

T.C.
SELÇUK ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

DOĞRUSAL OLMAYAN REGRESYONDA BOZULMA NOKTALARININ
HESABI VE BİR UYGULAMA

Ahmet PEKGÖR

DOKTORA TEZİ

MATEMATİK ANABİLİM DALI

KONYA, 2010

ÖZET

DOKTORA TEZİ

DOĞRUSAL OLMAYAN REGRESYONDA BOZULMA NOKTALARININ HESABI VE BİR UYGULAMA

Ahmet PEKGÖR

Selçuk Üniversitesi Fen Bilimleri Enstitüsü
Matematik Anabilim Dalı

Danışman: Doç.Dr. Aşır GENÇ
Yıl: 2010, 112 Sayfa

Jüri: Doç.Dr. Aşır GENÇ
Prof.Dr. Ayşen APAYDIN
Yrd.Doç.Dr. Hasan KÖSE
Doç.Dr. M.Fedai KAYA
Doç.Dr. Coşkun KUŞ

Bozulma noktası, bir tahmin edicinin sağlamlık ölçütlerinden birisidir. Sağlam tahmin edicilerde bu değer $1/2$, sağlam olmayan tahmin edicilerde de $1/n$ dir. Maalesef günümüz istatistiksel paket programlarda kullanıcıya, ele aldığı tahmin edicinin sağlamlık ölçütü olan bu değer yansıtılmamaktadır. Bu da kullanıcının, ele aldığı modelde tahmin edicinin ne derecede sağlam olacağının hesaplanması için bu işlemi bizzat elle ve teorik olarak çözmesi gerekmektedir. Ancak, modelin zorluğu ve işlem hacminin büyüklüğünden ötürü her zaman bu mümkün olmamaktadır. Bundan dolayı kullanıcının ele almış olduğu modelde ve tahmin ediciye dayanan sonuçlarda, bozulma noktalarının kullanıcıya yansıtılması gerekmektedir. Bu noktadan yola çıkarak, bozulma noktasının istatistiksel bir paket programıyla hesaplanması düşüncesi bu tezin ana temasını oluşturmaktadır. Bu çalışmada bozulma noktası yeni bir yaklaşımla (*Tespit yöntemi*) tanımlanarak bilgisayar ortamında hesaplanması mümkün kılınmış, ayrıca bu yeni tanım sayesinde tahmin edicilerin bozulma noktaları yönünden sıralanabilmesi de mümkün olmuştur.

Anahtar Kelimeler: Doğrusal olmayan regresyon, Sağlam tahmin ediciler, Aykırı değer tespiti, Bozulma noktası, *Selçuk STAT*, Monte Carlo simülasyonu.

ABSTRACT**Ph.D. Thesis****Computation of Breakdown Points in Nonlinear Regreession
and an Application****Ahmet PEKGÖR****Selcuk University****Graduate School of Natural and Applied Sciences****Department of Mathematics****Supervisor: Assoc.Prof.Dr. Aşır GENÇ****2010, 112 Page**

Breakdown point is one of the robustness measures of an estimator. In robust estimators, the value of this measure is $1/2$ on the other hand in nonrobust estimators, the value of this measure is $1/n$. Unfortunately in nowadays statistical packet programmes, the robustness measure of handed estimator can not be provided to the user. This impossibility causes the user to compute the robustness of an estimator of his interested model by hand and theoretically. But on the other hand by this way, computing of the robustness of an estimator is always impossible because of the difficulty of the model and the big size of computing. Therefore in the user's interested model and in the results of an estimator that the user is interested, computing breakdown point is very important. Starting from this point, computing breakdown point by a statistical packet programme is the main theme of this thesis study. In this study, by defining an alternative new breakdown point (*Tespit yöntemi*), computing of breakdown point is provided to the user in computer occasion and also by courtesy of this new definition, sorting of estimators in terms of breakdown points is allowed.

Keywords: Nonlinear regression, Robust estimators, Outlier detection, Breakdown point, *Selçuk STAT*, Monte Carlo simulation.

TEŞEKKÜR

Bu çalışmanın hazırlanması sırasında göstermiş olduğu yakın ilgisinden ve her türlü yönlendirici yardımlarından dolayı tez danışmanım Sayın Doç.Dr. Aşır GENÇ' e, tez izleme komitesi üyesi Sayın Prof. Dr. Ayşen APAYDIN ve Sayın Yrd.Doç.Dr. Hasan KÖSE'ye, ilgi ve desteklerini her zaman yanımda hissettiğim Sayın Doç.Dr. M.Fedai KAYA, Sayın Doç.Dr.Coşkun KUŞ ve Sayın İsmail KINACI' ya, uygulamadaki veri setini temin eden Sayın Yrd.Doç.Dr. Ufuk KARADAVUT'a ve diğer mesai arkadaşlarıma çok teşekkür ederim.

Beni bugünlere getirdiği, hiçbir fedakârlıktan kaçınmadığı, tüm dertlerime ortak olduğu, zorlukları benimle paylaştığı, bana her konuda destek olduğu ve can yoldaşı olduğu için canım Anneme sonsuz teşekkür ediyorum, hakkını nasıl öderim? İşlerimin yoğun olduğu günlerde bana büyük bir sabır gösteren sevgili eşim Dr.Selma PEKGÖR'e en içten teşekkürlerimi sunarım.

İÇİNDEKİLER

ÖZET.....	i
ABSTRACT.....	ii
TEŞEKKÜR.....	iii
İÇİNDEKİLER.....	iv
TABLO DİZİNİ.....	vii
ŞEKİLLER DİZİNİ.....	ix
SİMGE VE KISALTMALAR.....	x
1. GİRİŞ.....	1
2. TEMEL KAVRAMLAR.....	6
2.1. Doğrusal Olmayan Regresyon.....	8
2.1.1. Parametre tahmini.....	9
2.1.1.1. En küçük kareler yöntemi.....	9
2.1.1.2. En çok olabilirlik yöntemi.....	11
2.1.2. Yinelemeli yöntemler.....	12
2.1.2.1. Gauss-Newton yöntemi.....	12
2.1.2.2. En hızlı iniş yöntemi.....	14
2.1.2.3. Levenberg-Marquardt yöntemi.....	15
2.2. En Büyük Yan.....	17
2.2.1. Jackknife yöntemi.....	18
2.2.2. Bootstrap yöntemi.....	19
2.3. Gözlem Etkilerinin İncelenmesi.....	21
2.4. Sağlamlık Ölçütleri.....	22
2.4.1. Etki fonksiyonu.....	23
2.4.2. Bozulma noktası.....	23
3. DOĞRUSAL OLMAYAN REGRESYONDA SAĞLAM TAHMİN EDİCİLER.....	25
3.1. En Küçük Mutlak Sapma Yöntemi.....	25
3.2. En Küçük Kırpılmış Kareler Yöntemi.....	28
3.3. M Tahmin Edicisi.....	29
3.3.1. Huber fonksiyonu.....	32
3.3.2. Ramsay fonksiyonu.....	33
3.3.3. Andrews'in dalga fonksiyonu.....	35

3.3.4. Hampel fonksiyonu.....	36
3.3.5. Rousseeuw ve Yohai fonksiyonu.....	38
3.4. Genelleştirilmiş M Tahmin Edicileri.....	40
3.5. S Tahmin Edicisi.....	44
4. AYKIRI GÖZLEMLERİN İNCELENMESİ.....	47
4.1. Regresyonda Tek Aykırı Gözlemin Belirlenmesi.....	47
4.1.1. Gözlem uzaklığı matrisi.....	47
4.1.2. Cook uzaklığı.....	48
4.1.3. Mahalanobis uzaklığı.....	49
4.1.4. DFBETAS.....	50
4.1.5. DFFITS.....	50
4.1.6. Standart ve Student türü artıklar.....	51
4.1.7. COVRATIO.....	51
4.2. Regresyonda Birden Fazla Aykırı Gözlemlerin Belirlenmesi	52
4.2.1. En küçük hacimli elips yöntemi.....	52
4.2.2. En küçük kovaryans determinantı yöntemi.....	53
4.2.3. Aykırı değer sayısı yöntemi.....	55
4.2.4. Sağlam uzaklık yöntemi.....	62
5. BOZULMA NOKTASI VE ETKİ FONKSİYONU.....	64
5.1. Bozulma Noktası.....	64
5.1.1. Değiştirmeli bozulma noktası.....	65
5.1.1.1. Sonlu örneklem değiştirmeli bozulma noktası.....	65
5.1.1.2. Sınırlı sonlu örneklem değiştirmeli bozulma noktası.....	66
5.1.2. Eklemeli bozulma noktası.....	67
5.1.2.1. Eklemeli sonlu örneklem bozulma noktası.....	67
5.1.2.2. Sınırlı eklemeli sonlu örneklem bozulma noktası.....	67
5.2. Etki Fonksiyonu	68
6. BOZULMA NOKTASININ HESAPLANMASI VE BİR UYGULAMA.....	71
6.1. Bozulma Noktası İçin Yeni Bir Yaklaşım.....	71
6.2. Uygulama.....	74
6.2.1. Senaryo-I.....	86
6.2.2. Senaryo-II.....	89
6.2.3. Senaryo-III.....	91

6.2.4. Senaryo-IV.....	93
6.2.5. Senaryo-V.....	95
6.2.6. Senaryo-VI.....	96
6.2.7. Senaryo-VII.....	98
6.2.8. Senaryo-VIII.....	100
7. SONUÇ VE ÖNERİLER.....	103
8. KAYNAKLAR.....	105

TABLO DİZİNİ

Tablo 6.1. Bazı tahmin edicilerin asimptotik bozulma noktaları.....	72
Tablo 6.2. Uygulamada kullanılan aykırı değer türleri.....	74
Tablo 6.3. Simülasyonlar da kullanılan 48 farklı durum.....	76
Tablo 6.4. ADS yönteminin senaryolardaki ortalama başarı oranları	77
Tablo 6.5. RD yönteminin senaryolardaki ortalama başarı oranları	78
Tablo 6.6. Mısır bitkisi verilerinden elde edilen tahmin edicilerin parametre tahminleri ile HKT değerleri.....	79
Tablo 6.7. EKK artıkları ile ADS yönteminde hesaplanan parametre değerleri.....	80
Tablo 6.8. Huber'ın M artıkları ile ADS yönteminde hesaplanan parametre değerleri.....	81
Tablo 6.9. EKK ve Huber'ın M artıklarından RD yöntemiyle elde edilen uzaklıklar.....	82
Tablo 6.10. Tahmin edicilerin gerçek mısır bitkisi verilerinden 3., 8. ve 12. gözlem değerlerinin bozdurulmasıyla parametre tahminleri ve HKT değerleri ...	83
Tablo 6.11. Bozdurulmuş verilerde EKK artıkları ile ADS yönteminde hesaplanan parametre değerleri.....	84
Tablo 6.12. Bozdurulmuş verilerde Huber'ın M artıkları ile ADS yönteminde hesaplanan parametre değerleri.....	85
Tablo 6.13. EKK ve Huber'ın M artıklarından RD yöntemiyle elde edilen uzaklıklar	86
Tablo 6.14. Senaryo-I için EKK, M ve GM tahmin edicilerin tespit yöntemiyle bozulma noktaları.....	87
Tablo 6.15. Senaryo-I için LTS ve S tahmin edicilerin tespit yöntemiyle bozulma noktaları.....	88
Tablo 6.16. Senaryo-II için EKK, M ve GM tahmin edicilerin tespit yöntemiyle bozulma noktaları.....	89
Tablo 6.17. Senaryo-II için LTS ve S tahmin edicilerin tespit yöntemiyle bozulma noktaları.....	90
Tablo 6.18. Senaryo-III için EKK, M ve GM tahmin edicilerin tespit yöntemiyle bozulma noktaları.....	91
Tablo 6.19. Senaryo-III için LTS ve S tahmin edicilerin tespit yöntemiyle bozulma noktaları.....	92

Tablo 6.20. Senaryo-IV için EKK, M ve GM tahmin edicilerin tespit yöntemiyle bozulma noktaları.....	93
Tablo 6.21. Senaryo-IV için LTS ve S tahmin edicilerin tespit yöntemiyle bozulma noktaları.....	94
Tablo 6.22. Senaryo-V için EKK, M ve GM tahmin edicilerin tespit yöntemiyle bozulma noktaları.....	95
Tablo 6.23. Senaryo-V için LTS ve S tahmin edicilerin tespit yöntemiyle bozulma noktaları.....	96
Tablo 6.24. Senaryo-VI için EKK, M ve GM tahmin edicilerin tespit yöntemiyle bozulma noktaları.....	97
Tablo 6.25. Senaryo-VI için LTS ve S tahmin edicilerin tespit yöntemiyle bozulma noktaları.....	98
Tablo 6.26. Senaryo-VII için EKK, M ve GM tahmin edicilerin tespit yöntemiyle bozulma noktaları.....	99
Tablo 6.27. Senaryo-VII için LTS ve S tahmin edicilerin tespit yöntemiyle bozulma noktaları.....	100
Tablo 6.28. Senaryo-VIII için EKK, M ve GM tahmin edicilerin tespit yöntemiyle bozulma noktaları.....	101
Tablo 6.29. Senaryo-VIII için LTS ve S tahmin edicilerin tespit yöntemiyle bozulma noktaları.....	102
Tablo 7.1. Tüm Senaryolardaki tahminin edicilerin tespit yöntemiyle bozulma noktaları.....	103

ŞEKİLLER DİZİNİ

Şekil 2.1. İyi leverage nokta, dikey aykırı değer, kötü leverage nokta gösterimi.....	21
Şekil 3.1. Huber'in $\rho(x)$ ve $\psi(x)$ fonksiyonlarının grafiği.....	33
Şekil 3.2. Huber'in $w(x)$ ağırlık fonksiyonu grafiği.....	33
Şekil 3.3. Ramsay'ın $\rho(x)$ ve $\psi(x)$ fonksiyonlarının grafiği.....	34
Şekil 3.4. Ramsay'ın $w(x)$ ağırlık fonksiyonu grafiği.....	34
Şekil 3.5. Andrews'ın $\rho(x)$ ve $\psi(x)$ fonksiyonlarının grafiği.....	35
Şekil 3.6. Andrews'ın $w(x)$ ağırlık fonksiyonu grafiği.....	36
Şekil 3.7. Hampel'in $\rho(x)$ ve $\psi(x)$ fonksiyonlarının grafiği.....	37
Şekil 3.8. Hampel'in $w(x)$ ağırlık fonksiyonu grafiği.....	37
Şekil 3.9. Hampel'in a, b ve c 'nin farklı seviyeleri için $w(x)$ ağırlık fonksiyonlarının grafikleri.....	38
Şekil 3.10. Rousseeuw ve Yohai' nin $\rho(x)$ ve $\psi(x)$ fonksiyonlarının grafiği.....	39
Şekil 3.11. Rousseeuw ve Yohai' nin c 'nin seviyelerine göre $w(x)$ ağırlık fonksiyonu grafiği.....	39
Şekil 3.12. Huber M tahmin edicisinin davranışı.....	40
Şekil 3.13 Mallows M tahmin edicisinin davranışı.....	41
Şekil 3.14 Schweppe M tahmin edicisinin davranışı.....	42
Şekil 4.1. Öklid ve Mahalanobis Uzaklığında iki noktanın merkeze olan uzaklığı.....	49
Şekil 4.2. En küçük hacimli elips ile güven elipsi görünümü.....	53
Şekil 5.1. Aritmetik ortalama ve varyansın etki fonksiyonlarının grafiği.....	69
Şekil 6.1. Yinelemeler sırasında 8 ayrı senaryodan rasgele üretilen verilerden bir tanecikinin saçılım grafikleri	75
Şekil 6.2. Mısır bitkisinin zamana bağlı boy uzunluğu.....	78
Şekil 6.3. Gerçek mısır bitkisi verilerinden 3., 8. ve 12. gözlem değerlerinin bozdurulmasıyla elde edilen saçılım grafiği.....	82
Şekil 6.4. Bozulmuş verilerdeki kestirimler.....	83

SİMGE VE KISALTMALAR

$\underline{\theta}$	: Parametre vektörü
$\hat{\theta}$	: Tahmin edici
ε^*	: Bozulma noktası
$\diamond$	: Ağırlıklandırılmış medyan
$\rho(.)$	: Kriter fonksiyonu
$\mathbf{X}$	: Gözlem matrisi
$\mathbf{B}'$	: $\mathbf{B}$ matrisinin transpozesi
$\mathbf{C}$	: Varyans-kovaryans matrisi
$\llbracket . \rrbracket$	: Tam değer fonksiyonu
ADS	: Aykırı değer sayısı
HKT	: Hata kareler toplamı
LAD	: En küçük mutlak sapma (Least Absolute Deviation)
LTS	: En küçük kırılmış kareler (Least Trimmed Square)
MAD	: Medyan mutlak sapma (Median Absolute Deviation)
MCD	: En küçük kovaryans determinanı (Minumum Covariance Determinant)
MD_i	: i. gözlemin Mahalanobis uzaklığı (Mahalanobis Distance)
MLE	: En çok olabilirlik tahmin edicisi (Maximum Likelihood Estimator)
MVE	: En küçük hacimli elips (Minumum Volume Ellipsoid)
SMAD	: Örneklem ortalama mutlak sapma (Sample Mean Absolute Deviation)
RD_i	: i. gözlemin sağlam uzaklığı (Robust Distance)

1. GİRİŞ

Legendre 1805 yılında gezegenlerin yörüngelerini belirlemek amacıyla En Küçük Kareler (EKK) olarak bilinen tekniği geliştirmiştir. Basit regresyon doğrusunun belirlenmesi problemine uygulanmasını ve Regresyon terimini de ilk kez 1886’ da Francis Galton kullanmıştır. Yule (1899), İngiltere’deki yoksulluğun nedenlerinin araştırılması adlı çalışmasında bir regresyon modelinde parametre tahmini yapmıştır. Regresyon analizinde yapılan çalışmalar Neyman (1923) tarafından devam ettirilmiştir.

Doğrusal olmayan modeller üzerine çalışmalar yoğun bir biçimde 1960 lı yıllarda başlamıştır. Marquardt (1963), Hartley ve Booker (1963) doğrusal olmayan modellerde parametre tahmini için EKK yöntemi ile ilgili algoritmalar hazırlamıştır. Gallant (1975), De Bruin (1971), Seber ve Wild (1989) alışılmış varsayımlar altındaki doğrusal olmayan modellerde sonuç çıkarım ve hipotez testi problemleri üzerine çalışmışlardır. Beale (1960) doğrusal olmayan modellerde güven bölgelerini çalışmış ve doğrusal olmayan modellerin parametre tahminlerinin yanlış olduğunu göstermiştir. Doğrusal olmayan modellerde tahmin edicilerin açık bir ifadesi elde edilemediğinden tahmin ve hipotez testi problemlerinde asimptotik ifadelerden ve doğrusal yaklaşımlardan faydalanılmaktadır. Bu konuda Jennrich (1969) parametre tahminlerinin asimptotik sonuçlarını Cook, Tsai ve Wei (1986), Box (1971) doğrusal olmayan modellerdeki yanları, Gallant ve Fuller (1973) sonuç çıkarım için asimptotik sonuçları, Malinvaud (1970) doğrusal olmayan modellerde tahminlerin tutarlılığı ile ilgili asimptotik ifadeleri vermişlerdir.

1960’ lı yıllarda bir diğer regresyon alanındaki gelişmede Sağlam (Robust) tahmin ediciler üzerine olmuştur. İlk kez 1953 yılında Box bu sözcüğe istatistiksel bir anlam kazandırmış ve varsayımlardan sapmalara karşı sağlam sonuçlar veren istatistiksel işlemleri bu kelime ile tanımlamıştır. İlk olarak, alanının öncülerinden olan Tukey tarafından tartışılmıştır. Daha sonra Huber 1964 ve 1973 yıllarında ilk kez sağlam konum tahmin edicileri ile ilgili çalışmalarını yayımlamış ve M-tahmin edicilerini geliştirmiştir. Huber bu alanda temel iki teori ileri sürmüştür: bunlar sağlam tahmin ediciler için *En küçüklerin en büyüğü yaklaşımı* (Minimax approach) ve sağlam testler ile güven aralıkları için *Kapasiteler yaklaşımı*

(Capacities approach) dır. Hampel 1971 yılında sađlamlık ile ilgili temel alıřmalar yapmıř ve sađlam tahminler iin etki fonksiyonlarına dayanan nc bir yaklařım geliřtirmiřtir. Bu fonksiyon daha sonra sađlamlıđın iyi bir lt olarak kabul edilmiřtir. Sađlam tahmin ediciler ile ilgili yine Mosteller ve Tukey (1977) ve Staudte ve Sheather (1990), dođrusal olmayan modellerde Stromberg (1993), Midi (1999) yođun olarak alıřmıřlardır.

Modeldeki hata terimlerinin dađılımı normal dađılımdan gelmediđinde ya da verilerin aykırı deđer (Outlier) iermesi durumunda sađlam regresyon tahmin edicileri, EKK yntemine alternatif olarak kullanılabilmektedir. Rousseeuw (1984) alıřmasında, bir tane aykırı deđerin bile diđer btn verilerin verdiđi bilgiye engel olduđunu ve btn istatistikleri gvenilmez yaptığını gstermiřtir. Aykırı deđer varlıđında EKK tahminlerinin bozulması byk bir tehlike yaratmaktadır. EKK ynteminin bozulma sebebi, EKK ynteminin regresyon analizinde btn verilere eřit ađırlık vermekte olması ve oluřabilecek artık deđerlerinin kareleri toplamını indirmeye alıřmasıdır.

Literatrde veri kmesinde sadece bir tane aykırı deđer bulunması durumunda bunları tespit etmek iin gvenilir ve kolay uygulanabilir teknikler mevcuttur. Bunlara rnek Weisberg (1985)' in nerdiđi ortalama deđiřim (mean-shift) testi verilebilir. Klasik aykırı deđer tespit yntemleri iin Hadi ve Simonoff 'un (1993) alıřmasında geniř bir literatr bulunabilir, ancak birden ok aykırı deđer mevcut olması durumunda bu yntemler yetersiz kalmaktadır. Birbirine yakın aykırı deđerler bazen birbirlerini maskeleyebilmekte (masking) veya bu aykırı deđerler klasik tahmin yntemlerinde gvenilir verilerin bile aykırı deđer grnmesine sebep olabilmektedirler (srkleme-swamping). Sađlam regresyon yntemlerini kullanmanın bir bařka nemli nedeni diđer yntemlerdeki gibi hataların normal dađılımdan gelmesi veya verinin belli bir dađılıma sahip olması varsayımını sađlamak zorunluluđunun olmamasıdır (Hampel, 1973).

EKK ynteminin varsayımlarının bozulması durumunda, sađlam regresyon iyi bir alternatif oluřturmakta, fakat bu konuda Reinmann (2005) alıřmasında yer alan tartıřmada olduđu gibi aykırı deđer tespiti iin evrensel bir yntem daha geliřtirilememiřtir. Melouna ve Militk (2001) kendilerinin semiř oldukları sađlam regresyonda yer alan 100'e yakın alıřmaya atıf yapmıřlardır, ancak bu sađlam

regresyon kullanılan yöntemlerin tamamını yansıtmamaktadır. Bu konuda net olarak ne kadar yöntemin olduğu ve bu yöntemlerin ne derecede başarılı olduğu tam olarak tespit edilememiştir (Reinmann 2005).

Sağlam istatistikte önemli diğer bir kavram da veri setindeki sonsuz karmaşaya sebebiyet verecek aykırı değerlere karşı istatistiğin yeteneğini ölçen Bozulma Noktasıdır (Breakdown Point - BN). Literatürde Bozulma noktalarının değiştirmeli (replacement) ve eklemeli (addition) kavramları vardır (Hampel 1968, 1971 ve Donoho ve Huber 1983). Bozulma noktası tahmin edicinin global sağlamlığını çalışmak için kullanılmış ve daha sonraları geliştirilerek test istatistiği ile regresyon gibi diğer istatistik yöntemlerine de uygulanmıştır (He ve ark. 1990, Coakley ve Hettmansperger 1992, Stromberg ve Ruppert 1992, Sakata ve White 1995, Zhang 1996 ve Cho ve ark. 2009).

Bozulma noktası, doğrusal olmayan regresyonda teorik olarak, sınırlı bir model fonksiyonunda hesaplanılabilmekte (Stromberg ve Ruppert 1992) ancak diğer model fonksiyonlarında bozulma noktasının hesaplanması her zaman mümkün olmamaktadır. Tahmin edicilerin büyük bir çoğunluğunda bozulma noktaları $1/n$ veya $1/2$ dir. Sağlam olmayan tahmin edicilerde bozulma noktası $1/n$, sağlam olan tahmin edicilerde ise bozulma noktası yaklaşık $1/2$ dir. Bu durumda iki sağlam tahmin ediciyi kendi aralarında bozulma noktası bakımından sağlamlık sıralaması yapma imkanı yoktur. Ayrıca literatürde benzer alanda çalışılan makalelerin hemen hemen hepsi doğrusal regresyonla ilgilidir (Cho ve ark. 2009, Zuo 2001, Sakata ve White 1998, Yohai 1987).

Regresyondaki bu ilgili literatürler de yer alan çalışmaların ışığı altında tez çalışmasında, istatistik paket programlarda hesaplanması mümkün olan ve tahmin edicileri de sağlamlık açısından sıralanmasına imkan veren yeni bir bozulma noktası (*Tespit Yöntemi*) tanımlanacaktır. Yeni tanımlanan bozulma noktasını özel bir algoritma vasıtasıyla, istatistiksel bir paket programına (*Selçuk STAT*) uyumu için zemin oluşturulacaktır.

Yapılacak olan bu tez çalışması aşağıda verilen bölümlerden oluşacaktır.

Çalışmanın İkinci Bölümün'de tez çalışmasının altyapısı için gerekli görülen temel tanım ve kavramlar verilecektir.

Üçüncü Bölüm’de doğrusal olmayan regresyonda parametre tahminleri konusu ele alınacaktır. Bu bölümde sırasıyla En küçük mutlak sapma, En küçük kırılmış kareler, M, Genelleştirilmiş M ve S tahmin edicilerinin özellikleri ele alınacaktır. Doğrusal regresyondaki algoritmalar uyarlanarak ve geliştirilerek, bu tahmin edicilerin doğrusal olmayan regresyonda nasıl hesaplanıldığı algoritmalar ile gösterilecektir.

Dördüncü Bölüm’de aykırı değerlerin tespitine yönelik çalışmalar irdelenecektir. Bunlar tek aykırı değer varlığında kullanılan yöntemler ve çoklu aykırı değerlerin varlığında kullanılan yöntemler olmak üzere iki durum için incelenecektir.

Çalışmanın Beşinci Bölümün’de bozulma noktası ve etki fonksiyonu kavramları ele alınarak mevcut olan tanımlar anlatılacaktır.

Bu çalışmanın Altıncı Bölümünde tez konusu olan bozulma noktasının hesaplanması konusu ele alınacaktır. Bu bölümde bozulma noktasına alternatif olarak yeni bir yöntem (*Tespit Yöntemi*) önerilerek bu yeni yöntem ile bilgisayar ortamında bozulma noktalarının nasıl hesaplanabileceği algoritma halinde verilecektir. Richard’ın sigmoidal büyüme modelinde bozulma noktası için önerilen yöntem (*Tespit yöntemi*)’in Monte Carlo simülasyon çalışmalarına yer verilecektir. Bu simülasyon çalışmalarının kurgulanmasında, Serbert ve ark. (1998)’nin senaryolarından faydalanılacaktır. Yapılacak olan Monte Carlo simülasyon çalışmaları, Delphi görsel programlama dilinde kodlanan (Pekgör 2004) ve Selçuk Üniversitesi istatistik bölümü tarafından geliştirilmekte olan *Selçuk STAT* istatistik paket programından faydalanarak yapılacaktır. Önerilen yöntem (*Tespit yöntemi*)’in hesaplanması sırasında kullanılacak olan aykırı değer tespit yöntemlerinden hangisinin kullanılması gerektiğinin anlaşılması için Aykırı Değer Sayısı ve Sağlam Uzaklık yöntemleri karşılaştırılacaktır. Sonrasında Konya bölgesindeki 2008 dönemine ait bir mısır bitkisinin haftalara göre boy uzunluğu analiz edilerek bu iki yöntem ile orijinal veri üzerinden aykırı gözlem tespiti yapılacaktır. Son olarak bu model için Monte Carlo simülasyon çalışmaları yapılarak, tahmin edicilerin yeni yöntem (*Tespit yöntemi*) ile bozulma noktaları hesaplanacaktır.

Çalışmanın son bölümü olan Yedinci Bölümde, tez çalışmasından elde edilen sonuçlara ve önerilere yer verilecektir.

2. TEMEL KAVRAMLAR

Bu bölümde öncelikli olarak genel tanımlar verilecek ve ileriki bölümlerdeki konuların desteklenmesi için doğrusal olmayan regresyonda parametre tahminlerinde kullanılan yöntemler, gözlem etkileri ve sağlamlık ölçütleri tanımlanacaktır.

Tanım 2.1.

Reel sayılar üzerinde P olasılık ölçüsü yardımı ile

$$F_X : \mathbb{R} \rightarrow [0,1] \\ :x \rightarrow F_X(x) = P\{w : X(w) \leq x\}$$

şeklinde tanımlanan fonksiyona, X rasgele değişkeninin dağılım fonksiyonu (birikimli dağılım fonksiyonu) denir.

Tanım 2.2.

X rasgele değişkeninin dağılım fonksiyonu $F_\theta(x) \in F = \{F_\theta(x) : F_\theta(x) = F(x - \theta); \theta \in \Theta, F \text{ biçimsel olarak biliniyor}\}$ olmak üzere θ parametresine F sınıfı için bir ***konum parametresi*** denir. θ parametresinin X rasgele değişkeninin konum parametresi olması için gerek ve yeter şart $X - \theta$ rasgele değişkeninin dağılımının θ dan bağımsız olmasıdır (Öztürk ve ark. 2006).

Tanım 2.3.

X rasgele değişkeninin dağılım fonksiyonu $F_\theta(x) \in F = \{F_\theta(x) : F_\theta\left(\frac{x}{\theta}\right); \theta \in \Theta, F \text{ biçimsel olarak biliniyor}\}$ olmak üzere θ parametresine F sınıfı için bir ***ölçek parametresi*** denir. θ parametresinin X rasgele değişkeninin ölçek parametresi olması için gerek ve yeter şart $\frac{X}{\theta}$ rasgele değişkeninin dağılımının θ dan bağımsız olmasıdır (Öztürk ve ark. 2006).

Tanım 2.4.

Boş olmayan bir X kümesi, reel ya da kompleks sayıların K cismi üzerinde, bir vektör uzayı olsun. Buna göre,

- i. $\forall x \in X$ için, $p(x) \geq 0$
- ii. $\forall x \in X$ için, $p(x) = 0 \Leftrightarrow x = 0$
- iii. $\forall a \in K$ ve $\forall x \in X$ için, $p(ax) = |a|p(x)$
- iv. $\forall x, y \in X$ için, $p(x + y) \leq p(x) + p(y)$

özelliklerini sağlayan $p : X \rightarrow \mathbb{R}$ fonksiyonuna, X vektör uzayı üzerinde bir **norm** ve (X, p) ikilisine de **normlu uzay** (normlu lineer vektör uzay) denir. Eğer yalnızca i, iii, iv aksiyomları sağlanıyorsa, $p : X \rightarrow \mathbb{R}$ fonksiyonuna, X vektör uzayı üzerinde bir **yarı norm** denir (Balcı 2000).

Tanım 2.5.

Boş olmayan bir X kümesi verilsin. $d : X \times X \rightarrow \mathbb{R}$ fonksiyonu,

- i. $\forall x, y \in X (x \neq y)$ için, $d(x, y) \geq 0$
- ii. $\forall x, y \in X$ için, $d(x, y) = 0 \Leftrightarrow x = y$
- iii. $\forall x, y \in X$ için, $d(x, y) = d(y, x)$
- iv. $\forall x, y, z \in X$ için, $d(x, y) \leq d(x, z) + d(z, y)$

özelliklerini sağlıyor ve bu fonksiyon sonlu değerler alıyorsa, d fonksiyonuna X kümesi üzerinde bir metrik ve (X, d) ikilisine de **metrik uzay** denir. Eğer yalnızca i, iii, iv metrik aksiyomları sağlanıyorsa $d : X \times X \rightarrow \mathbb{R}$ fonksiyonuna, X kümesi üzerinde bir **yarı metrik** denir. Eğer i, iii, iv metrik aksiyomlarını ve ii aksiyomunun yalnızca yeter şartını sağlıyorsa d metriğine X kümesi üzerinde bir **pseudo metrik** denir (Balcı 2000).

Tanım 2.6.

$A \subset \mathbb{R}$ olsun. $\forall x \in A$ için $x \geq a$ olacak şekilde bir a reel sayısı varsa A kümesine **alttan sınırlıdır** denir, a sayısına da A 'nın **alt sınırı** adı verilir. Benzer olarak, A 'nın her x elemanı için $x \leq b$ olacak şekilde bir b reel sayısı varsa A kümesi **üstten sınırlıdır** denir, b sayısına da A 'nın **üst sınırı** adı verilir (Balcı 2000).

Tanım 2.7.

Üstten sınırlı bir A kümesinin üst sınırlarının en küçüğüne A 'nın en **küçük üst sınırı** veya **supremumu** denir ve $\sup(A)$ ile gösterilir. Alttan sınırlı bir kümesinin alt sınırlarının en büyüğüne de A 'nın **en büyük alt sınırı** veya **infimumu** denir, $\inf(A)$ ile gösterilir (Balcı 2000).

2.1. Doğrusal Olmayan Regresyon

Doğrusal olmayan regresyon modelinde değişkenler arasındaki bağıntı, parametrelerin en az birinin doğrusal olmayan fonksiyonu biçimindedir. $\mathbf{Y}$ açıklanan değişken vektörü, $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n)'$ açıklayıcı değişkenlerin matrisi, k açıklayıcı değişken sayısı ve gözlem değerleri $\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{ik})$, $i = 1, 2, \dots, n$ olmak üzere doğrusal olmayan regresyon modeli,

$$Y_i = f(\mathbf{x}_i; \boldsymbol{\theta}) + \varepsilon_i, \quad i = 1, 2, \dots, n \quad (2.1)$$

biçiminde ifade edilir. Burada $\boldsymbol{\theta} = (\theta_1, \theta_2, \dots, \theta_p)'$ ($\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^p$) bilinmeyen parametre vektörü ve f , $\boldsymbol{\theta}$ bilinmeyen parametre vektörünün bileşenlerinin en az birine göre doğrusal olmayan bir fonksiyondur. $\mathbf{f}(\mathbf{x}_i; \boldsymbol{\theta})$, $\mathbf{f}(\boldsymbol{\theta})$ olarak gösterilecek olursa (2.1) modeli vektörel formda,

$$\mathbf{Y} = \mathbf{f}(\boldsymbol{\theta}) + \boldsymbol{\varepsilon} \quad (2.2)$$

yazılabilir. Burada

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \quad \mathbf{f}(\boldsymbol{\theta}) = \begin{bmatrix} f(\mathbf{x}_1, \boldsymbol{\theta}) \\ f(\mathbf{x}_2, \boldsymbol{\theta}) \\ \vdots \\ f(\mathbf{x}_n, \boldsymbol{\theta}) \end{bmatrix} = \begin{bmatrix} f_1(\boldsymbol{\theta}) \\ f_2(\boldsymbol{\theta}) \\ \vdots \\ f_n(\boldsymbol{\theta}) \end{bmatrix}, \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}, \quad \boldsymbol{\theta} = \begin{bmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_p \end{bmatrix}$$

biçimindedir. (2.2) modeli için bazı ön varsayımlar vardır. Bunlar;

- $\mathbf{f}(\boldsymbol{\theta})$ biçimsel olarak biliniyor.

- $\mathbf{f}(\boldsymbol{\theta})$ fonksiyonu parametre vektörüne göre en az iki kere türevlenebilir.
- $E(\boldsymbol{\varepsilon}) = \mathbf{0}$
- $Cov(\boldsymbol{\varepsilon}) = \sigma^2 \mathbf{I}$
- Normallik varsayımı $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$

olarak ifade edilir (Bates ve Watts 1988, Genç 1997).

2.1.1. Parametre tahmini

Gerçek dünyada rasgelelik olgusu içeren bir özellik ile ilgili ölçme işlemine karşılık gelen X rasgele değişkenin olasılık dağılımı, $F = \{f(., \boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$ parametrik olasılık (yoğunluk) fonksiyonu ailesinin bir elemanı olsun. $\boldsymbol{\theta} \in \Theta$ için $f(., \boldsymbol{\theta})$ o.(y).f. na sahip dağılımdan bir örneklem $\mathbf{X} = (X_1, X_2, \dots, X_n)'$ olmak üzere parametre tahmini, örneklemin kendisinin veya bir $\hat{\theta}(X_1, X_2, \dots, X_n)$ istatistiğinin hangi $\boldsymbol{\theta}$ değerini desteklediği, yani $\boldsymbol{\theta}$ parametresini tahmin etme (kestirme) işlemidir (Öztürk ve ark. 2006).

Bir tahmin edici, tahmin edicilerde aranan özelliklerin bazılarına göre diğerlerinden daha iyi, bazılarına göre daha kötü olabilmektedir. Bundan dolayı, önce tahmin edicide aranan özellikler tespit edilmeli ve daha sonra böyle bir tahmin edici ortaya çıkaracak yöntemle başvurulmalıdır. Ancak bazı durumlar dışında, önceden öne sürülen özelliklere sahip tahmin ediciler ortaya çıkaracak genel yöntemler mevcut değildir. Genellikle, sezgisel olarak veya belli bir yöntemle göre tahmin edici bulunmakta ve aranan özelliğe veya özelliklere göre en iyisi olup olmadığına bakılmaktadır (Öztürk ve ark. 2006). Burada denklem (2.1) için parametre tahmin yöntemlerinden sırasıyla En küçük kareler ve En çok olabilirlik yöntemleri ele alınacaktır.

2.1.1.1. En küçük kareler yöntemi

En küçük kareler yöntemi (EKK) genel olarak, gözlem noktalarına eğri uydurmada kullanılan bir yakınlaştırma yöntemi olarak bilinmektedir. Bununla

birlikte regresyon, doğrusal ve doğrusal olmayan modellerde parametre tahmininde çok kullanılan bir tahmin yöntemidir.

Doğrusal olmayan regresyonda model varsayımlarından $\mathbf{f}(\boldsymbol{\theta})$ biçimsel olarak bilindiğinden,

$$Q(\boldsymbol{\theta}) = \sum_{i=1}^n (y_i - f(\mathbf{x}_i; \boldsymbol{\theta}))^2 \quad (2.3)$$

hata kareler toplamı minimum olacak şekilde $\boldsymbol{\theta}$ parametre değerini belirlemek EKK yöntemine göre en iyi tepki fonksiyonunu bulmak demektir. Eşitlik (2.3) de k açıklayıcı değişken sayısı olmak üzere $\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{ik})$, $i = 1, 2, \dots, n$ olarak tanımlanmıştır. Eşitlik (2.3) vektörel formda

$$Q(\boldsymbol{\theta}) = (\mathbf{Y} - \mathbf{f}(\boldsymbol{\theta}))' (\mathbf{Y} - \mathbf{f}(\boldsymbol{\theta})) \quad (2.4)$$

biçiminde yazılabilir.

$\boldsymbol{\theta}_0$ ($\boldsymbol{\theta}_0 \in \Theta$) 'nın bir komşuluğunda $\mathbf{f}(\boldsymbol{\theta})$ fonksiyonu sürekli ve en az iki kere türevlenebilir bir fonksiyondur. Buna göre,

$$\mathbf{F}(\boldsymbol{\theta}) = \begin{bmatrix} \frac{\partial}{\partial \theta_1} f(\mathbf{x}_1, \boldsymbol{\theta}) & \frac{\partial}{\partial \theta_2} f(\mathbf{x}_1, \boldsymbol{\theta}) & \dots & \frac{\partial}{\partial \theta_p} f(\mathbf{x}_1, \boldsymbol{\theta}) \\ \frac{\partial}{\partial \theta_1} f(\mathbf{x}_2, \boldsymbol{\theta}) & \frac{\partial}{\partial \theta_2} f(\mathbf{x}_2, \boldsymbol{\theta}) & \dots & \frac{\partial}{\partial \theta_p} f(\mathbf{x}_2, \boldsymbol{\theta}) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial}{\partial \theta_1} f(\mathbf{x}_n, \boldsymbol{\theta}) & \frac{\partial}{\partial \theta_2} f(\mathbf{x}_n, \boldsymbol{\theta}) & \dots & \frac{\partial}{\partial \theta_p} f(\mathbf{x}_n, \boldsymbol{\theta}) \end{bmatrix}_{(n \times p)} \quad (2.5)$$

biçiminde tanımlanan $\mathbf{F}(\boldsymbol{\theta})$ matrisi $\mathbf{f}(\boldsymbol{\theta})$ fonksiyonunun türev (jakobiyen) matrisi olmak üzere (2.4) eşitliğinin en küçüklenmesi

$$\begin{aligned} \frac{\partial}{\partial \boldsymbol{\theta}} Q(\boldsymbol{\theta}) \Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} &= \mathbf{0} \\ -2\mathbf{F}'(\hat{\boldsymbol{\theta}})(\mathbf{Y} - \mathbf{f}(\hat{\boldsymbol{\theta}})) &= \mathbf{0} \end{aligned} \quad (2.6)$$

biçimindedir. $\boldsymbol{\theta}$ 'nın EKK tahmin edicisi, Θ parametre uzayı üzerinde $Q(\boldsymbol{\theta})$ 'nın,

$$\min_{\boldsymbol{\theta} \in \Theta} Q(\boldsymbol{\theta}) = Q(\hat{\boldsymbol{\theta}})$$

en küçüklenmesinden elde edilen $\hat{\theta}$ değeridir. Bu durumda (2.6) ifadesinden

$$\mathbf{F}'(\hat{\theta})(\mathbf{Y} - \mathbf{f}(\hat{\theta})) = \mathbf{0} \quad (2.7)$$

elde edilir. (2.7) eşitliğinden Jakobiyen matrisi ile $\mathbf{e} = \mathbf{Y} - \mathbf{f}(\hat{\theta})$ artık vektörünün birbirlerine dik olduğu anlaşılmaktadır (Genç 1997).

Eşitlik (2.7) de $\hat{\theta}$, tek başına eşitliğin bir tarafına alınamamaktadır. Bu nedenle $\hat{\theta}$ değeri yinelemeli yöntemlerle elde edilebilmektedir (Gallant 1975).

2.1.1.2. En çok olabilirlik yöntemi

En çok olabilirlik yöntemi (Maximum Likelihood Estimator - MLE), tahmin edicileri elde etme yöntemleri arasında en çok kullanılanıdır. Olabilirlik ilkesine dayanan bu yöntem ile elde edilen tahmin ediciler, tahmin edicilerde aranan özelliklerin bir çoğuna göre iyi olmakla birlikte elde edilmeleri sırasındaki en büyükleme probleminin çözümünde sıkıntılar çekilmektedir.

Model (2.1) de ε_i , $i = 1, 2, \dots, n$ hata terimlerinin bağımsız ve her birinin sıfır ortalamalı bilinmeyen σ^2 varyanslı normal dağılıma sahip olduğu varsayılın. Bu durumda,

$$Y_i \sim N(f_i(\boldsymbol{\theta}), \sigma^2) \quad , \quad i = 1, 2, \dots, n$$

dağılımına sahip olup, olasılık yoğunluk fonksiyonu

$$f_Y(y_i; \boldsymbol{\theta}, \sigma^2) = \frac{1}{\sqrt{2\pi} \sigma} e^{-\frac{1}{2\sigma^2}(y_i - f_i(\boldsymbol{\theta}))^2} \quad , \quad y_i \in \mathbb{R}$$

olacaktır. $Y_1, Y_2, \dots, Y_n$ rasgele değişkenlerinin olabilirlik fonksiyonu,

$\mathbf{Y} = (Y_1, Y_2, \dots, Y_n)'$ olmak üzere

$$\begin{aligned} L(\mathbf{Y}; \boldsymbol{\theta}, \sigma^2) &= \prod_{i=1}^n f_Y(y_i; \boldsymbol{\theta}, \sigma^2) \\ &= (2\pi)^{-n/2} (\sigma^2)^{-n/2} e^{-\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - f_i(\boldsymbol{\theta}))^2} \end{aligned}$$

biçimindedir. $L(\mathbf{Y}; \boldsymbol{\theta}, \sigma^2)$ fonksiyonunu en büyükleyen bilinmeyen $\boldsymbol{\theta}$ ve σ^2 parametrelerinin tahmin edilmesi gerekmektedir. Logaritma fonksiyonunun monotonluk özelliğinden $L(\mathbf{Y}; \boldsymbol{\theta}, \sigma^2)$ fonksiyonunun logaritması alınarak elde edilen $\log L(\mathbf{Y}; \boldsymbol{\theta}, \sigma^2)$ en büyüklenerek

$$\begin{cases} \sum_{i=1}^n (Y_i - f_i(\boldsymbol{\theta})) \frac{\partial f_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'} = 0 \\ -n\sigma^2 + \sum_{i=1}^n (Y_i - f_i(\boldsymbol{\theta}))^2 = 0 \end{cases} \quad (2.8)$$

denklem sistemine ulaşılır. Denklem (2.8) sistemindeki birinci denklemin sol tarafı (2.7) deki ifadenin aynısıdır. Dolayısıyla $\boldsymbol{\theta}$ 'nın EKK yöntemi ile elde edilen tahmini MLE yöntemi ile elde edilen tahminine eşittir. $\boldsymbol{\theta}$ 'nın MLE yöntemi ile elde edilen tahmini $\hat{\boldsymbol{\theta}}$ olmak üzere ikinci denklemden σ^2 'nın tahmin edicisi,

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n (Y_i - f_i(\hat{\boldsymbol{\theta}}))^2}{n}$$

olarak elde edilir.

Eşitlik (2.8), $\boldsymbol{\theta}$ 'nın doğrusal olmayan bir fonksiyonudur. Doğrusal olmayan bu fonksiyonun en küçüklenmesi ile ilgili kullanılan optimizasyon algoritmasının ismine bağlı olarak tahmin yöntemi de bu ismi almaktadır (Marquardt 1963, Gallant 1977).

2.1.2. Yinelemeli yöntemler

$\hat{\boldsymbol{\theta}}$ değeri eşitlik (2.7) de tek başına eşitliğin bir tarafına alınamadığından, yinelemeli yöntemlerden Gauss-Newton, En hızlı iniş ve Levenberg-Marquardt yöntemleriyle elde edilebilmektedir.

2.1.2.1. Gauss-Newton yöntemi

Bu yöntemde sayısal hesaplamalar açısından amaç, işlemlerin yakınsama hızını artırmak ve ardışık işlem sayısını azaltmaktır. İşlemlerin yakınsama hızını artırabilmek ve ardışık işlem sayısını da azaltabilmek için ardışık işlem adımlarına

iyi bir değerden başlamak önemlidir. $\mathbf{f}(\boldsymbol{\theta})$ fonksiyonunun $\boldsymbol{\theta}^0$ noktasındaki Taylor serisine açılımı,

$$\mathbf{f}(\boldsymbol{\theta}) \cong \mathbf{f}(\boldsymbol{\theta}^0) + \mathbf{F}(\boldsymbol{\theta}^0)(\boldsymbol{\theta} - \boldsymbol{\theta}^0) \quad (2.9)$$

biçiminde yazılabilir (Gallant 1977). EKK ve MLE ile elde edilen (2.7) denkleminde (2.9) denklemi yerine konulursa,

$$\begin{aligned} \mathbf{F}'(\boldsymbol{\theta}^0)(\mathbf{Y} - \mathbf{f}(\boldsymbol{\theta}^0) - \mathbf{F}(\boldsymbol{\theta}^0)(\boldsymbol{\theta} - \boldsymbol{\theta}^0)) &= \mathbf{0} \\ \mathbf{F}'(\boldsymbol{\theta}^0)(\mathbf{Y} - \mathbf{f}(\boldsymbol{\theta}^0)) - \mathbf{F}'(\boldsymbol{\theta}^0)\mathbf{F}(\boldsymbol{\theta}^0)(\boldsymbol{\theta} - \boldsymbol{\theta}^0) &= \mathbf{0} \\ \mathbf{F}'(\boldsymbol{\theta}^0)\mathbf{F}(\boldsymbol{\theta}^0)(\boldsymbol{\theta} - \boldsymbol{\theta}^0) &= \mathbf{F}'(\boldsymbol{\theta}^0)(\mathbf{Y} - \mathbf{f}(\boldsymbol{\theta}^0)) \\ \boldsymbol{\theta} - \boldsymbol{\theta}^0 &= (\mathbf{F}'(\boldsymbol{\theta}^0)\mathbf{F}(\boldsymbol{\theta}^0))^{-1} \mathbf{F}'(\boldsymbol{\theta}^0)(\mathbf{Y} - \mathbf{f}(\boldsymbol{\theta}^0)) \\ \boldsymbol{\theta} &= \boldsymbol{\theta}^0 + (\mathbf{F}'(\boldsymbol{\theta}^0)\mathbf{F}(\boldsymbol{\theta}^0))^{-1} \mathbf{F}'(\boldsymbol{\theta}^0)(\mathbf{Y} - \mathbf{f}(\boldsymbol{\theta}^0)) \end{aligned} \quad (2.10)$$

elde edilir. Eşitlik (2.10)'nın çözümünde kullanılan algoritmik adımlar şöyledir (Hartley ve Booker, 1963) :

A1. $\varepsilon > 0$ ve $\tau > 0$ keyfi sabitleri (genellikle $\varepsilon = 10^{-5}$ ve $\tau = 10^{-3}$ olarak önerilir) ve $\boldsymbol{\theta}^0$ parametre başlangıç değerleri verilir.

A2. i. adım için $\mathbf{d}^{i-1} = (\mathbf{F}'(\boldsymbol{\theta}^{i-1})\mathbf{F}(\boldsymbol{\theta}^{i-1}))^{-1} \mathbf{F}'(\boldsymbol{\theta}^{i-1})(\mathbf{Y} - \mathbf{f}(\boldsymbol{\theta}^{i-1}))$ değeri hesaplanır ve $\lambda_0 \mathbf{d}^{i-1}$ ($0 \leq \lambda_0 \leq 1$) olmak üzere sabit bir λ_0 değerinden başlanarak, $Q(\boldsymbol{\theta}^i) \leq Q(\boldsymbol{\theta}^{i-1})$ olacak biçimde seçilen bir λ ile,

$$\boldsymbol{\theta}^i = \boldsymbol{\theta}^{i-1} + \lambda \mathbf{d}^{i-1}, \quad i = 1, 2, \dots$$

yeni parametre değerleri hesaplanır.

A3. $\|\boldsymbol{\theta}^i - \boldsymbol{\theta}^{i-1}\| < \varepsilon (\|\boldsymbol{\theta}^{i-1}\| + \tau)$ ve $|Q(\boldsymbol{\theta}^i) - Q(\boldsymbol{\theta}^{i-1})| < \varepsilon (Q(\boldsymbol{\theta}^{i-1}) + \tau)$ koşulları aynı anda sağlanıncaya kadar A2' ye gidilir.

2.1.2.2. En hızlı iniş yöntemi

Amaç fonksiyonunun $\boldsymbol{\theta} \in \mathbb{R}^p$ noktasındaki gradyant vektörü,

$$\nabla Q(\boldsymbol{\theta}) = \begin{bmatrix} \frac{\partial Q(\boldsymbol{\theta})}{\partial \theta_1} \\ \frac{\partial Q(\boldsymbol{\theta})}{\partial \theta_2} \\ \vdots \\ \frac{\partial Q(\boldsymbol{\theta})}{\partial \theta_p} \end{bmatrix}$$

dir. En hızlı iniş yöntemi bir gradyant yöntemidir. Bu yöntemde, Gauss-Newton yöntemindeki $\mathbf{d}^{i-1}$ vektörünün yerine gradyant vektörünün ters yönünde olan $-\nabla Q(\boldsymbol{\theta}^{i-1})$ vektörü kullanılmaktadır. Buna göre algoritma, $\boldsymbol{\theta}^0$ başlangıç değeri olmak üzere,

$$\boldsymbol{\theta}^i = \boldsymbol{\theta}^{i-1} - \lambda^{i-1} \nabla Q(\boldsymbol{\theta}^{i-1}) \quad , \quad i = 1, 2, \dots$$

biçimindedir. λ^{i-1} değerleri, genel olarak $Q(\boldsymbol{\theta}^i) \leq Q(\boldsymbol{\theta}^{i-1})$ olacak biçimde ayrıca belirlenen kurallara göre seçilebilir olmasına rağmen $Q(\boldsymbol{\theta}^{i-1} + \lambda^{i-1} \mathbf{d}^{i-1})$ yi en küçük yapan değer λ^{i-1} değeri olarak da seçilebilir. İndisler gözardı edildiğinde $Q(\boldsymbol{\theta} + \lambda \mathbf{d})$ değerini en küçükleyen λ değerini bulmak için $Q(\boldsymbol{\theta} + \lambda \mathbf{d})$ fonksiyonu $\boldsymbol{\theta}$ noktası civarında Taylor serisine açıldığında,

$$Q(\boldsymbol{\theta} + \lambda \mathbf{d}) = Q(\boldsymbol{\theta}) + \lambda \frac{\partial Q(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \mathbf{d} + \frac{1}{2} \lambda^2 \mathbf{d}' \frac{\partial^2 Q(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \mathbf{d}$$

olur. Burada λ değeri

$$Q(\boldsymbol{\theta} + \lambda \mathbf{d}) \leq Q(\boldsymbol{\theta})$$

koşulunu sağlayacak biçimde seçilir. Bu durumda,

$$Q(\boldsymbol{\theta} + \lambda \mathbf{d}) - Q(\boldsymbol{\theta}) \leq 0$$

olup

$$Q(\boldsymbol{\theta} + \lambda \mathbf{d}) - Q(\boldsymbol{\theta}) = \lambda \frac{\partial Q(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \mathbf{d} + \frac{1}{2} \lambda^2 \mathbf{d}' \frac{\partial^2 Q(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \mathbf{d} \leq 0$$

dır. λ değerini elde etmek için $Q(\boldsymbol{\theta} + \lambda \mathbf{d}) - Q(\boldsymbol{\theta})$ ifadesi λ üzerinden en küçüklenirse,

$$\frac{\partial Q(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \mathbf{d} + \lambda \mathbf{d}' \frac{\partial^2 Q(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \mathbf{d} = 0$$

eşitliğinden

$$\lambda^* = - \frac{\frac{\partial Q(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \mathbf{d}}{\mathbf{d}' \frac{\partial^2 Q(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \mathbf{d}}$$

bulunur (Bates ve Watts 1988).

2.1.2.3. Levenberg-Marquardt yöntemi

Model (2.1) de $\boldsymbol{\theta}$ bir $\boldsymbol{\theta}^0$ başlangıç değerine bağlı olarak $\boldsymbol{\theta} = \boldsymbol{\theta}^0 + \mathbf{d}$ biçiminde yazılsın. Bu durumda model,

$$Y_i = f(\mathbf{x}_i, \boldsymbol{\theta}^0 + \mathbf{d}) + \varepsilon_i, \quad i = 1, 2, \dots, n$$

olur. $f(\mathbf{x}_i, \boldsymbol{\theta}^0 + \mathbf{d})$ fonksiyonu $\boldsymbol{\theta}^0$ noktası civarında birinci dereceden Taylor serisine açılırsa

$$Y_i = f(\mathbf{x}_i, \boldsymbol{\theta}^0) + \sum_{j=1}^p \left(\frac{\partial f(\mathbf{x}_i, \boldsymbol{\theta}^0)}{\partial \theta_j} \right) d_j + \varepsilon_i, \quad i = 1, 2, \dots, n$$

vektör gösterimi ile

$$\mathbf{Y} = \mathbf{f}(\boldsymbol{\theta}^0) + \mathbf{F}'(\boldsymbol{\theta}^0) \mathbf{d} + \boldsymbol{\varepsilon}$$

biçimindedir ve

$$\mathbf{u}^0 = \mathbf{Y} - \mathbf{f}(\boldsymbol{\theta}^0)$$

değişken değiştirmesi ile

$$\mathbf{u}^0 = \mathbf{F}(\boldsymbol{\theta}^0) \mathbf{d} + \boldsymbol{\varepsilon}$$

denklemini elde edilir. Amaç uygun bir $\mathbf{d}$ yön vektörü belirlemektir. Normal denklemler

$$\mathbf{F}'(\boldsymbol{\theta}^0) \mathbf{F}(\boldsymbol{\theta}^0) \mathbf{d} = \mathbf{F}'(\boldsymbol{\theta}^0) \mathbf{u}$$

dır. $\mathbf{C}^0 = \mathbf{F}'(\boldsymbol{\theta}^0)\mathbf{F}(\boldsymbol{\theta})$ matrisinin köşegen elemanlarının dışındakilerin yerine sıfır ve c_{ii} köşegen elemanlarının yerine $1/\sqrt{c_{ii}}$ yazılmasıyla elde edilen köşegen matris,

$$\mathbf{D}_0 = \begin{bmatrix} 1/\sqrt{c_{11}} & 0 & \dots & 0 \\ 0 & 1/\sqrt{c_{22}} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1/\sqrt{c_{pp}} \end{bmatrix}$$

olmak üzere,

$$\mathbf{C}_*^0 = \mathbf{D}_0' \mathbf{C}^0 \mathbf{D}_0$$

$$\mathbf{g}_*^0 = \mathbf{D}_0 \mathbf{F}'(\boldsymbol{\theta}^0) \mathbf{u}$$

ve

$$\mathbf{d}^* = \mathbf{D}_0^{-1} \mathbf{d}^0$$

dönüşümleri altında normal denklemler

$$\mathbf{C}_*^0 \mathbf{d}^* = \mathbf{g}_*^0$$

şeklinde yazılır. Zayıf koşulluluğa bir tedbir olarak $\mathbf{C}^0$ matrisinde köşegen elemanlarına λ ($\lambda > 0$) gibi küçük bir sayının eklenmesiyle

$$(\mathbf{C}_*^0 + \lambda \mathbf{I}) \mathbf{d}^* = \mathbf{g}_*^0$$

yazılır. Buradan

$$\mathbf{d}_*^0 = (\mathbf{C}_*^0 + \lambda \mathbf{I})^{-1} \mathbf{g}_*^0$$

olmak üzere,

$$\boldsymbol{\theta} = \boldsymbol{\theta}^0 + \mathbf{d}^0$$

iyileştirilmiş $\boldsymbol{\theta}$ değerine dönüştürülebilir. $\mathbf{d}^0 = \mathbf{D}_0 \mathbf{d}_*^0$ yön vektörü yani λ ($\lambda > 0$) değeri $Q(\boldsymbol{\theta}) < Q(\boldsymbol{\theta}^0)$ olacak şekilde seçilmelidir. λ 'nın seçiminde kullanılabilecek algoritma aşağıdaki gibi olabilir (Marquardt 1963):

A1. ϑ ($\vartheta > 1$) sayısı seçilir. (Örneğin, $\vartheta = 10$)

A2. $\lambda_0 = 10^{-2}$ alınır.

A3. $a = \lambda_0 / 9$, $b = \lambda_0$ ve

$$Q_a = Q(\boldsymbol{\theta}^0 + \mathbf{D}_0(\mathbf{C}_*^0 + \lambda^0 / 9 \mathbf{I})^{-1} \mathbf{g}_*^0)$$

$$Q_b = Q(\boldsymbol{\theta}^0 + \mathbf{D}_0(\mathbf{C}_*^0 + \lambda^0 \mathbf{I})^{-1} \mathbf{g}_*^0)$$

olmak üzere,

(i) eğer $Q_a < Q(\boldsymbol{\theta}^0)$ ise $\lambda = a$ alınır.

(ii) eğer $Q_a > Q(\boldsymbol{\theta}^0)$ ve $Q_b < Q(\boldsymbol{\theta}^0)$ ise $\lambda = b$ alınır.

(iii) eğer $Q_a > Q(\boldsymbol{\theta}^0)$ ve $Q_b > Q(\boldsymbol{\theta}^0)$ ise $\lambda = \lambda^0 \mathcal{G}^\omega$ alınır. Burada ω ,

$$Q(\boldsymbol{\theta} + \mathbf{D}_0(\mathbf{C}_*^0 + \lambda^0 \mathcal{G}^\omega \mathbf{I})^{-1} \mathbf{g}_*^0) < Q(\boldsymbol{\theta}^0)$$

eşitsizliğini sağlayan en küçük pozitif tamsayıdır.

2.2. En Büyük Yan

Yanın hesaplanması ilk defa Cox ve Snell (1968) tarafından ele alınmıştır. Doğrusal olmayan regresyonda yanın ilk hesaplanmasını Box (1971) yapmıştır. Birçok modelde yanın kullanışlı sayısal çözümü Gillis ve Ratkowsky (1978), Laurent ve Cook (1992) çalışmalarında mevcuttur. Bates ve Watts (1980), tek açıklanan değişkenin bulunduğu modellerde eğriliğin parametreye olan etkisini incelemiştir.

$X_1, X_2, \dots, X_n$ parametresi $\boldsymbol{\theta}$ olan bir kitleden alınan örneklem ve $\hat{\boldsymbol{\theta}}$ da $\boldsymbol{\theta}$ 'nın bir tahmin edicisi olsun. $E_{\boldsymbol{\theta}}(\hat{\boldsymbol{\theta}})$ 'nin var olması durumunda $E_{\boldsymbol{\theta}}(\hat{\boldsymbol{\theta}}) - \boldsymbol{\theta}$ değerine, $\hat{\boldsymbol{\theta}}$ tahmin edicisinin yanı (bias) denir ve

$$Bias_{\boldsymbol{\theta}}(\hat{\boldsymbol{\theta}}) = E_{\boldsymbol{\theta}}(\hat{\boldsymbol{\theta}}) - \boldsymbol{\theta}$$

ile gösterilir.

Yanlı tahmin edicilerin yanlılığı azaltılabilir. Tahmin edicilerin yanlılığını azaltmak için değişik teknikler kullanılır (Martin 1989 - Efron ve Tibshirani, 1993 - Davison ve Hinkley 1997- Berrendero ve ark. 2007).

Bu kısımda yanı azaltmak için önerilen Jackknife ve Bootstrap yöntemlerine değinilecektir.

2.2.1. Jackknife yöntemi

Bu yöntemin amacı, bir tahmin ediciye ait yan miktarını azaltmaktır. $X_1, X_2, \dots, X_n$ gerçel değerli θ parametresine sahip bir kitleden alınmış n büyüklüğünde bir rasgele örneklem olsun. Tüm gözlemlerden θ parametresinin tahmin edicisi $\hat{\theta}_n = \hat{\theta}_n(X_1, X_2, \dots, X_n)$ olarak gösterilsin. Örneklem $m = \left\lfloor \frac{n}{k} \right\rfloor$ hacminde k eşit gruba bölünür. Burada $\left\lfloor . \right\rfloor$ gösterimi, tam değeri ifade etmektedir. Her adımda bir grup çıkarılarak geriye kalan gözlemlerle θ parametresi için ele alınan $\hat{\theta}_i$ 'nin tahmini bulunur. Sırasıyla i . grup çıkartılarak geriye kalan $n - m$ gözlemden

$$\begin{array}{l} X_{(m+1)}, X_{(m+2)}, \dots, X_n \\ X_1, X_2, \dots, X_{(m)}, X_{(2m+1)}, X_{(2m+2)}, \dots, X_n \\ \vdots \\ X_1, X_2, \dots, X_{(m(i-1))}, X_{(im+1)}, \dots, X_n \quad i = 1, 2, \dots, k \\ \vdots \\ X_1, X_2, \dots, X_{(n-m)} \end{array}$$

θ parametresi için elde edilen $\hat{\theta}_i$ tahmin edicisine Jackknife tahmin edicisi denir (Quenouille 1949 ve Tukey 1958).

$$J_i = k\hat{\theta} - (k-1)\hat{\theta}_i, \quad i = 1, 2, \dots, k$$

Olmak üzere,

$$J(\hat{\theta}) = \frac{1}{k} \sum_{i=1}^k (k\hat{\theta} - (k-1)\hat{\theta}_i) = \frac{1}{k} \sum_{i=1}^k J_i \quad (2.12)$$

biçiminde tanımlanır. Özel olarak $k = n$ alındığında Jackknife tahmin edicisi,

$$J(\hat{\theta}) = \frac{1}{n} \sum_{i=1}^n (n\hat{\theta} - (n-1)\hat{\theta}_i)$$

olur ve (2.12) eşitliği başka bir gösterimle

$$\bar{\theta} = \frac{1}{n} \sum_{i=1}^n \hat{\theta}_i$$

olmak üzere

$$J(\hat{\theta}) = \hat{\theta} + (n-1)(\hat{\theta} - \bar{\theta})$$

biçiminde de yazılabilir.

Jackknife örneklem varyans-kovaryans matrisi

$$S_j = \frac{\sum_{i=1}^k (J_i - J(\hat{\theta})) (J_i - J(\hat{\theta}))'}{k-1}$$

olmak üzere, Jackknife tahmin edicisinin güven aralığı

$$P\left(\tilde{\theta}_i - t_{\frac{\alpha}{2}, k-p} \sqrt{\frac{S_{ii}}{k}} \leq \tilde{\theta}_i \leq \tilde{\theta}_i + t_{\frac{\alpha}{2}, k-p} \sqrt{\frac{S_{ii}}{k}}\right) = 1 - \alpha, i = 1, 2, \dots, p$$

olur (Duncan 1978). Burada, p bağımsız değişken sayısını, k grup sayısını ve $\tilde{\theta}_i$ de Jackknife tahmin edicisinin i . bileşenini göstermektedir.

2.2.2. Bootstrap yöntemi

Bootstrap yöntemi ilk defa Efron (1979) tarafından öne sürülmüştür. Bu yöntem mevcut örnek verilerin iadeli ve rasgele örnekleme yöntemiyle oluşturulmasına, bu şekilde oluşturulan her bir örneklem için ilgili istatistiklerin tahmin edilmesine ve bu işlemin defalarca tekrarına dayanan bir süreçtir. Bootstrap yöntemi, matematik formülleri içermeyen, çok basit bir yöntemdir. Veri dağılımı ile ilgili ağır varsayımlar gerektirmediğinden bilinen istatistiksel yöntemlerin ve varsayımların yetersiz kaldığı durumlarda parametrik ve parametrik olmayan istatistik analizlerde kullanılabilen basit ve güvenilir bir yöntemdir (Efron 1979, Efron ve Tibshirani 1993).

Uygulamada veri dağılımına dayanan varsayımların doğruluğundan kesin olarak emin olunamadığı durumlarda bu yaklaşım çok yararlı olabilmektedir. Bootstrap yöntemi, n adet gözlemden oluşan bir veri seti üzerinde, bir tahmin edicinin $\hat{\theta}_n = \hat{\theta}_n(X_1, X_2, \dots, X_n)$ varyansını, deneysel dağılımını ve güven bölgelerini

oluşturmada kullanılabilir (Efron 1979, Wu 1986, Efron ve Tibshirani 1993, Visscher ve ark. 1996, Davison ve Hinkley 1997).

Örneğin $V_F(\hat{\theta}_n)$, $\hat{\theta}_n$ tahmin edicisinin varyansını gösterebilir. Bu varyans F 'in bir fonksiyonu olsun. Eğer F biliniyorsa, $\hat{\theta}_n$ tahmin edicisinin varyansı hesaplanabilir. Buna bir örnek verilecek olursa

$$\hat{\theta}_n = n^{-1} \sum_{i=1}^n X_i \text{ olduğunda,}$$

$$V_F(\hat{\theta}_n) = \frac{\sigma^2}{n} = \frac{\int x^2 dF(x) - \left(\int x dF(x) \right)^2}{n}$$

dır. $V_F(\hat{\theta}_n)$ görüldüğü gibi F 'nin bir fonksiyonu olur.

Bootstrap yöntemiyle $V_F(\hat{\theta}_n)$, $V_{\hat{F}_n}(\hat{\theta}_n)$ ile tahmin edilebilir. Buna göre Bootstrap (V_B) tahmini adımsal olarak şu şekilde elde edilir:

A1. Gözlem değerlerinden iadeli olarak n kere rasgele çekiliş yapılır. Bunlar $X_1^*, X_2^*, \dots, X_n^*$ olarak gösterilsin.

A2. Her rasgele çekilen gruptan ele alınan $\hat{\theta}^* = \hat{\theta}^*(X_1^*, X_2^*, \dots, X_n^*)$ tahmin edici hesaplanır.

A3. A1 ve A2 adımları B kere tekrar edilsin. Elde edilen tahmin değerleri $\hat{\theta}_1^*, \hat{\theta}_2^*, \dots, \hat{\theta}_B^*$ olarak gösterilirse, Bootstrap tahmini

$$V_B = \frac{1}{B} \sum_{i=1}^B \left(\hat{\theta}_i^* - \frac{1}{B} \sum_{j=1}^B \hat{\theta}_j^* \right)^2$$

olur.

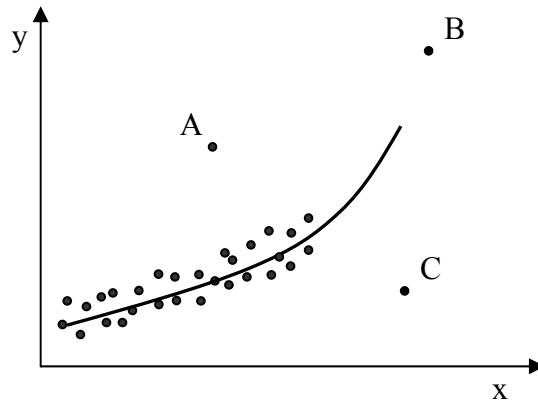
Büyük sayılar kanunundan V_B , B 'nin yeteri kadar büyük olması durumunda hemen hemen her yerde $V_{\hat{F}_n}(\hat{\theta}_n)$ ye yakınsamaktadır yani $V_B \xrightarrow[B \rightarrow \infty]{hhhy} V_{\hat{F}_n}(\hat{\theta}_n)$ olur.

$\hat{\theta}_n$ 'in standart hatasının tahmini de $\hat{s}_B = \sqrt{V_B}$ elde edilir (Efron 1979).

2.3. Gözlem Etkilerinin İncelenmesi

Doğrusal olmayan regresyonda yansız tahmin edicilere ulaşmadaki önemli önkoşullardan biri hataların sıfır ortalama ve sabit varyansla normal dağılıma sahip olduğu varsayımdır. Bu varsayımın sağlanmaması durumunda model parametrelerinin en iyi tahminlerine ulaşmak zorlaşacaktır. Bu sorun daha çok veride etkin (leverage) gözlemlerin ve aykırı değerlerin varlığında ortaya çıkar.

Çok bileşenli modellerde bir değer in açıklayıcı değişken yönünde aykırı değer olduğunu belirlemek oldukça zordur. Şekil 2.1. incelendiğinde B noktası **X** uzayında örneklemin diğer değerlerinden daha uzakta ancak neredeyse diğer değerlerin içinden geçen regresyon eğrisinin üstünde olduğu görülmektedir. Bu durumdaki gözlemlere *iyi etkin değer* denir. Ancak bunlar **X** uzayında iyi değerlere sahip değildir. Bu tip gözlemler regresyon katsayı tahminlerini etkilemezler fakat ilişki katsayısı R^2 , katsayıların standart hatası gibi model istatistikleri üzerinde çok önemli etkileri vardır. Şekil 2.1. deki C noktasındaki durum göz önüne alındığında, bu nokta **Y** uzayının içinde ancak **X** uzayında aykırı bir değerdir. Bu durumdaki gözlemlere *kötü etkin değer* veya *yatay aykırı değer* denir. Bu gözlemlerin model katsayıları üzerinde önemli etkileri vardır ve regresyon doğrusunun yönünü kendilerine doğru çekme etkisi yaparlar. Şekil 2.1. deki A noktasındaki durum göz önüne alındığında, bu gözlemin **Y** uzayından uzakta fakat **X** uzayında normal görünmektedir. Bu tip gözlemlere **Y** *uzayında aykırı değer* veya *dikey aykırı değer* denir. Bu gözlemler büyük artık değerlerine sahiptirler.



Şekil 2.1. İyi etkin nokta, dikey aykırı değer, kötü etkin nokta gösterimi.

Alışılmış çoklu doğrusal regresyonda $\mathbf{X}$ uzayında aykırı değer olan noktaları belirlemek için en çok kullanılan yöntem H şapka matrisinin köşegen elemanları h_{ii} 'leri incelemektir. Ek olarak Mahalanobis uzaklığı, Cook uzaklığı, DFFITS ölçüleri kullanılabilir (Adnan ve Mohamad 2003).

2.4. Sağlamlık Ölçütleri

Sağlam bir tahmin edici için istenen iki özellik dirençlilik ve sağlam etkinlik dir. Eğer bir tahmin edici az sayıdaki büyük hatadan ya da herhangi bir miktardaki yuvarlama-gruplama hatalarından sınırlı miktarda etkileniyorsa dirençlidir denir. Örneklemin küçük bir alt kümesi tahminler üzerinde gereğinden çok bir etkiye sahip değilse bu tahmin edici büyük hatalara karşı dirençlidir yorumu yapılır. Tahmin edici küçük hatalara sürekli olarak tepki gösteriyorsa hatta tahminler bu gözlemlerin küçük bir parçasının yuvarlanma ya da gruplanması ile belirleniyorsa bu tahmin edici gruplama-yuvarlama hatalarına karşı dirençli değildir. Bir tahmin edicinin varyansı her bir dağılım için en küçüğe yakın bir değer alıyorsa bu tahmin edici sağlam etkinliğe sahiptir denir.

Normal dağılımdan alınan örneklemeler üzerindeki tahmin edicilerin genellikle dirençli olmadığı yapılan çalışmalar sonucu gösterilmiştir (Hampel ve ark. 1986). Bu nedenle tahmin edicileri sadece bu dağılım altında incelemek yeterli değildir. Normal dağılıma ek olarak daha uzun kuyruklu dağılımlar da incelenmelidir.

Sağlam etkinlik özelliği, bir tahmin edicinin dağılımının belli olmadığı durumlarda tekrarlı olarak alınan örneklemeler içinde iyi sonuçlar verdiğini garanti altına alır. Örneklemin bozulduğu durumlarda tahminlerin çok az miktarda değişim göstermesi önemli bir durumdur (Motulsky ve Brown 2006, Huber ve Ronchetti 2009).

$\hat{\theta}(F)$, F dağılım fonksiyonları kümesi üzerinde gerçel değerli istatistiksel bir fonksiyonel olarak tanımlansın. F dağılımında ortaya çıkabilecek küçük değişimler $\hat{\theta}(F)$ üzerinde bir etki yaratmıyorsa, $\hat{\theta}(F)$ için nitel sağlamlığa sahiptir denir.

Sağlam bir yöntem kullanılarak elde edilen tahmin edicinin ne derecede sağlam olduğuna karar vermede Etki fonksiyonu (Influence Function/Curve – IF/IC) ve Bozulma noktasının önemi büyüktür.

Tezin bütünlüğünü bozmamak için bu iki özelliğin sadece temel tanımları alt kesimde verilecektir. Bu özellikler 5. bölümde daha ayrıntılı olarak ele alınacaktır.

2.4.1. Etki fonksiyonu

G yönündeki $\hat{\theta}$ ' nın F deki **Gateaux türevi**,

$$L_F(G) = \lim_{\varepsilon \rightarrow 0} \frac{\hat{\theta}((1-\varepsilon)F + \varepsilon G) - \hat{\theta}(F)}{\varepsilon}$$

olarak tanımlanır. Eğer x deki nokta değeri $G = \delta_x$ ise $L_F(x) \equiv L_F(\delta_x)$ yazılabilir ve

$$L_F(x) = \lim_{\varepsilon \rightarrow 0^+} \frac{\hat{\theta}((1-\varepsilon)F + \varepsilon \delta_x) - \hat{\theta}(F)}{\varepsilon} \quad (2.13)$$

elde edilen (2.13) eşitliğine **etki fonksiyonu** denir (Hampel 1974). Benzer biçimde

$\hat{L}(x) = L_{\hat{F}_n}(x)$ olarak tanımlanırsa **deneysel (empirical) etki fonksiyonu**,

$$\hat{L}(x) = \lim_{\varepsilon \rightarrow 0^+} \frac{\hat{\theta}((1-\varepsilon)\hat{F}_n + \varepsilon \delta_x) - \hat{\theta}(\hat{F}_n)}{\varepsilon}$$

olur.

$\hat{\theta}(F) = \int x dF(x)$ lineer bir fonksiyonel olmak üzere, δ_x diract dağılımına sahip olduğundan etki fonksiyonu $L_F(x) = x - \hat{\theta}(F)$ ve deneysel etki fonksiyonu $\hat{L}(x) = x - \hat{\theta}(\hat{F}_n)$ olur (Wasserman 2006).

2.4.2. Bozulma noktası

Bir tahmin edicinin sağlamlık ölçüsü olarak tanımlanan bozulma noktasını, Hampel (1971), Hodges (1967) in düşüncesini genelleştirerek ilk defa ortaya koymuştur. Donoho (1982) ve Donoho ve Huber (1983) de bu kavramın sonlu örneklem için durumunu ele almışlardır.

Bir tahmin edicinin bozulma noktası, veri kümesindeki gözlemlerin yer değiştirmesinde tahmin edicinin etkilenmesi sonucu hata teriminin davranışı olarak ve bir tahmin edicinin sağlamlık ölçüsü olarak tanımlanır.

$\mathbf{X}$ örneklem değerlerinden oluşan gözlem veri setini, $\mathbf{X}_m$ de $\mathbf{X}$ deki herhangi m noktalarıyla $(\mathbf{X}, \mathbb{R})$ deki herhangi noktaların yer değiştirilmesi ile elde edilen veri kümelerinin cümlesini $(m = 1, 2, \dots, n)$ gösterebilir. $\mathbf{X}$ de $\hat{\theta}$ tahmin edicisinin değiştirmeli sonlu örneklem bozulma noktası,

$$\varepsilon^*(\hat{\theta}, \mathbf{X}) = \min \left\{ \frac{m}{n} : \sup_{\mathbf{X}_m} \|\hat{\theta}(\mathbf{X}_m) - \hat{\theta}(\mathbf{X})\| = \infty \right\}$$

dır (Donoho ve Huber 1983).

Bozulma noktası $0 \leq \varepsilon^* \leq 1$ aralığında değerler almaktadır. Sağlam tahmin edicilerde bozulma noktasının $1/2$ olması arzu edilir. Bozulma noktasının $1/2$ den büyük olması, mevcut gözlem değerlerinin yarısından fazlasının tahmin değerleri için hesaba alınmadığını gösterir. Bozulma noktasının sıfır'a yakın olması da tahmin edicinin aykırı gözlem değerlerinden aşırı derecede etkileneceğini gösterir (Jureckova ve ark. 1990).

3. DOĞRUSAL OLMAYAN REGRESYONDA SAĞLAM TAHMİN EDİCİLER

Sağlam tahmin ediciler arasında L, R, S ve M tahmin edicileri önemli bir yer tutar. L tahmin ediciler sıra istatistiklerinin doğrusal bileşimidir (linear combination of order statistics). L tahmin edicileri örneklem ortalamasını, medyanını ve kırılmış (kesilmiş-trimmed) ortalamaları özel durum olarak içermektedir. R tahmin edicileri rank testlerinden elde ediliyor olmalarından dolayı bu ismi almışlardır. R tahmin ediciler tek örneklem durumunda yalnızca konum parametreleri için bulunur (Huber 1981). S tahmin ediciler yüksek bozulma noktasına (high breakdown point) sahip tahmin ediciler sınıfıdır. M tahmin ediciler MLE türü tahmin edicilere karşılık gelmektedir. M tahmin ediciler işlerlik, uygunluk bakımından diğer tahmin edicilere göre daha iyidir (Rocke ve Woodruff 1995).

Bu bölümde sırasıyla En küçük mutlak sapma, En küçük kırılmış kareler, M, Genelleştirilmiş M ve S tahmin edicilerinin özellikleri incelenmiştir. Bu tahmin edicilerin doğrusal olmayan regresyonda nasıl hesaplandığı algoritmalar ile gösterilmiştir.

3.1. En Küçük Mutlak Sapma Yöntemi

En Küçük Mutlak Sapma (Least Absolute Deviation – LAD) yöntemi ilk defa Boscovich (1757) tarafından önerilmiştir. Regresyon için amaç fonksiyonu,

$$\min_{\theta} \sum_{i=1}^n |y_i - f(\mathbf{x}_i; \theta)| \quad (3.1)$$

hataların mutlak sapmalarının en küçüklenmesidir. Bu (3.1) optimizasyon probleminin çözümünde bilinmeyen parametreler sırasıyla konum parametresi haline getirilir ve sonrasında medyan veya ağırlıklandırılmış medyan değerleri hesaplanılmaktadır (Ammemiya 1982, Li ve Arce 2004). Basit doğrusal regresyon modeli

$$Y_i = aX_i + b + \varepsilon_i, \quad i = 1, 2, \dots, n \quad (3.2)$$

olmak üzere (3.2) regresyon modelindeki a ve b parametrelerinin LAD ile çözümü şöyledir:

Basit doğrusal regresyon için LAD algoritması:

A1. a 'nın başlangıç değeri $a^{(0)}$ olarak ele alınır.

A2. k . yinelemede, $\min_{\theta} \sum_{i=1}^n |y_i - ax_i - b|$ probleminin çözümü için

$b^{(k)} = \text{med}(y_i - a^{(k-1)} x_i)_{i=1}^n$ olarak bulunur ve sonrasında

$\min_{\theta} \sum_{i=1}^n |y_i - b^{(k)} - ax_i| = \min_{\theta} \sum_{i=1}^n |x_i| \left| \frac{y_i - b^{(k)}}{x_i} - a \right|$ probleminin çözümü için

$a^{(k)} = \text{med} \left(|x_i| \diamond \frac{y_i - b^{(k)}}{x_i} \right)_{i=1}^n$ ağırlıklandırılmış medyan olarak bulunur.

A3. Parametreler aynı anda yakınsayana kadar A2 adımı tekrarlanır.

Bu algoritmada kullanılan $a^{(k)} = \text{med} \left(|x_i| \diamond \frac{y_i - b^{(k)}}{x_i} \right)_{i=1}^n$ ağırlıklandırılmış medyan

aşağıdaki adımlar ile hesaplanır:

Ağırlıklandırılmış medyan hesaplama algoritması:

B1. $W = \frac{1}{2} \sum_{i=1}^n |x_i|$ değeri bulunur.

B2. $\frac{y_i - b^{(k)}}{x_i}$, $i = 1, 2, \dots, n$ değerleri küçükten büyüğe $|x_i|$ değerleriyle birlikte sıralanır.

B3. Sıralama sonrası $|x_i|$ değerleri sırasıyla toplanır ve ilk $\sum_{j=1}^n |x_j| \geq W$ eşitsizliğini

sağlayan j . gözlemlerden hesaplanan $\frac{y_j - b^{(k)}}{x_j}$ değeri ağırlıklandırılmış medyan

değeri olur.

Doğrusal olmayan modellerde her model için ayrı ayrı tüm parametreler konum parametresi haline getirilir. Her bir parametre için doğrusal regresyondaki çözüm algoritması uygulanır. Örneğin model fonksiyonu olarak

$$Y_i = \frac{\theta_1}{(1 + \exp(\theta_2 - \theta_3 \mathbf{x}_i))^{\frac{1}{\theta_4}}} + \varepsilon_i, \quad i = 1, 2, \dots, n \quad (3.3)$$

biçimindeki Richard'ın sigmoidal modeli ele alındığında, bu modelin doğrusal olmayan regresyon da LAD çözüm adımları şöyledir:

A1. θ 'nın başlangıç değerleri $\theta^{(0)}$ olarak seçilsin.

A2. k . yinelemede ($k = 1, 2, 3, \dots$), $\min_{\theta} \sum_{i=1}^n \left| y_i - \frac{\theta_1}{(1 + \exp(\theta_2 - \theta_3 x_i))^{\frac{1}{\theta_4}}} \right|$ probleminin çözümü için,

$$\theta_1^{(k)} = \text{med} \left(\left| \frac{1}{\left(1 + \exp(\theta_2^{(k-1)} - \theta_3^{(k-1)} x_i)\right)^{\frac{1}{\theta_4^{(k-1)}}}} \right| \diamond \left[y_i \left(1 + \exp(\theta_2^{(k-1)} - \theta_3^{(k-1)} x_i)\right)^{\frac{1}{\theta_4^{(k-1)}}} \right]_{i=1}^n \right)$$

$$\theta_2^{(k)} = \text{med} \left(\left| \frac{y_i \left(1 + \exp(\theta_2^{(k-1)} - \theta_3^{(k-1)} x_i)\right)^{\frac{1}{\theta_4^{(k-1)}}} - \theta_1^{(k)}}{\left(1 + \exp(\theta_2^{(k-1)} - \theta_3^{(k-1)} x_i)\right)^{\frac{1}{\theta_4^{(k-1)}}} \left(\theta_3^{(k-1)} x_i + \ln \left(\exp \left(\ln \left(\frac{\theta_1^{(k)}}{y_i} \right) \theta_4^{(k-1)} \right) - 1 \right) - \theta_2^{(k-1)} \right)} \right| \diamond \left[\theta_3^{(k-1)} x_i + \ln \left(\exp \left(\ln \left(\frac{\theta_1^{(k)}}{y_i} \right) \theta_4^{(k-1)} \right) - 1 \right) \right]_{i=1}^n \right)$$

$$\theta_3^{(k)} = med \left(\left[\frac{x_i \left(y_i \left(1 + \exp(\theta_2^{(k)} - \theta_3^{(k-1)} x_i) \right)^{\frac{1}{\theta_4^{(k-1)}}} - \theta_1^{(k)} \right)}{\left(1 + \exp(\theta_2^{(k-1)} - \theta_3^{(k-1)} x_i) \right)^{\frac{1}{\theta_4^{(k-1)}}} \left(\theta_3^{(k-1)} x_i + \ln \left(\exp \left(\ln \left(\frac{\theta_1^{(k)}}{y_i} \right) \theta_4^{(k-1)} \right) - 1 \right) - \theta_2^{(k)} \right)} \right]_{i=1}^n \right) \diamond$$

$$\theta_4^{(k)} = med \left(\left[\frac{\ln \left(\frac{\theta_1^{(k)}}{y_i} \right) \left(y_i \left(1 + \exp(\theta_2^{(k)} - \theta_3^{(k)} x_i) \right)^{\frac{1}{\theta_4^{(k-1)}}} - \theta_1^{(k)} \right)}{\left(1 + \exp(\theta_2^{(k)} - \theta_3^{(k)} x_i) \right)^{\frac{1}{\theta_4^{(k-1)}}} \left(\ln \left(1 + \exp(\theta_2^{(k)} - \theta_3^{(k)} x_i) \right) - \ln \left(\frac{\theta_1^{(k)}}{y_i} \right) \theta_4^{(k-1)} \right)} \right]_{i=1}^n \right) \diamond$$

hesaplanır.

A3. Parametreler aynı anda yakınsayana kadar A2 adımı tekrarlanır.

3.2. En Küçük Kırpılmış Kareler Yöntemi

En Küçük Kırpılmış Kareler (Least Trimmed Square – LTS) yöntemi ilk kez Rousseeuw (1984) tarafından tanımlanmıştır.

$$\min_{\theta} \sum_{i=1}^k (y_i - f(\mathbf{x}_i; \theta))^2, \quad k = \lfloor pn \rfloor + 1, \quad p \in [1/2, 1) \quad (3.4)$$

LTS tahmin edicisi (3.4) deki amaç fonksiyonunu en küçükler. Burada p değeri örneklem hacminin oranını, $\lfloor . \rfloor$ gösterimi de kendinden küçük veya eşit olan en büyük tamsayıyı ifade etmektedir. Bu yöntemde önce EKK ile başlangıç tahminine bağlı olarak elde edilen artıkların karesi alınarak küçükten büyüğe sıralanır. Daha

sonra bu sıralanmış artıkların ilk k tanesinin toplamının en küçüklenmesi sağlanır (Rousseeuw ve Leroy 1987, Hawkins 1994, Atkinson ve Cheng 1999).

Doğrusal olmayan regresyonda LTS tahmin edicisi üzerine Chen (1996)' ın çalışması mevcuttur. Bu çalışmada LTS tahmin edicisinin bozulma noktası da elde edilmiştir.

3.3. M Tahmin Edicisi

M tahmin edicileri, mutlak sapmalar toplamından veya sapmaların kareleri toplamından daha genel olan tahminlerden elde edilen gözlem değerlerinin sapmalarının fonksiyonlarını en küçükler. M tahmin edicileri belirtilmiş bir dağılımda konum parametresinin MLE tahmin edicisi fikrini genelleştirmektedir.

M tahmin edicisi $\mathbf{Y}$ uzayındaki aykırı değerlere karşı sağlam sonuçlar vermesine rağmen, $\mathbf{X}$ uzayındaki aykırı değerlere karşı çok duyarlı olup etki altında kalmaktadır. M tahmini önce ağırlıklandırılmış EKK yı kullanarak ilk tahmini oluşturmaktadır. Bu tahmin sonucuna göre yinelemeli yöntemlerle yeni ağırlıklar hesaplanmakta ve bunları bir sonraki adımda kullanmaktadır. Durdurma kuralları sağlanıncaya kadar yinelemeye devam edilmektedir.

M tahmin edicileri, EKK da kullanılan artık değerlerin karelerinin en küçüklenmesi yerine artık değerlerin başka bir fonksiyonuyla en küçüklemeyi elde etmektedir. EKK tahmininin bozulma duyarlılığını azaltmak için artıkların simetrik bir fonksiyonu olan ρ (criterion - kriter) fonksiyonu kullanılmaktadır (Huber 1977). Buna göre,

$$\min_{\theta} \sum_{i=1}^n \rho(y_i - f(\mathbf{x}_i; \theta)) \quad (3.5)$$

şeklindeki hataların bir fonksiyonunu en küçükleyen tahmin edici M tahmin edicisi olarak tanımlanır. Burada kullanılan $\rho(\cdot)$ kriter fonksiyonu,

- i. $[0, \infty)$ aralığında azalmayan
- ii. Çift fonksiyon
- iii. Sürekli
- iv. $\rho(0) = 0$

özelliklerine sahip bir fonksiyondur (Lopuha 1992).

ψ fonksiyonu ρ fonksiyonunun türev fonksiyonu ve $\mathbf{F}(\boldsymbol{\theta})$ da parametrelere göre $n \times p$ lik türev (jakobiyen) matrisi olmak üzere doğrusal olmayan regresyonda normal denklem sistemi

$$\mathbf{F}'(\boldsymbol{\theta})\psi(\mathbf{e}) = \mathbf{0} \quad (3.6)$$

olarak elde edilir. Eşitlik (3.6) da ölçek parametresi invaryant değildir. İnvaryant ölçek parametresi elde etmek için,

$$\mathbf{F}'(\boldsymbol{\theta})\psi\left(\frac{\mathbf{e}}{\hat{\sigma}}\right) = \mathbf{0} \quad (3.7)$$

kullanılır (Huber 1981). Eşitlik (3.7) de kullanılan $\hat{\sigma}$ değeri sağlam ölçek parametresi olup

$$\hat{\sigma} = 1.4826 \times MAD \quad (3.8)$$

biçiminde bulunur. Eşitlik (3.8) deki MAD değeri mutlak sapmaların medyanıdır (Median Absolute Deviation - MAD). Bu değer en yaygın olarak,

$$MAD = \text{medyan}\{|e_i - e_{0.5}|\}, \quad i = 1, 2, \dots, n$$

biçiminde hesaplanmaktadır (Gather ve Hilker 1997). Burada e_i i. gözleme ait artık değerini, $e_{0.5}$ de artıkların medyan değerini göstermektedir. Eşitlik (3.8) deki

$\frac{1}{\Phi^{-1}(3/4)} = 1.4826$ sabiti, n 'nin yeteri kadar büyük olması ve hata dağılımının

normal olması durumunda $\hat{\sigma}$ 'yı, σ 'nın yansız bir tahmin edicisi yapar. Eşitlik (3.7) de $e_i^s = e_i / \hat{\sigma}$ ve $\mathbf{W}$ de köşegen elemanları

$$w_{ii} = \frac{\psi(e_i^s)}{e_i^s} \quad i = 1, 2, \dots, n$$

olan $n \times n$ lik köşegen matris olmak üzere (3.7) eşitliği

$$\mathbf{F}'(\boldsymbol{\theta})\mathbf{W}(\mathbf{e}^s) = \mathbf{0} \quad (3.9)$$

biçiminde yazılabilir. $\mathbf{W}$ ağırlık matrisi için kriter fonksiyonlarından farklı hesaplamalar yapılabilir. $f(\boldsymbol{\theta})$ fonksiyonu $\boldsymbol{\theta}^0$ noktasındaki Taylor serisine açılırsa (3.9) denklemi,

$$\mathbf{F}'(\boldsymbol{\theta}^0) \mathbf{W} (\mathbf{e}^s - \mathbf{F}'(\boldsymbol{\theta}^0)(\boldsymbol{\theta} - \boldsymbol{\theta}^0)) = \mathbf{0} \quad (3.10)$$

biçiminde yazılabilir ve (3.10) denkleminde işlemlere devam edildiğinde

$$\begin{aligned} \mathbf{F}'(\boldsymbol{\theta}^0) \mathbf{W} \mathbf{e}^s &= \mathbf{F}'(\boldsymbol{\theta}^0) \mathbf{W} \mathbf{F}(\boldsymbol{\theta}^0)(\boldsymbol{\theta} - \boldsymbol{\theta}^0) \\ \boldsymbol{\theta} - \boldsymbol{\theta}^0 &= \left(\mathbf{F}'(\boldsymbol{\theta}^0) \mathbf{W} \mathbf{F}(\boldsymbol{\theta}^0) \right)^{-1} \mathbf{F}'(\boldsymbol{\theta}^0) \mathbf{W} \mathbf{e}^s \\ \boldsymbol{\theta} &= \boldsymbol{\theta}^0 + \left(\mathbf{F}'(\boldsymbol{\theta}^0) \mathbf{W} \mathbf{F}(\boldsymbol{\theta}^0) \right)^{-1} \mathbf{F}'(\boldsymbol{\theta}^0) \mathbf{W} \mathbf{e}^s \end{aligned} \quad (3.11)$$

elde edilir. Denklem (3.11), denklem (2.10) 'nun ağırlıklandırılmış biçimidir. Doğrusal regresyonda $\boldsymbol{\theta}^0$ ve $\hat{\sigma}$ birinci adım olarak EKK tahmininden elde edilir ancak, doğrusal olmayan regresyonda $\hat{\sigma}$ ilk EKK dan elde edilmesine rağmen $\boldsymbol{\theta}^0$ keyfi bir başlangıç noktası olmalıdır. Bunun nedeni (3.11) denkleminde ağırlık fonksiyonunun yinelemede etkisi olmaksızın yineleme sona ermektedir. Doğrusal olmayan regresyonda M algoritması şöyledir;

Doğrusal olmayan regresyonda M algoritması:

A1. Uygun bir başlangıç değeri alınarak EKK yöntemi ile $\hat{\boldsymbol{\theta}}$ 'nin tahmin değerleri hesaplanır.

A2. EKK dan hesaplanan $\hat{\boldsymbol{\theta}}$ değerlerine göre $\hat{\sigma}$ için, sağlam tahmin edici (3.8) eşitliği yardımıyla hesaplanır ve sonraki adımlarda sabit olarak kullanılır.

A3. Sağlam kriter fonksiyonuna bağlı olarak $\mathbf{W}$ ağırlık matrisinin köşegen elemanları, $w_{ii} = \frac{\psi(e_i^s)}{e_i^s}$ $i = 1, 2, \dots, n$ olarak belirlenir.

A4. $\boldsymbol{\theta}^0$ için uygun başlangıç değerleri verilir.

A5. $0 \leq \lambda_0 \leq 1$ olmak üzere bir λ_0 değerinden başlanarak

$$\boldsymbol{\theta}^i = \boldsymbol{\theta}^{i-1} + \lambda_0 \left(\left(\mathbf{F}'(\boldsymbol{\theta}^{i-1}) \mathbf{W} \mathbf{F}(\boldsymbol{\theta}^{i-1}) \right)^{-1} \mathbf{F}'(\boldsymbol{\theta}^{i-1}) \mathbf{W} \mathbf{e}^s \right)$$

$Q(\boldsymbol{\theta}^i) \leq Q(\boldsymbol{\theta}^{i-1})$ olacak biçimde seçilen bir λ ile,

$$\boldsymbol{\theta}^i = \boldsymbol{\theta}^{i-1} + \lambda \left(\left(\mathbf{F}'(\boldsymbol{\theta}^{i-1}) \mathbf{W} \mathbf{F}(\boldsymbol{\theta}^{i-1}) \right)^{-1} \mathbf{F}'(\boldsymbol{\theta}^{i-1}) \mathbf{W} \mathbf{e}^s \right)$$

eşitliğinden yeni parametre değerleri elde edilir.

A6. $\varepsilon > 0$ ve $\tau > 0$ keyfi sabitler olmak üzere (genellikle $\varepsilon = 10^{-5}$ ve $\tau = 10^{-3}$) yinelemelere son verilebilmesi için,

$$\|\boldsymbol{\theta}^i - \boldsymbol{\theta}^{i-1}\| < \varepsilon (\|\boldsymbol{\theta}^{i-1}\| + \tau) \quad \text{ve} \quad |Q(\boldsymbol{\theta}^i) - Q(\boldsymbol{\theta}^{i-1})| < \varepsilon (Q(\boldsymbol{\theta}^{i-1}) + \tau)$$

eşitsizliklerinin aynı anda sağlanması gerekmektedir. Eğer her ikisi aynı anda sağlanmazsa A5 adımına gidilir.

M tahmin edicisinde yer alan en yaygın kullanılan kriter $\rho(x)$ fonksiyonları; Huber, Ramsay, Andrews, Hampel, Rousseeuw ve Yohai alt kesimlerde incelenecektir.

3.3.1. Huber fonksiyonu

Huber'in M tahmin edicisi olarak da bilinen Huber fonksiyonu, hataların karesi ile hataların mutlak değerinin yani EKK ile LAD'ın birleşiminden hataların bir fonksiyonu olarak

$$\rho(x) = \begin{cases} \frac{x^2}{2} & , |x| \leq t \\ |x|t - \frac{1}{2}t^2 & , |x| > t \end{cases} \quad (3.12)$$

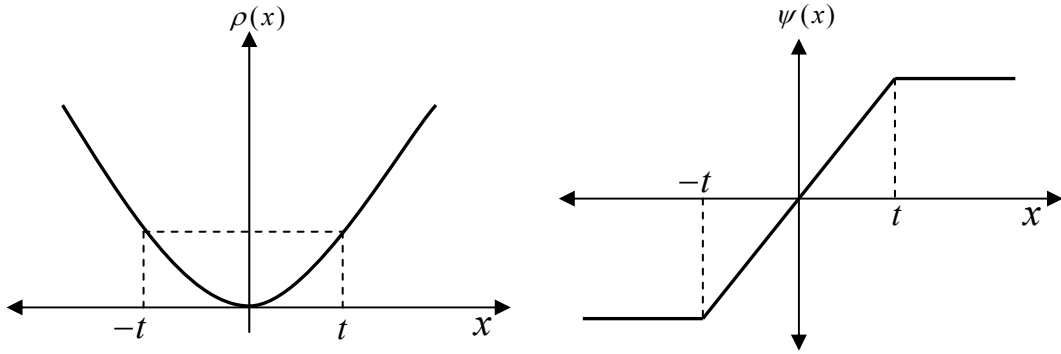
biçiminde elde edilmiştir (Huber 1964, 1973). Eşitlik (3.12) deki $\rho(x)$ fonksiyonunda t düzeltme terimi, hataların normal dağılımdan geldiği durumlarda $t \cong 2$ kullanılmaktadır. $\rho(x)$ ' in x 'e göre türevi olan $\psi(x)$ etki fonksiyonu,

$$\psi(x) = \begin{cases} x & , |x| \leq t \\ t \operatorname{sgn}(x) & , |x| > t \end{cases}$$

olarak elde edilir. Ağırlık fonksiyonu da etki fonksiyonunun x 'e oranıyla

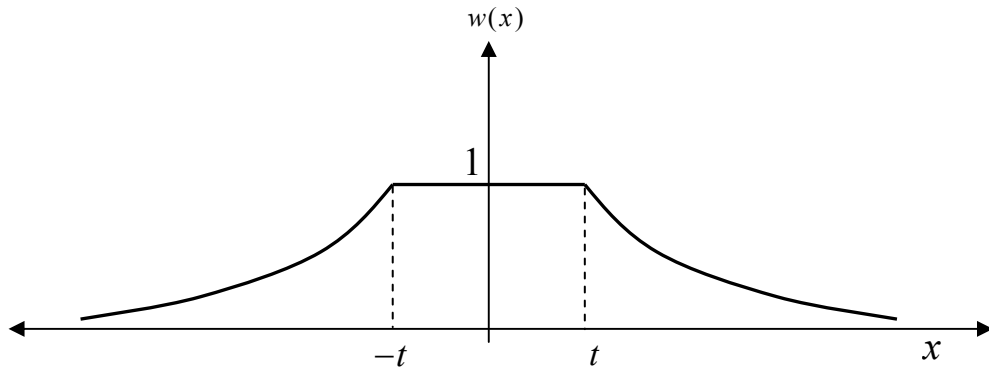
$$w(x) = \begin{cases} 1 & , |x| \leq t \\ \frac{t}{|x|} & , |x| > t \end{cases}$$

biçiminde bulunur.



Şekil 3.1. Huber'in $\rho(x)$ ve $\psi(x)$ fonksiyonlarının grafiği.

Huber fonksiyonunun $\rho(x)$ ve $\psi(x)$ fonksiyonlarının grafiği Şekil 3.1 de, $w(x)$ ağırlık fonksiyonu da Şekil 3.2 de verilmiştir.



Şekil 3.2. Huber'in $w(x)$ ağırlık fonksiyonu grafiği.

3.3.2. Ramsay fonksiyonu

Ramsay'ın $\rho(x)$ fonksiyonu,

$$\rho(x) = a^{-2} \left[1 - \exp(-a|x|)(1 + a|x|) \right] \quad , \quad |x| \leq \infty$$

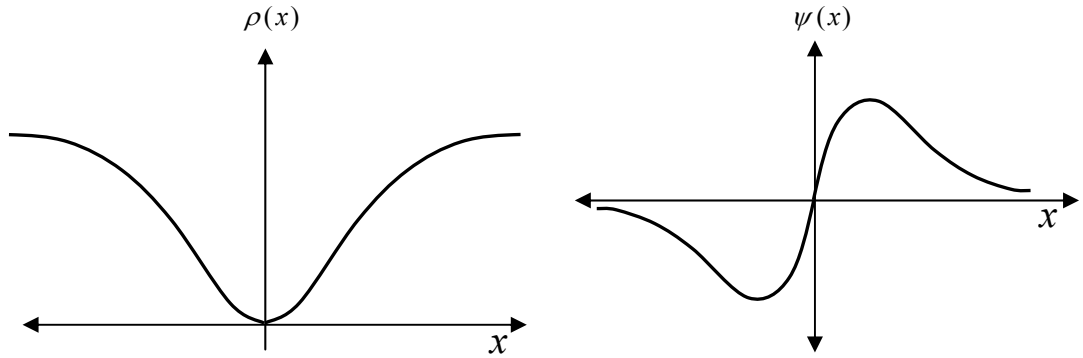
biçimindedir (Ramsay 1977). Ramsay'ın fonksiyonunda a düzeltme terimi genellikle $a = 0.3$ kullanılmaktadır. $\rho(x)$ fonksiyonunun x 'e göre türevi olan $\psi(x)$ etki fonksiyonu,

$$\psi(x) = x \exp(-a|x|) \quad , |x| \leq \infty$$

ve ağırlık fonksiyonu da etki fonksiyonuyla

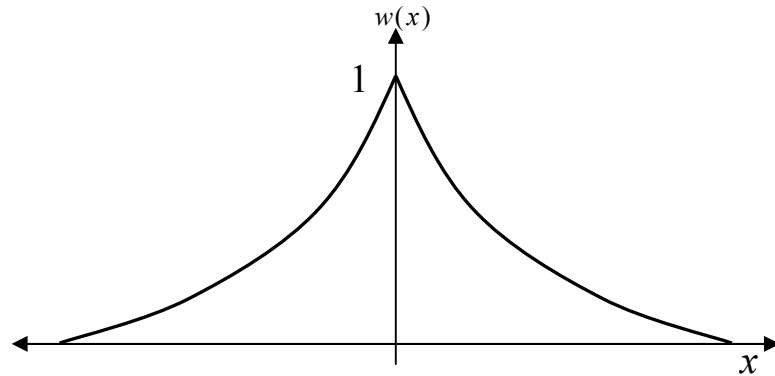
$$w(x) = \exp(-a|x|) \quad , |x| \leq \infty$$

olarak elde edilir.



Şekil 3.3. Ramsay'ın $\rho(x)$ ve $\psi(x)$ fonksiyonlarının grafiği.

Ramsay fonksiyonunun $\rho(x)$ ve $\psi(x)$ fonksiyonlarının grafiği Şekil 3.3 de, $w(x)$ ağırlık fonksiyonu da Şekil 3.4 de verilmiştir.



Şekil 3.4. Ramsay'ın $w(x)$ ağırlık fonksiyonu grafiği.

3.3.3. Andrews'in dalga fonksiyonu

Andrews'in $\rho(x)$ fonksiyonu,

$$\rho(x) = \begin{cases} a^2 [1 - \cos(x/a)] & , \quad |x| \leq a\pi \\ 2a^2 & , \quad |x| > a\pi \end{cases}$$

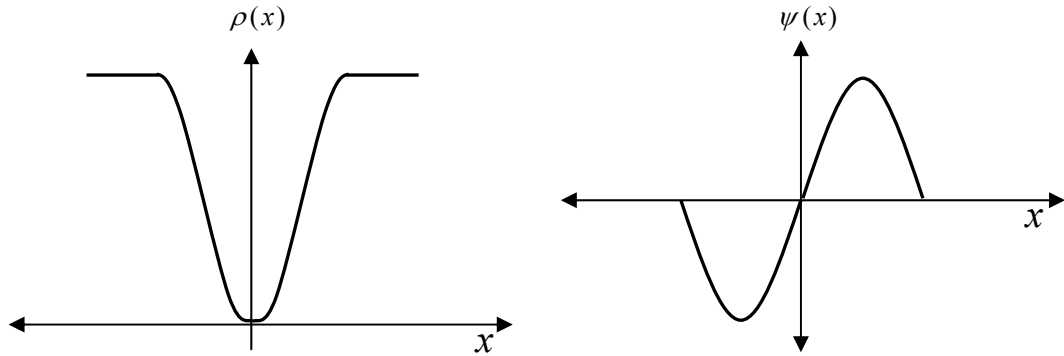
biçimindedir (Andrews ve ark. 1972). Andrews'in fonksiyonunda a düzeltme terimi için genellikle $a = 1.339$ kullanılmaktadır. $\rho(x)$ fonksiyonunun x 'e göre türevi olan $\psi(x)$ etki fonksiyonu,

$$\psi(x) = \begin{cases} a \sin(x/a) & , \quad |x| \leq a\pi \\ 0 & , \quad |x| > a\pi \end{cases}$$

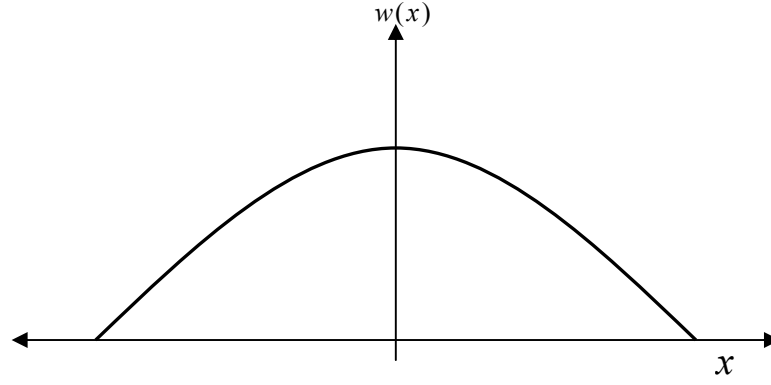
ve ağırlık fonksiyonu da etki fonksiyonunun x 'e oranıyla

$$w(x) = \begin{cases} \frac{\sin(x/a)}{x} & , \quad |x| \leq a\pi \\ 0 & , \quad |x| > a\pi \end{cases}$$

elde edilir. Andrews fonksiyonunun $\rho(x)$ ve $\psi(x)$ fonksiyonlarının grafiği Şekil 3.5' te, $w(x)$ ağırlık fonksiyonu da Şekil 3.6' te verilmiştir.



Şekil 3.5. Andrews'in $\rho(x)$ ve $\psi(x)$ fonksiyonlarının grafiği.



Şekil 3.6. Andrews'ın $w(x)$ ağırlık fonksiyonu grafiği.

3.3.4. Hampel fonksiyonu

Hampel'ın kayıp fonksiyonu olan $\rho(x)$ fonksiyonu,

$$\rho(x) = \begin{cases} x^2/2 & , \quad |x| \leq a \\ a|x| - \frac{1}{2}a^2 & , \quad a < |x| \leq b \\ \frac{a\left(c|x| - \frac{1}{2}x^2\right)}{c-b} - \frac{7a^2}{6} & , \quad b < |x| \leq c \\ a(b+c-a) & , \quad |x| > c \end{cases}$$

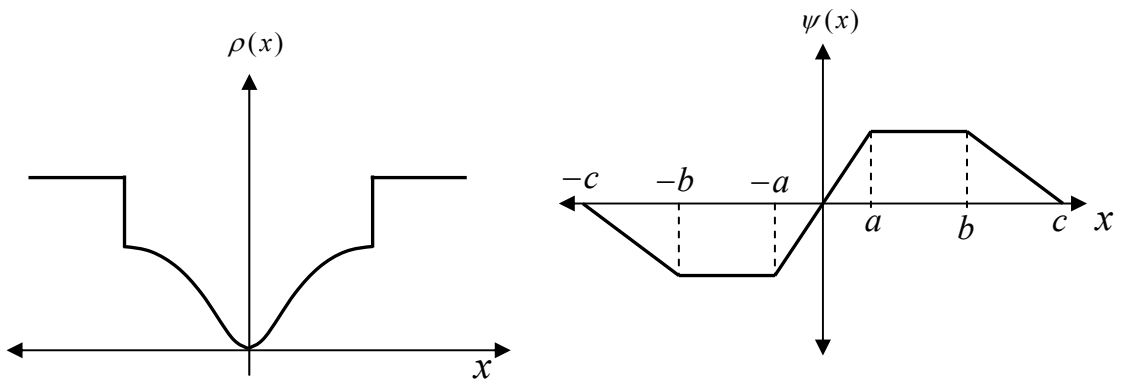
biçimindedir (Hampel 1971). Hampel'ın $\rho(x)$ fonksiyonunda a, b, c düzeltme terimleri için genellikle $a = 1.7$, $b = 3.4$, $c = 8.5$ kullanılmaktadır. $\rho(x)$ fonksiyonunun x 'e göre türevi olan $\psi(x)$ etki fonksiyonu,

$$\psi(x) = \begin{cases} x & , \quad |x| \leq a \\ a \operatorname{sgn}(x) & , \quad a < |x| \leq b \\ \frac{a \operatorname{sgn}(x)(c - |x|)}{c - b} & , \quad b < |x| \leq c \\ 0 & , \quad |x| > c \end{cases}$$

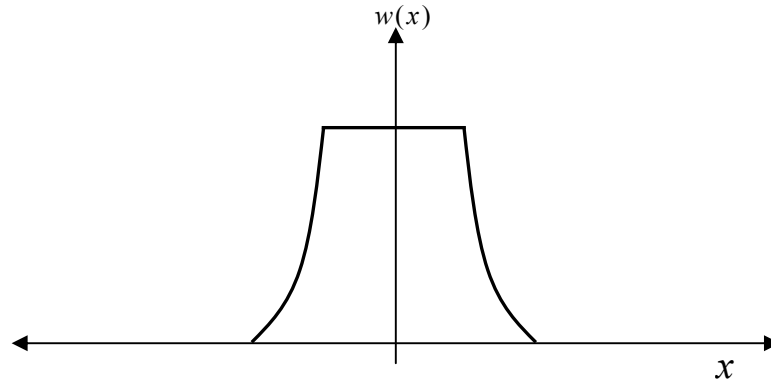
ve ağırlık fonksiyonu da etki fonksiyonuyla

$$w(x) = \begin{cases} 1 & , \quad |x| \leq a \\ \frac{a}{|x|} & , \quad a < |x| \leq b \\ \frac{a(c-|x|)}{|x|(c-b)} & , \quad b < |x| \leq c \\ 0 & , \quad |x| > c \end{cases}$$

olarak elde edilir. Hampel fonksiyonunun $\rho(x)$ ve $\psi(x)$ fonksiyonlarının grafiği Şekil 3.7' te, $w(x)$ ağırlık fonksiyonu da Şekil 3.8' te verilmiştir.

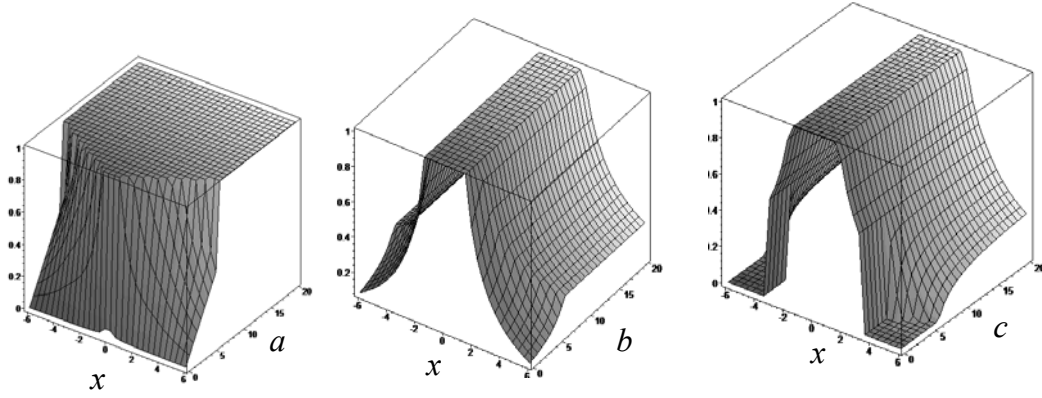


Şekil 3.7. Hampel'in $\rho(x)$ ve $\psi(x)$ fonksiyonlarının grafiği.



Şekil 3.8. Hampel'in $w(x)$ ağırlık fonksiyonu grafiği.

a, b ve c ' nin farklı seviyeleri için Hampel'in ağırlık fonksiyonunun grafikleri Şekil 3.9 da verilmiştir.



Şekil 3.9. Hampel'in a, b ve c 'nin farklı seviyeleri için $w(x)$ ağırlık fonksiyonlarının grafikleri.

3.3.5. Rousseeuw ve Yohai fonksiyonu

Rousseeuw ve Yohai'nin fonksiyonu olan $\rho(x)$ fonksiyonu,

$$\rho(x) = \begin{cases} \frac{x^2}{2} - \frac{x^4}{2c^2} + \frac{x^6}{6c^4} & , \quad |x| \leq c \\ \frac{c^2}{6} & , \quad |x| > c \end{cases}$$

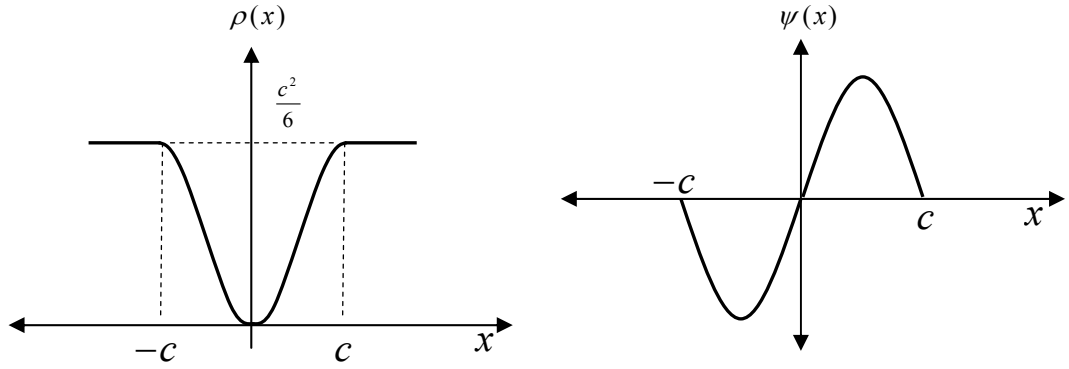
biçimindedir (Rousseeuw ve Yohai 1984). Rousseeuw ve Yohai'nin $\rho(x)$ fonksiyonunda c düzeltme terimi için genellikle $c = 1.547$ kullanılmaktadır. $\rho(x)$ fonksiyonunun x 'e göre türevi olan $\psi(x)$ etki fonksiyonu,

$$\psi(x) = \begin{cases} x - \frac{2x^3}{c^2} + \frac{x^5}{c^4} & , \quad |x| \leq c \\ \frac{c}{3} & , \quad |x| > c \end{cases}$$

ve ağırlık fonksiyonu da etki fonksiyonuyla

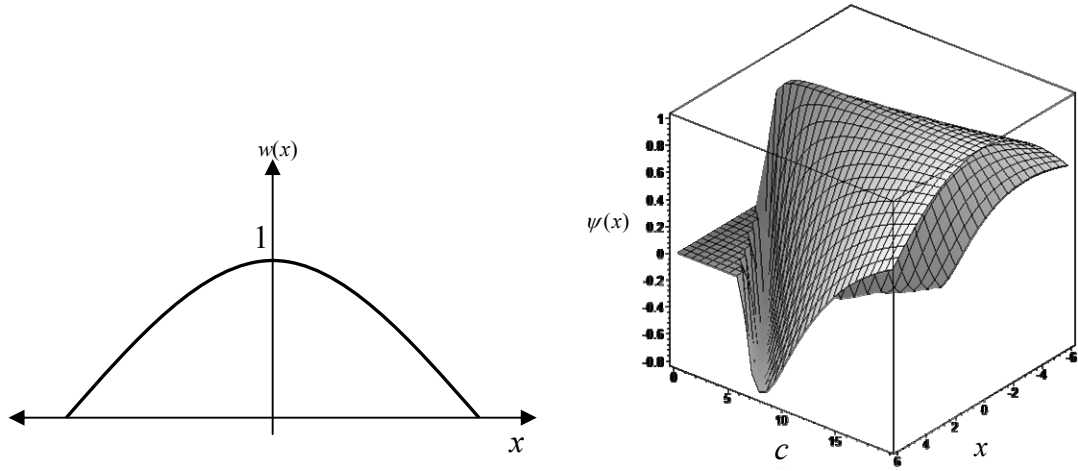
$$w(x) = \begin{cases} \left[1 - \left(\frac{x}{c} \right)^2 \right]^2 & , \quad |x| \leq c \\ 0 & , \quad |x| > c \end{cases}$$

olarak elde edilir.



Şekil 3.10. Rousseeuw ve Yohai' nin $\rho(x)$ ve $\psi(x)$ fonksiyonlarının grafiği.

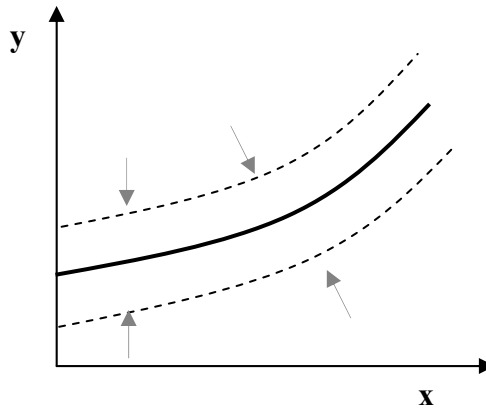
Rousseeuw ve Yohai'nin $\rho(x)$ ve $\psi(x)$ fonksiyonlarının grafiği Şekil 3.10 da, $w(x)$ ağırlık fonksiyonu da Şekil 3.11 de verilmiştir.



Şekil 3.11. Rousseeuw ve Yohai' nin c ' nin seviyelerine göre $w(x)$ ağırlık fonksiyonu grafiği.

3.4. Genelleştirilmiş M Tahmin Edicileri

M tahmin edicileri $\mathbf{X}$ uzayındaki aykırı değerlerin varlığından etkilenmektedir. Bunun nedeni artıklar, gözlemlere karşılık gelen tasarım noktası ile ilişkisiz olarak ağırlıklandırılır. Bir gözleme karşılık gelen artık değeri çok küçük ancak karşılık gelen gözlem uzaklığı matrisinin köşegen elemanı h_{ii} değeri çok büyük olabilir. Bu durumda M tahmin edicileri ile yüksek etki noktaları 1'ye yakın (hatta 1) ağırlığına sahip olacaktır. Böylece bu tip aykırı gözlemler tahmin ediciyi kötü etkileyecektir. M tahmin edicileri açıklanan değişkendeki aykırı değerlere karşı etkili ancak Şekil 3.12 de görüldüğü gibi açıklayıcı değişkendeki aykırı değerlere karşı hassas olan tahmin edicilerdir. M tahmin edicilerinin bu dezavantajı sınırlı etkili regresyon tahmin edicilerinin (bounded influential regression estimators) geliştirilmesini sağlamıştır. Bu tahmin edicilere Genelleştirilmiş M (GM) tahmin edicileri denir. GM tahmin edicilerinin M tahmin edicilerine göre avantajı $\mathbf{X}$ uzayındaki aykırı değerlere karşı hassas olmamalarıdır. GM tahmin edicilerinin dezavantajı ise bozulma noktalarının $1/p$ 'den büyük olmamalarıdır. Sonuç olarak, çok parametrelili bir model için bozulma noktası $1/n$ 'den daha iyi olmayacaktır. Ancak p parametre sayısı küçük ise bu tahmin ediciler çok iyi sonuçlar üretecektir.

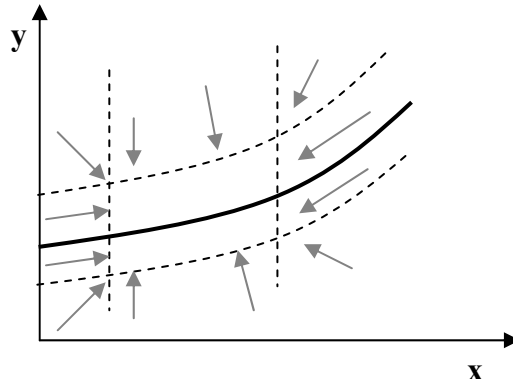


Şekil 3.12. Huber M tahmin edicisinin davranışı.

Doğrusal olmayan regresyonda genel olarak GM tahmin edicileri,

$$\min_{\theta} \sum_{i=1}^n \rho \left(\frac{y_i - f(\mathbf{x}_i; \theta)}{\hat{\sigma} u_i^k} \right) u_i^{1+k} \quad (3.13)$$

biçiminde tanımlanan (3.13) fonksiyonun en küçüklenmesiyle bulunur. Burada u_i 'ler $\mathbf{x}_i$ 'lere bağlı ağırlıkları göstermektedir. Dikkat edilirse GM tahmin edicileri M tahmin edicilerine $\mathbf{X}$ 'e bağlı ağırlıkların eklenmesi ile elde edilmektedir. Bu şekilde tahmin edicinin hem $\mathbf{X}$ hem de $\mathbf{Y}$ uzayında sınırlandırılması sağlanmıştır. Şekil 3.13 de Mallows M tahmin edicisinin davranışı gösterilmiştir.



Şekil 3.13 Mallows M tahmin edicisinin davranışı.

Sağlam istatistikte seçilen ρ fonksiyonunun türevinin yani ψ etki fonksiyonunun sınırlı olması gerekir. Eşitlik (3.13) ile tanımlanan en küçükleme probleminde k 'nın seçimi önemlidir. GM tahmin edicileri için iki türlü yaklaşım vardır. Bunlardan birincisi $k = 0$ olarak seçilmesiyle elde edilen

$$\min_{\theta} \sum_{i=1}^n \rho \left(\frac{y_i - f(\mathbf{x}_i; \theta)}{\hat{\sigma}} \right) u_i$$

biçimindeki Mallows yaklaşımıdır (Mallows 1975). Mallows tipi M tahmin edicisi olarak da bilinen bu yaklaşımda uygun Mallows ağırlığı, $\bar{\mathbf{x}} = (\bar{\mathbf{x}}_1, \bar{\mathbf{x}}_2, \dots, \bar{\mathbf{x}}_k)$ sağlam konum değerleri, $\mathbf{C}$ sağlam varyans-kovaryans matrisi ve c 'de sabit bir sayı olmak üzere,

$$MD_i = (\mathbf{x}_i - \bar{\mathbf{x}}) \mathbf{C}^{-1} (\mathbf{x}_i - \bar{\mathbf{x}})'$$

$$u_i = \min \left\{ 1, \sqrt{\frac{c}{MD_i}} \right\}$$

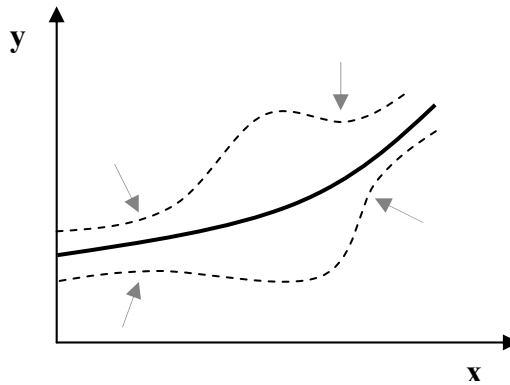
biçiminde verilir (Simpson ve ark., 1992). c sayısı $\alpha = 0.1$ veya $\alpha = 0.05$ olacak şekilde $1 - \alpha$ anlam seviyesine sahip, $p-1$ serbestlik dereceli χ^2 dağılımı olarak seçilir. İkinci yaklaşım ise eşitlik (3.13) de $k = 1$ olarak seçilmesiyle elde edilen

$$\min_{\theta} \sum_{i=1}^n \rho \left(\frac{y_i - f(\mathbf{x}_i; \theta)}{\hat{\sigma} u_i} \right) u_i^2$$

biçimindeki Schweppe yaklaşımıdır (Schweppe 1975). Schweppe tipi M tahmin edicisinin Mallows tipi M tahmin edicinin aksine Şekil 3.14 de gösterildiği gibi, etkin gözleme karşılık gelen artık değeri büyükse etkin gözlemin etkilerini azaltmaya çalışır. Uygun Mallows ağırlığı, $\bar{\mathbf{x}}$ sağlam konum değerleri vektörü ve $\mathbf{C}$ sağlam varyans-kovaryans matrisi olmak üzere,

$$u_i = \frac{1}{\sqrt{MD_i}}$$

biçiminde verilir (Hampel ve ark. 1986). Burada görülüyor ki ağırlıklar, $\mathbf{x}_i$ noktasının orijinden metrik uzaklığının bir fonksiyonudur.



Şekil 3.14 Schweppe M tahmin edicisinin davranışı

Metrik, bilinen sabit konum ile sağlam kovaryans matrisi olarak verilebilir. Mallows formu $\mathbf{X}$ uzayındaki konum ile artıklar için farklı sınırlar koyar ve $\mathbf{X}$ uzayındaki etkili gözlemler için ağırlıkları açıklanan değişkenin durumunu dikkate

almaksızın belirler. Schweppe formu ise, $\mathbf{x}_i$ 'lerin ağırlıklarına karşılık gelen artık değeri büyükse, gözlemlerin toplam etkisine sınır koyar. Buna göre, Schweppe formu bu tip durumlarda Mallows formuna göre daha iyi sonuçlar üretecektir denilebilir. Şekil 3.12 deki grafik Huber'in M tahmin edicisinin ne türlü sınırlandırıldığını göstermektedir. Bu grafikte tahmin edicinin $\mathbf{X}$ bağımsız değişken yönündeki aykırı değerlerin etkilerini sınırlamadığı görülür. Bu tahmin edici sadece y açıklanan değişken yönündeki aykırı değerlerin etkilerini sınırlamaktadır. Şekil 3.13 deki grafikte ise Mallows tipi M tahmin edicisinin hem açıklayan değişken yönündeki etkili gözlemlerin hem de açıklanan değişken yönündeki aykırı değerlerin etkilerini sınırladığı görülmektedir. Şekil 3.14 deki grafik de Schweppe tipi M tahmin edicisinin açıklayan değişken yönündeki etkili gözlemlere karşılık gelen artık büyükse, bu gözlemlerin etkilerini sınırlamaktadır.

ψ fonksiyonu ρ fonksiyonunun türev fonksiyonu, $\mathbf{F}(\boldsymbol{\theta})$ da parametrelere göre $n \times p$ lik türev matrisi, $\hat{\sigma}$ MAD değeri, $e_i^{su} = \frac{e_i}{\hat{\sigma} u_i^k}$, $i = 1, 2, \dots, n$ ve $\psi(\mathbf{e}^{su})$ de köşegen elemanları $\psi(e_i^{su})$ olan $n \times n$ lik köşegen matris olmak üzere (3.7) eşitliğinden,

$$\mathbf{F}'(\boldsymbol{\theta})\psi(\mathbf{e}^{su})\mathbf{u} = \mathbf{0} \quad (3.14)$$

şeklindeki amaç fonksiyonunu en küçükleyen tahmin edici GM tahmin edicisi olarak tanımlanır. Eşitlik (3.14) den işlemlere devam edildiğinde

$$\mathbf{F}'(\boldsymbol{\theta})\mathbf{W}\mathbf{u} = \mathbf{0} \quad (3.15)$$

formuna ulaşılır. Eşitlik (3.15) deki $\mathbf{W}$ $n \times n$ lik köşegen ağırlık matrisi olmak üzere, ρ için farklı kriterler hesaplanmıştır ve bunlar Kesim 3.3 de verilmiştir. Bundan sonraki hesaplama işlemleri M tahmin edicilerinin yineleme yöntemiyle yapılan adımları ile aynıdır.

3.5. S Tahmin Edicisi

S tahmin edicisi ölçek istatistiğinin en küçüklenmesine dayanır. Bu tahmin ediciyi ilk defa regresyonda Rousseeuw ve Yohai (1984) kullanmışlardır. S tahmin

edicisi EKK, LAD, LTS ve M tahmin edicilerini kapsar (Hampel ve ark. 1986). S tahmin edicisi $0 \leq b \leq 1$ olmak üzere,

$$s_M(e) = \inf \left\{ \hat{\sigma} > 0 : \frac{1}{n} \sum_{i=1}^n \rho \left(\frac{e_i}{\hat{\sigma}} \right) \leq b \right\} \quad (3.16)$$

denkleminin optimizasyon çözümünden elde edilir (Rousseeuw ve Yohai 1984, Croux ve ark. 1994). Yaygın olarak kullanılan algoritmalarda, n elemanlı bir veri kümesinin tüm mümkün m elemanlı alt kümelerine $\binom{n}{m}$ M yöntemi uygulanır.

Küçük veri kümeleri için S tahminlerinin kesin değerlerini hesaplamak bu şekilde mümkün olsa da, büyük veri kümelerinde mümkün olan tüm altkümelerin taranması ve M tahminin uygulanması işlem yükü açısından oldukça zor olacaktır. Bu durumda veri içerisinde bazı altkümelerin rasgele olarak çekilmesi ve amaç fonksiyonunun bu altkümelerde uygulanması düşünülebilir. Belirli kısıtlar altında veriden rasgele olarak çekilecek en az bir altkümenin istenilen sonucu verme olasılığının 1 (bir)'e yakın olduğu gösterilmiştir (Rousseeuw ve Leroy 1987). Buna göre $\binom{n}{m}$ 'nin çok büyük değerleri için bir veri kümesinden m elemanlı k tane altküme seçildiğinde, içerisinde 1 (bir) tane aykırı olmayan değer içeren, en az bir altküme rastlama olasılığını

$$P = 1 - \left[1 - (1 - \varepsilon)^m \right]^k \quad (3.17)$$

eşitliği ile tanımlanmıştır (Rousseeuw ve Leroy 1987). Eşitlik (3.17) da veri kümesinin bozulma oranı (tüm gözlemler içerisindeki aykırı gözlemlerin oranı) ε ile gösterilmektedir. Buradan arzu edilen alt örneklerdeki gözlem sayısı ve alt kümeler içerisinde aykırı değer içermeyen en az 1 (bir) kümenin olması olasılığına göre en az kaç kere alt örneklem çekilmesi gerektiği

$$k \geq \frac{\ln(1 - P)}{\ln(1 - (1 - \varepsilon)^m)} \quad (3.18)$$

ifadesiyle hesaplanır. Ancak gerçekte mevcut verinin ne derecede bozulmuş olduğunun bilinmesi zor bir problemdir. Literatürde bu problemin doğrusal regresyonda normallik varsayımı altında çözümü mevcuttur. Wu ve Lee (2006)'nin çalışmalarında doğrusal regresyonda LAD tahmin edicisine bağlı olarak olası en yüksek aykırı gözlem sayısını tespit etmişlerdir.

Her ne kadar (3.18) eşitliğinde mümkün k sayısı elde edilse de, büyük ε ve m değerlerine bağlı olarak hesaplanılan k sayısı çok yüksek olabilir. Örneğin $\varepsilon = 0.2$, $m = 50$ ve alt kümeler içerisinde aykırı değer içermeyen en az 1 (bir) kümenin olması olasılığı 0.99 olduğunda $k \geq 322659$ elde edilir. $m = 100$ olduğunda ise $k \geq 22.607.209.544$ tane alt örneklemden hesaplama yapmak gerekir. Bu sayılar çok yüksek olduğundan mevcut algoritmalar yetersiz kalmaktadır (Barrera ve Yohai 2006).

S tahmin edicisi kullanıldığında hesaplama süresinin fazla olması bu tahmin edicinin en büyük dezavantajıdır. Doğrusal regresyonda **Fast-S** ismini verdikleri yeni algoritmalarıyla Barrera ve Yohai (2006), bu süreyi bir parça kısaltabilmişlerdir. Buna göre doğrusal olmayan regresyondaki S tahmini şu şekilde hesaplanılabilir:

Doğrusal olmayan regresyonda S tahmin algoritması:

Adım1:

A1. Eşitlik (3.16) daki ölçek tahmini (3.8) denklemi ile hesaplanır.

A2. $\min_{\theta \in \mathbb{R}^p} \left\{ \sum_{i=1}^n \rho \left(\frac{e_i}{\hat{\sigma}} \right) \right\}$ çözümü için doğrusal olmayan regresyonda seçilen bir kriter fonksiyonu ile M tahmin edicisi hesaplanır.

A3. A2 den elde edilen artıklardan, Eşitlik (3.16) kullanılarak $\hat{\sigma}$ hesaplanır.

A4. Bulunan $\hat{\sigma}$ değeri ile $\min_{\theta \in \mathbb{R}^p} \left\{ \sum_{i=1}^n \rho \left(\frac{e_i}{\hat{\sigma}} \right) \right\}$ çözümünden elde edilen parametre tahmin değerleri, S tahmin edicisi için mümkün tahmin değerlerinden birisi olacaktır.

Adım2:

- B1.** Mevcut veri setinden p boyutlu k kere rasgele örneklem çekilerek, her bir örneklem için Adım1 deki işlemler uygulanır ve elde edilen tahmin değerleri θ_i , $i = 1, 2, \dots, k$ muhafaza edilir.
- B2.** B1 den elde edilen θ_i , $i = 1, 2, \dots, k$ değerlerine bağlı olarak M ölçek tahminleri hesaplanır. İçlerinden en küçük ölçek tahminini veren alt örneklem seçilir ve bu örneklemden M tahminleri hesaplanır. Bu tahmin değerleri S tahmin değerleri olacaktır.

4. AYKIRI GÖZLEMLERİN İNCELENMESİ

Farklılık gösteren gözlemlerin belirlenmesinde $\mathbf{H}$ gözlem uzaklığı matrisinin köşegen elemanları h_{ii} 'ler, Mahalanobis uzaklığı, Cook uzaklığı, DFFITS, DFBETAS, Standart artıklar ve Student türü artıkların ölçüleri kullanılabilir. Ancak bu istatistikler veride sadece bir tek aykırı değer olması durumunda etkilidirler. Veride birden çok aykırı değer olması durumunda bu gözlemlerin etkileri maskelenebilir yani açıklayıcı değişkenler bakımından aykırı değerler büyük h_{ii} veya büyük Mahalanobis uzaklığına sahip olmayabilirler. Adnan ve ark. (2003) bu durumda bu tip gözlemlere karşı duyarlı olan sağlam tahmin edicilerin kullanılmasının daha güvenilir sonuçlar üreteceğini belirtmişlerdir.

Sağlam tahmin edicilerden elde edilen sonuçlar üzerinden, farklılık gösteren gözlemlerin tespitinde kullanılan bu yöntemler tek aykırı gözlemin varlığında ve çoklu aykırı gözlemlerin varlığında olmak üzere iki kısımda incelenmiştir.

4.1. Regresyonda Tek Aykırı Gözlemin Belirlenmesi

Bir regresyon modelinde, verinin çoğunluğunun uyduğu yapıdan farklı bir şekilde oluşmuş veri olabilir. Bu veri kayıt hatası, üretim aşamasında bir bozukluk veya insanlarla ilişkili sebeplerden meydana gelebilir. Bozuk bir veri model kurma hatasına, yanlış parametre tahminlerine veya yanlış analiz sonuçlarına sebep olur (Li ve ark. 2004).

4.1.1. Gözlem uzaklığı matrisi

$\mathbf{X}$ matrisi, $n \times p$ boyutunda gözlemlenen açıklayıcı değişken matrisini, $\mathbf{Y}$ açıklanan değişken vektörünü ve $\boldsymbol{\varepsilon}$ hata vektörünü göstermek üzere doğrusal regresyon modeli

$$\mathbf{Y}_{n \times 1} = \mathbf{X}_{n \times p} \boldsymbol{\theta}_{p \times 1} + \boldsymbol{\varepsilon}_{n \times 1}$$

olarak ifade edilmektedir. Doğrusal regresyonda parametrelerin EKK tahmin edicisi

$$\hat{\boldsymbol{\theta}}_{EKK} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{Y}$$

biçimindedir ve kestirim vektörü $\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\theta}}$ olmak üzere doğrusal regresyonda $\mathbf{H}$ matrisi

$$\mathbf{H}_{n \times n} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}' \quad (4.1)$$

olup

$$\hat{\mathbf{Y}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{Y} = \mathbf{H}\mathbf{Y}$$

şeklindedir. Denklem (4.1) de $\mathbf{H}$ matrisi simetrik ($\mathbf{H}' = \mathbf{H}$), eşgüçlü (idempotent) ($\mathbf{H}'\mathbf{H} = \mathbf{H}$), esas köşegen üzerinde bulunan elemanlar $0 < h_{ii} < 1$ ve parametre sayısı p olmak üzere matrisin izi (trace) $\text{iz}(\mathbf{H}) = p$ ve $\text{rank}(\mathbf{H}) = p$ dır.

Doğrusal olmayan regresyonda (2.5) denklemindeki $\mathbf{F}$ matrisi, model fonksiyonunun türev (jakobiyen) matrisi olmak üzere $\mathbf{H}$ matrisi

$$\mathbf{H} = \mathbf{F}(\mathbf{F}'\mathbf{F})^{-1} \mathbf{F}' \quad (4.2)$$

biçimindedir (Cook ve ark. 1986). $\mathbf{H}$ matrisinin esas köşegen elemanları, açıklayıcı değişken uzayında yer alan aykırı gözlemlerin teşhisinde kullanılmaktadır. $\mathbf{H}$ matrisinin esas köşegen üzerindeki i . elemanı h_{ii} olmak üzere $h_{ii} > 2p/n$ olan gözlemler aykırı gözlem olarak tespit edilebilir (Hoaglin ve Welsch 1978). Aykırı gözlemlerin tespitinde başka bir sınır $h_{ii} > 3p/n$ olarak ele alınabilmektedir (Rousseeuw ve Leroy 1987).

Gözlem veri setinde birden çok aykırı değer olması durumunda bazı aykırı değerler diğerleri tarafından gizlenebileceği için h_{ii} 'de bu gözlemler normal değermiş gibi görülebilir.

4.1.2. Cook uzaklığı

Doğrusal olmayan regresyonda $\hat{\boldsymbol{\theta}}$ tüm gözlemlerden elde edilecek tahmin edici, $\hat{\boldsymbol{\theta}}_{(-i)}$ i . gözlemin veriden çıkartılmasıyla elde edilen tahmin edici ve $\mathbf{F}$ matrisi tüm verilerden elde edilen tahminde yineleme sonrası hesaplanan Jakobiyen matrisini ifade etmek üzere

$$D_i = \frac{(\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_{(-i)})' \mathbf{F}'\mathbf{F}(\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_{(-i)})}{ps^2}, \quad i = 1, 2, \dots, n$$

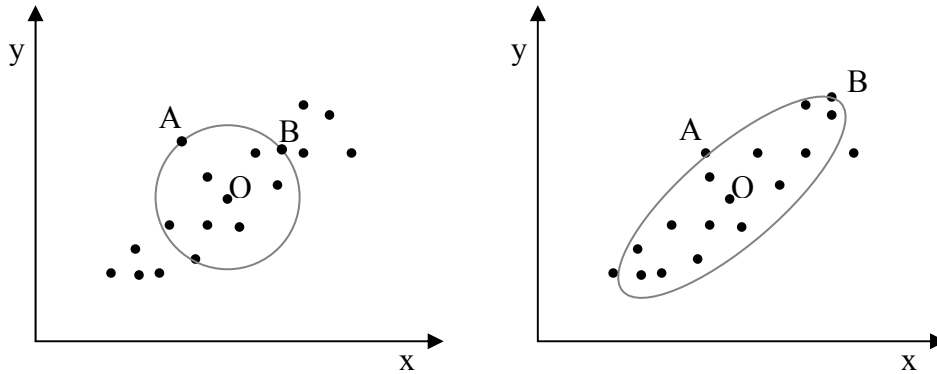
olarak hesaplanır (Cook 1979). D_i istatistiği p ve $n-p$ serbestlik dereceli F dağılımına $D_i \sim F_{p,n-p}$ uymaktadır. α anlam seviyesinde $D_i > F_{1-\alpha;p,n-p}$ olan gözlemler aykırı gözlem olarak tespit edilebilir.

4.1.3. Mahalanobis uzaklığı

Mahalanobis uzaklığı, $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_k]$ veri matrisindeki herhangi bir gözlemin, verinin merkezinden (ortalama vektöründen) uzaklığının bir ölçüsüdür. Bu veri setinde $\mathbf{C}$ varyans-kovaryans matrisi, $\bar{\mathbf{x}} = (\bar{x}_1, \bar{x}_2, \dots, \bar{x}_k)$ konum tahmin edicisinin ortalama değerlerinin vektörü, $\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{ik})$, $i = 1, 2, \dots, n$ de gözlem değerleri olmak üzere, Mahalanobis uzaklığı

$$MD_i = \sqrt{(\mathbf{x}_i - \bar{\mathbf{x}}) \mathbf{C}^{-1} (\mathbf{x}_i - \bar{\mathbf{x}})'} \quad , i = 1, 2, \dots, n \quad (4.3)$$

olarak hesaplanmaktadır (Mahalanobis 1936). İki gözlem değerinin merkeze olan uzaklıkları, Mahalanobis uzaklığı ve Öklid uzaklığı ile Şekil 4.1 de gösterilmiştir. Burada y açıklanan değişkeni, x de açıklayan değişkeni ifade etmektedir.



Şekil 4.1. Öklid ve Mahalanobis Uzaklığında iki noktanın merkeze olan uzaklığı.

Şekil 4.1.'in soldaki grafiği Öklid uzaklığını, sağdaki grafiği de Mahalanobis uzaklığını göstermektedir. Her iki grafikte de A ve B noktalarının bulundukları bölgelerinde, O merkezine olan uzaklıkları eşit çıkmaktadır.

$\mathbf{H}$ matrisinin köşegen elemanları ile Mahalanobis uzaklığı birebir ilişkilidir. Bu ilişki

$$MD_i = (n-1) \left(h_{ii} - \frac{1}{n} \right), \quad i = 1, 2, \dots, n$$

şeklindedir.

Mahalanobis uzaklığının karesi $MD_i^2 \sim \chi_p^2$ dağılımına uymaktadır. Mahalanobis uzaklığı tek başına aykırı gözlemlerin tespiti için kullanılsa da, diğer aykırı gözlem tespit algoritmalarının hemen hemen hepsinde ara hesap olarak kullanılmaktadır (Rousseeuw ve Leroy 1987, Serbert ve ark. 1998).

4.1.4. DFBETAS

$\hat{\theta}$ tüm gözlemlerden hesaplanacak olan tahmin ediciyi, $\hat{\theta}_{(-i)}$ i. gözlemin veriden çıkartılmasıyla hesaplanacak tahmin ediciyi ve $\mathbf{F}_{(-i)}$ matrisi i. gözlemin çıkarılmasıyla elde edilen tahmindeki Jakobiye matrisi olmak üzere

$$DFBETAS_i = \frac{\|\hat{\theta} - \hat{\theta}_{(-i)}\|}{s_{(-1)} \sqrt{c_{ii}}}, \quad i = 1, 2, \dots, n$$

dir. c_{ii} değeri $(\mathbf{F}'\mathbf{F})^{-1}$ matrisinin esas köşegenindeki i. değerdir. $DFBETAS_i > 2/\sqrt{n}$ olan gözlemler aykırı gözlem olarak tespit edilebilir (Billingsley ve ark. 1980).

4.1.5. DFFITS

$\hat{\mathbf{Y}}$ tüm gözlemlerden elde edilen öngörü değerlerini, $\hat{\mathbf{Y}}_{(-i)}$ i. gözlemin veriden çıkartılmasıyla elde edilen öngörü değerlerini göstermek üzere DFFITS değerleri

$$DFFITS_i = \frac{\hat{y}_i - \hat{y}_{i(-i)}}{s_{(-i)} \sqrt{h_{ii}}} = \frac{(\hat{\theta} - \hat{\theta}_{(-i)})' \mathbf{F}_{(-i)}' \mathbf{F}_{(-i)} (\hat{\theta} - \hat{\theta}_{(-i)})}{p s_{(-i)}^2}, \quad i = 1, 2, \dots, n$$

olarak hesaplanmaktadır. $DFFITS_i > 2\sqrt{p/n}$ olan gözlemler aykırı gözlem olarak tespit edilebilir (Rousseeuw ve Leroy 1987).

4.1.6. Standart ve Student türü artıklar

Regresyon modelinde ε_i hata teriminin tahmin edicisi $e_i = y_i - \hat{y}_i$ eşitliği ile tanımlanan artık terimidir (Cook ve Tsai 1985). Hata teriminin varyans-kovaryans matrisinin $V(\varepsilon) = \sigma^2 \mathbf{I}$ tahmin edicisi olarak artıkların varyans-kovaryans matrisi $V(\mathbf{e})$ kullanılır. σ^2 'nin tahmin edicisi ise

$$s^2 = \frac{\sum_{i=1}^n e_i^2}{n-p}$$

dır. Buna göre standart artıklar

$$t_i = \frac{e_i}{s\sqrt{1-h_{ii}}} \quad , i=1,2,\dots,n$$

olarak tanımlanır. $s_{(-i)}$, i. gözlem ihmal edilerek hesaplanan regresyonun hata teriminin varyansının bir tahmin edicisi olmak üzere,

$$t_{(-i)} = \frac{e_i}{s_{(-i)}\sqrt{1-h_{ii}}} \quad , i=1,2,\dots,n$$

ifadesine student türü artıklar denir. $t_{(-i)} \sim t_{n-p}$ dağılımından gelmektedir.

4.1.7. COVRATIO

COVRATIO değeri ile i. gözlemin veri setinden çıkartılmasıyla hesaplanan kovaryans matrisinin determinantının, tüm gözlemler üzerinden hesaplanan kovaryans matrisinin determinantına oranı olarak hesaplanır. Bu istatistik genel varyansın bir gözlemin çıkartılmasıyla nasıl etkilendiğini gösterir. Bu istatistiğin hesabı

$$COVRATIO_i = \frac{\det\left(s_{(-i)}^2 (\mathbf{F}'_{(-i)} \mathbf{F}_{(-i)})^{-1}\right)}{\det\left(s^2 (\mathbf{F}' \mathbf{F})^{-1}\right)} \quad , i=1,2,\dots,n$$

biçiminde olup, $1-3p/n < COVRATIO_i < 1+3p/n$ aralığında olmayan gözlemler aykırı gözlem olarak tespit edilebilir (Billingsley ve ark. 1980).

4.2. Regresyonda Birden Fazla Aykırı Gözlemlerin Belirlenmesi

Regresyondan hesaplanan tahmin değerleri ve artıklardan yapılan algoritmalara dolaysız yöntemler denilmektedir (Rousseeuw ve Leroy 1987). Bu yöntemlerin hemen hemen hepsinde tek bir aykırı gözlemin tespitinde bu gözlemin veri setinden çıkartılması temeline dayanmaktadır. Ancak bazı durumlarda bir aykırı gözlem bir başka gözlem değerini çeşitli biçimlerde etkisi altına alabilmektedir. Bu aykırı gözlem ya diğer bir normal gözlemi aykırı gözlemmiş gibi gösterebilir (sürüklenme-swamping) ya da aykırı bir gözlemi aykırı gözlem değilmiş gibi (gizleme-masking) gösterebilir (Campbell ve ark. 1998). Bu iki etkiden dolayı aykırı gözlem tespitinde, sağlam bir tahmin edici kullanılarak elde edilen artıklar üzerinden geliştirilmiş dolaylı yöntemler kullanılmaktadır. En çok kullanılan bu dolaylı yöntemler alt kesimlerde incelenecektir.

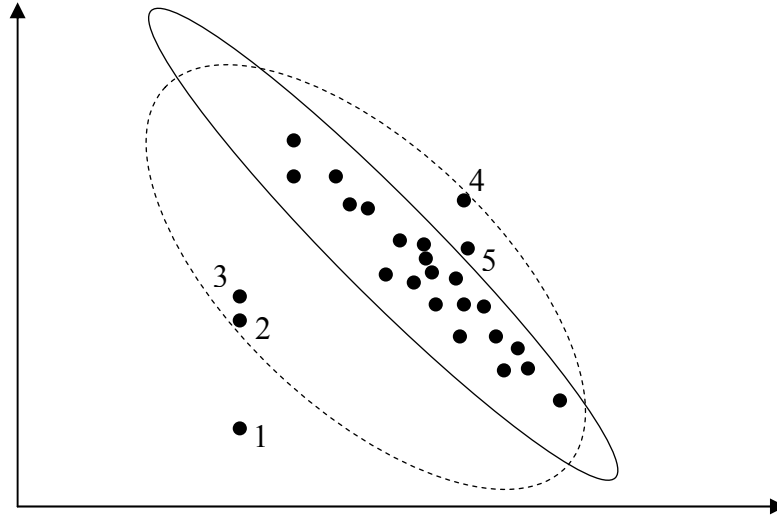
4.2.1. En küçük hacimli elips yöntemi

En küçük hacimli elips (Minimum Volume Ellipsoid - MVE) yöntemi Rousseeuw'un 1985'te tanımladığı, aykırı gözlemlerin tespitindeki problemleri ortadan kaldıracak sağlam bir yöntemdir. Bu yöntemde klasik Mahalanobis uzaklığı kullanılmaktadır ancak (4.3) eşitliğindeki Mahalanobis uzaklığı, konum ve ölçek tahminleri sağlam istatistiksel yöntemler kullanılarak hesaplanmaktadır. Mahalanobis uzaklığının sağlam yöntemlerle hesaplanmış haline sağlam uzaklıklar (Robust Distance) denilmektedir (Rousseeuw ve Zomeren 1990).

$\hat{\theta}$, p parametrelili regresyon modelinde gözlemlerden en az k tanesini içeren elips merkezinin tahmin edicisi olsun. Burada k , $\left\lfloor \frac{n}{2} + 1 \right\rfloor$ veya $\left\lfloor \frac{n+p}{2} \right\rfloor$ olarak alınır. $\left\lfloor . \right\rfloor$ gösterimi kendinden küçük yada eşit en büyük tamsayı olduğunu ifade etmektedir. C varyans-kovaryans matrisi, k bağımsız değişken sayısı, $\bar{\mathbf{x}} = (\bar{x}_1, \bar{x}_2, \dots, \bar{x}_k)$ konum tahmin edicisinin ortalama değerlerinin vektörü ve $\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{ik})$, $i = 1, 2, \dots, n$ de gözlem değerleri olmak üzere,

$$\# \left\{ i : (\mathbf{x}_i - \bar{\mathbf{x}}) C^{-1} (\mathbf{x}_i - \bar{\mathbf{x}})' \geq k \right\}$$

ifadesinden C ' nin en küçüklenmesi, En küçük hacimli elips'in tahmin edicisidir. En küçük hacimli elips ile güven elipsinin karşılaştırılması Şekil 4.2. de gösterilmektedir.



Şekil 4.2. En küçük hacimli elips ile güven elipsi görünümü.

Şekil 4.2. deki kesikli çizgili elips, α anlam seviyesinde ortalama ve varyans-kovaryans matrisi ile elde edilmiş $1-\alpha$ lık güven elipsini göstermektedir. Güven elipsi beş gözlem değerinden aşırı etkilenmektedir. Sürekli çizgiyle gösterilen elips de medyan ve sağlam varyans-kovaryans tahmin edicisiyle hesaplanan En küçük hacimli elipstir. En küçük hacimli elips' in bu beş gözlem değerinden etkilenmediği görülmektedir.

4.2.2. En küçük kovaryans determinanı yöntemi

En küçük kovaryans determinanı (Minimum Covariance Determinant - MCD) yöntemi, gözlemlerin varyans-kovaryans matrisinin determinantını en küçükleyecek m tane gözlemin $\left(\left\lceil \frac{n+p}{2} \right\rceil \leq m \leq n\right)$ tespit edilmesiyle elde edilir.

Genelde m değeri $m = \left\lceil \frac{n+p}{2} \right\rceil$ olarak alınır. Burada n tüm gözlemlerin sayısını, p de regresyon modelindeki parametre sayısını göstermektedir.

En küçük kovaryans determinantı yönteminde amaç fonksiyonu, m elemanlı en küçük determinantlı varyans-kovaryans matrisine sahip alt kümeyi araştırdığından bütün m elemanlı $\binom{n}{m}$ tane alt kümede varyans-kovaryans determinantının hesaplanması gerekmektedir. Gözlem sayısının çok büyük olması durumunda bu hesaplamalar imkansız hale gelebilmektedir. Bu durumdan kurtulabilmek için Rousseeuw ve Driessen (1999)' in çalışmalarında **Fast-MCD** ismini verdikleri algoritma kullanılmaktadır. Bu algoritma iki farklı durum için düzenlenmiştir. Birinci durum da gözlem sayısı en fazla 600 olanlarla ilgilidir. Birinci durumun algoritması şu şekilde çalışmaktadır:

A1. m değeri $\left\lceil \frac{n+p}{2} \right\rceil \leq m \leq n$ olacak şekilde alınır. Burada n tüm gözlemlerin sayısını, p de regresyon modelindeki parametre sayısını göstermektedir.

A2. Bu adım iki aşamadan oluşmaktadır. İlk aşamada gözlemler seçilir ve ikinci aşamada da ortalama ile varyans-kovaryans matrislerinden Mahalanobis uzaklıkları hesaplanır. Bu uzaklıklardan en küçük olan m tanesinden elde edilen varyans kovaryans matrisinin determinantı hesaplanır. Bu adımlar 500 kere tekrar edilir.

A2.a. Veri setinden rasgele p elemanlı bir alt küme çekilir. Çekilen bu alt kümenin varyans-kovaryans matrisinin determinantı sıfır çıkacak olursa, determinantı sıfır olmayana kadar p elemanlı yeni bir alt küme çekilir. Bu alt kümeden ortalama ile varyans-kovaryans matrisleri bulunur. Bu alt kümeden elde edilen değerlerden tüm gözlemlerin Mahalanobis uzaklıkları hesaplanır ve bu uzaklıklara göre m tanesi alınır. Bu gözlemlerden elde edilen varyans-kovaryans matrisinin determinantı $\det(\mathbf{S}_1^{(i)})$ olarak tutulur.

A2.b. Adım A2.a dan elde edilen m tane gözlem üzerinden ortalama ve varyans-kovaryans matrisleri hesaplandıktan sonra tüm gözlemlerin kareli Mahalanobis uzaklıkları hesaplanır ve bu uzaklıklara göre sıralanarak ilk m tanesi alınır. Bu

yeni tespit edilen m tane gözlemden varyans-kovaryans matrisinin determinantı hesaplanır ve $\det(\mathbf{S}_2^{(i)})$ olarak tutulur. Tekrar m tane gözlemden Mahalanobis uzaklıkları hesaplanır ve bu uzaklıklar küçükten büyüğe sıralanarak ilk m tane gözlemden ortalama vektörü ile varyans-kovaryans matrisinin determinantı hesaplanır ve bu da $\det(\mathbf{S}_3^{(i)})$ olarak tutulur. Her zaman $\det(\mathbf{S}_1^{(i)}) \geq \det(\mathbf{S}_2^{(i)}) \geq \det(\mathbf{S}_3^{(i)}) \geq \dots$ olarak belli bir değere yakınsamaktadır. Bu kısım **C-adım** olarak isimlendirilmektedir.

A3. A2' nin 500 tekrarından sonra tüm bulunan $\det(\mathbf{S}_3^{(i)})$, $i = 1, 2, \dots, 500$ determinantları içerisinde en küçük olan m gözlemlili alt kümelerden 10 tanesi tespit edilir.

A4. A3 den elde edilen 10 en iyi alt kümeler içerisinde varyans-kovaryans determinantını en küçükleyen veriden hesaplanan ortalama vektörü ile varyans-kovaryans matrisi en küçük kovaryans determinantının tahmini olacaktır.

4.2.3. Aykırı değer sayısı yöntemi

Wu ve Lee (2006) normallik varsayımı altında doğrusal regresyonda LAD çözümü ile bir örneklemden olası en yüksek aykırı gözlem sayısını tespit etmişlerdir. *Aykırı değer sayısı* (ADS) adı verilen bu yöntemle göre, mevcut örneklemden aykırı değerlerin sayısı en fazla m tane ve $N(\mu + \lambda\sigma, \sigma^2)$, $\lambda > 0$ dağılımına sahip, $n - m$ gözlem de $N(\mu, \sigma^2)$ dağılımına sahip olduğu varsayılır. Standart normal dağılımdan,

$$F(x; \mu, \sigma) = \Phi\left(\frac{x - \mu}{\sigma}\right), \quad -\infty < x, \mu < \infty, \quad \sigma > 0$$

ters dönüşüm yapıldığında

$$U(x) = \phi^{-1} [F(x; \mu, \sigma)] = \frac{x - \mu}{\sigma}$$

elde edilir. Agostino ve Stephens (1986)' nin çalışmalarında, deneysel normal dağılım fonksiyonu,

$$\hat{F}(\mathbf{X}; \mu, \sigma) \cong \frac{i - c}{n - 2c + 1}, i = 1, 2, \dots, n, \quad 0 \leq c < 1 \quad (4.4)$$

eşitliği ile hesaplanılabilmektedir. Örneklemdeki değerlerin sıralı hali $X_{(1)}, X_{(2)}, \dots, X_{(n)}$ olarak gösterilsin. Eşitlik (4.4) deki c değeri sırasıyla $c = 0, 0.3, \frac{1}{3}, 0.5$ alınarak ve $i = 1, 2, \dots, n - m$ olmak üzere

$$\hat{U}_c(x_i) = \frac{x_{(i)} - \mu}{\sigma}, \quad -\infty < \mu < \infty, \quad \sigma > 0$$

hesaplanır, sonrasında $i = n - m + 1, n - m + 2, \dots, n$ olmak üzere

$$\hat{U}_c(x_i) = \frac{x_{(i)} - (\mu + \lambda\sigma)}{\sigma}, \quad -\infty < \mu < \infty, \quad \sigma, \lambda > 0$$

hesaplanır. $m = 0$ için,

$$Q_1(\theta) = \left[\sum_{i=1}^n \rho \left(\hat{U}_c(x_i) - \frac{x_i - \mu}{\sigma} \right) \right] \quad (4.5)$$

ve $m = 1, 2, \dots, \left\lfloor \frac{n}{2} \right\rfloor - 1$ için,

$$Q_2(\theta) = \sum_{i=1}^{n-m} \rho \left(\hat{U}_c(x_i) - \frac{x_i - \mu}{\sigma} \right) + \sum_{i=n-m+1}^n \rho \left(\hat{U}_c(x_i) - \frac{x_i - (\mu + \lambda\sigma)}{\sigma} \right) \quad (4.6)$$

amaç fonksiyonlarını en küçükleyen μ , σ ve λ değerleri tahmin edildikten sonra $m = 0$ için,

$$\text{SMAD} = \frac{1}{n-2} \left[\sum_{i=1}^n \rho \left(\hat{U}_c(x_i) - \frac{x_i - \mu}{\sigma} \right) \right] \quad (4.7)$$

ve $m = 1, 2, \dots, \left\lfloor \frac{n}{2} \right\rfloor - 1$ için,

$$\text{SMAD} = \frac{1}{n-3} \left[\sum_{i=1}^{n-m} \rho \left(\hat{U}_c(x_i) - \frac{x_i - \mu}{\sigma} \right) + \sum_{i=n-m+1}^n \rho \left(\hat{U}_c(x_i) - \frac{x_i - (\mu + \lambda\sigma)}{\sigma} \right) \right] \quad (4.8)$$

değerleri hesaplanır. Tüm c' ler içerisinde en küçük SMAD değerine sahip m değeri olası en yüksek aykırı değer sayısını vermektedir (Wu ve Lee 2006).

Bu yöntem doğrusal olmayan regresyona uyarlanıldığında, artıklar üzerinden $e_1, e_2, \dots, e_n$ değerlendirilecektir. Bu artıklarda olabilecek aykırı değerlerin sayısı en fazla m tane olsun.

$$f(\boldsymbol{\theta}) = \frac{e_i - \mu}{\sigma} \quad i = 1, 2, \dots, n-m \quad -\infty < \mu < \infty, \quad \sigma > 0$$

$$g(\boldsymbol{\theta}) = \frac{e_i - (\mu + \lambda\sigma)}{\sigma} \quad i = n-m+1, n-m+2, \dots, n \quad -\infty < \mu < \infty, \quad \sigma, \lambda > 0$$

Biçiminde tanımlanan $f(\boldsymbol{\theta})$ ve $g(\boldsymbol{\theta})$ fonksiyonlarında $\boldsymbol{\theta}' = (\lambda, \mu, \sigma)$ olmak üzere

Eşitlik (4.5) ve Eşitlik (4.6) da tanımlanan amaç fonksiyonları en küçüklenir. $f(\boldsymbol{\theta})$

fonksiyonunun türev matrisi,

$$\mathbf{F}(\boldsymbol{\theta}) = \begin{bmatrix} \frac{\partial}{\partial \theta_1} f(e_1, \boldsymbol{\theta}) & \frac{\partial}{\partial \theta_2} f(e_1, \boldsymbol{\theta}) & \frac{\partial}{\partial \theta_3} f(e_1, \boldsymbol{\theta}) \\ \frac{\partial}{\partial \theta_1} f(e_2, \boldsymbol{\theta}) & \frac{\partial}{\partial \theta_2} f(e_2, \boldsymbol{\theta}) & \frac{\partial}{\partial \theta_3} f(e_2, \boldsymbol{\theta}) \\ \vdots & \vdots & \vdots \\ \frac{\partial}{\partial \theta_1} f(e_{n-m}, \boldsymbol{\theta}) & \frac{\partial}{\partial \theta_2} f(e_{n-m}, \boldsymbol{\theta}) & \frac{\partial}{\partial \theta_3} f(e_{n-m}, \boldsymbol{\theta}) \end{bmatrix}_{(n-m) \times 3}$$

ve $g(\boldsymbol{\theta})$ fonksiyonunun türev matrisi,

$$\mathbf{G}(\boldsymbol{\theta}) = \begin{bmatrix} \frac{\partial}{\partial \theta_1} g(e_{n-m+1}, \boldsymbol{\theta}) & \frac{\partial}{\partial \theta_2} g(e_{n-m+1}, \boldsymbol{\theta}) & \frac{\partial}{\partial \theta_3} g(e_{n-m+1}, \boldsymbol{\theta}) \\ \frac{\partial}{\partial \theta_1} g(e_{n-m+2}, \boldsymbol{\theta}) & \frac{\partial}{\partial \theta_2} g(e_{n-m+2}, \boldsymbol{\theta}) & \frac{\partial}{\partial \theta_3} g(e_{n-m+2}, \boldsymbol{\theta}) \\ \vdots & \vdots & \vdots \\ \frac{\partial}{\partial \theta_1} g(e_n, \boldsymbol{\theta}) & \frac{\partial}{\partial \theta_2} g(e_n, \boldsymbol{\theta}) & \frac{\partial}{\partial \theta_3} g(e_n, \boldsymbol{\theta}) \end{bmatrix}_{m \times 3}$$

biçimindedir. ψ fonksiyonu da ρ fonksiyonunun türev fonksiyonu ve $\mathbf{e} = e_{(1)}, e_{(2)}, \dots, e_{(n)}$ olmak üzere

$$\frac{\partial}{\partial \boldsymbol{\theta}'} Q(\boldsymbol{\theta}) \Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} = \mathbf{0}$$

$$\begin{aligned} \frac{\partial}{\partial \boldsymbol{\theta}'} Q(\boldsymbol{\theta}) &= \psi \left(\hat{U}_c(\mathbf{e}) - \mathbf{f}(\boldsymbol{\theta}) \right)' \mathbf{F}(\boldsymbol{\theta}) + \psi \left(\hat{U}_c(\mathbf{e}) - \mathbf{g}(\boldsymbol{\theta}) \right)' \mathbf{G}(\boldsymbol{\theta}) \\ &= \mathbf{F}(\boldsymbol{\theta})' \psi \left(\hat{U}_c(\mathbf{e}) - \mathbf{f}(\boldsymbol{\theta}) \right) + \mathbf{G}(\boldsymbol{\theta})' \psi \left(\hat{U}_c(\mathbf{e}) - \mathbf{g}(\boldsymbol{\theta}) \right) \end{aligned} \quad (4.9)$$

normal denklemi elde edilir. Kesim 3.3 deki dönüştürme (4.9) denkleminde uygulandığında

$$\mathbf{F}(\boldsymbol{\theta})' \left(\frac{\psi \left(\frac{\hat{U}_c(\mathbf{e}) - \mathbf{f}(\boldsymbol{\theta})}{s} \right)}{\left(\frac{\hat{U}_c(\mathbf{e}) - \mathbf{f}(\boldsymbol{\theta})}{s} \right)} \right) \left(\frac{\hat{U}_c(\mathbf{e}) - \mathbf{f}(\boldsymbol{\theta})}{s} \right) + \mathbf{G}(\boldsymbol{\theta})' \left(\frac{\psi \left(\frac{\hat{U}_c(\mathbf{e}) - \mathbf{g}(\boldsymbol{\theta})}{s} \right)}{\left(\frac{\hat{U}_c(\mathbf{e}) - \mathbf{g}(\boldsymbol{\theta})}{s} \right)} \right) \left(\frac{\hat{U}_c(\mathbf{e}) - \mathbf{g}(\boldsymbol{\theta})}{s} \right) = \mathbf{0}$$

olmak üzere

$$\mathbf{W}_1 = \left(\psi \left(\frac{\hat{U}_c(\mathbf{e}) - \mathbf{f}(\boldsymbol{\theta})}{s} \right) \right) / \left(\frac{\hat{U}_c(\mathbf{e}) - \mathbf{f}(\boldsymbol{\theta})}{s} \right)$$

$$\mathbf{W}_2 = \left(\psi \left(\frac{\hat{U}_c(\mathbf{e}) - \mathbf{g}(\boldsymbol{\theta})}{s} \right) \right) / \left(\frac{\hat{U}_c(\mathbf{e}) - \mathbf{g}(\boldsymbol{\theta})}{s} \right)$$

köşegen ağırlık matrisleri elde edilir. $\boldsymbol{\theta}^0$ ($\boldsymbol{\theta}^0 \in \Theta$) 'ın bir komşuluğunda $\mathbf{f}(\boldsymbol{\theta})$ ve $\mathbf{g}(\boldsymbol{\theta})$ fonksiyonları sürekli ve en az iki kere türevlenebilir bir fonksiyon olmak üzere, $\mathbf{f}(\boldsymbol{\theta})$ ve $\mathbf{g}(\boldsymbol{\theta})$ fonksiyonlarının $\boldsymbol{\theta}^0$ noktasındaki Taylor serisine açılımı,

$$\mathbf{f}(\boldsymbol{\theta}) \cong \mathbf{f}(\boldsymbol{\theta}^0) + \mathbf{F}(\boldsymbol{\theta}^0)(\boldsymbol{\theta} - \boldsymbol{\theta}^0)$$

$$\mathbf{g}(\boldsymbol{\theta}) \cong \mathbf{g}(\boldsymbol{\theta}^0) + \mathbf{G}(\boldsymbol{\theta}^0)(\boldsymbol{\theta} - \boldsymbol{\theta}^0)$$

olacak biçimde (4.9) denkleminde yerine yazıldığında

$$\begin{aligned} & \mathbf{F}(\boldsymbol{\theta}^0)' \mathbf{W}_1 \left(\frac{\hat{U}_c(\mathbf{e}) - \mathbf{f}(\boldsymbol{\theta}^0) + \mathbf{F}(\boldsymbol{\theta}^0)(\boldsymbol{\theta} - \boldsymbol{\theta}^0)}{s} \right) \\ & + \mathbf{G}(\boldsymbol{\theta}^0)' \mathbf{W}_2 \left(\frac{\hat{U}_c(\mathbf{e}) - \mathbf{g}(\boldsymbol{\theta}^0) + \mathbf{G}(\boldsymbol{\theta}^0)(\boldsymbol{\theta} - \boldsymbol{\theta}^0)}{s} \right) = \mathbf{0} \end{aligned}$$

elde edilir. İşlemlere devam edilirse,

$$\begin{aligned} & \mathbf{F}(\boldsymbol{\theta}^0)' \mathbf{W}_1 \left(\frac{\hat{U}_c(\mathbf{e}) - \mathbf{f}(\boldsymbol{\theta}^0)}{s} \right) + \mathbf{F}(\boldsymbol{\theta}^0)' \mathbf{W}_1 \left(\frac{\mathbf{F}(\boldsymbol{\theta}^0)(\boldsymbol{\theta} - \boldsymbol{\theta}^0)}{s} \right) \\ & + \mathbf{G}(\boldsymbol{\theta}^0)' \mathbf{W}_2 \left(\frac{\hat{U}_c(\mathbf{e}) - \mathbf{g}(\boldsymbol{\theta}^0)}{s} \right) + \mathbf{G}(\boldsymbol{\theta}^0)' \mathbf{W}_2 \left(\frac{\mathbf{G}(\boldsymbol{\theta}^0)(\boldsymbol{\theta} - \boldsymbol{\theta}^0)}{s} \right) = \mathbf{0} \end{aligned}$$

olup düzenleme yapılırsa,

$$\begin{aligned} & \mathbf{F}(\boldsymbol{\theta}^0)' \mathbf{W}_1 (\hat{U}_c(\mathbf{e}) - \mathbf{f}(\boldsymbol{\theta}^0)) + \mathbf{F}(\boldsymbol{\theta}^0)' \mathbf{W}_1 \mathbf{F}(\boldsymbol{\theta}^0)(\boldsymbol{\theta} - \boldsymbol{\theta}^0) \\ & + \mathbf{G}(\boldsymbol{\theta}^0)' \mathbf{W}_2 (\hat{U}_c(\mathbf{e}) - \mathbf{g}(\boldsymbol{\theta}^0)) + \mathbf{G}(\boldsymbol{\theta}^0)' \mathbf{W}_2 \mathbf{G}(\boldsymbol{\theta}^0)(\boldsymbol{\theta} - \boldsymbol{\theta}^0) = \mathbf{0} \\ & \left(\mathbf{F}(\boldsymbol{\theta}^0)' \mathbf{W}_1 \mathbf{F}(\boldsymbol{\theta}^0) + \mathbf{G}(\boldsymbol{\theta}^0)' \mathbf{W}_2 \mathbf{G}(\boldsymbol{\theta}^0) \right) (\boldsymbol{\theta} - \boldsymbol{\theta}^0) \\ & = - \left(\mathbf{F}(\boldsymbol{\theta}^0)' \mathbf{W}_1 (\hat{U}_c(\mathbf{e}) - \mathbf{f}(\boldsymbol{\theta}^0)) + \mathbf{G}(\boldsymbol{\theta}^0)' \mathbf{W}_2 (\hat{U}_c(\mathbf{e}) - \mathbf{g}(\boldsymbol{\theta}^0)) \right) \end{aligned}$$

dır. Buradan

$$\boldsymbol{\theta} - \boldsymbol{\theta}^0 = - \left(\mathbf{F}(\boldsymbol{\theta}^0)' \mathbf{W}_1 \mathbf{F}(\boldsymbol{\theta}^0) + \mathbf{G}(\boldsymbol{\theta}^0)' \mathbf{W}_2 \mathbf{G}(\boldsymbol{\theta}^0) \right)^{-1} \\ \left(\mathbf{F}(\boldsymbol{\theta}^0)' \mathbf{W}_1 (\hat{U}_c(\mathbf{e}) - \mathbf{f}(\boldsymbol{\theta}^0)) + \mathbf{G}(\boldsymbol{\theta}^0)' \mathbf{W}_2 (\hat{U}_c(\mathbf{e}) - \mathbf{g}(\boldsymbol{\theta}^0)) \right)$$

olacaktır. Dolayısıyla

$$\boldsymbol{\theta} = \boldsymbol{\theta}^0 - \left(\mathbf{F}(\boldsymbol{\theta}^0)' \mathbf{W}_1 \mathbf{F}(\boldsymbol{\theta}^0) + \mathbf{G}(\boldsymbol{\theta}^0)' \mathbf{W}_2 \mathbf{G}(\boldsymbol{\theta}^0) \right)^{-1} \\ \left(\mathbf{F}(\boldsymbol{\theta}^0)' \mathbf{W}_1 (\hat{U}_c(\mathbf{e}) - \mathbf{f}(\boldsymbol{\theta}^0)) + \mathbf{G}(\boldsymbol{\theta}^0)' \mathbf{W}_2 (\hat{U}_c(\mathbf{e}) - \mathbf{g}(\boldsymbol{\theta}^0)) \right)$$

dır. Burada

$$\mathbf{d}^0 = \left(\mathbf{F}(\boldsymbol{\theta}^0)' \mathbf{W}_1 \mathbf{F}(\boldsymbol{\theta}^0) + \mathbf{G}(\boldsymbol{\theta}^0)' \mathbf{W}_2 \mathbf{G}(\boldsymbol{\theta}^0) \right)^{-1} \\ \left(\mathbf{F}(\boldsymbol{\theta}^0)' \mathbf{W}_1 (\hat{U}_c(\mathbf{e}) - \mathbf{f}(\boldsymbol{\theta}^0)) + \mathbf{G}(\boldsymbol{\theta}^0)' \mathbf{W}_2 (\hat{U}_c(\mathbf{e}) - \mathbf{g}(\boldsymbol{\theta}^0)) \right)$$

gösterimi ile

$$\boldsymbol{\theta} = \boldsymbol{\theta}^0 - \mathbf{d}^0$$

yazılır. $\boldsymbol{\theta}^0$ bilinmediği için elde edilen $\boldsymbol{\theta}$ 'nın bir anlamı yoktur. Ancak, $\boldsymbol{\theta}^0$ yerine bir $\boldsymbol{\theta}_0$ başlangıç değerinden başlayıp $\mathbf{d}^0$ vektörü yönünde $\lambda_0 \mathbf{d}^0$ ($0 \leq \lambda_0 \leq 1$) olmak üzere sabit bir λ_0 değerinden başlanarak, $Q(\boldsymbol{\theta}^i) \leq Q(\boldsymbol{\theta}^{i-1})$ olacak biçimde seçilen bir λ ile,

$$\boldsymbol{\theta}^i = \boldsymbol{\theta}^{i-1} - \lambda \mathbf{d}^{i-1} \quad , \quad i = 1, 2, \dots$$

elde edilir. Daha sonra ardışık işlem adımlarına son verilebilmesi için durdurma kuralı $\varepsilon > 0$ ve $\tau > 0$ keyfi sabitler olmak üzere (genellikle $\varepsilon = 10^{-5}$ ve $\tau = 10^{-3}$ olarak önerilir),

$$\|\boldsymbol{\theta}^i - \boldsymbol{\theta}^{i-1}\| < \varepsilon (\|\boldsymbol{\theta}^{i-1}\| + \tau)$$

ve

$$|Q(\boldsymbol{\theta}^i) - Q(\boldsymbol{\theta}^{i-1})| < \varepsilon (Q(\boldsymbol{\theta}^{i-1}) + \tau)$$

eşitsizlikleri aynı anda sağlanana kadar yinelemeye devam edilir. Doğrusal olmayan regresyonda aykırı değer sayısının bulunabilmesi için yinelemeli çözüm algoritması şöyledir;

A1. Uygun başlangıç değerleri alınarak M yöntemi ile θ değerleri hesaplanır ve e artıklar vektörü bulunur.

A2. σ için, A1 adımında hesaplanan θ değerlerine göre sağlam tahmin edici

$$\hat{\sigma} = \frac{\text{medyan}\{|e_i - e_{0.5}|\}}{0.6745}, \quad i = 1, 2, \dots, n$$

biçiminde hesaplanır ve sonraki adımlarda sabit olarak kullanılır.

A3. Sırasıyla $c = 0, 0.3, \frac{1}{3}, 0.5$ alınır.

A4. Sırasıyla $m = 0, 1, \dots, \left\lceil \frac{n}{2} \right\rceil$ alınır.

A5. θ^0 için başlangıç değerleri verilir.

A6. i. iterasyonda ağırlıklar

$$\mathbf{W}_1^i = \frac{\psi\left(\frac{\hat{U}_c(\mathbf{e}) - \mathbf{f}(\theta^{i-1})}{s}\right)}{\left(\frac{\hat{U}_c(\mathbf{e}) - \mathbf{f}(\theta^{i-1})}{s}\right)} \quad \text{ve} \quad \mathbf{W}_2^i = \frac{\psi\left(\frac{\hat{U}_c(\mathbf{e}) - \mathbf{g}(\theta^{i-1})}{s}\right)}{\left(\frac{\hat{U}_c(\mathbf{e}) - \mathbf{g}(\theta^{i-1})}{s}\right)}$$

olacak biçimde sağlam kriter fonksiyonu belirlenir.

$$\mathbf{d}^{i-1} = \left(\mathbf{F}(\theta^{i-1})' \mathbf{W}_1^i \mathbf{F}(\theta^{i-1}) + \mathbf{G}(\theta^{i-1})' \mathbf{W}_2^i \mathbf{G}(\theta^{i-1}) \right)^{-1} \\ \left(\mathbf{F}(\theta^{i-1})' \mathbf{W}_1^i (\hat{U}_c(\mathbf{e}) - \mathbf{f}(\theta^{i-1})) + \mathbf{G}(\theta^{i-1})' \mathbf{W}_2^i (\hat{U}_c(\mathbf{e}) - \mathbf{g}(\theta^{i-1})) \right)$$

değeri hesaplandıktan sonra $0 \leq \lambda_0 \leq 1$ olmak üzere bir λ_0 değerinden

başlanarak

$$\theta^i = \theta^{i-1} - \lambda_0 \mathbf{d}^{i-1}$$

$Q(\theta^i) \leq Q(\theta^{i-1})$ olacak biçimde seçilen bir λ ile,

$$\boldsymbol{\theta}^i = \boldsymbol{\theta}^{i-1} - \lambda \mathbf{d}^{i-1}$$

yeni parametre değerleri elde edilir.

A7. $\varepsilon > 0$ ve $\tau > 0$ keyfi sabitler olmak üzere (genellikle $\varepsilon = 10^{-5}$ ve $\tau = 10^{-3}$ alınabilir),

$$\|\boldsymbol{\theta}^i - \boldsymbol{\theta}^{i-1}\| < \varepsilon (\|\boldsymbol{\theta}^{i-1}\| + \tau) \quad \text{ve} \quad |Q(\boldsymbol{\theta}^i) - Q(\boldsymbol{\theta}^{i-1})| < \varepsilon (Q(\boldsymbol{\theta}^{i-1}) + \tau)$$

Eşitsizlikler aynı anda sağlanıncaya kadar A6 adımına gidilir.

A8. $m = 0$ için,

$$\text{SMAD} = \frac{1}{n-2} \left[\sum_{i=1}^n \rho \left(\hat{U}_c(\mathbf{e}_i) - \frac{\mathbf{e}_i - \boldsymbol{\mu}}{\sigma} \right) \right]$$

ve $m = 1, 2, \dots, \left\lfloor \frac{n}{2} \right\rfloor$ için,

$$\text{SMAD} = \frac{1}{n-3} \left[\sum_{i=1}^{n-m} \rho \left(\hat{U}_c(\mathbf{e}_i) - \frac{\mathbf{e}_i - \boldsymbol{\mu}}{\sigma} \right) + \sum_{i=n-m+1}^n \rho \left(\hat{U}_c(\mathbf{e}_i) - \frac{\mathbf{e}_i - (\boldsymbol{\mu} + \lambda \sigma)}{\sigma} \right) \right]$$

değerleri hesaplanır.

A9. $m = \left\lfloor \frac{n}{2} \right\rfloor$ değerine ulaşıncaya kadar A5 adımına gidilir.

A10. $c = 0.5$ değerine ulaşıncaya kadar A4 adımına gidilir.

A11. SMAD değerlerinden en küçüğüne sahip olan m değeri, olası en yüksek aykırı gözlem sayısını verecektir.

4.2.4. Sağlam uzaklık yöntemi

Doğrusal olmayan regresyonda modelin eğriselliğinden dolayı çoklu aykırı gözlem tespitinde MVE ve MCD yöntemlerinin elipsoidal bölge kullanmaları uygun

düşmemektedir. Diğer yandan ADS yönteminde normallik varsayımının olması, sadece bu dağılım altında hesaplanabileceğinden dolayı genele yönelik çözüm oluşturmaz. Bu çalışmada Rousseeuw ve Zomeren (1990)'ın çalışmaları dikkate alınmıştır. Çoklu aykırı gözlem tespiti için Rousseeuw ve Zomeren (1990)'ın çalışmalarında, MVE yönteminde kullanılan konum ve ölçek parametrelerinin sağlam tahmin ediciler ile hesaplanması fikrini ortaya koymuşlardır. Buna göre doğrusal olmayan regresyonda bir tahmin edici ile elde edilen artıklar üzerinden, 4.3 eşitliğindeki klasik mahalanobis uzaklığı yerine,

$$RD_i = \frac{e_i - e_{0.5}}{\hat{\sigma}} \quad (4.10)$$

kullanılması daha sağlam sonuçların verilmesi açısından önemli olduğu belirtilmektedir (Rousseeuw ve Zomeren 1990). Eşitlik 4.10'a **Sağlam uzaklık** (Robust Distance - RD) denilmektedir. Mahalanobis uzaklığının karesi ki-kare dağılımına gitmekte idi, benzer biçimde sağlam uzaklıkta da $RD_i^2 \sim \chi_{p,1-\alpha}^2$ olur. $RD_i^2 > \chi_{p,1-\alpha}^2$ olan gözlemler aykırı değerler olarak kabul edilir (Rousseeuw ve Zomeren 1990).

5. BOZULMA NOKTASI VE ETKİ FONKSİYONU

Hipotez modellerinde, hata vektörünün normal dağıldığı varsayılır. Böyle bir regresyon modelinde model yeterliliği için ya modele yeni parametreler eklemek ya da mevcut parametrelerden azaltma yapmak veya farklı yeni modeller ortaya koyarak model yeterliliği incelenebilir. Tüm bunlara hata terimi incelenerek karar verilmektedir. Gözlemlerin parametre tahminleri üzerine etkilerini incelemek için de hata terimleri ele alınmak durumundadır. Ayrıca, modelin doğrusal olup olmadığı, parametre tahmin edicilerinin yanlı olup olmadığı vb. gibi problemlerde yine hata terimi ve hata teriminin türevleri ele alınarak karar verilmektedir (Cook ve ark. 1986).

5.1. Bozulma Noktası

Bir tahmin edicinin bozulma noktası, veri kümesindeki gözlemlerin yer değiştirmesinde tahmin edicinin etkilenmesi sonucu hata teriminin davranışı olarak ve bir tahmin edicinin sağlamlık ölçüsü olarak tanımlanabilir. Hampel (1971), Hodges (1967) in sağlamlık düşüncesini genelleştirerek ilk defa bozulma noktası kavramını ortaya koymuştur. Donoho (1982) ve Donoho ve Huber (1983) de bu kavramın sonlu örneklem için olanını ele almışlardır. Donoho ve Huber (1983)'ün çalışmaları hala güncelliğini korumaktadır ve sağlamlık kavramının geçtiği makalelerde referans olarak gösterilmektedir (Genton ve Lucas 2003, Giloni ve Simonoff 2005, Cho ve ark. 2009).

Doğrusal olmayan modellerde yeniden parametrelendirmede bozulma noktasının değişmezlik (invariant) özelliğinin olmamasından dolayı literatürde hep doğrusal regresyonda bozulma noktası tanımlanmıştır. Bozulma noktasının orijinal tanımı doğrusal olmayan problemler için çok özeldir. Her bir regresyon modeli için ele alınan tahmin edicinin ayrı bozulma noktası vardır.

Bozulma noktası $0 \leq \varepsilon^* \leq 1$ aralığında değerler almaktadır. Sağlam tahmin edicilerde bozulma noktasının $1/2$ olması arzu edilir. Şayet bozulma noktası $1/2$ den büyük olursa, mevcut gözlem değerlerinin yarısından fazlasının tahmin değerleri için hesaba alınmadığını gösterir.

Bozulma noktaları, deęiřtirmeli (replacement) bozulma noktası ve eklemeli (addition) bozulma noktası olmak üzere iki grupta incelenecektir.

5.1.1. Deęiřtirmeli bozulma noktası

$\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n)'$ ve $\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{ik})$, $i = 1, 2, \dots, n$ örneklem deęerlerinden oluşan gözlem veri seti olsun. $\mathbf{X}_m$, $\mathbf{X}$ deki herhangi m tane deęerin $(\mathbf{X}, \mathfrak{R})$ uzayındaki herhangi deęerler ile yer deęiřtirmesiyle elde edilen veri kümelerinin cümlesini ($m = 1, 2, \dots, n$) gösterebilir. $\hat{\theta}$ ifadesi de ilgilenilen tahmin edici olsun. Bu bilgilere göre sırasıyla sonlu örneklem ve sınırlı sonlu örneklem deęiřtirmeli bozulma noktaları tanıtılacaktır.

5.1.1.1. Sonlu örneklem deęiřtirmeli bozulma noktası

$\mathbf{X}$ de $\hat{\theta}$ tahmin edicisinin sonlu örneklem deęiřtirmeli bozulma noktası,

$$\varepsilon^*(\hat{\theta}, \mathbf{X}) = \min \left\{ \frac{m}{n} : \sup_{\mathbf{x}_m} \|\hat{\theta}(\mathbf{X}_m) - \hat{\theta}(\mathbf{X})\| = \infty \right\} \quad (5.1)$$

olarak tanımlanır (Donoho ve Huber 1983).

Konum parametresinin tahmin edicisi aritmetik ortalama olduęunda, sadece bir gözlemin deęiřtirilerek sonsuzluęa ulařtırılması, elde edilecek yeni tahmin deęeri ile orijinal verilerden elde edilen tahmin deęerinin sonsuza gittięi görülür. Bu da (5.1) eřitlięindeki $\sup_{\mathbf{x}_1} \|\hat{\theta}(\mathbf{X}_1) - \hat{\theta}(\mathbf{X})\| = \infty$ bozuluma sebebiyet vereceęinden $m = 1$

olur ve aritmetik ortalamanın bozulma noktası $\varepsilon^*(\hat{\theta}, \mathbf{X}) = 1/n$ olur. n deęeri yeteri kadar büyürse bozulma noktası da 0'a yakınsıyacaktır. Medyan için mevcut gözlem deęerlerinin ancak yarısı sonsuz bozuluma uğratılırsa yeni elde edilen medyan deęeri de sonsuz bozuluma uğrayacaktır. Bu nedenle medyanın bozulma noktası $1/2$ ye yakınsar.

5.1.1.2. Sınırlı sonlu örneklem değıştirmeli bozulma noktası

İlgilenilen regresyon model fonksiyonu $f(\mathbf{x}, \boldsymbol{\theta})$ olsun. Her bir gözleme ilişkin ayrı ayrı $\sup_{\boldsymbol{\theta}} f(\mathbf{x}, \boldsymbol{\theta})$ ve $\inf_{\boldsymbol{\theta}} f(\mathbf{x}, \boldsymbol{\theta})$ değerleriyle, alt ve üst bozulma değerleri bulunarak bu değerlerden en küçük olan değer o gözlemin bozulma noktasını oluşturur. Tüm gözlemlerin bozulma noktalarının infimum değeri de tahmin edicinin sınırlı sonlu örneklem değıştirmeli bozulma noktasıdır (Stromberg ve Ruppert 1992).

Bozulma fonksiyonları kullanılarak,

$$\begin{aligned} \varepsilon_+(\mathbf{x}_i, f, \hat{\boldsymbol{\theta}}, \mathbf{X}) \\ = \begin{cases} \min_{0 \leq m \leq n} \left\{ \frac{m}{n}; \sup_{\mathbf{x}_m} f(\mathbf{x}_i, \hat{\boldsymbol{\theta}}(\mathbf{X}_m)) = \sup_{\boldsymbol{\theta}} f(\mathbf{x}_i, \boldsymbol{\theta}) \right\} & , \sup_{\boldsymbol{\theta}} f(\mathbf{x}_i, \boldsymbol{\theta}) > f(\mathbf{x}_i, \hat{\boldsymbol{\theta}}) \\ 1 & , \text{dy} \end{cases} \end{aligned}$$

ve

$$\begin{aligned} \varepsilon_-(\mathbf{x}_i, f, \hat{\boldsymbol{\theta}}, \mathbf{X}) \\ = \begin{cases} \min_{0 \leq m \leq n} \left\{ \frac{m}{n}; \inf_{\mathbf{x}_m} f(\mathbf{x}_i, \hat{\boldsymbol{\theta}}(\mathbf{X}_m)) = \inf_{\boldsymbol{\theta}} f(\mathbf{x}_i, \boldsymbol{\theta}) \right\} & , \inf_{\boldsymbol{\theta}} f(\mathbf{x}_i, \boldsymbol{\theta}) < f(\mathbf{x}_i, \hat{\boldsymbol{\theta}}) \\ 1 & , \text{dy} \end{cases} \end{aligned}$$

olup tüm gözlemlere ilişkin

$$\varepsilon(\mathbf{x}_i, f, \hat{\boldsymbol{\theta}}, \mathbf{X}) = \min \left\{ \varepsilon_-(\mathbf{x}_i, f, \hat{\boldsymbol{\theta}}, \mathbf{X}), \varepsilon_+(\mathbf{x}_i, f, \hat{\boldsymbol{\theta}}, \mathbf{X}) \right\} \quad i = 1, 2, \dots, n \quad (5.2)$$

ifadeleri hesaplanır ve buradan

$$\varepsilon(f, \hat{\boldsymbol{\theta}}, \mathbf{X}) = \inf_{\mathbf{x}_i \in \mathbf{X}} \left\{ \varepsilon(\mathbf{x}_i, f, \hat{\boldsymbol{\theta}}, \mathbf{X}) \right\}$$

elde edilir. Çoğu doğrusal olmayan regresyon model fonksiyonları için EKK tahmin edicisinin bozulma noktası $1/n$ dir. Sınırsız regresyon fonksiyonlarının büyük bir sınıfı için LMS ve LTS tahmin edicilerinin bozulma noktaları 0.5'e yakınsamaktadır (Stromberg ve Ruppert 1992).

5.1.2. Eklemeli bozulma noktası

Bu kesimde $\mathbf{Y}^m$ yeni örneklem değerlerinden oluşan gözlem veri seti olmak üzere, eklemeli sonlu örneklem ve sınırlı sonlu örneklem bozulma noktaları sırasıyla tanıtılacaktır.

5.1.2.1. Eklemeli sonlu örneklem bozulma noktası

$\mathbf{X}$ de $\hat{\theta}$ ' nın eklemeli sonlu örneklem bozulma noktası,

$$\varepsilon^*(\hat{\theta}, \mathbf{X}) = \min \left\{ \frac{m}{m+n} : \sup_{\mathbf{Y}^m} \|\hat{\theta}(\mathbf{X} \cup \mathbf{Y}^m) - \hat{\theta}(\mathbf{X})\| = \infty \right\}$$

olur. Dolayısıyla tahmin edicinin veri kümesinden nedenli etkilenip etkilenmediği tahmin edicilerin elde edilmesinde kullanılan tahmin yöntemlerine bağlı olacaktır (Donoho ve Huber 1983).

5.1.2.2. Sınırlı eklemeli sonlu örneklem bozulma noktası

Uyum değerlerinin bozulma ölçüsü olan bozulma fonksiyonları kullanılarak,

$$\begin{aligned} & \varepsilon_+(\mathbf{x}_i, f, \hat{\theta}, \mathbf{X}) \\ &= \begin{cases} \min_{0 \leq m \leq n} \left\{ \frac{m}{m+n} ; \sup_{\mathbf{x}_m} f(\mathbf{x}_i, \hat{\theta}(\mathbf{X} \cup \mathbf{Y}^m)) = \sup_{\theta} f(\mathbf{x}_i, \theta) \right\} & , \sup_{\theta} f(\mathbf{x}_i, \theta) > f(\mathbf{x}_i, \hat{\theta}) \\ 1 & , dy \end{cases} \end{aligned}$$

ve

$$\begin{aligned} & \varepsilon_-(\mathbf{x}_i, f, \hat{\theta}, \mathbf{X}) \\ &= \begin{cases} \min_{0 \leq m \leq n} \left\{ \frac{m}{m+n} ; \inf_{\mathbf{x}_m} f(\mathbf{x}_i, \hat{\theta}(\mathbf{X} \cup \mathbf{Y}^m)) = \inf_{\theta} f(\mathbf{x}_i, \theta) \right\} & , \inf_{\theta} f(\mathbf{x}_i, \theta) < f(\mathbf{x}_i, \hat{\theta}) \\ 1 & , dy \end{cases} \end{aligned}$$

olup tüm gözlemlere ilişkin

$$\varepsilon(\mathbf{x}_i, f, \hat{\theta}, \mathbf{X}) = \min \left\{ \varepsilon_-(\mathbf{x}_i, f, \hat{\theta}, \mathbf{X}), \varepsilon_+(\mathbf{x}_i, f, \hat{\theta}, \mathbf{X}) \right\} \quad i = 1, 2, \dots, n \quad (5.3)$$

ifadeleri hesaplanır ve buradan

$$\varepsilon(f, \hat{\theta}, \mathbf{X}) = \inf_{\mathbf{x}_i \in \mathbf{X}} \left\{ \varepsilon(\mathbf{x}_i, f, \hat{\theta}, \mathbf{X}) \right\}$$

elde edilir.

5.2. Etki Fonksiyonu

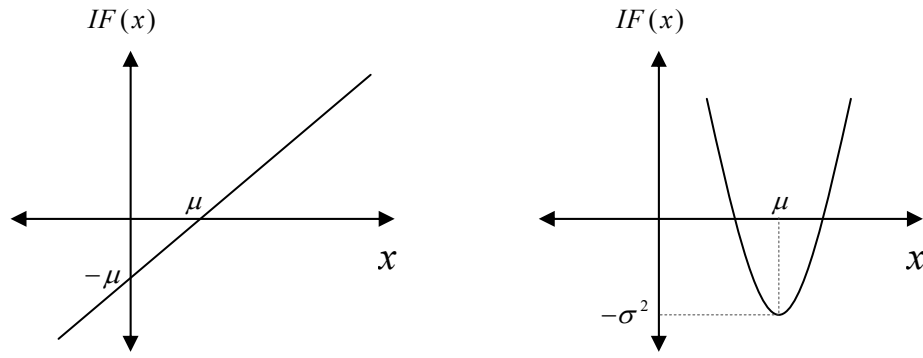
$\hat{\theta}(F)$, F dağılım fonksiyonları kümesi üzerinde gerçel değerli istatistiksel bir fonksiyonel olarak tanımlansın. F dağılımında ortaya çıkabilecek küçük değişimler $\hat{\theta}(F)$ üzerinde bir etki yaratmıyorsa $\hat{\theta}(F)$ için nitel sağlamlığa sahiptir denir. Sonsuz küçük sağlamlık ise en iyi etki fonksiyonu ile değerlendirilmektedir. $\hat{\theta}$ 'nın F deki etki fonksiyonu,

$$IF(x) = \lim_{\varepsilon \rightarrow 0^+} \frac{\hat{\theta}((1-\varepsilon)F + \varepsilon\delta_x) - \hat{\theta}(F)}{\varepsilon} \quad (5.4)$$

olarak tanımlanır (Hampel 1974). Eşitlik (5.4) deki δ_x , dirac dağılımından gelmektedir. $(1-\varepsilon)F + \varepsilon\delta_x$ gösterimi $0 < \varepsilon < 1$ olmak üzere, F ve δ_x 'in karışımını ifade etmektedir.

Bir tahmin edicinin etki fonksiyonu (Influence Function ya da Influence Curve), büyük örneklerde ölçülemeyecek kadar örneklerdeki küçük bozulma oranının parametre tahmin değerlerini nasıl etkilediğini gösterir.

Örneğin birinci momenti mevcut tüm olasılık ölçülerinde aritmetik ortalama, $\hat{\theta} = \int x dF(x)$ olarak tanımlanır. F 'nin ortalaması mevcut ve μ 'ye eşit olmak üzere, $\hat{\theta}$ 'nın F deki etki fonksiyonu $IF(x) = x - \mu$ olur. Benzer biçimde varyans için $\hat{\theta} = \int (x - \mu)^2 dF(x)$ olup F 'nin varyansı σ^2 ve ortalaması μ 'ye eşit olmak üzere, etki fonksiyonu $IF(x) = (x - \mu)^2 - \sigma^2$ olur.



Şekil 5.1. Aritmetik ortalama ve varyansın etki fonksiyonlarının grafiği.

Şekil 5.1' in sol grafiği, aritmetik ortalamanın etki fonksiyonunu göstermektedir. Bu grafikten n birimlik örneklemin, örneklemdaki bozulumlara karşı ortalamanın ne derecede etkilendiğini göstermektedir. $\lim_{x \rightarrow +\infty} (IF(x))$ limit değeri $+\infty$ ve $\lim_{x \rightarrow -\infty} (IF(x))$ limit değeri de $-\infty$ olduğundan, aritmetik ortalamanın sınırsız bir biçimde etkilenmekte olduğu görülmektedir. Sağdaki grafikte de $\lim_{x \rightarrow +\infty} (IF(x))$ veya $\lim_{x \rightarrow -\infty} (IF(x))$ limit değerleri $+\infty$ 'a gitmektedir. Bu durumda etki fonksiyonu sınırsız bir şekilde etkilenmekte ve varyansın örneklemdaki bozulumlara karşı çok daha fazla tepki verdiği görülmektedir.

$IF(x)$ yardımıyla bir tahmin edici için diğer sağlamlık ölçütleri olan Duyarlılık Eğrisi (Sensitivity Curve - SC), Değer Değişim Eğrisi (Change of Value Curve - CV), Yerel Kayma Duyarlılığı (Local Shift Sensitivity - LSS), Büyük Hata Duyarlılığı (Gross Error Sensitivity - GES), Reddetme Noktası (Rejection Point - RP) elde edilebilir (Maronna ve ark. 2006).

Etki fonksiyonundan hesaplanılan önemli bir kriter de duyarlılık eğrisidir. Duyarlılık eğrisi, mevcut veri setine yeni bir gözlem değeri ilave edildiğinde tahmin değerlerinde olabilecek maksimum etkiyi ölçmektedir.

Duyarlılık eğrisi, etki fonksiyonunun dönüştürülmüş ve yeniden ölçeklenmiş bir durumudur. Duyarlılık eğrisi,

$$\begin{aligned} SC_n(x_0) &= \frac{\hat{\theta}_{n+1}(x_1, x_2, \dots, x_n, x_0) - \hat{\theta}_n(x_1, x_2, \dots, x_n)}{1/(n+1)} \\ &= (n+1) \left(\hat{\theta}_{n+1}(x_1, x_2, \dots, x_n, x_0) - \hat{\theta}_n(x_1, x_2, \dots, x_n) \right) \end{aligned}$$

olarak ifade edilir. Bu eşitlikten

$$SC_n(x_0) \xrightarrow{hhhy} IF_{\hat{\theta}}(x_0, F)$$

duyarlılık eğrisinin etki fonksiyonuna hemen hemen her yerde yakınsamakta olduğu görülmektedir. Ayrıca,

$$\frac{SC_n(x_0)}{\max_x |SC_n(x)|} \xrightarrow{hhhy} \frac{IF(x_0)}{\max_x |IF(x)|} = \text{sgn}(x - \text{med}(x))$$

değerinde eşit olmaktadır (Croux 1998).

6. BOZULMA NOKTASININ HESAPLANMASI VE BİR UYGULAMA

Doğrusal olmayan regresyonda teorik olarak, sınırlı bir model fonksiyonu ve modeldeki parametrelerin tahmininde kullanılacak tahmin ediciyi ele alarak, bu tahmin edicinin modeldeki sağlamlığının bir ölçüsü olan bozulma noktası hesaplanılabilmektedir (Stromberg ve Ruppert 1992). İstatistiksel paket programlarda tüm modellerin ve tahmin edicilerin ayrı ayrı bozulma noktalarının değerleri program içerisine konması bir çözüm oluşturabilir. Ancak, kullanıcının oluşturacağı modellerde bu istatistiksel paket programlar vasıtasıyla bozulma noktasının hesaplanması mümkün olamayacaktır. Bu da kullanıcının, ele aldığı modelde tahmin edicinin ne derecede sağlam olacağının hesaplanması için bu işlemi bizzat elle ve teorik olarak çözmesi gerekmektedir. Modelin zorluğu ve işlem hacminin büyüklüğünden ötürü her zaman bu mümkün kılınmamaktadır. Bundan dolayı bir algoritma yardımıyla kullanıcının ele almış olduğu modelde ve tahmin ediciye sonuçlarda, bozulma noktalarının kullanıcıya yansıtılması önemlidir. Maalesef günümüz istatistiksel paket programlarında bu mümkün değildir. Bu noktadan yola çıkarak, özel bir algoritma vasıtasıyla, bozulma noktasının istatistiksel bir paket programıyla hesaplanması düşüncesi bu tezin ana temasını oluşturmaktadır.

6.1. Bozulma Noktası için Yeni bir Yaklaşım

Önceki tanımlarda bozulma noktası, veri kümesindeki gözlemlerin yer değiştirmesinde veya eklenmesinde, tahmin edicinin sınırsız bir şekilde etkilenmesi sonucu hata teriminin davranışı olarak ele alınmıştı. Doğrusal olmayan regresyonda teorik olarak, sınırlı bir model fonksiyonunda tahmin edicinin bozulma noktası hesaplanılabilmektedir (Stromberg ve Ruppert 1992). Ancak diğer model fonksiyonlarında bozulma noktasının hesaplanması mümkün olmamaktadır.

Tablo 6.1. Bazı tahmin edicilerin asimptotik bozulma noktaları.

	Asimptotik Bozulma noktaları
EKK	$1/n$
M (Huber)	0,5
M (Hampel)	0,5
GM ($\alpha = 0$)	0,5
GM ($\alpha = 1$)	0,5
LTS (%25)	0,25
LTS (%50)	0,5
S ($b = 0.1$)	0,1
S ($b = 0.5$)	0,5

Tahmin edicilerin genelde bozulma noktaları Tablo 6.1 de görüldüğü gibi $1/n$ veya $1/2$ dir. Sağlam olan tahmin edicilerde bozulma noktası $1/2$, sağlam olmayan tahmin edicilerde ise bozulma noktası $1/n$ olur. Bu durumda iki sağlam tahmin ediciyi kendi aralarında bozulma noktası bakımından sağlamlık sıralaması yapma imkanı yoktur. Bundan dolayı İstatistik paket programlarda hesaplanması mümkün olabilecek ve tahmin edicileri de sağlamlık açısından sıralanmasına imkan veren yeni bir bozulma noktasının tanımlanması gereklidir.

Bozulma noktası kavramı, bir tahmin edicinin aykırı gözlemleri ne derecede dışlayabildiği (bu gözlemlere karşı dirençli olduğu) düşüncesiyle paralel bir kavramdır. Bu düşünceye göre bozulma noktası, bir tahmin edicinin aykırı gözlemleri dışlayabildiği orana denktir.

Bozulma Noktası için yeni bir yöntem (Tespit Yöntemi) :

Veri setinde m tane aykırı gözlem olsun. $\hat{\theta}$ tahmin edicisine dayalı olarak bu m tane aykırı gözlemde doğru tespit edilebilenlerin sayısı da k olsun.

$$A_i = \begin{cases} n, & k_i = m \\ m, & k_i \neq m \end{cases} \quad (6.1)$$

Olmak üzere, $\hat{\theta}$ tahmin edicisinin bozulma noktası,

$$\varepsilon^*(\hat{\theta}, \mathbf{X}) = \frac{1}{n} \text{Min} \left\{ \bigcup_{m=1}^n \left\{ \text{Min} \left\{ \bigcup_{i=1}^{\binom{n}{m}} \{A_i\} \right\} \right\} \right\} \quad (6.2)$$

şeklinde önerilmektedir.

Buna göre önerilen **Tespit yöntemi** ile bozulma noktasının (BN) hesaplanması şöyledir:

Yeni Yöntem ile Bozulma Noktası Hesaplama Algoritması (Tespit Yöntemi):

- A1.** $N(0,1)$ dağılımından rasgele n tane sayı üretilir. Üretilen değerler e_i ler olsun.
- A2.** Tanım bölgesinden belli bir $[a, b]$ $a, b \in \mathbb{R}$ aralığından rasgele n tane sayı üretilir ve bu üretilen değerler $\mathbf{x}_i$ üretilir ve $\boldsymbol{\theta} = (\theta_1, \theta_2, \dots, \theta_p)'$ ($\boldsymbol{\theta} \in \Theta \subset R^p$) parametre değerlerine $\boldsymbol{\theta} = \boldsymbol{\theta}^*$ sabit değerler verilir. $\boldsymbol{\theta}^* = (\theta_1^*, \theta_2^*, \dots, \theta_p^*)$
- A3.** $y_i = f(\mathbf{x}_i; \boldsymbol{\theta}^*) + e_i$, $i = 1, 2, \dots, n$
Olacak biçimde y_i değerleri hesaplanır ve $BN = 1$ alınır.
- A4.** $m = 0$
- A5.** $m = m + 1$
- A6.** $i = 0$
- A7.** $i = i + 1$
- A8.** A3 den hesaplanan normal verileri, Şekil 6.1 deki senaryolardan bir tanesi ile m tanesi bozdurulur. Bu m tane gözlemin sırası muhafaza edilir.
- A9.** $\hat{\theta}$ tahmin edicisine bağlı olarak elde edilen artıklardan, Kesim 4.2.4 deki algoritma kullanılarak aykırı gözlemler tespit edilir.
- A10.** Eğer eşitlik (6.1) de yer alan $k_i \neq m$ ise $BN = \frac{m}{n}$ olur ve A13 adımına gidilir.
- A11.** $i < \binom{n}{m}$ ise A7 adımına git.
- A12.** $m < n$ ise A5 adımına git.
- A13.** BN değeri $\hat{\theta}$ tahmin edicisinin bozulma noktası olur.

6.2. Uygulama

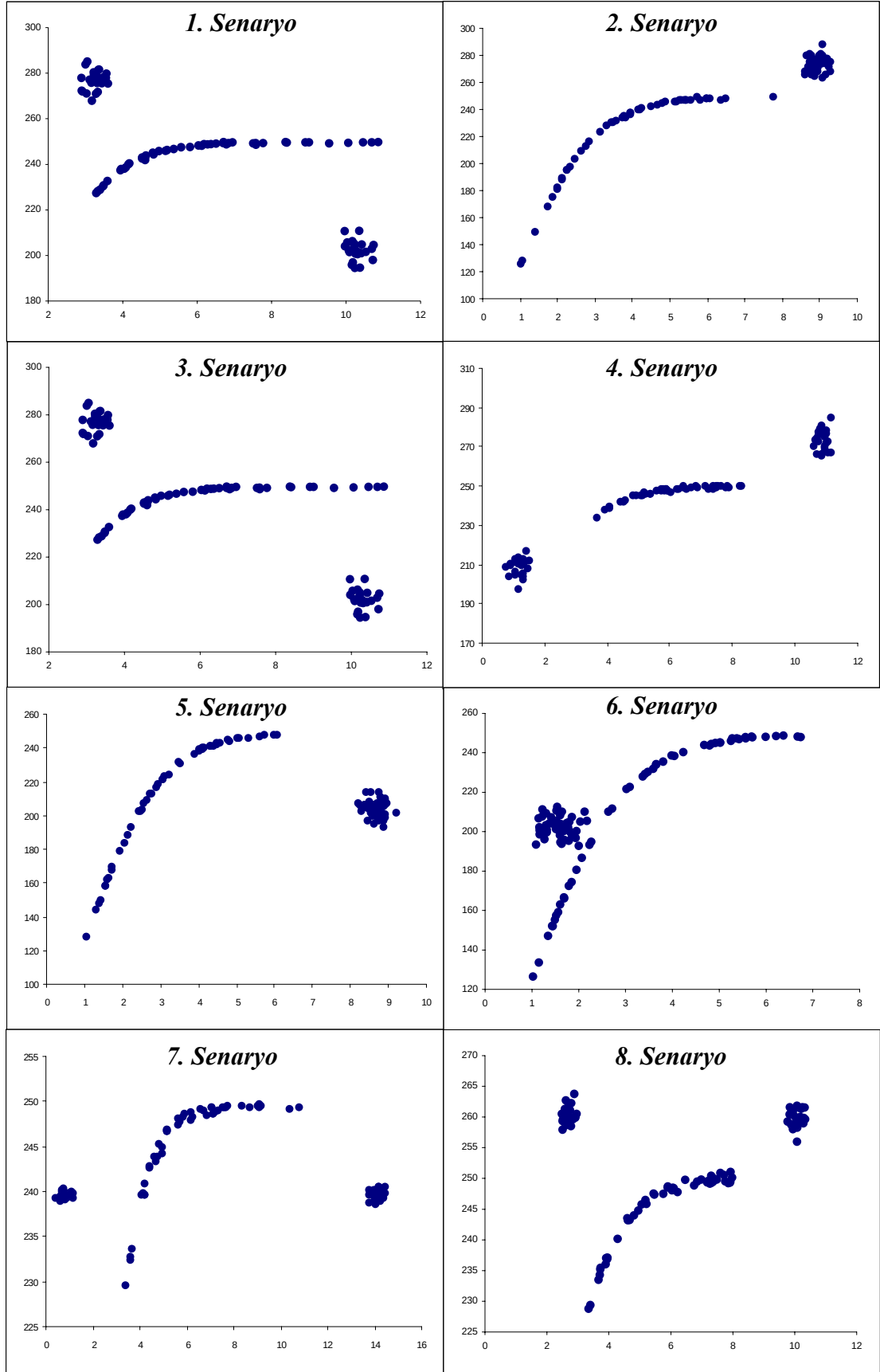
Bu bölümde doğrusal olmayan regresyonda Richard'ın sigmoidal modelinde Şekil 6.1 deki senaryolara uygun rasgele üretilen veri setleri ile tahmin edicilerin bozulma noktaları hesaplanarak bu farklı senaryolardaki sağlamlık dereceleri incelenecektir. Monte Carlo simülasyon çalışmaları Delphi görsel programlama dilinde kodlanan *Selçuk STAT* istatistik paket programından faydalanarak yapılacaktır.

Şekil 6.1 deki ilk altı senaryo Serbert ve ark. (1998)'nin çalışmalarından uyarlanmıştır. Bu senaryoların hangi tür aykırı değerler içerdikleri Tablo 6.2 de verilmiştir. Tablo 6.2 deki '*' işaretleri bulunduğu konumda aykırı değer olduğunu, '-' işareti ise aykırı değer olmadığını göstermektedir.

Tablo 6.2. Uygulamada kullanılan aykırı değer türleri

Senaryolar	Y uzayında aykırı değer	X uzayında aykırı değer	Artıklarda aykırı değer
1.Senaryo	*	-	*
2.Senaryo	*	*	*
3.Senaryo	*	-	*
4.Senaryo	*	*	*
5.Senaryo	-	*	*
6.Senaryo	-	-	*
7.Senaryo	-	*	*
8.Senaryo	*	*	*

Yapılacak olan simülasyon çalışmalarında, her bir senaryo için gözlem sayıları 20, 40 ve 100 alınarak farklı gruplarda bozulmuş değerler için standart sapmalar 5 ve 10 alınacaktır. Tablo 6.3. deki 48 durumdan her bir durum için Kesim 6.1 de tanımlanan *Tespit yöntemi* ile bozulma noktaları hesaplanacaktır.



Şekil 6.1. Yinelemeler sırasında 8 ayrı senaryodan rasgele üretilen verilerden bir tanesinin saçılım grafikleri.

Tablo 6.3. Simülasyonlar da kullanılan 48 farklı durum

Durum	Senaryo	Aykırı Değer Grupları	Gözlem Sayısı (n)	Standart Sapma
1	1	1	20	5
2	1	1	20	10
3	1	1	40	5
4	1	1	40	10
5	1	1	100	5
6	1	1	100	10
7	2	1	20	5
8	2	1	20	10
9	2	1	40	5
10	2	1	40	10
11	2	1	100	5
12	2	1	100	10
13	3	2	20	5
14	3	2	20	10
15	3	2	40	5
16	3	2	40	10
17	3	2	100	5
18	3	2	100	10
19	4	2	20	5
20	4	2	20	10
21	4	2	40	5
22	4	2	40	10
23	4	2	100	5
24	4	2	100	10
25	5	1	20	5
26	5	1	20	10
27	5	1	40	5
28	5	1	40	10
29	5	1	100	5
30	5	1	100	10
31	6	1	20	5
32	6	1	20	10
33	6	1	40	5
34	6	1	40	10
35	6	1	100	5
36	6	1	100	10
37	7	2	20	5
38	7	2	20	10
39	7	2	40	5
40	7	2	40	10
41	7	2	100	5
42	7	2	100	10
43	8	2	20	5
44	8	2	20	10
45	8	2	40	5
46	8	2	40	10
47	8	2	100	5
48	8	2	100	10

Önerilen *Tespit yöntemi* ile bozulma noktasının hesaplanmasında aykırı değerlerin tespitinde kullanılabilecek olan ADS ve RD yöntemlerinden hangisinin daha uygun olduğunun anlaşılması için senaryolarda Monte Carlo simülasyonları yapıldı. Bu simülasyonlarda gözlem sayısı 20, 40 ve 100, standart sapma 5 ve 10, bozulma oranı %10 ve %25 alınarak ADS ve RD yöntemleri uygulandı. Çıkan sonuçlar Tablo 6.4 ve Tablo 6.5 de verilmiştir.

Tablo 6.4. ADS yönteminin senaryolardaki ortalama başarı oranları

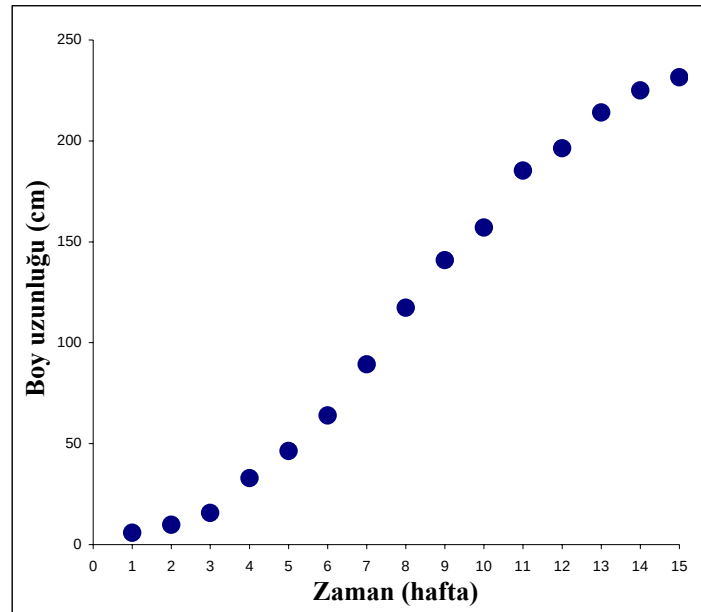
Gözlem Sayısı (n)	Std. Sapma	Bozulma Oranı (%)	Senaryolardaki Ortalama Başarı Oranları							
			S1	S2	S3	S4	S5	S6	S7	S8
20	5	10	0,98	0,92	0,78	0,77	0,98	0,97	0,98	0,93
		25	0,89	0,85	0,72	0,71	0,89	0,88	0,83	0,81
	10	10	0,99	0,89	0,80	0,79	0,84	0,83	0,98	0,97
		25	0,92	0,88	0,74	0,74	0,76	0,76	0,84	0,81
40	5	10	0,95	0,89	0,83	0,82	0,99	0,99	0,98	0,95
		25	0,88	0,82	0,70	0,69	0,87	0,87	0,86	0,85
	10	10	0,96	0,88	0,84	0,83	0,83	0,82	0,96	0,95
		25	0,90	0,83	0,72	0,72	0,75	0,74	0,87	0,87
100	5	10	0,90	0,80	0,78	0,77	1	1	0,91	0,90
		25	0,82	0,79	0,69	0,68	1	1	0,87	0,86
	10	10	0,93	0,87	0,80	0,78	0,85	0,83	0,92	0,92
		25	0,87	0,84	0,75	0,74	0,71	0,73	0,89	0,88

RD yönteminin, senaryolardan elde edilen değerlerin Senaryo 3 ve 4 dışında hemen hemen hepsinde bozuk olanları doğru tespit ettiği Tablo 6.5 de görülmektedir. Bundan dolayı Tablo 6.4 ve Tablo 6.5 deki sonuçlar ışığı altında önerilen *Tespit yöntemi* ile bozulma noktasının hesaplanmasın da RD yöntemi tercih edilecektir.

Tablo 6.5. RD yönteminin senaryolardaki ortalama başarı oranları

Gözlem Sayısı (n)	Std. Sapma	Bozulma Oranı (%)	Senaryolardaki Ortalama Başarı Oranları (%)							
			S1	S2	S3	S4	S5	S6	S7	S8
20	5	10	1	1	0,99	1	1	0,98	1	1
		25	1	1	0,98	1	1	0,98	1	1
	10	10	1	0,99	0,98	0,99	1	0,97	1	1
		25	1	0,99	0,97	0,99	1	0,96	1	0,99
40	5	10	1	1	0,99	1	1	0,96	1	1
		25	1	1	0,98	1	1	0,96	1	1
	10	10	1	1	0,96	1	1	0,95	1	1
		25	1	0,99	0,96	1	1	0,94	1	0,99
100	5	10	1	1	0,99	1	1	0,96	1	1
		25	1	1	0,98	1	1	0,96	1	1
	10	10	1	1	0,97	1	1	0,96	1	0,99
		25	1	0,99	0,96	0,99	1	0,96	1	0,99

Önerilen *Tespit yöntemi* ile bozulma noktasının hesaplanmasına geçmeden önce gerçek veri seti üzerinden aykırı gözlemlerin tespiti yapılacaktır. Örnek veri seti olarak Konya bölgesindeki 2008 dönemine ait bir mısır bitkisinin haftalara göre boy uzunluğu kullanılacaktır. Örnek veri setinin saçılım grafiği Şekil 6.2 de verilmiştir.

**Şekil 6.2.** Mısır bitkisinin zamana bağlı boy uzunluğu

Mısır bitkisi verilerinden elde edilen EKK, LTS, M, GM ve S tahmin edicilerinin parametre tahminleri ile HKT değerleri Tablo 6.6 de verilmiştir. Bu verilerin sigmoidal modeldeki LAD tahminleri, yinelemeler sırasında logaritma fonksiyonunun negatif değerler almasından dolayı hesaplanamamıştır.

Tablo 6.6. Mısır bitkisi verilerinden elde edilen tahmin edicilerin parametre tahminleri ile HKT değerleri

	θ_1	θ_2	θ_3	θ_4	HKT
EKK	259,7353	1,2072	0,30191	0,31998	76,6730
M (Huber)	272,6421	-1,4579	0,24737	0,03687	104,9560
M (Hampel)	272,6423	-1,4561	0,24837	0,03564	105,2856
GM ($\alpha = 0$)	274,5661	-1,7812	0,24340	0,02747	105,4806
GM ($\alpha = 1$)	269,6919	-1,7739	0,24976	0,02747	132,1959
LTS (%25)	270,7628	-1,7365	0,25072	0,02812	105,5107
LTS (%50)	272,8394	-1,7447	0,24899	0,02752	105,5099
S ($b = 0.1$)	265,5244	0,6577	0,29845	0,01645	106,3422
S ($b = 0.5$)	267,5634	-0,0143	0,23172	0,02149	107,1231

ADS yöntemiyle bu verilerden elde edilecek olası en yüksek aykırı gözlem sayısı

Tablo 6.7 deki değerlerden sonra belli olacaktır. Burada $m \in \left\{0, 1, 2, \dots, \left\lceil \frac{15}{2} \right\rceil\right\}$

alınacaktır.

Tablo 6.7. EKK artıkları ile ADS yönteminde hesaplanan parametre değerleri

	m	SMAD	$\hat{\mu}$	$\hat{\sigma}$	$\hat{\lambda}$
$c = 0$	0	0,1193	0,1138	2,7092	0
	1	*0,0970	0,2706	3,0780	-0,2134
	2	0,1163	0,0927	2,8432	-0,1496
	3	0,1187	0,2162	3,0155	-0,1874
	4	0,1031	0,2308	3,0104	-0,1903
	5	0,0984	0,1212	2,7904	-0,3806
	6	0,1115	0,2747	3,0826	-0,2039
	7	0,1120	0,2750	3,0830	-0,1498
$c = 0.3$	0	0,1327	0,1206	2,4981	0
	1	*0,0962	0,0866	2,5981	-0,5816
	2	0,1249	0,0986	2,6288	-0,1885
	3	0,1298	0,2241	2,7610	-0,2327
	4	0,1160	0,2378	2,7688	-0,2115
	5	0,1105	0,2450	2,7727	-0,2068
	6	0,1277	0,2419	2,7710	-0,1895
	7	0,1286	0,2563	2,7789	-0,1415
$c = 1/3$	0	0,1267	0,1236	2,5465	0
	1	*0,0983	0,1216	2,6543	-0,4781
	2	0,1213	0,0968	2,7113	-0,3805
	3	0,1250	0,2140	2,8639	-0,2097
	4	0,1098	0,2347	2,8816	-0,2175
	5	0,1041	0,2533	2,9057	-0,2110
	6	0,1200	0,2645	2,9191	-0,1926
	7	0,1207	0,2679	2,9228	-0,1495
$c = 0.5$	0	0,1277	0,1214	2,5493	0
	1	*0,0980	0,1180	2,6430	-0,4926
	2	0,1217	0,0934	2,7197	-0,3848
	3	0,1266	0,1789	2,8109	-0,1872
	4	0,1106	0,2358	2,8657	-0,2257
	5	0,1049	0,2505	2,8871	-0,2105
	6	0,1211	0,2613	2,8987	-0,2039
	7	0,1218	0,2673	2,9054	-0,1496

Tablo 6.7 deki değerler incelendiğinde tüm c gruplarında SMAD değerini en küçük yapan m değerinin 1 olduğu görülmektedir. Huber'ın M tahmin edicisi ile hesaplanan artıklara ADS yöntemi uygulandığında Tablo 6.8 deki değerler elde edilmiştir. Tablo 6.7 ve Tablo 6.8 deki sonuçlar incelendiğinde tüm c ler için m değerleri 1 olmuştur. Bu sonuçlara göre veri setinde 1 tane aykırı değer tespit edilmiştir.

Tablo 6.8. Huber'ın M artıkları ile ADS yönteminde hesaplanan parametre değerleri

	m	SMAD	$\hat{\mu}$	$\hat{\sigma}$	$\hat{\lambda}$
$c = 0$	0	0,1175	0,5246	3,1768	0
	1	*0,1029	0,6369	3,6312	-0,2083
	2	0,1253	0,5501	3,2134	-0,2127
	3	0,1122	0,6464	3,3955	-0,2237
	4	0,1267	0,5517	3,2150	-0,0221
	5	0,1079	0,5471	3,1856	-0,2413
	6	0,1118	1,0403	3,7660	-0,3143
	7	0,1049	0,8341	3,6255	-0,2411
$c = 0.3$	0	0,1528	0,4962	2,8737	0
	1	*0,1324	0,5320	2,9266	-0,4215
	2	0,1601	0,5588	2,9658	-0,2109
	3	0,1453	0,5767	3,0250	-0,1996
	4	0,1359	0,6922	3,2338	-0,2555
	5	0,1400	0,9098	3,3307	-0,3000
	6	0,1521	0,9847	3,3959	-0,3271
	7	0,1644	0,5388	2,9350	-0,0413
$c = 1/3$	0	0,1354	0,5140	2,9932	0
	1	*0,1195	0,6898	3,3308	-0,1942
	2	0,1427	0,5548	3,0604	-0,2090
	3	0,1288	0,5712	3,1169	-0,1864
	4	0,1205	0,5336	3,0349	-0,3291
	5	0,1223	0,7812	3,4787	-0,2476
	6	0,1322	0,9974	3,5336	-0,3170
	7	0,1458	0,5440	3,0466	-0,0313
$c = 0.5$	0	0,1378	0,5041	2,9793	0
	1	*0,1218	0,6931	3,3059	-0,2610
	2	0,1452	0,5542	3,0469	-0,2058
	3	0,1311	0,6334	3,2130	-0,2336
	4	0,1220	0,5424	3,0282	-0,3500
	5	0,1247	0,7795	3,4426	-0,2459
	6	0,1350	1,0580	3,5775	-0,3438
	7	0,1484	0,5433	3,0292	-0,0333

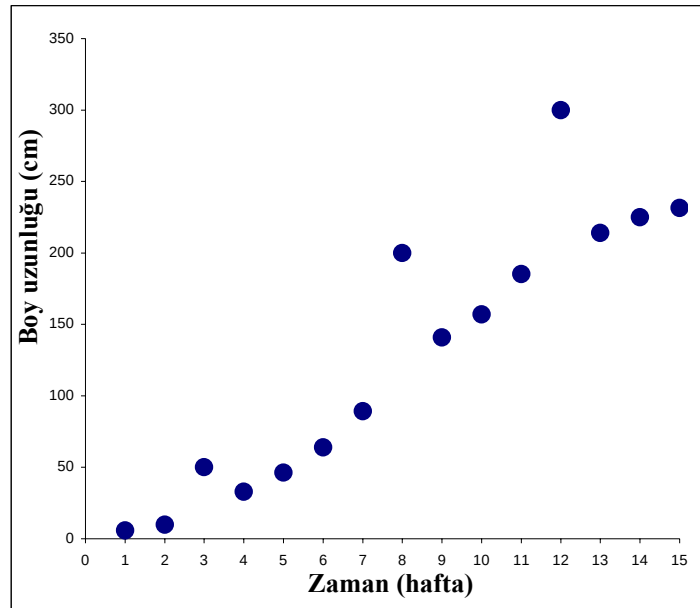
Mısır bitkisi verisinde RD yöntemiyle aykırı gözlemlerin tespiti yapıldığında Tablo 6.9 daki sonuçlar elde edilmiştir.

Tablo 6.9. EKK ve Huber'ın M artıklarından RD yöntemiyle elde edilen uzaklıklar.

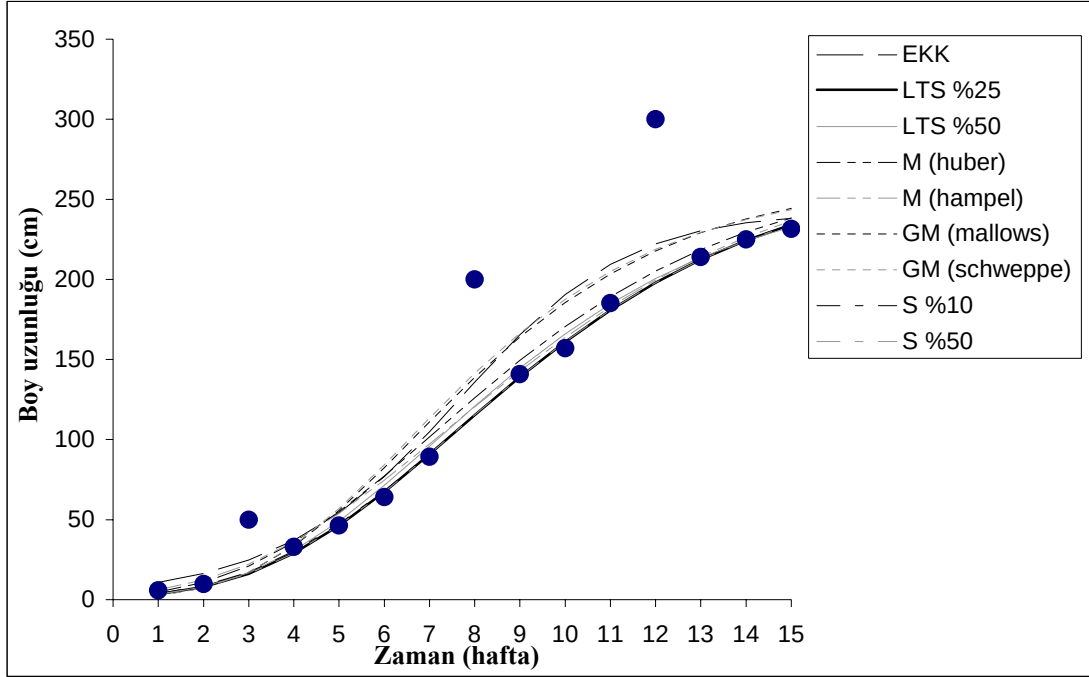
Gözlem No (<i>i</i>)	EKK RD_i^2	Huber M RD_i^2
1	0,028671	1,189104
2	0,101834	0,628280
3	2,779599	0,531221
4	3,420888	3,467028
5	0	0,396372
6	3,184222	7,993786
7	0,415693	3,526321
8	2,282854	0,179877
9	0,843219	0,153877
10	11,11840	9,317507
11	3,971413	4,645694
12	3,301495	1,932104
13	0,458742	0,380806
14	0,351225	0
15	0,750249	4,179223

Tablo 6.9 incelendiğinde $\chi^2_{4,0,975} = 11.143$ tablo değerinden büyük değer olmadığı için, mevcut veri setinde hiç aykırı gözlem olmadığı sonucuna varılır.

Gerçek mısır bitkisi verilerinden sonra bu verilerden 3., 8. ve 12. gözlem değerleri keyfi olarak bozdurularak (contamination) elde edilen saçılım grafiği Şekil 6.3 de verilmiştir.



Şekil 6.3. Gerçek mısır bitkisi verilerinden 3., 8. ve 12. gözlem değerlerinin bozdurulmasıyla elde edilen saçılım grafiği



Şekil 6.4. Bozulmuş verilerdeki kestirimler

Şekil 6.4 deki EKK sonuçları, diğer sağlam tahmin edicilere göre bu 3 bozulmuş değerden fazlasıyla etkilenmiştir.

Tablo 6.10. Tahmin edicilerin gerçek mısır bitkisi verilerinden 3., 8. ve 12. gözlem değerlerinin bozdurulmasıyla parametre tahminleri ve HKT değerleri

	θ_1	θ_2	θ_3	θ_4	HKT
EKK	242,2268	4,6040	0,5608	1,3022	14111,3573
M (Huber)	271,7340	-0,6635	0,2566	0,0834	16064,4534
M (Hampel)	279,8662	-0,7448	0,2342	0,0834	17517,5948
GM ($\alpha = 0$)	263,6584	-0,4382	0,3077	0,0834	14739,1286
GM ($\alpha = 1$)	260,6587	-0,4156	0,3167	0,0834	14771,8116
LTS (%25)	271,7924	-0,5638	0,2546	0,0834	18847,3861
LTS (%50)	262,4510	-0,5457	0,2700	0,0834	17474,1219
S ($b = 0.1$)	274,4902	-0,6032	0,2484	0,0835	18793,4607
S ($b = 0.5$)	271,4473	-0,5633	0,2550	0,0835	18798,4116

ADS yönteminde ilk tahmin değerlerinden sonra elde edilen artıklara sırasıyla $c = 0, 0.3, 1/3, 0.5$ değerleri verildiğinde,

Tablo 6.11. Bozdurulmuş verilerde EKK artıkları ile ADS yönteminde hesaplanan parametre değerleri

	m	SMAD	$\hat{\mu}$	$\hat{\sigma}$	$\hat{\lambda}$
$c = 0$	0	0,4067	2,5910	49,0696	0
	1	0,3956	-2,7893	31,6869	1,0114
	2	*0,2537	-7,6319	16,8889	3,3814
	3	0,3373	-7,6798	18,1030	2,8217
	4	0,4118	12,9922	81,2149	-0,7353
	5	0,3545	19,7273	80,0352	-0,8077
	6	0,3093	22,8676	78,9931	-0,8379
	7	0,2886	29,4918	71,8624	-0,8609
$c = 0.3$	0	0,4479	-0,4241	42,6923	0
	1	0,4423	-2,8384	29,0049	0,9486
	2	*0,2765	-7,6684	15,3995	3,6421
	3	0,3660	-7,5126	16,8021	2,9879
	4	0,4567	7,7733	60,6415	-0,6780
	5	0,3989	16,4811	62,6906	-0,8547
	6	0,3500	16,3366	61,1749	-0,8280
	7	0,3259	17,0762	58,8293	-0,8005
$c = 1/3$	0	0,4308	0,0463	46,0318	0
	1	0,4210	-2,8643	30,1988	0,9827
	2	*0,2662	-7,6624	15,8602	3,6613
	3	0,3526	-7,6361	16,7292	3,0708
	4	0,4396	12,6002	69,3952	-0,7500
	5	0,3770	18,2485	69,7573	-0,8359
	6	0,3302	19,6126	69,3913	-0,8511
	7	0,3104	26,0391	68,6400	-0,9355
$c = 0.5$	0	0,4337	-0,1910	45,6050	0
	1	0,4241	-2,8797	30,0382	0,9766
	2	*0,2678	-7,6695	15,7711	3,6296
	3	0,3566	-7,4910	17,6405	2,8320
	4	0,4398	9,7481	67,9911	-0,7098
	5	0,3801	19,3328	68,9077	-0,8622
	6	0,3331	19,0722	68,0061	-0,8471
	7	0,3130	25,7451	67,8421	-0,9434

Tablo 6.11 ve Tablo 6.12 deki değerler elde edilmiştir. Bu tablo değerleri incelendiğinde, tüm c değerlerin de SMAD değerini en küçük yapan m değerinin 2 olduğu görülmektedir. Bu da eldeki veri setinde olabilecek aykırı değer sayısının 2 olduğunu göstermektedir.

Tablo 6.12. Bozdurulmuş verilerde Huber'ın M artıkları ile ADS yönteminde hesaplanan parametre değerleri

	m	SMAD	$\hat{\mu}$	$\hat{\sigma}$	$\hat{\lambda}$
$c = 0$	0	0,3087	7,8929	48,0437	0
	1	0,3200	6,1703	40,9148	0,6315
	2	*0,1457	43,8432	81,2212	-0,9485
	3	0,3339	6,7409	44,8603	0,1284
	4	0,2893	22,3543	76,8223	-0,7059
	5	0,2313	32,0822	84,5090	-0,8587
	6	0,1829	38,8103	84,2646	-0,9173
	7	0,2253	0,4890	21,0895	2,6330
$c = 0.3$	0	0,3899	7,8929	41,6737	0
	1	0,4104	6,4827	36,6115	0,5778
	2	*0,2005	40,3747	67,6756	-1,0285
	3	0,4206	6,1314	37,4514	0,2352
	4	0,3733	20,2694	63,0375	-0,7363
	5	0,3050	29,0896	69,3909	-0,9164
	6	0,2467	35,4162	69,6430	-0,9880
	7	0,2877	0,6085	18,6707	2,9261
$c = 1/3$	0	0,3515	7,8929	44,3995	0
	1	0,3672	6,3238	38,3983	0,6130
	2	*0,1733	41,9978	73,4865	-0,9945
	3	0,3798	6,4287	40,6602	0,1801
	4	0,3330	21,2524	68,9687	-0,7264
	5	0,2692	30,4883	75,8783	-0,8934
	6	0,2154	37,0008	75,9139	-0,9586
	7	0,2579	0,5517	19,7008	2,7948
$c = 0.5$	0	0,3572	7,8929	43,9664	0
	1	0,3736	6,3460	38,1081	0,6089
	2	*0,1772	41,7549	72,5645	-1,0000
	3	0,3858	6,3855	40,1544	0,1877
	4	0,3389	21,1061	68,0294	-0,7284
	5	0,2744	30,2788	74,8491	-0,8972
	6	0,2199	36,7633	74,9187	-0,9634
	7	0,2623	0,5601	19,5367	2,8150

Tablo 6.11 ve Tablo 6.12 deki bu sonuçlar doğru tespitte bulunmamaktadır. Aynı veri setinden EKK ve Huber'ın M tahminleri ile elde edilen artıklardan RD yöntemiyle aykırı gözlemlerin tespiti yapıldığında Tablo 6.13 deki sonuçlar elde edilmiştir.

Tablo 6.13. EKK ve Huber'ın M artıklarından RD yöntemiyle elde edilen uzaklıklar.

Gözlem No (<i>i</i>)	EKK RD_i^2	Huber M RD_i^2
1	0,24379	0,8581
2	0,04732	0,3589
3	*33,29631	*32,0081
4	0,42508	0,0879
5	0	0,4479
6	0,82382	2,2004
7	1,97990	1,7734
8	*157,97760	*174,1607
9	8,75025	0,4507
10	*20,40990	2,3176
11	8,25329	0,0066
12	*223,63330	*279,4052
13	2,23898	0,0000
14	0,20025	0,0000
15	0,02767	0,1676

Tablo 6.13 deki sonuçlar incelendiğinde, Huber'ın M artıklarının RD_i^2 değerlerinin, $\chi_{4,0.975}^2 = 11.143$ tablo değerinden büyük olan gözlemler 3., 8. ve 12. gözlemlerdir. Huber'ın M ile elde edilen artıklardan RD yöntemi, bozdurulan bu 3 gözlem değerini doğru tespit edebilmiştir. EKK artıklarının RD_i^2 değerlerine bakıldığında 3., 8., 10. ve 12. gözlemler aykırı değer olarak gözükmemektedir. Bu değerler 10. gözlem değerinden de etkilenmiştir.

Şimdi sırasıyla sekiz ayrı senaryo da ayrı ayrı RD yöntemiyle aykırı gözlemler tespit edilerek Bölüm 3' te ele alınan tahmin edicilerin, doğrusal olmayan regresyonda Richard'ın sigmoidal modelinin **Tespit yöntemiyle** bozulma noktaları hesaplanacaktır.

6.2.1. Senaryo-I

Bu senaryo, $\mathbf{Y}$ vektör uzayında (bağımlı değişken yönünde) tek guruplanmış aykırı değerleri ele almaktadır. Tahmin edicilerin **Tespit yöntemiyle** simülasyonlar sonrası hesaplanan bozulma noktaları (ε^*) Tablo 6.14 ve Tablo 6.15 de verilmiştir.

Tablo 6.14. Senaryo-I için EKK, M ve GM tahmin edicilerin tespit yöntemiyle bozulma noktaları.

Tahmin Ediciler	Gözlem Sayısı (n)	Std. Sapma	m	ε^*
EKK	20	5	1	0,050
		10	1	0,050
	40	5	1	0,025
		10	2	0,050
	100	5	5	0,050
		10	4	0,040
M (Huber)	20	5	8	0,400
		10	8	0,400
	40	5	18	0,450
		10	20	0,500
	100	5	49	0,490
		10	50	0,500
M (Hampel)	20	5	10	0,500
		10	10	0,500
	40	5	18	0,450
		10	19	0,475
	100	5	50	0,500
		10	50	0,500
GM ($\alpha = 0$)	20	5	8	0,400
		10	8	0,400
	40	5	19	0,475
		10	19	0,475
	100	5	50	0,500
		10	50	0,500
GM ($\alpha = 1$)	20	5	9	0,450
		10	8	0,400
	40	5	20	0,500
		10	19	0,475
	100	5	50	0,500
		10	50	0,500

Tablo 6.14 incelendiğinde, EKK tahmin edicisi sağlam olmayan bir tahmin edici olduğundan bozulma noktası değerleri bakımından en düşük değerlere sahiptir. M tahmin edicilerinden Huber tahmin edicisinin $t = 2$ iken bozulma noktası, n değeri arttıkça 0.5'e yakınsamaktadır. Hampel tahmin edicisinde düzeltme katsayıları $a = 2.17$, $b = 3.4$, $c = 8.5$ olduğunda bozulma noktası 0.5 olmaktadır. GM tahmin edicilerinden Mallows ve Schweppe tahmin edicilerinin bozulma noktaları da 0.5 olarak elde edilmiştir.

Tablo 6.15. Senaryo-I için LTS ve S tahmin edicilerin tespit yöntemiyle bozulma noktaları.

Tahmin Ediciler	Gözlem Sayısı (n)	Std. Sapma	m	ε^*
LTS (%25)	20	5	6	0,300
		10	5	0,250
	40	5	10	0,250
		10	10	0,250
	100	5	25	0,250
		10	24	0,240
LTS (%50)	20	5	7	0,350
		10	6	0,300
	40	5	13	0,325
		10	13	0,325
	100	5	34	0,340
		10	30	0,300
S ($b = 0.1$)	20	5	2	0,100
		10	2	0,100
	40	5	5	0,125
		10	6	0,150
	100	5	13	0,130
		10	15	0,150
S ($b = 0.5$)	20	5	7	0,350
		10	8	0,400
	40	5	16	0,400
		10	18	0,450
	100	5	41	0,410
		10	46	0,460

LTS (%25) tahmin edicisinin bozulma noktası 0.25, LTS (%50) tahmin edicisinin simülasyon sonrası bozulma noktasıda 0.5 olarak elde edildi. S tahmin edicilerinde standart sapmalar arttıkça bozulma noktaları da artmaktadır. S($b = 0.1$) tahmin edicisinin bozulma noktası yaklaşık 0.15, S($b = 0.5$) tahmin edicisinin bozulma noktası da yaklaşık 0.45 olarak bulundu.

Senaryo I de, yalnız $\mathbf{Y}$ vektör uzayındaki aykırı gözlemlerin var olması durumunda M, GM, LTS ve S tahmin edicileri bu aykırı değerlere karşı dirençli olduğundan bozulma noktaları yüksek çıkmaktadır. EKK tahmin edicisi de aykırı değerlerden aşırı derecede etkilendiğinden dolayı bozulma noktası düşük çıkmıştır.

6.2.2. Senaryo-II

Bu senaryoda, hem Y vektör uzayı hemde X uzayında tek guruplanmış aykırı değerler incelenmiştir. Tahmin edicilerin *Tespit yöntemiyle* simülasyonlar sonrası hesaplanan bozulma noktaları Tablo 6.16 ve Tablo 6.17 de verilmiştir.

Tablo 6.16. Senaryo-II için EKK, M ve GM tahmin edicilerin tespit yöntemiyle bozulma noktaları.

Tahmin Ediciler	Gözlem Sayısı (n)	Std. Sapma	m	ε^*
EKK	20	5	1	0,050
		10	1	0,050
	40	5	1	0,025
		10	5	0,125
	100	5	2	0,020
		10	11	0,110
M (Huber)	20	5	9	0,450
		10	9	0,450
	40	5	19	0,475
		10	20	0,500
	100	5	50	0,500
		10	50	0,500
M (Hampel)	20	5	9	0,450
		10	10	0,500
	40	5	20	0,500
		10	20	0,500
	100	5	50	0,500
		10	50	0,500
GM ($\alpha = 0$)	20	5	8	0,400
		10	8	0,400
	40	5	16	0,400
		10	16	0,400
	100	5	41	0,410
		10	42	0,420
GM ($\alpha = 1$)	20	5	9	0,450
		10	9	0,450
	40	5	17	0,425
		10	16	0,400
	100	5	43	0,430
		10	45	0,450

Bu senaryoda, EKK tahmin edicisinin düşük standart sapmalar da bozulma noktası düşük çıkmaktadır, ancak standart sapmalardaki büyüme bozulma noktasını yükseltmektedir. Bunun nedeni ele alınan modelin X örneklem uzayındaki aykırı

gözlemlerden fazla etkilenmemesidir. Yine Huber ve Hampel tahmin edicilerinde de bozulma noktaları 0.5 olmaktadır. GM tahmin edicileri M tahmin edicilerinden daha düşük bozulma noktasına sahip olmuştur.

Tablo 6.17. Senaryo-II için LTS ve S tahmin edicilerin tespit yöntemiyle bozulma noktaları.

Tahmin Ediciler	Gözlem Sayısı (n)	Std. Sapma	m	ε^*
LTS (%25)	20	5	3	0,150
		10	5	0,250
	40	5	8	0,200
		10	9	0,225
	100	5	22	0,220
		10	25	0,250
LTS (%50)	20	5	6	0,300
		10	8	0,400
	40	5	13	0,325
		10	16	0,400
	100	5	30	0,300
		10	33	0,330
S ($b = 0.1$)	20	5	1	0,050
		10	2	0,100
	40	5	3	0,075
		10	4	0,100
	100	5	7	0,070
		10	10	0,100
S ($b = 0.5$)	20	5	5	0,250
		10	6	0,300
	40	5	11	0,275
		10	12	0,300
	100	5	28	0,280
		10	30	0,300

LTS(%25) tahmin edicisinin bozulma noktası 0.25 olmasına karşın, LTS(%50) tahmin edicisinin simülasyon sonrası bozulma noktası yaklaşık 0.33 olarak elde edildi. S($b = 0.1$) tahmin edicisinin bozulma noktası 0.10, ancak S($b = 0.5$) tahmin edicisinin bozulma noktası I. senaryoya göre düşüş gösterdi ve yaklaşık 0.3 olarak tespit edildi.

6.2.3. Senaryo-III

Bu senaryoda, Y vektör uzayında ters yönlerde gruplanmış aykırı değerler ele alınmıştır. Tahmin edicilerin *Tespit yöntemiyle* simülasyonlar sonrası hesaplanan bozulma noktaları Tablo 6.18 ve Tablo 6.19 da verilmiştir.

Tablo 6.18. Senaryo-III için EKK, M ve GM tahmin edicilerin tespit yöntemiyle bozulma noktaları.

Tahmin Ediciler	Gözlem Sayısı (n)	Std. Sapma	m	ε^*
EKK	20	5	1	0,050
		10	1	0,050
	40	5	1	0,025
		10	1	0,025
	100	5	1	0,010
		10	1	0,010
M (Huber)	20	5	2	0,100
		10	3	0,150
	40	5	14	0,350
		10	15	0,375
	100	5	41	0,410
		10	43	0,430
M (Hampel)	20	5	3	0,150
		10	3	0,150
	40	5	15	0,375
		10	16	0,400
	100	5	44	0,440
		10	45	0,450
GM ($\alpha = 0$)	20	5	4	0,200
		10	5	0,250
	40	5	16	0,400
		10	16	0,400
	100	5	45	0,450
		10	47	0,470
GM ($\alpha = 1$)	20	5	5	0,250
		10	6	0,300
	40	5	17	0,425
		10	18	0,450
	100	5	48	0,480
		10	49	0,490

EKK tahmin edicisi diğer senaryolarda olduğu gibi aykırı değerlerden fazlasıyla etkilenmiştir. Huber ve Hampel tahmin edicilerinde de bozulma noktaları Senaryo I ve II deki değerlerine göre biraz daha düşüktür. Yine GM tahmin

edicilerinin ilk iki senaryoya göre bozulma noktalarında biraz düşüş gözle çarpılmaktadır.

Tablo 6.19. Senaryo-III için LTS ve S tahmin edicilerin tespit yöntemiyle bozulma noktaları.

Tahmin Ediciler	Gözlem Sayısı (n)	Std. Sapma	m	ε^*
LTS (%25)	20	5	1	0,050
		10	1	0,050
	40	5	3	0,075
		10	4	0,100
	100	5	7	0,070
		10	8	0,080
LTS (%50)	20	5	2	0,100
		10	3	0,150
	40	5	5	0,125
		10	7	0,175
	100	5	14	0,140
		10	15	0,150
S ($b = 0.1$)	20	5	1	0,050
		10	1	0,050
	40	5	3	0,075
		10	4	0,100
	100	5	7	0,070
		10	11	0,110
S ($b = 0.5$)	20	5	5	0,250
		10	7	0,350
	40	5	14	0,350
		10	18	0,450
	100	5	42	0,420
		10	46	0,460

LTS(%25) ve LTS(%50) tahmin edicilerinin bozulma noktaları daha önceki senaryolardan oldukça düşüktür. Bu senaryodaki genel olarak S tahmin edicisi hariç bozulma noktalarındaki düşüşün sebebi aynı doğrultuda manivela etkisi yapan iki grubun var olmasıdır. Bundan dolayı bu aykırı değerler tahmin edicilerin tahmin değerlerini etkisi altına almaktadır. S tahmin edicilerinin bozulma noktaları yaklaşık 0.1 ve 0.46 olarak elde edildi.

6.2.4. Senaryo-IV

Bu senaryoda, hem Y vektör uzayı hemde X uzayında ters yönlerde guruplanmış aykırı değerler incelenmektedir. Tahmin edicilerin *Tespit yöntemiyle* simülasyonlar sonrası hesaplanan bozulma noktaları Tablo 6.20 ve Tablo 6.21 de verilmiştir.

Tablo 6.20. Senaryo-IV için EKK, M ve GM tahmin edicilerin tespit yöntemiyle bozulma noktaları.

Tahmin Ediciler	Gözlem Sayısı (n)	Std. Sapma	m	ε^*
EKK	20	5	1	0,050
		10	1	0,050
	40	5	1	0,025
		10	1	0,025
	100	5	1	0,010
		10	1	0,010
M (Huber)	20	5	8	0,400
		10	9	0,450
	40	5	20	0,500
		10	20	0,500
	100	5	50	0,500
		10	50	0,500
M (Hampel)	20	5	9	0,450
		10	9	0,450
	40	5	19	0,475
		10	20	0,500
	100	5	50	0,500
		10	50	0,500
GM ($\alpha = 0$)	20	5	9	0,450
		10	10	0,500
	40	5	17	0,425
		10	18	0,450
	100	5	48	0,480
		10	49	0,490
GM ($\alpha = 1$)	20	5	8	0,400
		10	8	0,400
	40	5	19	0,475
		10	20	0,500
	100	5	49	0,490
		10	50	0,500

Bu senaryodaki sonuçlar Senaryo I deki sonuçlarla benzerlik göstermektedir. n örneklem sayısı arttıkça M ve GM tahmin edicileri 0.5 bozulma noktası değerine yakınsamaktadır.

Tablo 6.21. Senaryo-IV için LTS ve S tahmin edicilerin tespit yöntemiyle bozulma noktaları.

Tahmin Ediciler	Gözlem Sayısı (n)	Std. Sapma	m	ε^*
LTS (%25)	20	5	2	0,100
		10	3	0,150
	40	5	7	0,175
		10	7	0,175
	100	5	14	0,140
		10	15	0,150
LTS (%50)	20	5	6	0,300
		10	7	0,350
	40	5	13	0,325
		10	15	0,375
	100	5	32	0,320
		10	34	0,340
S ($b = 0.1$)	20	5	2	0,100
		10	2	0,100
	40	5	4	0,100
		10	5	0,125
	100	5	10	0,100
		10	12	0,120
S ($b = 0.5$)	20	5	3	0,150
		10	4	0,200
	40	5	9	0,225
		10	10	0,250
	100	5	27	0,270
		10	32	0,320

LTS(%50) tahmin edicisinin simülasyon sonrası bozulma noktası yaklaşık 0.34 olmasına karşın, LTS(%25) tahmin edicisinin bozulma noktası 0.15 olarak elde edildi. LTS(%25) deki bozulma noktaları Senaryo I'in değerlerine göre daha düşüktür. S($b = 0.1$) tahmin edicisinin bozulma noktası yaklaşık 0.12, S($b = 0.5$) tahmin edicisinin bozulma noktası da yaklaşık 0.3 olarak tespit edildi.

6.2.5. Senaryo-V

Bu senaryoda, X uzayında gruplanmış aykırı değerler incelenmektedir. Tahmin edicilerin *Tespit yöntemiyle* simülasyonlar sonrası hesaplanan bozulma noktaları Tablo 6.22 ve Tablo 6.23 de verilmiştir.

Tablo 6.22. Senaryo-V için EKK, M ve GM tahmin edicilerin tespit yöntemiyle bozulma noktaları.

Tahmin Ediciler	Gözlem Sayısı (n)	Std. Sapma	m	ε^*
EKK	20	5	5	0,250
		10	5	0,250
	40	5	9	0,225
		10	9	0,225
	100	5	21	0,210
		10	20	0,200
M (Huber)	20	5	10	0,500
		10	10	0,500
	40	5	20	0,500
		10	20	0,500
	100	5	50	0,500
		10	50	0,500
M (Hampel)	20	5	10	0,500
		10	10	0,500
	40	5	20	0,500
		10	20	0,500
	100	5	50	0,500
		10	50	0,500
GM ($\alpha = 0$)	20	5	10	0,500
		10	10	0,500
	40	5	20	0,500
		10	20	0,500
	100	5	50	0,500
		10	50	0,500
GM ($\alpha = 1$)	20	5	10	0,500
		10	10	0,500
	40	5	20	0,500
		10	20	0,500
	100	5	50	0,500
		10	50	0,500

Bu senaryoda tüm tahmin edicilerin bozulma noktalarında yükselme göze çarpmaktadır. Bunun nedeni modelin X örneklem uzayındaki aykırı gözlemlerden fazla etkilenmemesidir. EKK tahmin edicisinin bozulma değerleri, diğer tahmin

edicilere göre daha fazla artış göstermektedir. M ve GM tahmin edicileri 0.5 bozulma noktası değerine ulaşmışlardır.

Tablo 6.23. Senaryo-V için LTS ve S tahmin edicilerin tespit yöntemiyle bozulma noktaları.

Tahmin Ediciler	Gözlem Sayısı (n)	Std. Sapma	m	ε^*
LTS (%25)	20	5	4	0,200
		10	5	0,250
	40	5	9	0,225
		10	9	0,225
	100	5	24	0,240
		10	25	0,250
LTS (%50)	20	5	8	0,400
		10	9	0,450
	40	5	15	0,375
		10	17	0,425
	100	5	46	0,460
		10	48	0,480
S ($b = 0.1$)	20	5	3	0,150
		10	4	0,200
	40	5	6	0,150
		10	7	0,175
	100	5	13	0,130
		10	15	0,150
S ($b = 0.5$)	20	5	8	0,400
		10	9	0,450
	40	5	14	0,350
		10	17	0,425
	100	5	42	0,420
		10	48	0,480

LTS (%25) tahmin edicisinde bozulma noktası 0.25, LTS (%50) tahmin edicisinde bozulma noktaları 0.5 olarak elde edilmiştir. S tahmin edicilerinin bozulma noktaları sırasıyla yaklaşık olarak 0.15 ve 0.45 olarak bulundu.

6.2.6. Senaryo-VI

Bu senaryoda, hem Y vektör uzayı hemde X uzayında normal ancak artıklarda tek gruplanmış aykırı değerler incelenmektedir. Tahmin edicilerin *Tespit yöntemiyle* simülasyonlar sonrası hesaplanan bozulma noktaları Tablo 6.24 ve Tablo 6.25 de verilmiştir.

Tablo 6.24. Senaryo-VI için EKK, M ve GM tahmin edicilerin tespit yöntemiyle bozulma noktaları.

Tahmin Ediciler	Gözlem Sayısı (n)	Std. Sapma	m	ε^*
EKK	20	5	1	0,050
		10	1	0,050
	40	5	1	0,025
		10	1	0,025
	100	5	1	0,010
		10	1	0,010
M (Huber)	20	5	2	0,100
		10	3	0,150
	40	5	4	0,100
		10	6	0,150
	100	5	10	0,100
		10	12	0,120
M (Hampel)	20	5	2	0,100
		10	2	0,100
	40	5	4	0,100
		10	5	0,125
	100	5	10	0,100
		10	11	0,110
GM ($\alpha = 0$)	20	5	2	0,100
		10	2	0,100
	40	5	4	0,100
		10	4	0,100
	100	5	10	0,100
		10	10	0,100
GM ($\alpha = 1$)	20	5	2	0,100
		10	2	0,100
	40	5	5	0,125
		10	5	0,125
	100	5	11	0,110
		10	11	0,110

Tahmin ediciler bu senaryo da diğer senaryolara göre aykırı değerlerden fazlasıyla etkilenmişlerdir. Klasik bozulma noktası kavramına göre bu tahmin edicilerin EKK da dahil olmak üzere bozulma değerleri 1 olur. Ancak yeni tanımlanan *Tespit yöntemiyle* sadece **Y** ve **X** uzaylarında aykırı olanlarla değil ayrıca artıklardaki aykırı gözlemlere karşı tahmin edicilerin sağlamlığını tespit edebilmektedir.

Burada M ve GM tahmin edicilerin bozulma noktaları benzerlik göstermekte ve yaklaşık 0.11 değerine sahiptirler. Yine benzer biçimde LTS (%25) tahmin

edicisinde bozulma noktası 0.1, LTS (%50) tahmin edicisin de bozulma noktası yaklaşık olarak 0.14 olarak elde edilmiştir.

Tablo 6.25. Senaryo-VI için LTS ve S tahmin edicilerin tespit yöntemiyle bozulma noktaları.

Tahmin Ediciler	Gözlem Sayısı (n)	Std. Sapma	m	ε^*
LTS (%25)	20	5	2	0,100
		10	2	0,100
	40	5	4	0,100
		10	4	0,100
	100	5	10	0,100
		10	10	0,100
LTS (%50)	20	5	2	0,100
		10	2	0,100
	40	5	4	0,100
		10	5	0,125
	100	5	13	0,130
		10	14	0,140
S ($b = 0.1$)	20	5	1	0,050
		10	2	0,100
	40	5	2	0,050
		10	3	0,075
	100	5	10	0,100
		10	10	0,100
S ($b = 0.5$)	20	5	5	0,250
		10	6	0,300
	40	5	11	0,275
		10	13	0,325
	100	5	26	0,260
		10	29	0,290

S tahmin edicilerinin bozulma noktaları S($b = 0.1$) için 0.10, S($b = 0.5$) için yaklaşık olarak 0.30 bulundu.

6.2.7. Senaryo-VII

Bu senaryoda, $\mathbf{X}$ uzayında ters yönlerde gruplanmış aykırı değerler incelenmiştir. Tahmin edicilerin *Tespit yöntemiyle* simülasyonlar sonrası hesaplanan bozulma noktaları Tablo 6.26 ve Tablo 6.27 de verilmiştir.

Tablo 6.26. Senaryo-VII için EKK, M ve GM tahmin edicilerin tespit yöntemiyle bozulma noktaları.

Tahmin Ediciler	Gözlem Sayısı (n)	Std. Sapma	m	ε^*
EKK	20	5	1	0,050
		10	1	0,050
	40	5	1	0,025
		10	1	0,025
	100	5	1	0,010
		10	1	0,010
M (Huber)	20	5	9	0,450
		10	10	0,500
	40	5	20	0,500
		10	20	0,500
	100	5	50	0,500
		10	50	0,500
M (Hampel)	20	5	10	0,500
		10	10	0,500
	40	5	20	0,500
		10	20	0,500
	100	5	50	0,500
		10	50	0,500
GM ($\alpha = 0$)	20	5	10	0,500
		10	10	0,500
	40	5	20	0,500
		10	20	0,500
	100	5	50	0,500
		10	50	0,500
GM ($\alpha = 1$)	20	5	10	0,500
		10	10	0,500
	40	5	20	0,500
		10	20	0,500
	100	5	50	0,500
		10	50	0,500

Doğrusal regresyon da M tahmin edicileri, **X** uzayındaki aykırı değerlerin etkisi altında kalmaktaydı. Ancak doğrusal olmayan regresyonda model fonksiyonuna bağlı olarak bu durum farklılık göstermektedir. Nitekim bu senaryoda olduğu gibi **X** uzayındaki aykırı değerlerin olmasına karşın **Tespit yöntemiyle** M ve GM tahmin edicilerinin bozulma noktaları 0.5 olmuştur. LTS (%25) tahmin edicisi 0.25, LTS (%50) tahmin edicisinin bozulma noktası yaklaşık 0.48 bulunmuştur.

Tablo 6.27. Senaryo-VII için LTS ve S tahmin edicilerin tespit yöntemiyle bozulma noktaları.

Tahmin Ediciler	Gözlem Sayısı (n)	Std. Sapma	m	ε^*
LTS (%25)	20	5	4	0,200
		10	5	0,250
	40	5	9	0,225
		10	10	0,250
	100	5	25	0,250
		10	25	0,250
LTS (%50)	20	5	9	0,450
		10	10	0,500
	40	5	17	0,425
		10	18	0,450
	100	5	47	0,470
		10	48	0,480
S ($b = 0.1$)	20	5	2	0,100
		10	2	0,100
	40	5	4	0,100
		10	5	0,125
	100	5	9	0,090
		10	11	0,110
S ($b = 0.5$)	20	5	7	0,350
		10	9	0,450
	40	5	15	0,375
		10	18	0,450
	100	5	41	0,410
		10	45	0,450

S ($b = 0.1$) tahmin edicisinin bozulma noktası yaklaşık 0.11, S ($b = 0.5$) tahmin edicisinin bozulma noktası da yaklaşık olarak 0.45 hesaplanıldı.

6.2.8. Senaryo-VIII

Bu senaryoda, hem Y vektör uzayı hemde X uzayında iki grup aykırı değerler ele alınmıştır. Tahmin edicilerin *Tespit yöntemiyle* simülasyonlar sonrası hesaplanan bozulma noktaları Tablo 6.28 ve Tablo 6.29 de verilmiştir.

Tablo 6.28. Senaryo-VIII için EKK, M ve GM tahmin edicilerin tespit yöntemiyle bozulma noktaları.

Tahmin Ediciler	Gözlem Sayısı (n)	Std. Sapma	m	ε^*
EKK	20	5	1	0,050
		10	1	0,050
	40	5	1	0,025
		10	1	0,025
	100	5	2	0,020
		10	2	0,020
M (Huber)	20	5	9	0,450
		10	10	0,500
	40	5	18	0,450
		10	19	0,475
	100	5	48	0,480
		10	48	0,480
M (Hampel)	20	5	10	0,500
		10	10	0,500
	40	5	18	0,450
		10	18	0,450
	100	5	48	0,480
		10	49	0,490
GM ($\alpha = 0$)	20	5	10	0,500
		10	10	0,500
	40	5	19	0,475
		10	20	0,500
	100	5	49	0,490
		10	49	0,490
GM ($\alpha = 1$)	20	5	10	0,500
		10	10	0,500
	40	5	20	0,500
		10	20	0,500
	100	5	49	0,490
		10	50	0,500

Bu senaryodaki sonuçlar Senaryo I ve IV deki sonuçlarla benzerlik göstermektedir. M ve GM tahmin edicilerinin bozulma noktaları yaklaşık olarak 0.5 elde edilmiştir.

Tablo 6.29. Senaryo-VIII için LTS ve S tahmin edicilerin tespit yöntemiyle bozulma noktaları.

Tahmin Ediciler	Gözlem Sayısı (n)	Std. Sapma	m	ε^*
LTS (%25)	20	5	2	0,100
		10	2	0,100
	40	5	7	0,175
		10	8	0,200
	100	5	20	0,200
		10	21	0,210
LTS (%50)	20	5	3	0,150
		10	4	0,200
	40	5	12	0,300
		10	13	0,325
	100	5	28	0,280
		10	30	0,300
S ($b = 0.1$)	20	5	2	0,100
		10	3	0,150
	40	5	5	0,125
		10	7	0,175
	100	5	17	0,170
		10	19	0,190
S ($b = 0.5$)	20	5	7	0,350
		10	9	0,450
	40	5	15	0,375
		10	18	0,450
	100	5	39	0,390
		10	46	0,460

LTS (%25) tahmin edicisi 0.20, LTS (%50) tahmin edicisinin bozulma noktası da yaklaşık olarak 0.30 olmuştur. S($b = 0.1$) tahmin edicisinin bozulma noktası yaklaşık 0.19, S($b = 0.5$) tahmin edicisinin bozulma noktası da yaklaşık olarak 0.46 olarak hesaplanıldı.

7. SONUÇ VE ÖNERİLER

Bu çalışmada mevcut bozulma noktası tanımlarının, bilgisayarda hesaplanması yönünde yetersiz kalmasından dolayı bozulma noktasına yeni bir yaklaşım getirilmiştir. *Tespit yöntemi* adı verilen bu yöntem sayesinde, bilgisayar ortamında bozulma noktasının hesaplanması mümkün kılındı, ayrıca tahmin edicilerin bozulma noktaları yönünden sıralanabilmesine de olanak tanınmıştır.

Tablo 7.1. Tüm Senaryolardaki tahminin edicilerin tespit yöntemiyle bozulma noktaları.

Tahmin Ediciler	Senaryolardaki Bozulma Noktaları							
	S1	S2	S3	S4	S5	S6	S7	S8
EKK	0,040	0,110	0,010	0,010	0,200	0,010	0,010	0,020
M (Huber)	0,500	0,500	0,430	0,500	0,500	0,120	0,500	0,480
M (Hampel)	0,500	0,500	0,450	0,500	0,500	0,110	0,500	0,490
GM ($\alpha = 0$)	0,500	0,420	0,470	0,490	0,500	0,100	0,500	0,490
GM ($\alpha = 1$)	0,500	0,450	0,490	0,500	0,500	0,110	0,500	0,500
LTS (%25)	0,240	0,250	0,080	0,150	0,250	0,100	0,250	0,210
LTS (%50)	0,300	0,330	0,150	0,340	0,480	0,140	0,480	0,300
S ($b = 0.1$)	0,150	0,100	0,110	0,120	0,150	0,100	0,110	0,190
S ($b = 0.5$)	0,460	0,300	0,460	0,320	0,480	0,290	0,450	0,460

Yapılan simülasyon çalışmalarından *Tespit yöntemi* ile tahmin edicilerin bozulma noktaları, gözlem sayıları 100 ve standart sapmaları 10 olan durumları, Tablo 7.1 de verilmiştir. Senaryo III ve Senaryo VI daki bozulma noktalarının diğer senaryolardakinden $\alpha = 0.05$ anlam seviyesinde farklı olduğu görülmüştür. Bunun nedeni aykırı gözlem tespitinde kullanılan Tablo 7.1 deki RD yönteminin, Senaryo III ve VI da başarı oranının diğer senaryolara göre farklılık göstermesidir.

Simülasyon sonuçlarında en çok göze çarpan bir durum da Senaryo VI daki sonuçlar olmuştur. **Tespit yöntemi** tüm aykırı gözlemlere karşı tahmin edicinin sağlamlığını ölçebilmesinin yanında, sadece artıklardaki aykırı gözlemlere karşı da tahmin edicilerin sağlamlığını ölçebilmektedir.

Bu tez çalışması esnasında Delphi görsel programlama diliyle kodlanan **Selçuk Stat** istatistik paket programının kodları kullanılarak, bu paket programda **Analiz – Regresyon - Doğrusal Olmayan Regresyon** arayüzünde özel opsiyonlar ilave edilmiştir. Bu opsiyonlar sayesinde kullanıcı dilediği regresyon modelini, tahmin ediciyi ve bozulma noktaları opsiyonunu seçerek bu model ve bu tahmin ediciye göre bozulma noktalarını hesaplayabilecektir.

Simülasyonlarda **Tespit yöntemi** ile $k_i \neq m$ değerinin hesaplanması sırasında bir düzenleme yapılmıştır. $\hat{\theta}$ tahmin edicisine bağlı olarak hesaplanan artıklara RD yöntemiyle aykırı gözlemlerin tespiti yapıldığında elde edilen aykırı gözlemlerin sayısı bazen m değerinden büyük olmuştur. Bundan dolayı RD yöntemiyle bulunan aykırı gözlemlerin sayısı λ olmak üzere $k_i \neq m$ ifadesi için %10 luk bir tölans tanınmıştır. $\lambda > k$ durumlarında yöntem tüm aykırı değerleri doğru tespit etmiş ancak fazladan bazı gözlemleride beraberinde aykırı gözlem olarak tespit ettiğinden $k_i = m$ yerine $m \leq \lambda \leq \lceil m + m/10 \rceil$ olduğu durumlar başarı olarak alınmıştır.

Şu an için kullanılan yöntemler ve işlemlerin fazla olmasından dolayı bilgisayarda **Tespit yöntemi** ile bozulma noktasının hesaplama süresi çok uzun olmaktadır. Tezdeki 8 ayrı senaryodaki tüm simülasyonlar aynı anda 4 farklı bilgisayar tarafından yaklaşık olarak bir haftada hesaplanmıştır. Daha sonra yapılacak olan çalışmalarla mevcut algoritmaların iyileştirilmesi ve geliştirilmesi bu süreyi azaltabilir.

8. KAYNAKLAR

1. Adnan, R., Mohamad, M. N. ve Setan, H. (2003), *Multiple Outliers Detection Procedures in Linear Regression*, Matematika, vol. 19, no. 1, 29-45.
2. Agostino, R. B. D. ve Stephens, M. A. (1986), *Goodness of Fit Techniques*, Marcel Dekker Inc., New York.
3. Ammemiya, T. (1982), *Two Stage Least Absolute Deviations Estimators*, Econometrica, vol. 50, n0. 3, pp. 689-710.
4. Andrews, D. F., Bickel, P. J., Hampel, F. R., Huber, P. J., Rogers, W. H. ve Tukey, J. W. (1972), *Robust Estimates of Location: Survey and Advances*, Princeton University Press, Princeton, New Jersey.
5. Atkinson, A.C. ve Cheng, T.C. (1999), *Computing Least Trimmed Squares Regression with the Forward Search*, Statistics and Computing, vol. 9, pp. 251-263.
6. Balcı, M. (2000), *Reel Analiz*, Bilim Kitap Kırtasiye LTD. ŞTİ., Ankara.
7. Barrera, M.S. ve Yohai V.J. (2006), *A Fast Algorithm for S-Regression Estimates*, Journal of Computational and Graphical Statistics, vol. 15, no.2, pp 1-14.
8. Bates, D. M. ve Watts, D. G. (1980), *Relative Curvature Measures of Nonlinearity*”, J. R. Statist. Soc. B 42, pp. 1-25.
9. Bates, D. M. ve Watts, D. G. (1988), *Nonlinear Regression Analysis and Its Applications*, John Willey&Sons, Inc., USA.
10. Berrendero, J. R., Mendes, B. V. M. ve Tyler, D. E. (2007), *On the Maximum Bias Functions of MM-Estimates and Constrained M Estimates of Regression*, The Annals of Statistics, vol. 35, no. 1, 13-40.
11. Billingsley, P., Belsley, D. A., Kuh, E. ve Welsch, R. E. (1980), *Regression diagnostics: identifying influential data and sources of collinearity*, John Willey&Sons, Inc., USA.
12. Box, M. J. (1971), *Bias in Nonlinear Estimation*, J. R. Statist. Soc. Vol. 32, pp. 171-201.

13. Campbell, N.A., Lopuha, H.P. ve Rousseeuw, P.J. (1998), *On the Calculation of A Robust S-Estimator of a Covariance Matrix*, Statist. Med. vol. 17, pp. 2685-2695.
14. Chen, Y. L. (1996), *The Nonlinear Least Trimmed Squares Regression Estimator*, PhD thesis, University of Kentucky.
15. Cho, J. J., Chen, Y. ve Ding, Y. (2009), *Calculating the Breakdown Point of Sparse Linear Models*, American Statistical Association and the American Society for Quality, Technometrics, vol. 51, no. 1, 34-46.
16. Coakley, C. W. ve Hettmansperger, T. P. (1992), *Breakdown Bounds and Expected Test Resistance*, J. Nonparametr. Statist. 1 267–276.
17. Cook, R. D. (1979), *Influential Observations in Linear Regression*, Journal of the American Statistical Association, vol. 74, 169-174.
18. Cook, R. D. ve Tsai, C.L. (1985), *Residuals in Nonlinear Regression*, Biometrika, 72, 1, 23-29.
19. Cook, R. D., Tsai, C.L. ve Wei, B.C. (1986), *Bias in Nonlinear Regression*, Biometrika, 73, 3, 615-623.
20. Cox, D. R. ve Snell, E. J. (1968), *A General Definition of Residuals*, J.R. Statist. Soc. B 30, 248-275.
21. Croux, C., Rousseeuw, P.J. ve Hössjer, O. (1994), *Generalized S-Estimators*, Journal of the American Statistical Association, vol. 89, no. 428, 1271-1281.
22. Croux, C. (1998), *Limit Behavior of the Empirical Influence Function of the Median*, Statistics and Probability Letters, 37, 331–340.
23. Davison, A. C. ve Hinkley, D. V. (1997), *Bootstrap Methods and Their Application*, Cambridge: Cambridge University Pres.
24. Donoho, D.L. (1982), *Breakdown Properties of Multivariate Location Estimators*, Ph.D. qualifying paper, Dept. Statistics, Harvard Uni.
25. Donoho, D.L. ve Huber, P.J. (1983), *The Notion of Breakdown Point*, in A Festschrift for Erich Lehmann (P.J. Bickel, K. A. Doksum ve J. L. Hodges, eds.) 157-184. Wadsworth, Belmont, CA.
26. Duncan, G. T. (1978), *An Empirical Study of Jackknife Constructed Confidence Regions in Nonlinear Regression*, Technometrics, vol. 20, 123-129.

27. Efron, B. (1979), *Bootstrap Methods: Another Look at the Jackknife*, The Annals of Statistics, vol. 7, 1-26.
28. Efron, B. ve Tibshirani, R. J. (1993), *An Introduction to the Bootstrap*, Chapman & Hall, London.
29. Gallant, A. R. ve Fuller, W. A. (1973), *Fitting Segmented Polynomial Regression Models whose Join Points have to be Estimated*, Journal of the American Statistical Association, 68, 144-147.
30. Gallant, A. R. (1975), *Nonlinear Regression*, Journal of the American Statistical Association, 29, 73-81.
31. Gallant, A. R. (1977), *Testing a Nonlinear Regression Specification: a Nonregular Case*, Journal of the American Statistical Association, 72, 523-530.
32. Gather, U. ve Hilker, T. (1997), *A Note on Tyler's Modification of The Mad for the Stahel-Donoho Estimator*, The Annals of Statistics, vol. 25, no. 5, 2024-2026.
33. Genç A. (1997), Çok Değişkenli Lineer Olmayan Modeller: Parametre Tahmini ve Hipotez Testleri, Ankara Üniversitesi Fen Bilimleri Enstitüsü, Doktora Tezi (basılmamış), Ankara.
34. Genton, M. G. ve Lucas, A. (2003), *Comprehensive Definitions of Breakdown Points for Independent and Dependent Observations*, J. R. Statist. Soc. vol. 65, 81-94.
35. Gillis, P. R. ve Ratkowsky, D. A. (1978), *The Behavior of Estimators of the Parameters of Various Yield-Density Relationships*, Biometrics vol. 34, 191-198.
36. Giloni, A. ve Simonoff, J. S. (2005), *The Conditional Breakdown Properties of Least Absolute Value Local Polynomial Estimators*, Nonparametric Statistics, vol. 17, no. 1, 15-30.
37. Hadi, A., and Simonoff, J. (1993), *Procedures for the Identification of Multiple Outliers in Linear Models*, Journal of the American Statistical Association, 88, 1264-1272.
38. Hampel, F. R. (1968), *Contributions to the Theory of Robust Estimation*, Ph.D. dissertation, Dept. Statistics, Univ. California, Berkeley.

39. Hampel, F. R. (1971), *A general Qualitative Definition of Robustness*, Annals of Mathematical Statistics, vol. 42, 1887–1896.
40. Hampel, F. R. (1973), *Robust Estimation: A Condensed Partial Survey*, Z. Wahrscheinlichkeitstheorie Verw. Gebiete 27, 87-104.
41. Hampel, F. R. (1974), *The Influence Curve and Its Role in Robust Estimation*, Journal of the American Statistical Association, vol. 69, no. 346, 383-393.
42. Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J. ve Stahel, W. A. (1986), *Robust Statistics: The Approach Based on Influence Functions*, John Willey&Sons, Inc., New York.
43. Hartley, H. O. ve Booker, A. (1963), *Non-linear Least Squares Estimation*, Unpublished Report, Iowa State University.
44. Hawkins, D.M. (1994), *The Feasible Solution Algorithm for Least Trimmed Squares Regression*, Computatioanal Statistics and Data analysis, vol. 17, 185-196.
45. He, X., Simpson, D. G. ve Portnoy, S. L. (1990), *Breakdown Robustness of Tests*, J. Amer. Statist. Assoc. 85 446–452.
46. Hoaglin, D. C. ve Welsch, R. E. (1978), *The Matrix in Regression and ANOVA*, Am. Statist. 32, 17-22.
47. Hodges, J. L. (1967), *Efficiency in Normal Samples and Tolerance of Extreme Values for Some Estimates of Location*, Proc. Fifth Berkeley Symp. Math. Statist. Probab. 1 163-186, Univ. California Press, Berkeley
48. Huber, P. J. (1964), *Robust Estimation of a Location Paremater*, Ann. Math. Statist. vol 36, 73-101.
49. Huber, P. J. (1973), *Robust regression: Asymptotics Conjectures and Monte Carlo*, The Annals of Statistics, 1, 799-821.
50. Huber, P. J. (1977), *Robust Statistical Procedures*, Society for Industrial and Applied Mathematics, Philadelphia.
51. Huber, P. J. (1981), *Robust Statistics*, John Willey&Sons, Inc., USA.
52. Huber, P. J. ve Ronchetti, E. M. (2009), *Robust Statistics*, John Willey&Sons, Inc., USA.

53. Jureckova, J., Koenker, R. ve Portnoy, S. L. (1990), *Tail Behavior of Regression Estimators and Their Breakdown Points*, *Econometrica* 58 1195–1214.
54. Laurent, R. T. S. ve Cook, R. D. (1992), *Leverage and Superleverage in Nonlinear Regression*, *American Statistical Association Journal of the American Statistical Association*, vol. 87, no. 420, pp. 985-990.
55. Li, Y. ve Arce, G. R. (2004), *A Fast Maximum Likelihood Estimation Approach to LAD Regression*, *ICASSP'04 IEEE International Conference*, vol. 2, 889-892.
56. Lopuha, H. P. (1992), *Highly Efficient Estimators of Multivariate Location with High Breakdown Point*, *Ann. Statist.* 20 398–413.
57. Mahalanobis, P. C. (1936), *On the Generalized Distance in Statistics*, *The National Institute of Science of India*, vol. 2, no. 1, 49-55.
58. Mallows, C. L. (1975), *Some comments on Cp*, *Technometrics*, 15:661.
59. Maronna, R.A., Martin, R.D. ve Yohai, V.J. (2006), *Robust Statistics Theory and Methods*, John Wiley&Sons, Inc., USA.
60. Marquardt, D.W. (1963), *An Algorithm for Least-Squares Estimation of Nonlinear Parameters*, *Journal Soc. Indust. Appl. Math.*, vol. 11, no. 2, 431-441.
61. Martin, R. D., Yohai, V. J. ve Zamar, R. H. (1989), *Min-Max Bias Robust Regression*, *The Annals of Statistics*, vol. 17, 1608-1630.
62. Melouna, M. ve Militky, J. (2001), *Tutorial: Detection of Single Influential Points in OLS Regression Model Building*, *Analytica Chimica Acta* 439, 169–191.
63. Midi, H. (1999), *Preliminary Estimators for Robust Non-linear Regression Estimation*, *Journal of Applied Statistics*, vol. 26, no. 5, pp. 591-600.
64. Mosteller, F. ve Tukey, J. W. (1977), *Data Analysis and Regression*, Addison-Wesley, Reading, MA.
65. Motulsky, H.J ve Brown, R.E. (2006), *Detecting Outliers when Fitting Data with Nonlinear regression – A new Method Based on Robust Nonlinear Regression and the False Discovery Rate*, *BMC, Bioinformatics*, 1471-2105/7/123.

66. Okafor, R. (1990), *A Biweight Approach to Estimate Currency Exchange Rate: The Nigerian Example*, Journal of the American Statistical Association, 17 73–82.
67. Öztürk, F., Akdi, Y., Aydoğdu, H. ve Karabulut, İ. (2006), Parametre Tahmini ve Hipotez Testi, Bıçaklar kitabevi, Ankara.
68. Pekgör A. (2004), Delphi ile İstatistiksel Hesaplamalar ve Veri Analizi: **Selçuk STAT**, Selçuk Üniversitesi Fen Bilimleri Enstitüsü, Yüksek Lisans Tezi (basılmamış), Konya.
69. Quenouille, M.H., (1949), *Approximate Tests of Correlation in Time Series*, Journal of The Royal Statistical Society, Series B, 11, 68-84.
70. Ramsay, J. O. (1977), *Maximum Likelihood Estimation in Multi Dimensional Scaling*, psychometrika, 42, 241-266.
71. Rocke D. M., Woodruff D.L. (1995), *Robust Estimation of Multivariate Location and Shape*, Journal of Statistical Planning and Inference vol. 57, 245-255.
72. Rousseeuw, P. J. (1984), *Least Median of Squares Regression*, Journal of the American Statistical Association, vol. 79, 871-880.
73. Rousseeuw, P. J. (1994), *Unconventional Features of Positive-Breakdown Estimators*, Statist. Probab. Lett. 19, 417–431.
74. Rousseeuw, P. J. ve Yohai, V. J. (1984), *Robust Regression by Means of S-Estimator in Robust and Nonlinear Time Series Analysis*, eds. J. Franke, W. Hardle and R. D. Martin, (Lecture Notes in Statistics), Springer - Verlag, New York, pp. 256-272.
75. Rousseeuw, P. J. ve Leroy, A.M. (1987), *Robust Regression and Outlier Detection*, John Willey&Sons, Inc., USA.
76. Rousseeuw, P. J. ve Zomeren, B. C. (1990), *Unmasking Multivariate Outliers and Leverage Points*, Journal of the American Statistical Association, vol. 85, no. 411, 633-639.
77. Rousseeuw, P. J. ve Driessen, K.V. (1999), *A Fast Algorithm for the Minimum Covariance Determinant Estimator*, Technometrics, vol. 41, no. 3, 213-223.

78. Ruppert, D. (1992), *Computing S Estimator for Regression and Multivariate Location/Dispersion*, American Statistical Association, Institute of Mathematical Statistics, vol. 1, no. 3, pp. 253-270.
79. Sakata, S. ve White, H. (1995), *An Alternatif Defination of Finite Sample Breakdown Point with Applications to Regression Model Estimators*, Journal of the American Statistical Association, vol. 90, no. 431, pp. 1099-1106.
80. Sakata, S. ve White H. (1998), *High Breakdown Point Conditional Dispersion Estimation with Application to S&P 500 Daily Returns Volatility*, Econometrica, vol. 66, no. 3, pp. 529-567.
81. Seber, G. A. F. ve Wild, C. J. (1989), *Nonlinear Regression*, John Willey&Sons, Inc., USA.
82. Serbert, D.M., Montgomery, D.C. ve Rollier, D. (1998), *A Clustering Algorithm for Identifying Multiple Outliers in Linear Regression*, Computational Statistics & Data Analysis, vol. 27, pp. 461-484.
83. Staudte, R. G. ve Sheather, S. J. (1990), *Robust Estimation and Testing*, John Wiley & Sons, Inc., New York.
84. Stromberg, A.J. ve Ruppert, D. (1992), *Breakdown in Nonlinear Regression*, Journal of the American Statistical Association, vol. 87, no. 420, pp. 991-997.
85. Stromberg, A.J. (1993), *Computation of High Breakdown Nonlinear Regression Parameters*, American Statistical Association Journal of the American Statistical Association, vol. 88, no. 421, pp. 237-244.
86. Tukey, J.W., (1958), *Bias and confidence in not quite large samples* (Abstract), Annals of Mathematical Statistics, 29, 614.
87. Visscher P.M., Thompson R. ve Haley C.S. (1996). *Confidence Intervals in QTL Mapping by Bootstrapping*, Genetics, Vol:143 (2), 1013-1020.
88. Wasserman, L. (2006), *All of Nonparametric Statistics*, Springer Science, Business Media, Inc., USA.
89. Weisberg, S. (1985), *Applied Linear Regression*, Second Edition, John Willey&Sons, Inc., USA.
90. Wu, C. F. J. (1986), *Jackknife, Bootstrap and Other Resampling Methods in Regression Analysis*, (Invited Paper), The Annals os Statistics, vol. 14, no 4, 1261- 1295.

91. Wu, J. W. ve Lee, W.C. (2006), *Computational Algorithm of Least Absolute Deviation Method for Determining Number of Outliers Under Normality*, Applied Mathematics and Computation, vol. 175, pp. 609-617.
92. Yohai, V.J. (1987), *High Breakdown Point and High Efficiency Robust Estimates for Regression*, The Annals of Statistics, vol. 15, no. 20, 642-656.
93. Zhang, J. (1996), *The Sample Breakdown Points of Tests*, J. Statist. Plann. Inference 52, 161–181.
94. Zuo, Y. (2001), *Some Quantitative Relationships Between two Types of Finite Sample breakdown Point*, Statistics&Probability Letters, vol. 51, 369-375.