İstatistik Araştırma Dergisi Cilt 04 No 03 Sayfa 76-91 Aralık 2005

EKK VE BAZI DAYANIKLI TAHMİNCİLERİN DERİNLİKLERİNİN KİRLENMEYE KARŞI DEĞİŞİMLERİNİN İNCELENMESİ

Enis SINIKSARAN* M. Hakan SATMAN* Y. Barış ALTAYLIGİL*

ÖZET

Bu çalışmada, veri kirliliği karşısında EKK ve bazı dayanıklı (robust) regresyon yöntemlerinin derinliklerinin davranışları incelendi.Yapılan Monte Carlo simülasyonları sonucunda bazı tahmincilere ilişkin örüntüler tespit edildi. Çalışmada ayrıca Bootstrap yöntemi ile elde edilen derinlik dağılımlarına dayanarak, EKK tahmincilerinin hipotez testlerinin derinliğe dayalı olarak yapılabileceği gösterildi.

Anahtar Kelimeler : *Aykırı Değerler, Bootstrap, Dayanıklı Yöntemler, En Derin Regresyon, P-Değeri Regresyon Derinliği.*

1. GİRİŞ

Derinlik, DR (Deepest regression) dışında herhangi bir dayanıklı yöntemin ve EKK'nın amaç fonksiyonunda yer alan bir kavram değildir. Bu anlamda regresyon derinliğinin, DR dışındaki tahmincilerin kalitelerine ilişkin bir ölçüt olup olamayacağı konusunda, literatürde net bir yanıt bulunmamaktadır. Bu çalışmada temel olarak bu sorunun yanıtı arandı.

Çalışma ana hatlarıyla şöyledir: Bu bölümün ardından gelen bölümde regresyon derinliği kavramı ve DR tahmincisi tanıtıldı. 3. bölümde aykırı değerlere karşı dayanıklı olan DR, LMS (Least Median Squares), LTS (Least Trimmed Squares) ve Huber'in M tahmincileriyle birlikte EKK ve LAD (Least Absolute Deviation) tahmincilerinin çeşitli oranlardaki kirlenmeler karşısında derinliklerinin nasıl değiştiği Monte Carlo simülasyonlarına dayanarak incelendi. Kirlenmenin yönü ve derecesine göre değişik örüntüler saptandı. Çalışmanın 4. bölümünde regresyon derinliğinin bootstrap dağılımları elde edildi ve bu dağılımlara dayanarak EKK parametrelerine ilişkin hipotez testlerinin yapılabilme olanakları araştırıldı. Klasik hipotez testlerinin p-değerleri ile karşılaştırıldığında, oldukça yakın sonuçlar elde edildiği görüldü.

* İstanbul Üniversitesi İktisat Fakültesi, Ekonometri Bölümü, İstanbul, TÜRKİYE
esiniksaran@istanbul.edu.tr

2. REGRESYON DERİNLİĞİ VE EN DERİN REGRESYON

Derinlik kavramının, ilk kez "yarı-uzay derinliği" olarak Tukey tarafından ortaya atılmasının ardından (Tukey, 1975) sorgulayıcı istatistikte konum, basıklık ve çarpıklık ölçüsünden (Liu, Parelius, Singh, 1990), çanta grafiği (bag plot) ile aykırı değer teşhisçisine (Rousseeuw, Ruts ve Tukey, 1999) ve kalite kontrol grafiklerinden (Liu ve Singh, 1993) asimptotik p-değerlerinin hesabına (Liu ve Singh, 1997) kadar pek çok alanda uygulanmıştır. (Karabulut ve Öztürk, 2003) ise yarı uzay derinliği kavramını dengeli bootstrap güven aralıklarının oluşturulmasında kullanmışlardır.

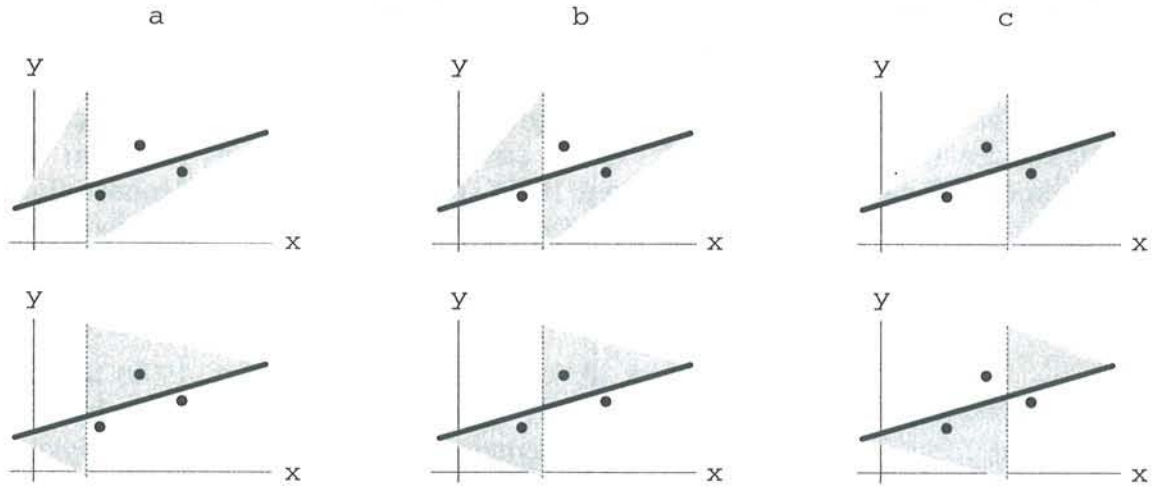
Tukey'in tanımına dayanarak ilk kez (Rouseeuw ve Hubert, 1999) tarafından sunulan regresyon derinliği kavramı (rdepth) ise Daniels'ın (Daniels, 1954) çalışması ile yakından ilişkilidir. Regresyon derinliği kavramı bir tahmincinin (θ) bir veriyle (Z_n) olan ilişkisini temsil eder. Burada θ, n birimlik bir örneklemden herhangi bir regresyon yöntemi ile elde edilmiş parametre tahmincilerinin oluşturduğu bir vektördür. Örneğin, $y = \beta_0 + \beta_1 x + \varepsilon$ basit doğrusal regresyon modelinin tahmin denklemi $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ ise, $\theta = (\hat{\beta}_0, \hat{\beta}_1)$ ve $Z_n = \{(x_i , y_i), i = 1, 2,\ldots,n\} \subset R^2$ olacaktır. Bu uzay için regresyon derinliğinin hesabında (Rousseeuw ve Hubert, 1999) bir algoritma önermişlerdir. Buna göre; verideki gözlemler x değerleri aynı olanlardan sadece biri alınarak $x_1 \leq x_2 \leq \ldots \leq x_n$ şeklinde dizildikten sonra n tane referans doğrusu x-eksenini $\{x_1\text{-}1, (x_1+x_2)/2, \ldots, (x_{n-1}+x_n)/2\}$ değerlerinden dik olarak kesecek şekilde belirlenir. Her bir v referans doğrusunun solunda kalan pozitif kalıntılar L^+, sağında kalan negatif kalıntılar R^- olarak gösterilmek üzere

$$L^+(v) = \#\{j; x_j \leq v \text{ ve } r_j \geq 0\} \quad ve \quad R^-(v) = \#\{j; x_j > v \text{ ve } r_j \leq 0\} \qquad (1)$$

değerleri hesaplanır. Benzer şekilde L^- ve R^+ değerleri de hesaplandıktan sonra bir doğrunun regresyon derinliği:

$$rdepth(\theta, Z_n) = \min_{1 \leq i \leq n} (\min\{L^+(x_i) + R^-(x_i), L^-(x_i) + R^+(x_i)\}) \qquad (2)$$

eşitliğini sağlayacak şekilde bulunur.



Şekil 1. Regresyon Derinliğinin Araştırılması

(2) eşitliği başka bir biçimde açıklanırsa; bir doğrunun regresyon derinliği, doğruyu uyumsuz (non-fit) yapabilmek için veriden atılması gereken minimum gözlem sayısıdır. Burada uyumsuz bir doğrunun geometrik yeri $x = c$ doğrusudur. Geometrik olarak açıklanırsa, bir doğru gözlemlerle çakışmayan her bir kaldıraç noktasına göre ve uyumsuz bir doğruya (x eksenine dik bir doğruya) dönüşecek şekilde çevrildiğinde üzerinden geçtiği nokta sayılarından minimum olanı bu doğrunun derinliğini verecektir denilebilir. Şekil 1'de 3 gözlemden oluşan bir verinin serpilme diyagramı ve bu veriye ilişkin bir regresyon doğrusu görmekteyiz. Kesikli olarak gösterilen v referans doğrularının, regresyon doğrusu ile kesim noktaları kaldıraç noktası olarak düşünülürse, regresyon doğrusunun her bir kaldıraç noktasına göre saat yönünde (üsttekiler) ve saatin tersi yönünde (alttakiler) döndürüldüğü durumlar a, b ve c olarak 3 bölgede gösterilmiştir. Regresyon doğrusu dönerken taradığı alanlar da Şekil 1'de görülmektedir. Eşitlik (2) uygulanırsa;

a: $(L^{+} + R^{-}) = 0 + 2 = 2$, $(L^{-} + R^{+}) = 0 + 1 = 1 \Rightarrow \min(2, 1) = 1$

b : $(L^{+} + R^{-}) = 0 + 1 = 1$, $(L^{-} + R^{+}) = 1 + 1 = 2 \Rightarrow \min(1, 2) = 1$

c : $(L^{+} + R^{-}) = 1 + 1 = 2$, $(L^{-} + R^{+}) = 1 + 0 = 1 \Rightarrow \min(2, 1) = 1$

$rdepth(\theta, Z_n) = \min(1, 1, 1) = 1$ yazılabilir.

Bir verideki gözlemlerin tamamı regresyon doğrusunun üzerinde yer alırsa, regresyon doğrusu döndürüldüğünde bu noktaların üzerinden geçmiş olarak düşünüleceği için regresyon derinliği "n" olacaktır. Dolayısıyla bir doğrunun regresyon derinliği için $0 \leq \text{rdepth}(\boldsymbol{\theta}, Z_n) \leq n$ eşitsizliği yazılabilir.

Bağımsız değişken sayısı arttırıldığı zaman ise regresyon derinliği tanımlanırken doğruların yerini düzlemler (iki bağımsız değişkende) ve hiperdüzlemler (ikiden fazla bağımsız değişkende) alacaktır. Bu çalışmada regresyon derinliğini hesaplamada 2 boyut için yazılan ve ekte sunulan Mathematica kodları, 3 ve 4 boyut içinse Rousseeuw vd.'nin yazdığı Fortran yazılımları olan Medsweep, Rdepth3 ve Rdepth4[†] kullanılmıştır. (Bakınız Ek.1)

Bir regresyon tahmincisi θ'nın derinliği, uyum kalitesi olarak düşünülürse, maksimum derinliği verecek olan tahminciyi aramak da doğal olacaktır. Amaç fonksiyonu derinliğe dayanan ve en derin regresyon (DR) olarak isimlendirilen bu regresyon Rousseeuw ve Hubert (1999) tarafından aşağıdaki gibi tanımlanmış ve hesaplanması için algoritmalar sunulmuştur:

$$DR(Z_n) = \arg\max_{\theta} \text{rdepth}(\theta, Z_n)$$

Burada θ, maksimum derinliğe sahip tahmincidir. θ, birden fazla olduğu zaman ise bu tahmincilerin ortalamasının alınması önerilmektedir. Maksimum derinliğe ilişkin alt ve üst sınır ise $Z_n \subset R^p$ için

$$\left\lceil \frac{n}{p+1} \right\rceil \le \max_{\theta} \text{rdepth}(\theta, Z_n) \le \left\lceil \frac{n+p}{2} \right\rceil$$

eşitsizliği ile belirlenir.

Amaç fonksiyonunda derinliğin yer almadığı EKK, LTS, LAD tahmincilerinin derinliklerine ilişkin alt sınırların ise birincisi için 1, diğerleri için p olduğu yukarıda söz edilen çalışmada ispatlanmıştır. Ancak literatürde bu tahmincilerin ve LMS, M gibi diğer dayanıklı tahmincilerin regresyon derinliklerine, özellikle kirlenme altındaki davranışlarına ilişkin ampirik bulgular henüz yer almamıştır. Bundan sonraki bölümde bu konuya ilişkin Monte Carlo simülasyonlarından elde edilen sonuçlar ve yorumları sunulacaktır.

3. DERİNLİK HESAPLAMALARINA İLİŞKİN MONTE CARLO SİMÜLASYONLARI

Çalışma tek, iki ve üç bağımsız değişkenli lineer regresyon modelleri için düşünüldü:

Model 1 : $y = \beta_0 + \beta_1 x + \varepsilon$ (3)

Model 2 : $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$ (4)

[†] Söz konusu programlar http://www.agoras.ua.ac.be adresinden indirilebilir.

Model 3 : $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$ (5)

Simülasyonlarda temiz veri şöyle üretildi: $\boldsymbol{X}$ matrisinin sütunları; ortalamaları 7, varyansları 16 ve kovaryansları 0 olan çok değişkenli normal dağılımdan; hata vektörü ε ise standart normal dağılımdan elde edildi. Parametre vektörü $\boldsymbol{\beta} = [\beta_0, \beta_1, ..., \beta_k]' = [5, 5, ..., 5]'$ şeklinde seçilirken y değerleri de $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$ denkleminden elde edildi. Bu şekilde üretilen regresyon verisi klasik varsayımları karşılamaktadır.

x yönünde veri kirletmek için, kirlenme oranı kadar seçilen rastlantısal gözlemlere şu dönüşüm uygulandı:

$$x_{ij}^{yeni} = \mu_j^x + \delta\sigma_j^x + Uniform(0,2) \quad (6)$$

Eşitlik (6)'da x_{ij}, $\boldsymbol{X}$ matrisinin i. satır ve j. sütun elemanını; δ, x yönünde kaç standart sapma sapılacağını, μ_j^x kirletilecek olan gözleme ait j. değişkenin ortalamasını, σ_j^x ise standart sapmasını göstermektedir. 0 ve 2 parametreli Uniform dağılımdan çekilecek rastlantısal sayıları simgeleyen *Uniform (0, 2)* terimi ise simülasyonlarda başvurulan tekrarlı yeniden örnekleme süreçlerinde $\boldsymbol{X}$ matrisinin herhangi bir alt kümesinin determinantının sıfırdan farklı çıkması için eklenmiştir.

Benzer şekilde y yönünde (dikey) kirlenme için Eşitlik (7)'deki dönüşüm uygulandı:

$$y_i^{yeni} = \boldsymbol{x}_i\boldsymbol{\beta} + \delta\sigma_\varepsilon = \boldsymbol{x}_i\boldsymbol{\beta} + \delta \quad (7)$$

Burada y_i, $\boldsymbol{y}$ vektörünün i. elemanını; $\boldsymbol{x}_i$, $\boldsymbol{X}$ matrisinin i. satırını; δ, i. kalıntının kaç standart sapma sapacağını göstermektedir. Eşitlik (6) ve Eşitlik (7), aynı gözlemler için uygulandığında hem x hem de y yönünde kirlenmeler elde edildi.

Simülasyonlar 3 farklı şekilde uygulandı. Birincisinde, derinliğin, örneklem büyüklüğü ve parametre sayısı ile ilişkisini çıkarmak için n = 10 birimden, n = 100 birime kadar örneklemler m = 25 kez çekildi ve her 3 model için EKK tahminlerinin derinlik medyanları hesaplandı. Bu deney temiz veriye uygulandı ve elde edilen sonuçlar Tablo 1'de verildi.

İkinci deneyde Model 2 ve Model 3 için n = 100 birimlik örneklemler m = 25 kez çekildi. Her bir örneklem için EKK, LMS, LTS, LAD, Huber'in M ve DR tahmincileri ve karşılık gelen derinlikleri ve bu derinliklerin medyan ($B_{1/2}$) ve MAD ($\sum_{i=1}^{m} |x_i - B_{1/2}| / m$) değerleri hesaplandı. Bu süreç temiz veri (c = 0) ve c = {0.10, 0.20, 0.40} düzeylerinde kirletilmiş veri için tekrarlandı. Her bir kirlenme düzeyi x ve y

yönünde ve hem x, hem y yönünde ortalamadan 3 ve 5 standart sapma olacak şekilde tekrarlandı. Böylece derinliğin x yönündeki aykırı değerler (kaldıraç değerler) ile y yönündeki aykırı değerlere (dikey aykırı değerler) ve her iki yöndeki aykırı değerlere (bir kısmı potansiyel iyi huylu kaldıraç olabilecek) karşı değişimini inceleme olanakları elde edildi. Bu simülasyonlardan elde edilen sonuçlar Tablo 2 ve Tablo 3'te verilmiştir.

Üçüncü deneyde Model 2 yukarıda betimlenen kısıtlarla c = 0.01' den c = 0.49' a kadar 0,01'lik kirlenme artışlarıyla uygulandı ve her bir tahminci için hesaplanan derinliklerin grafikleri elde edildi. Bu grafikler Şekil 2, Şekil 3 ve Şekil 4'te görülmektedir. Şekillerdeki her bir grafiğin başlığında yer alan iki rakamdan ilki x yönündeki, ikincisi y yönündeki sapma miktarını diğer bir deyişle Eşitlik (6) ve (7)'deki δ değerini göstermektedir.

Üç deneyden elde ettiğimiz sonuçlara dayanarak aşağıdaki çıkarsamalar yapılabilir:

1) Bekleneceği ve Tablo 1'den açıkça görüleceği gibi, örneklem büyüklüğü arttırıldıkça derinlik artmaktadır. Parametre sayısı ile derinlik arasında ise ters yönlü bir ilişki olduğu yine anlaşılmaktadır.

2) Temiz veride derinliği en yüksek regresyon bekleneceği gibi DR'dir. İkinci derin regresyon ise LAD çıkmıştır. Bu sonuç kanımızca sürpriz sayılmamalıdır. Bilindiği gibi DR ve LAD tek değişkenli veriler için verilen 2 farklı medyan tanımının genelleştirilmiş halleridir. DR, medyanın gözlemler sıralandığında ortaya gelen terim olma tanımına, LAD ise $\sum_{i=1}^{n}|x_i - B_{1/2}|$ toplamını minimum yapma tanımına dayanmaktadır. Tek değişkenli verilerde en derin konum parametresi olan medyan, çok değişkenli versiyonlarında bu özelliğini korumuştur. Derinlik sıralamasında EKK ve Huber'in M tahmincileri DR ve LAD'dan sonra gelirken, LMS ve LTS daha sonra yer almışlardır. LMS ve LTS ayrıca en büyük saçılım gösteren (MAD'ı en büyük) tahminciler olarak görülmektedir.

3) Tüm tahminciler kirlenmeden, yönüne ve derecesine bağlı olarak farklı şekillerde etkilenmişlerdir. DR ve LAD, y yönünde kirlenmelerden hemen hiç etkilenmezlerken, x yönünde bir yerel minimum yapacak şekilde etkilenmektedirler. LMS, LTS ve Huber 'in M regresyonu kirlenmeye karşı birbirlerine benzer davranışlar sergilemişlerdir. Her üçü de gerek x, gerekse y yönünde kirlenme altında derinliklerini monoton azalan bir karakterde kaybetmektedir, EKK ise gerek x, gerekse y yönünde derinliğini çok küçük yüzdelerde hızla kaybetmekte; yerel bir minimumdan sonra kirlenme yüzdesi arttırıldıkça derinliği de artmaktadır. Bu özellik EKK'deki kadar belirgin olmasa da x yönündeki kirlenmelerde LAD'da ve bir miktar da DR'de görülmektedir. Bu sonuçlar, derinliği her tahminci için bir uyum iyiliği ölçüsü olarak kullanmanın sakıncalı olabileceğine işaret etmektedir.

Tablo 1. Model 1, 2 ve 3 için Örneklem Büyüklüğü ve Derinliklerin Medyanları

n	p = 2	p = 3	p = 4
10	3	2	2
15	5	3	2
20	7	5	4
25	9	7	6
30	11	9	8
35	13	11	10
40	15	13	12
45	17	15	13
50	19	18	15
55	22	19	17
60	24	21	20
65	26	24	21
70	28	25	23
75	30	28	25
80	33	29	27
85	36	32	31
90	38	35	32
95	40	36	34
100	42	38	37

Tablo 2. Model 2 için Tahmincilerin Derinliklerinin Medyan ve MAD Değerleri

Kısaltmalar: Kirlenme Yüzdesi - X Yönünde Sapma - Y Yönünde Sapma

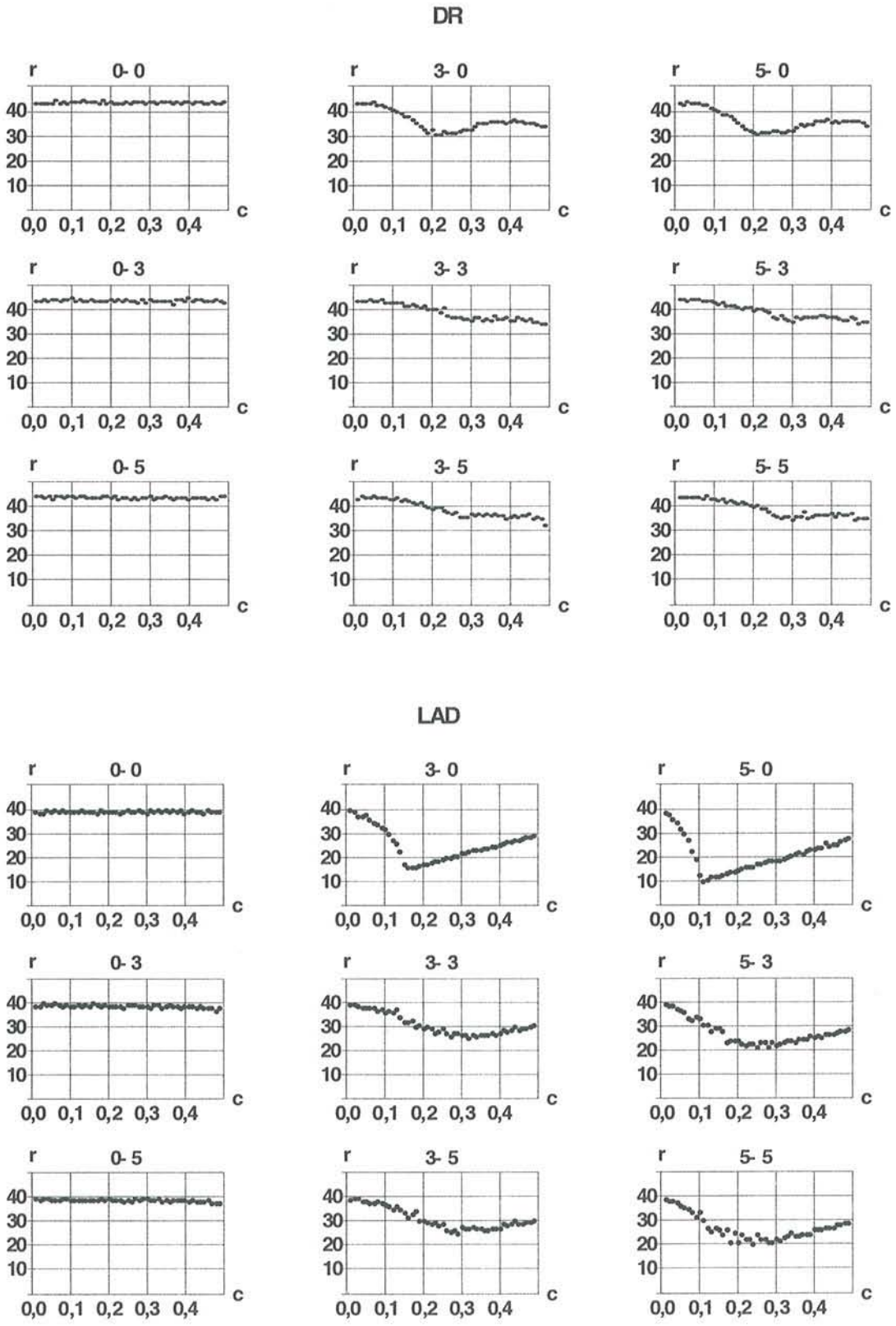
n=100	%10 - 3 - 0		%10 - 5 - 0		%10 - 0 - 3		%10 - 0 - 5		%10 - 3 - 3		%10 - 3 - 5		%10 - 5 - 3		%10 - 5 - 5	
m=25	Med.	MAD	Med.	MAD	Med.	MAD	Med.	MAD	Med.	MAD	Med.	MAD	Med.	MAD	Med.	MAD
OLS	9	0,64	9	0,76	10	0,96	10	0,96	7	1,48	7	1,12	8	1,72	7	1,92
DR	40	1,48	40	0,96	43	1,60	43	1,64	43	1,44	42	1,84	42	1,72	43	1,44
LMS	31	2,52	32	2,68	34	2,88	36	2,92	37	2,52	34	2,84	34	3,08	35	3,08
LTS	32	2,00	31	2,92	34	2,96	35	2,64	34	3,20	35	3,00	34	2,56	34	2,60
LAD	31	2,64	10	2,84	39	1,64	39	1,40	35	2,44	37	2,52	30	4,36	31	5,16
HuberM	31	3,52	32	2,72	37	2,88	36	3,12	34	2,44	34	2,68	34	3,48	35	2,16
n=100	%20 -3 - 0		%20 - 5 - 0		%20 - 0 - 3		%20 - 0 - 5		%20 - 3 - 3		%20 - 3 - 5		%20 - 5 - 3		%20 - 5 - 5	
m=25	Med.	MAD	Med.	MAD	Med.	MAD	Med.	MAD	Med.	MAD	Med.	MAD	Med.	MAD	Med.	MAD
OLS	16	1,08	14	1,04	17	1,48	18	1,36	12	0,92	12	1,20	12	2,32	11	2,48
DR	31	1,12	31	1,60	44	1,12	43	1,52	41	2,48	40	2,40	39	2,56	38	2,84
LMS	28	3,48	29	2,48	35	3,00	31	2,72	34	2,80	33	2,88	33	3,72	32	2,36
LTS	28	2,88	29	2,08	34	2,84	35	3,00	34	2,92	35	2,36	34	2,32	34	2,52
LAD	17	0,84	15	1,00	39	1,72	39	1,76	33	4,60	30	5,24	20	5,24	21	5,60
HuberM	29	2,24	26	1,92	32	3,28	35	2,68	33	2,56	33	2,96	35	3,80	33	3,96
n=100	%40 - 3 - 0		%40 - 5 - 0		%40 - 0 - 3		%40 - 0 - 5		%40 - 3 - 3		%40 - 3 - 5		%40 - 5 - 3		%40 - 5 - 5	
m=25	Med.	MAD	Med.	MAD	Med.	MAD	Med.	MAD	Med.	MAD	Med.	MAD	Med.	MAD	Med.	MAD
OLS	23	1,72	23	1,48	36	1,28	33	1,.88	22	1,48	21	1,80	21	1,64	22	1,68
DR	37	2,00	35	2,28	44	1,20	43	1,72	36	2,80	37	1,88	36	2,04	36	2,52
LMS	20	1,96	20	1,80	28	3,48	27	3,32	28	2,24	31	3,32	30	2,92	30	2,40
LTS	22	1,52	22	1,40	27	2,16	28	2,60	30	1,92	31	2,36	32	2,08	31	2,64
LAD	26	1,64	22	1,48	39	1,64	39	1,16	27	1,12	27	1,84	25	1,28	25	0,96
HuberM	21	1,80	21	1,68	28	3,32	27	3,12	31	2,88	31	2,64	30	2,88	30	2,60
n=100	%0 - 0 - 0															
m=25	Med.	MAD														
OLS	39	1,12														
DR	43	1,60														
LMS	35	2,56														
LTS	35	1,92														
LAD	40	1,40														
HuberM	37	2,12														

Tablo 3. Model 3 için Tahmincilerin Derinliklerinin Medyan ve MAD Değerleri

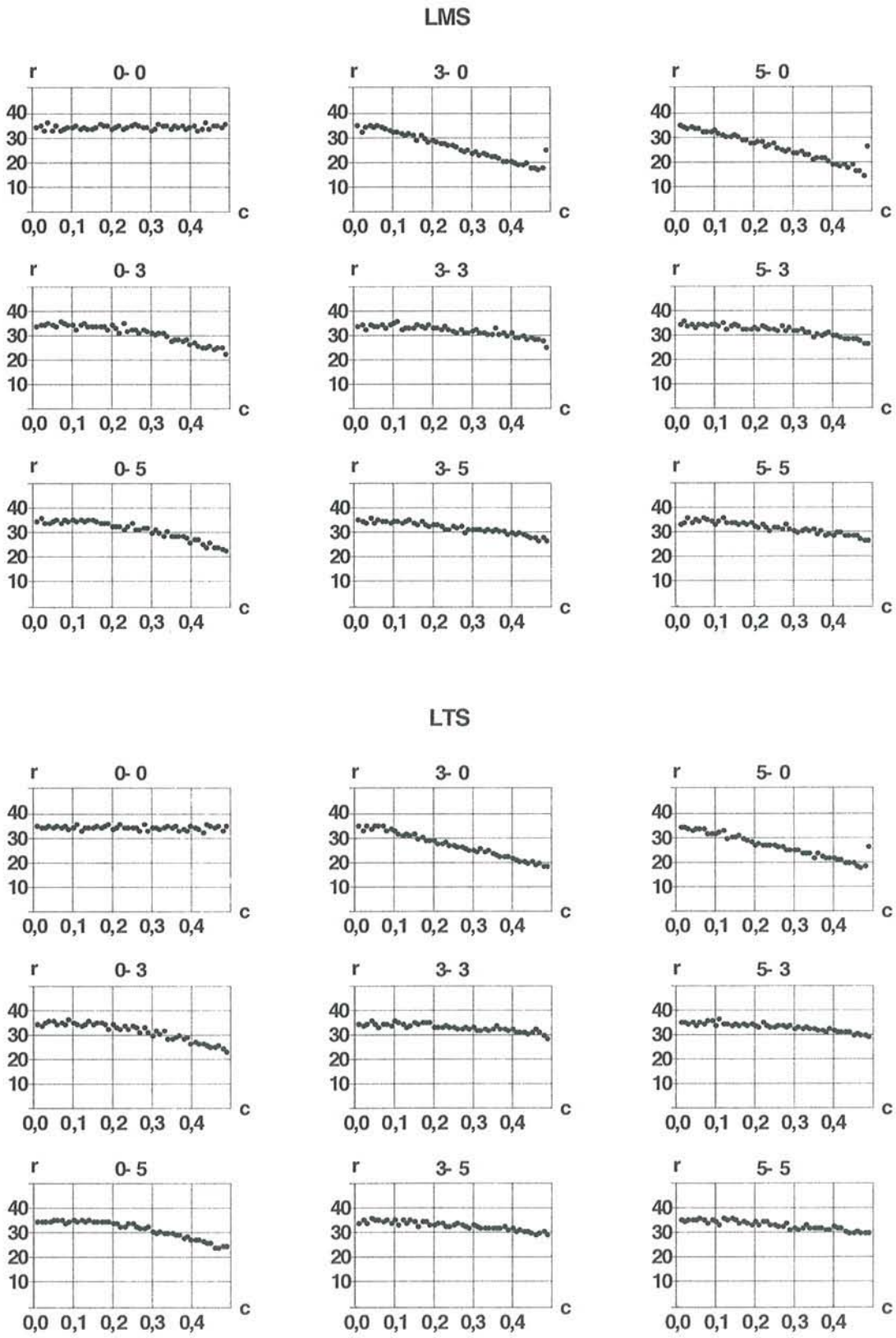
Kısaltmalar: Kirlenme Yüzdesi - X Yönünde Sapma - Y Yönünde Sapma

n=100	%10 - 3 - 0		%10 - 5 - 0		%10 - 0 - 3		%10 - 0 - 5		%10 - 3 - 3		%10 - 3 - 5		%10 - 5 - 3		%10 - 5 - 5	
m=25	Med.	MAD	Med.	MAD	Med.	MAD	Med.	MAD	Med.	MAD	Med.	MAD	Med.	MAD	Med.	MAD
OLS	9	0,56	8	0,92	9	1,08	9	0.92	6	0,96	6	1,32	6	1,40	6	1,80
DR	40	1,12	39	1,32	43	1,36	43	1,28	42	1,84	42	1,72	42	2,08	42	1,28
LMS	29	2,44	32	2,36	32	3,48	32	2,60	31	2,56	29	2,88	31	1,92	30	2,28
LTS	29	2,04	27	2,76	32	2,32	31	2,96	32	2,40	30	2,76	31	2,12	31	2,56
LAD	24	2,84	9	1,04	35	1,60	35	1,68	32	2,12	32	2,76	26	7,56	22	6,80
HuberM	28	2,60	29	2,64	31	3,28	31	2,40	29	3,68	32	3,04	32	2,80	30	3,08
n=100	%20 - 3 - 0		%20 - 5 - 0		%20 - 0 - 3		%20 - 0 - 5		%20 - 3 - 3		%20 - 3 - 5		%20 - 5 - 3		%20 - 5 - 5	
m=25	Med.	MAD	Med.	MAD	Med.	MAD	Med.	MAD	Med.	MAD	Med.	MAD	Med.	MAD	Med.	MAD
OLS	15	1,28	14	1,20	17	0,88	18	0,80	11	1,12	11	1,00	10	1,32	11	1,40
DR	34	1,64	34	1,52	43	1,44	43	0,96	39	2,64	38	2,92	38	2,52	36	2,48
LMS	25	2,76	26	1,64	31	2,84	31	3,44	31	2,64	28	2,12	30	3,40	28	2,76
LTS	25	2,40	27	1,56	32	2,12	31	2,16	31	2,32	30	1,88	31	2,60	30	3,00
LAD	16	1,28	13	1,56	35	1,36	35	1,40	26	4,76	20	4,60	17	3,28	16	2,32
HuberM	27	2,72	26	2,36	32	3,00	30	3,04	30	2,00	32	2,96	30	3,48	30	2,88
n=100	%40 - 3 - 0		%40 - 5 - 0		%40 - 0 - 3		%40 - 0 - 5		%40 - 3 - 3		%40 - 3 - 5		%40 - 5 - 3		%40 - 5 - 5	
m=25	Med.	MAD	Med.	MAD	Med.	MAD	Med.	MAD	Med.	MAD	Med.	MAD	Med.	MAD	Med.	MAD
OLS	24	1,64	23	1,76	33	1,56	32	1,60	21	0,84	21	1,40	20	1,12	22	1,48
DR	37	2,00	37	2,20	43	2,00	43	1,48	38	2,04	37	1,88	38	2,04	38	2,08
LMS	18	2,08	18	1,56	24	2,72	25	2,96	27	2,76	28	3,36	25	3,08	26	1,92
LTS	19	1,68	18	2,00	25	2,20	26	2,28	27	2,64	28	2,76	27	2,40	26	2,04
LAD	24	2,32	23	2,56	35	1,52	34	1,80	25	1,52	25	1,60	23	1,32	24	1,48
HuberM	17	2,28	18	1,92	27	2,88	26	4,12	27	2,04	26	2,56	27	2,52	27	2,84

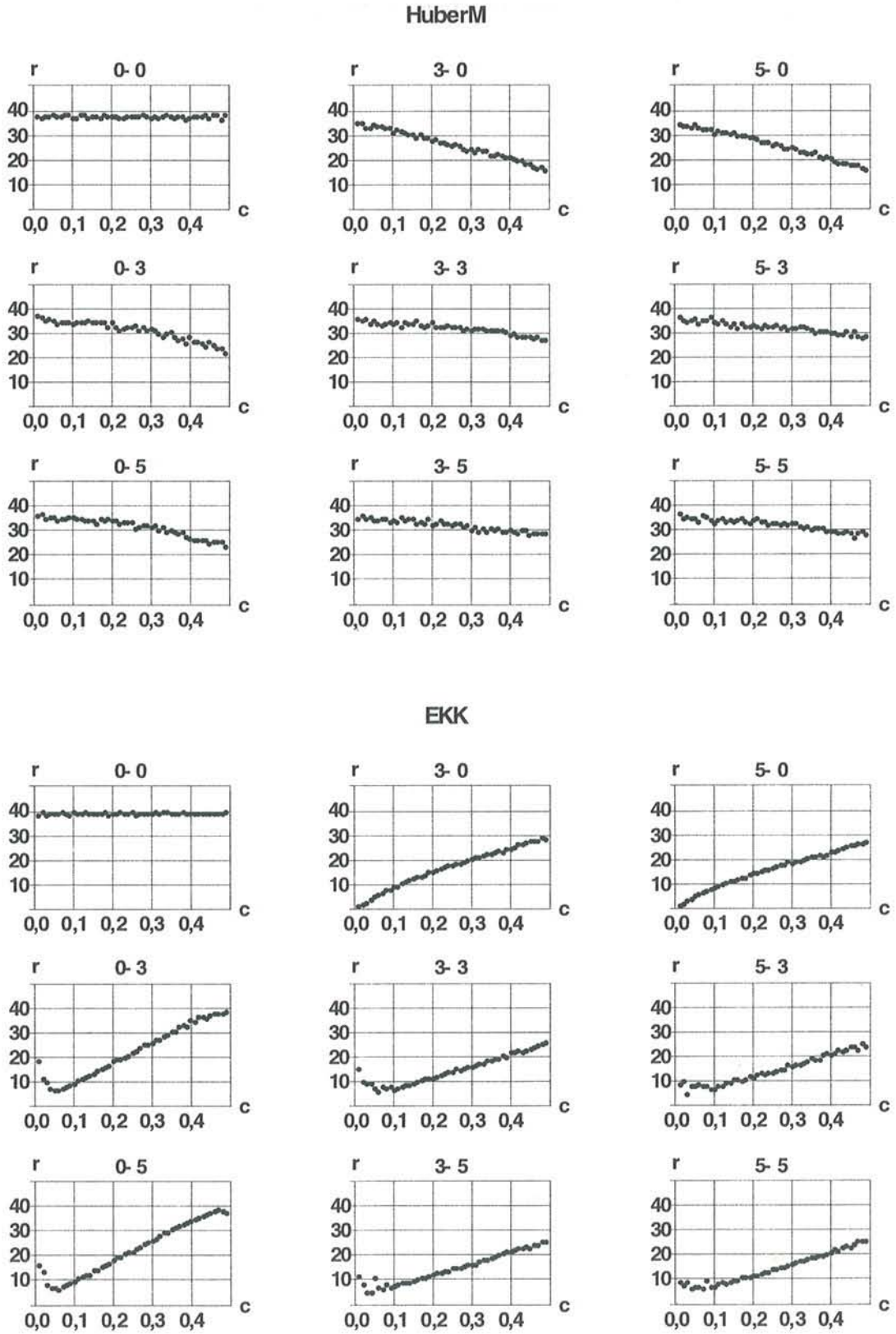
n=100	%0 - 0 - 0	
m=25	Med.	MAD
OLS	37	1,28
DR	44	1,48
LMS	32	3,36
LTS	32	2,96
LAD	35	1,44
HuberM	34	1,84



Şekil 2. DR ve LAD Tahmincilerinin Çeşitli Kirlenme Yüzdelerinde (c) Derinlikleri (r)



Şekil 3. LMS ve LTS Tahmincilerinin Çeşitli Kirlenme Yüzdelerinde (c) Derinlikleri (r)



Şekil 4. Huber'in M ve EKK Tahmincilerinin Çeşitli Kirlenme Yüzdelerinde (c) Derinlikleri (r)

4. DERİNLİK DAĞILIMLARI İLE HİPOTEZ TESTLERİ

Van Aelst vd., (2002) DR'nin parametrelerinin anlamlılık testlerini gerçekleştirmede maksimum derinliğin dağılımının kullanılabileceğini önermişlerdir. 2 boyut için (tek bağımsız değişkenli model) böyle bir dağılım Daniels (1954)'te elde edilmiştir. 2'den büyük boyut için ise dağılımların simülasyonlarla elde edilebileceği yine Aelst'in çalışmasında önerilmiştir.

Bir önceki bölümde tahmincilerin derinliklerinin kirlenmeden etkilendikleri, açık örüntüler sergiledikleri görüldü. Dolayısıyla Van Aelst vd. (2002)'de DR'ye uygulanan parametrelerin anlamlılık testlerine EKK uygulanabileceği düşüncesi akla gelebilir. Düşüncenin araştırılması amacıyla Eşitlik (4)'teki 2 bağımsız değişkenli model için n = 100 birimlik örneklemler çekilip, her bir örneklemden de B=1000 bootstrap örneklemi elde ederek, H_0: $\beta_i = 0$ hipotezi altındaki derinliklerin dağılımları bulundu. Dağılımların kantillerinden ilgilenilen hipotezlerin p-değerleri hesaplandı ve bunlar klasik yolla yani t dağılımından hesaplanan p-değerleriyle karşılaştırıldı. m = 25 kez uygulanan bu sürecin sonunda % 90' ın üzerinde tutarlılıklar elde edildi. Diğer bir deyişle klasik yöntem ve derinlik dağılımları ile uygulanan süreçle hipotez testi sürecinde aynı sonuca varma olasılığı 0,90'nın üzerindedir. Sonuçlar Tablo 4'te verilmiştir.

Uygulanan süreç, EKK'nın dağılımsal varsayımlarının ihlal edildiği ve kaldıraç etkisi yapabilecek aşırı değerlerin olduğu verilerde dayanıklı bir yöntem olarak önerilebilir.

Tablo 4. Klasik Süreçten ve Derinliklerin Bootstrap Dağılımlarından Bulunan p-Değerleri

$H_0:\beta_0=0$, Klasik	$H_0:\beta_0=0$, Derinlik	$H_0:\beta_1=0$, Klasik	$H_0:\beta_1=0$, Derinlik	$H_0:\beta_2=0$, Klasik	$H_0:\beta_2=0$, Derinlik
0.35607	0.25	0.41051	1.	0.0014605	0.
0.90508	0.99	0.28324	0.	1.1564×10^{-10}	0.
0.016943	0.	0.064383	0.	0.15362	0.96
0.47282	0.6	6.7121×10^{-6}	0.	0.012055	0.
0.10975	0.	0.094342	0.06	0.72028	0.92
0.0013437	0.	0.026793	0.	0.43660	0.08
0.0060544	0.	0.050335	0.	0.061832	0.01
0.46853	0.91	9.8772×10^{-10}	0.	0.000097474	0.
0.46265	0.95	0.23471	0.8	0.084624	0.57
0.0032205	0.	0.16828	0.1	0.86380	1.
0.0065012	0.	0.0090960	0.	0.041791	0.25
0.11605	0.05	0.79132	0.56	0.0013340	0.
0.18035	0.09	3.2195×10^{-9}	0.	0.0067276	0.
0.19853	0.14	0.00071839	0.	0.54661	0.42
0.22275	0.17	0.0066673	0.	0.52604	0.43
0.91249	0.88	0.00051203	0.05	0.18752	0.42
0.037204	0.	0.00031639	0.	0.12748	0.05
0.26202	0.5	0.17236	0.37	0.24331	0.68
0.75644	0.76	0.043975	0.04	0.12043	0.33
0.66189	0.27	0.91368	0.93	0.0014065	0.
0.77260	0.99	0.040525	0.01	0.043852	0.03
0.51636	0.4	0.027517	0.01	0.0018272	0.01
0.24201	0.01	0.0033037	0.	0.16190	0.37
0.0013644	0.	0.010118	0.	0.0078881	0.
0.29193	0.15	0.0056876	0.	0.99860	0.92

5. SONUÇ

Bu çalışmada çeşitli tahmincilerin derinliklerinin kirlenme karşındaki değişimleri incelendi ve derinliğin, tahmincilerin uyum iyiliklerinin (goodness of fit) bir ölçüsü olabilme olanakları araştırıldı. Çalışmadaki simülasyonlar, kirlenme karşısında, ilgilenilen 6 tahmincinin derinliklerinin belirgin örüntüler sergilediğini gösterdi. LMS, LTS ve Huber'in M tahmincileri kirlenme arttırıldıkça derinliklerini kaybederken, EKK'de sezgiye ters gelebilecek ters yönlü bir ilişki görüldü. EKK'deki kadar belirgin olmasa da yalnızca x yönündeki kirlenme karşısında LAD ve hafif ölçüde DR'de benzer bir özellik görüldü. Bu sonuçlar, her tahminci için derinliğin bir uyum ölçüsü olamayacağına ancak derinlik ve uyumun bazı durumlarda, birlikte incelenmeye değer olduğuna işaret edebilir.

Çalışmada ayrıca bootstrap yöntemiyle elde edilen derinlik dağılımları ile hipotez testi sürecinin uygulanma olanakları araştırılmış, klasik hipotez testi süreciyle karşılaştırıldığında anlamlı sonuçlar elde edilmiştir.

Ek 1 . 2 Boyutlu Veride Regresyon Derinlik Hesabının Mathematica Programı

```
RDepth2::usage =
  "RDepth2[data,parameters], nx2 boyutlu bir matris ile verilen veri seti ve 2x1
    boyutlu bir parametre vektörü için regresyon derinliği hesaplar";
Off[General::"spell1"];
RDepth2[data_, parameters_] :=
  Module[{ind, dep, n, distx, ones, xmatrix, e, vj, j, kaldirac, splus, sminus, gplus,
    gminus, rdepths, depth2},
   (* Design Control *)
   If [ MatrixQ[data] == False, {Print["Veri seti nx2 boyutlu bir matris olmalıdır."]; Abort[];}];
   If [Dimensions[data][[2]] ≠ 2 , {Print["Veri matrisi 2 sütundan oluşmalıdır."]; Abort[];}];
   If [Length[parameters] ≠ 2 || VectorQ[parameters] == False,
    {Print["Parametreler 2x1 boyutlu bir vektör olmalıdır"]; Abort[];}];
   If [Length[data] < Length[parameters],
    {Print["Parametre sayısı, gözlem sayısından fazla olamaz."]; Abort[];}];
   (* Variable Definition *)
   ind = Transpose[data][[1]];
   dep = Transpose[data][[2]];
   n = Length[ind];
   distx = Union[ind];
   ones = Table[1, {Length[ind]}];
   xmatrix = Transpose[Table[{ones, ind}]];
   (* Residuals and Leverages Calculation *)
   e = dep - xmatrix.parameters;
   vj = Table[0, {Length[distx]}];
   vj[[1]] = distx[[1]] - 1;
   For [j = 2, j ≤ Length[distx], j++,
    vj[[j]] = (distx[[j - 1]] + distx[[j]]) / 2;
   ];
   rdepths = Table[0, {0}];
   (* All Possible Depths for All Leverages *)
   For [j = 1, j ≤ Length[vj], j++,
    kaldirac = vj[[j]];
    splus = 0; sminus = 0; gplus = 0; gminus = 0;
    For [i = 1, i ≤ n, i++,
     If [ind[[i]] < kaldirac && e[[i]] ≥ 0 , {splus++;}];
     If [ind[[i]] < kaldirac && e[[i]] < 0 , {sminus++;}];
     If [ind[[i]] > kaldirac && e[[i]] ≤ 0 , {gminus++;}];
     If [ind[[i]] > kaldirac && e[[i]] > 0 , {gplus++;}];
    ];
    rdepths = Append[rdepths, Min[splus + gminus, sminus + gplus]];
   ];
   (* Minimum of all depths is Rdepth *)
   depth2 = Min[rdepths];
   Return [depth2];
  ];
```

KAYNAKLAR

DANIELS, H.E. (1954), *A Distribution-Free Test for Regression Parameters*, Annals of Mathematical Statistics, 25,499-513.

KARABULUT, İ., ÖZTÜRK, F. (2003), *Lineer Modellerde Yarı Uzay Derinliğine Dayalı Dengeli Bootstrap Güven Bölgeleri*, İstatistik Araştırma Dergisi, Cilt 2, No:3, 63-72.

LIU, R., PARELIUS, J. , SINGH, K. (1990), *Multivarite Analysis By Data Depth : Descriptive Statistics, Graphics and Inference*, the Annals of Statistics, Vol.27, No.3, 783-858.

LIU, R., SINGH, K. (1993), *A Quality Index Based on Data Depth and Multivariate Rank Tests*, Journal of the American Statistical Association, Vol.88, 257-260.

LIU, R., SINGH, K. (1997), *Notions of Limiting P-Values on Data Depth and Bootstrap*, Journal of the American Statistical Association, Vol.91, 266-277.

ROUSSEEUW, P. J., HUBERT, M. (1999), *Regression Depth*, Journal of the American Statistical Association, Vol.94, 388-402.

ROUSSEEUW, P.J., RUTS, I., TUKEY, J.W. (1999), *The Bagplot: A Bivarite Boxplot*, the American Statistican, Vol.53, No.4, 382-387.

TUKEY, J.W. (1975), *Mathematics and Picturing Data*, Proceedings of the 1974 International Congress of Mathematics, 523-531.

VAN AELST, S., ROUSSEEUW, P.J., HUBERT, M., STRUYF, A. (2002), *The Deepest Regression Method*, Journal of Multivariate Analysis, Vol.81, 138-166.

A STUDY FOR EXAMINING THE BEHAVIORS OF REGRESSION DEPTHS OF OLS AND SOME ROBUST REGRESSION METHODS UNDER CONTAMINATION

ABSTRACT

In this paper, we examine the behaviors of the regression depths of OLS and some robust regression methods under contamination. Upon Monte Carlo simulations, we determine some patterns for some estimators. Using bootstrap method, we also show that the distributions of regression depth can be utilized in hypothesis testing of regression parameters of OLS.

Key Words: *Bootstrap, Outliers, P-Value, Regression Depth, Robust Methods, The Deepest Regression.*