleri önerilen yöntem ve literatürde yer alan En Küçük Kareler yöntemi ile tahmin edilerek, belirlenerek modellerin performansları verdikleri tahmin değerleri ve bu değerlere ilişkin hata miktarları ile karşılaştırılmıştır. Elde edilen tahminlere ilişkin hata miktarlarının düşük olması, önerilen yöntemin, karşılaşılacak çoklu doğrusal regresyon modeli kurma problemlerinde alternatif olarak kullanılabilir olduğunu göstermektedir. Bağımsız değişkenlerden herhangi birinin Pareto dağılımı dışında başka dağılımlara sahip olmaları durumu irdelenerek bu durumlara ilişkin yöntemler geliştirilebilir.

Kaynakça

Baykal, Nazife ve Timur Beyan. *Bulanık Mantık İlke ve Temelleri*. Ankara: Bıçaklar Kitabevi, 2004.

Chen, M. S. ve Wang, S. W. "Fuzzy Clustering Analaysis for Optimizing Fuzzy Membership Functions". *Fuzzy Sets and Systems* 103. (1999): 239-254.

Civanlar, M. R. ve Trussell, H. J. "Tructing Membership Functions Using Statistical Data". *Fuzzy Sets and Systems* 18. (1986): 1-13.

Dombi, J. "Membership Functions As An Evalution". *Fuzzy Sets and Systems* 35. (1990): 1-21.

Erbay, Dalkılıç, Türkan. "Switching Regresyon'da Bulanık Sinir Ağları Yaklaşımı ile Parametre Tahmini". Doktora Tezi, Ankara Üniversitesi, Fen Bilimleri Enstitüsü, 2005.

Gamgam, Hamza ve Bülent Altunkaynak. S*PSS Uygulamalı Parametrik Olmayan Yöntemler*. Ankara: Seçkin Yayıncılık, 2012.

Klir, G. J. ve Yuan, B. *Fuzzy Sets and Fuzzy Logic*, USA: Prentice-Hall, 1995.

Köseoğlu, Mustafa ve Rahmi Yamak. *Uygulamalı İstatistik ve Ekonometri*. Trabzon: Celepler Matbaacılık, 2004.

Zadeh, L. A. "Fuzzy Sets". *Informationand Control* 8. (1965): 338-353.

Parameter Estimation Based on Membership Function When an Independent Variable Has Pareto Distribution

Türkan Erbay Dalkılıç / Tuğba Karan

Abstract: *If classical methods are being employed in simple and multiple linear regression analysis, estimation of parameters related to the model is based on some assumptions. If the data have any distribution other than normal distribution in estimation of unknown parameters of linear regression models, the analysis has to go beyond classical analysis. In such cases analysis methods based on fuzzy logic manifest themselves as alternative ways. In this paper, an algorithm has been proposed in order to be able to estimate the unknown parameters of multiple linear regression model in the case that any independent variable has a Pareto distribution and incompatible observations exist in data set and estimations obtained from this algorithm have been compared with estimations obtained from the methods existing in the literature.*

Keywords: *Parameter Estimation, Fuzzy Membership Function, Pareto Distribution.*