

**ÇOKLU REGRESYONDA ETKİLİ GÖZLEMLERİN SAPTANMASI İÇİN
KULLANILAN TANI YÖNTEMLERİNİN KARŞILAŞTIRILMASI**

Irmak ACARLAR

**YÜKSEK LİSANS TEZİ
İSTATİSTİK**

**GAZİ ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ**

**HAZİRAN 2009
ANKARA**

Irmak ACARLAR tarafından hazırlanan ÇOKLU REGRESYONDA ETKİLİ GÖZLEMLERİN SAPTANMASI İÇİN KULLANILAN TANI YÖNTEMLERİNİN KARŞILAŞTIRILMASI adlı bu tezin Yüksek Lisans tezi olarak uygun olduğunu onaylarım.

Prof. Dr. Hamza GAMGAM

Tez Danışmanı, İstatistik Anabilim Dalı

Bu çalışma, jürimiz tarafından oy birliği ile İstatistik Anabilim Dalında Yüksek Lisans tezi olarak kabul edilmiştir.

Prof. Dr. Semra ERBAŞ

İstatistik Anabilim Dalı, Gazi Üniversitesi

Prof. Dr. Hamza GAMGAM

İstatistik Anabilim Dalı, Gazi Üniversitesi

Prof. Dr. Aydın ÜNSAL

Ekonometri Anabilim Dalı, Gazi Üniversitesi

Tarih: 09/06/2009

Bu tez ile G.Ü. Fen Bilimleri Enstitüsü Yönetim Kurulu Yüksek Lisans derecesini onamıştır.

Prof. Dr. Nail ÜNSAL

Fen Bilimleri Enstitüsü Müdürü

TEZ BİLDİRİMİ

Tez içindeki bütün bilgilerin etik davranış ve akademik kurallar çerçevesinde elde edilerek sunulduğunu, ayrıca tez yazım kurallarına uygun olarak hazırlanan bu çalışmada bana ait olmayan her türlü ifade ve bilginin kaynağına eksiksiz atıf yapıldığını bildiririm.

Irmak ACARLAR

**ÇOKLU REGRESYONDA ETKİLİ GÖZLEMLERİN SAPTANMASI İÇİN
KULLANILAN TANI YÖNTEMLERİNİN KARŞILAŞTIRILMASI
(Yüksek Lisans Tezi)**

Irmak ACARLAR

**GAZİ ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ
Haziran 2009**

ÖZET

Regresyonda etkili gözlem ve gözlem grupları, tahmin değerlerinde önemli derecede farklılaşmalara neden olabilir. Bu farklılaşmalar modelin açıklanabilirliğini azalttığı için verideki etkili gözlem veya gözlem gruplarının saptanması regresyon analizinin verimliliği açısından önemlidir. Bu çalışmada etkili gözlem ve gözlem gruplarının saptanması için kullanılan DFFITS, DFBETAS, COVRATIO, Cook Uzaklığı, S tanı istatistikleri ve grafik yöntemi incelenmiştir. Bu yöntemler etkili bir gözlem veya iki gözlemden oluşan etkili bir gözlem grubu içeren veride bu gözlem ve gözlem grubunu etkili gözlem olarak saptama oranı bakımından karşılaştırılmıştır.

Bilim Kodu : 205.1.066
Anahtar Kelimeler : Etkili gözlem, tanı istatistikleri
Sayfa Adedi : 68
Tez Yöneticisi : Prof. Dr. Hamza GAMGAM

**COMPARISON OF THE DIAGNOSTIC METHODS USED FOR
DETECTING INFLUENTIAL OBSERVATIONS IN MULTIPLE
REGRESSION
(M.Sc. Thesis)**

Irmak ACARLAR

**GAZİ UNIVERSITY
INSTITUTE OF SCIENCE AND TECHNOLOGY
June 2009**

ABSTRACT

In regression, an influential observation and influential sets would cause noticeable differentiations on fitted values. Since these differentiations decrease explainable of model, detecting the influential observation or the influential sets in data is important for efficiency of regression analysis. In this study DFFITS, DFBETAS, COVRATIO, Cook Distance, S statistics and graphical technique used for detecting influential observation or influential sets are examined. These methods are compared with regard to ratios of detecting influential observation and influential set in data which includes an influential observation or influential sets.

**Science Code : 205.1.066
Key Words : Influential observation, diagnostics
Page Number : 68
Adviser : Prof. Dr. Hamza GAMGAM**

TEŞEKKÜR

Çalışmalarım boyunca değerli yardım ve katkılarıyla beni yönlendiren hocam Prof. Dr. Hamza GAMGAM' a, yardımlarını esirgemeyen hocam Yrd. Doç. Dr. Fikri GÖKPINAR' a, programa başlamam konusunda beni yüreklendiren, maddi ve manevi her türlü imkânı sağlayan aileme teşekkürlerimi sunarım.

İÇİNDEKİLER

Sayfa

ÖZET	iv
ABSTRACT	v
TEŞEKKÜR	vi
İÇİNDEKİLER.....	vii
ÇİZELGELERİN LİSTESİ	ix
ŞEKİLLERİN LİSTESİ	x
SİMGELER VE KISALTMALAR	xi
1. GİRİŞ	1
2. TEMEL KAVRAMLAR.....	6
2.1. Doğrusal Regresyon Modeli	6
2.2. Hat Matrisi	7
2.3. Etki Eğrisi ve Etki Eğrisinin Örnek Uyarlamaları	8
2.3.1. Etki eğrisi	8
2.3.2. Etki eğrisinin örnek uyarlamaları	10
2.4. Studentleştirilmiş Artıklar	12
2.4.1. Dâhili studentleştirilmiş artıklar	12
2.4.2. Harici studentleştirilmiş artıklar (RSUDENT).....	13
3. ETKİLİ GÖZLEMLERİN SAPTANMASI İÇİN TANI İSTATİSTİKLERİ	15
3.1. Bir Etkili Gözlemi Saptamaya Dayalı Tanı İstatistikleri.....	16
3.1.1. DFBETAS tanı istatistiği	16

Sayfa

3.1.2. DFFITS tanı istatistiği.....	18
3.1.3. Pena' nın S_i tanı istatistiği.....	20
3.2. Hem Bir Etkili Gözlemi Hem de Etkili Gözlem Gruplarını Saptamaya Dayalı Tanı İstatistikleri	22
3.2.1. Cook uzaklığı	23
3.2.2. Kovaryans oranı (COVRATIO)	27
3.3. Maskeleye ve Yanılgıya-Düşürme.....	29
4. ÇOKLU REGRESYONDA ETKİLİ GÖZLEMLERİN SAPTANMASI İÇİN GRAFİKSEL YÖNTEM.....	33
5. SİMÜLASYON ÇALIŞMASI	45
5.1. Bağımsız Değişken Sayısı ve Örnek Hacminin Farklı Değerleri için Tanı Yöntemlerinin Karşılaştırılması	48
5.1.1. Bir etkili gözlem için tanı yöntemlerinin karşılaştırılması.....	48
5.1.2. Etkili gözlem grupları için tanı yöntemlerinin karşılaştırılması.....	53
5.2. Bağımsız Değişken Sayısı ve Örnek Hacmi Sabitken Tanı Yöntemlerinin Karşılaştırılması.....	57
6. SONUÇ VE ÖNERİLER	60
KAYNAKLAR.....	64
EKLER.....	66
EK-1 Grafik yönteminde $w_l^{(j)}$ vektörünün hesaplanması için algoritma	67
ÖZGEÇMİŞ.....	68

ÇİZELGELERİN LİSTESİ

Çizelge	Sayfa
Çizelge 4.1. Fare verisi (Rat data)	41
Çizelge 4.2. Tüm tanı grafikleri için gösterge değişkenlerinin değerleri	43
Çizelge 5.1. $p = 3$ iken bir etkili gözlem için tanı yöntemlerinin etkili bir gözlem içeren veride etkili gözlemi saptama oranları	48
Çizelge 5.2. $p = 4$ iken bir etkili gözlem için tanı yöntemlerinin etkili bir gözlem içeren veride etkili gözlemi saptama oranları	49
Çizelge 5.3. $p = 5$ iken bir etkili gözlem için tanı yöntemlerinin etkili bir gözlem içeren veride etkili gözlemi saptama oranları	50
Çizelge 5.4. $p = 3$ iken etkili gözlem grupları için tanı yöntemlerinin iki gözlemde oluşan etkili gözlem grubu içeren veride etkili gözlem grubunu saptama oranları.....	53
Çizelge 5.5. $p = 4$ iken etkili gözlem grupları için tanı yöntemlerinin iki gözlemde oluşan etkili gözlem grubu içeren veride etkili gözlem grubunu saptama oranları	54
Çizelge 5.6. $p = 5$ iken etkili gözlem grupları için tanı yöntemlerinin iki gözlemde oluşan etkili gözlem grubu içeren veride etkili gözlem grubunu saptama oranları.....	55
Çizelge 5.7. $n = 20$ ve $p = 3$ iken bir etkili gözlem için tanı yöntemlerinin farklı hata değişken değerlerine göre üretilen etkili gözlemi saptama oranları.....	58
Çizelge 5.8. $n = 20$ ve $p = 3$ iken etkili gözlem grupları için tanı yöntemlerinin farklı hata değişken değerlerine göre üretilen iki etkili gözlemin oluşturduğu etkili gözlem grubunu saptama oranları ...	59

ŞEKİLLERİN LİSTESİ

Şekil	Sayfa
Şekil 3.1. İki regresyon katsayısının bulunduğu bir model için $\hat{\beta}_{1(i)}$ ve $\hat{\beta}_{2(i)}$ tahminlerine ait grafik.....	26
Şekil 3.2. Maskelemeye ilişkin iki örnek durum.....	30
Şekil 3.3. Yanılgıya-düşürmeye ilişkin bir örnek durum.....	31
Şekil 4.1. Fare verisi için serpmeye diyagramları.....	42
Şekil 4.2. Fare verisi için tanı grafikleri.....	43

SİMGELER VE KISALTMALAR

Bu çalışmada kullanılmış bazı simge ve kısaltmalar, açıklamaları ile birlikte aşağıda sunulmuştur.

Simgeler	Açıklama
n	Örnek hacmi
p	Tahmin edilecek parametre sayısı
H	Hat (Projeksiyon) matrisi
h_{ii}	Hat matrisinin i . Köşegen elemanı
r_i	Dahili studentleştirilmiş artıklar
t_i	Harici studentleştirilmiş artıklar
	Veriden i . gözlemin silinmesiyle elde edilen X matrisi
	Veriden i . gözlemin silinmesiyle elde edilen Y matrisi
$\hat{\beta}$	Veriden i . gözlemin silinmesiyle elde edilen $\hat{\beta}$ vektörü
Kısaltmalar	Açıklama
EKK	Enküçük kareler
EIC	Deneysel etki eğrisi (Empirical influence curve)
SIC	Örnek etki eğrisi (Sample influence curve)
EIC_(i)	Gözlem silme tekniğine dayalı deneysel etki eğrisi
RSF	Görelî duyarlılık faktörü (relative sensitivity factor)

1. GİRİŞ

Regresyonda, gözlemlerden biri veya birkaçı verinin geneline uymayabilir. Bu tip gözlemler aykırı gözlemler (outliers) olarak adlandırılır. Bazı aykırı gözlemler ise mutlak değerce anormal büyüklükte artıklara sahip olabilir ve bunlar regresyon sonuçlarını olumsuz yönde etkileyebilir. Bir veride, regresyon parametrelerinin en küçük kareler, EKK, tahminlerinde önemli derecede farklılaşmalara neden olan gözlemler, etkili gözlemler olarak tanımlanır[1-3].

Etkili gözlemlerin artıkları mutlak değerce oldukça büyük oldukları için, bunlar artık kareler toplamı değerinin büyümesine neden olur. Bundan dolayı bu gözlemler, belirleme katsayısı ve regresyon katsayılarının standart hataları gibi modele ilişkin istatistikler üzerinde olumsuz yönde etkide bulunur. Bu yüzden etkili gözlemlerin saptanması ve saptandıktan sonra bu gözlemlerin etkisinin azaltılması için gerekli ağırlıklandırmaların yapılması, yapılan regresyon analizinin daha verimli olması açısından önemlidir.

Etkili gözlemlerin incelenmesi ilk kez Cook (1977) tarafından çalışılmıştır. Son otuz yılda bu alanda birçok çalışma yapılmıştır. Bu süreç içerisinde etkili gözlemlerin saptanması için birimlerin tek tek incelenmesinin yanı sıra birimlerin gruplar halinde incelenmesinin de önemi ortaya çıkmıştır. Literatürde etkili gözlemlerin saptanması için önerilen tanı istatistikleri beş başlık altında toplanabilir. Bunlar

- Artıklara dayalı tanı istatistikleri,
- Hat (projeksiyon) matrisine dayalı tanı istatistikleri,
- Güven elipsoidlerinin hacmine dayalı tanı istatistikleri,
- Etki eğrisine dayalı tanı istatistikleri ve
- Kısmi etkililiğe dayalı tanı istatistikleri

biçiminde ifade edilmiştir[4]. Bu tanı istatistiklerinin çoğunda gözlem silme tekniği kullanılmaktadır. Gözlem silme tekniğinde, bir gözlem veya gözlem kümesi veriden

çıkartıldıktan sonra regresyon modeline ilişkin istatistiksel sonuçların nasıl etkilendiği incelenir.

Verideki i . birimin bağımlı değişken bakımından ölçüm değeri Y_i ve bu birimin regresyon modeli ile elde edilen tahmin değeri $\hat{Y}_i$ olmak üzere $e_i = Y_i - \hat{Y}_i$ olarak tanımlanan, bilinen e_i artıklarına dayalı olan ve aykırı gözlemlerin saptanmasında kullanılan tanı istatistiklerinden biri Studentleştirilmiş Artıklar' dır. Studentleştirilmiş Artıklar, aykırı gözlemleri belirlemenin yanında etkili gözlemlerin belirlenmesi için de kullanılır. Bu yöntemde önemli derecede büyük değerli Studentleştirilmiş artıklara sahip gözlemler etkili gözlemler olarak değerlendirilebilir. Dahili ve harici olarak ikiye ayrılan studentleştirilmiş artıklar Margolin (1977) ve David (1981) tarafından tartışılmıştır[5,6].

Hoaglin ve Welsch (1978) aykırı gözlemlerin ve etkili gözlemlerin belirticisi olan yüksek dereceli kaldıraç noktalarını (high leverage points) saptamak için projeksiyon matrisi olarak da bilinen Hat matrisinin köşegen elemanlarının kullanılabileceğini belirtmişlerdir. Bu matrisin köşegen elemanları h_{ii} ile gösterilir ve kaldıraç değeri olarak bilinir. Ayrıca Hoaglin ve Welsch (1978), tahmin edilecek parametre sayısı p ve örnek çapı n olmak üzere, h_{ii} değeri $2p/n$ 'den büyük olan gözlemleri yüksek dereceli kaldıraç noktası olarak tanımlamışlardır. Buna ek olarak h_{ii} değerleriyle beraber Studentleştirilmiş Artıklar da incelenerek ilgili gözlemin etkili olup olmadığına daha iyi karar verilebildiğini de belirtmişlerdir[7]. Hat matrisi ikinci bölümde ayrıntılı bir biçimde tanıtılacaktır.

Güven elipsoitlerinin hacmine dayalı tanı istatistiklerine Andrews-Pergibon (1978) tarafından önerilen Andrews-Pergibon istatistiği örnek verilebilir [8]. Ayrıca Belsley ve diğ. (1980) tarafından önerilen kovaryans oranlarına dayalı *COVRATIO* istatistiği oldukça kullanışlıdır [9]. *COVRATIO* istatistiğiyle, regresyon parametrelerinin tahminleri üzerinde hem tek başına etkili olan gözlemler hem de ortak bir etkililiğe sahip gözlemler incelenebilir. Cook ve Weisberg (1982) tarafından geliştirilen iki

tanı istatistiği olan Ençok Olabilirlik Uzaklığı ve Cook-Weisberg İstatistiği de güven elipsoitlerin hacmine dayalı istatistiklerdendir[10].

Uygulamada sıkça kullanılan ve Cook (1977) tarafından önerilen Cook Uzaklığı, etki eğrisi (influence curve/function) kavramının örnek versiyonu olan örnek etki eğrisi (sample influence curve/function) kavramına dayalı bir istatistiktir[2]. Etki eğrisi ve etki eğrisinin örnek versiyonları ikinci bölümde ayrıntılı bir biçimde anlatılmıştır. Gözlem silme tekniğine dayalı olan Cook Uzaklığı hem tek başına etkili olan gözlemleri hem de ortak etkililiğe sahip gözlem kümelerini saptamada kullanılır. Ayrıca bu tanı istatistiğinin değiştirilmiş bir biçimi olan Düzeltilmiş Cook Uzaklığı da etkili gözlemlerin tespit edilmesinde kullanılan diğer bir istatistiktir[2,10]. Bunlara ek olarak etkili gözlemlerin saptanmasında Belsley ve diğ. (1980) tarafından geliştirilen ve gözlem silme tekniğine dayalı iki kullanışlı tanı istatistiği olan *DFBETAS* ve *DFFITS* istatistikleri de etki eğrisinden türetilmiş istatistiklerdir[9,11,12].

Yüksek dereceli kaldıraç noktalarının ve etkili gözlemlerin incelenmesinde, Hat matrisinin ayrıştırılmasıyla elde edilen j . değişkenin i . gözlemin h_{ii} değerine katkısını ölçmeye yarayan kısmi kaldıraç değeri (partial leverage) ve bu katkının görsel olarak incelenebildiği kısmi artık grafiği (partial residuals plot) de kullanılmaktadır. Hoaglin ve Welsch (1978) ve Chatterjee ve Hadi (1986) kısmi kaldıraç değerleri ve kısmi artık grafiği ile etkili gözlemlerin bulunması konusu üzerine çalışmışlardır[4,7].

Hem bir etkili gözlem hem de etkili gözlem gruplarını saptamak için kullanılan bir başka istatistik de Hadi (1992) tarafından önerilen H_i^2 tanı istatistiğidir. Bu istatistik x -uzayında, y -uzayında veya xy -uzayında potansiyel olarak etkili olan gözlemlerin aykırı değer olup olmadığını saptamada kullanılır[13,14]. Ayrıca tek başına etkili olan gözlemlerin saptanması için kullanışlı bir başka istatistik Pena (2005) tarafından önerilmiştir. Bu tanı istatistiği ise tahmin değerlerine dayalı Cook uzaklığının

geliştirilmiş bir biçimdir ve verideki her bir gözlemin silinmesiyle i . gözleme ilişkin tahminin duyarlılığını ölçmektedir[15].

Etkili gözlemlerin saptanması için bir başka yöntem de bağımlı değişkene karşı bağımsız değişken için oluşturulan serpmeye diyagramının incelenmesidir. Basit doğrusal regresyonda bu serpmeye diyagramı iki boyutlu olduğu için etkili gözlemler açık bir şekilde belirlenir. Fakat bağımsız değişken sayısı 2 olduğunda, bağımlı değişkene karşı bağımsız değişken için oluşturulan üç boyutlu serpmeye diyagramından, gözlemlerin genel eğilimine uymayan noktalar görsel olarak saptanamayabilir. Çünkü bu üç boyutlu diyagramda bir hacim söz konusudur. Bağımsız değişken sayısının ikiden fazla olduğu durumlarda ise, serpmeye diyagramı boyut sorunu olduğu için oluşturulamaz.

Bu zorluluğu gidermek için Li ve diğ. (2001) çoklu regresyonda etkili gözlemlerin saptanması için bir grafiksel yöntem geliştirmişlerdir. Bu yöntemdeki ana fikir yüksek boyutlu bir regresyon problemini iki boyutlu tanı grafiklerinin bir setine indirgeyerek, bu grafiklerin görsel olarak incelenmesine dayanır. Li ve diğ. (2001) bu metodolojiyi hem daha kolay bir yorumlamayı elde etmek, hem de hesaplamalarla diğer yöntemlere göre daha az uğraşmak amacıyla geliştirmişlerdir[16].

Etkili gözlem grupları incelenirken veride regresyon tahminleri üzerinde ortak etkisi olan gözlemler mevcut olabileceği gibi bu tahminler üzerinde koşullu olarak etkili olan gözlemler de bulunabilir. Son yıllarda ortak etkililik (joint influence) ve koşullu etkililik (conditional influence) kavramlarına bağlı olarak maskeleye (masking) ve yanlışya-düşürme (swamping) sorunları üzerinde durulmaktadır. Lawrance (1995) etkili gözlemlerin analizinde maskelemeyi ve yanlışya düşürme sorunlarını incelemiştir ve etkili gözlemler için bu sorunların aykırı gözlemlerdeki orijinal tanımlarıyla aynı olmadığını vurgulamıştır[17]. Maskeleye ve yanlışya düşürme sorunları üçüncü bölümde ayrıntılı olarak anlatılmıştır.

Regresyon katsayıları üzerinde ortak etkisi olan gözlem grupları araştırılırken karşılaşılabilecek bir başka sorun da gözlemlerin yüksek boyutlu kombinasyonlarını

incelerken hesaplama zorluğunun oluşmasıdır. Bundan dolayı uygulamada gözlemlerin en çok 5' li kombinasyonları dikkate alınarak etkili gözlemler tespit edilmektedir.

Çalışmanın ikinci bölümünde bu çalışmada kullanılan tanı istatistikleri için temel kavramlara yer verilmiştir. Üçüncü bölümde etkili gözlemlerin saptanması için kullanılan Cook uzaklığı, *DFBETAS*, *DFITS*, *COVRATIO*, Pena' nın S_i tanı istatistiği ve maskeleye (masking) ile yanılgıya düşürme (swamping) hakkında bilgi verilmiştir. Dördüncü bölümde ise çoklu regresyonda etkili gözlemlerin saptanması için kullanılan ve yeni bir yöntem olan grafik yöntemi tanıtılmıştır. Beş tanı istatistiği ve grafik yönteminin simülasyon çalışmasıyla karşılaştırılması beşinci bölümde verilmiştir. Son olarak altıncı bölümde de sonuç ve öneriler sunulmuştur.

2. TEMEL KAVRAMLAR

Bu bölümde doğrusal regresyon modeli ile tanı istatistiklerine ilişkin temel kavramlardan Etki Eğrisi, Etki Eğrisinin Sonlu Örnek Uyarlamaları, Hat Matrisi ve Studentleştirilmiş Artıklar kısaca tanıtılacaktır.

2.1. Doğrusal Regresyon Modeli

Verideki gözlem sayısı n ve regresyon modelindeki parametre sayısı da p olmak üzere, $n \times 1$ boyutlu yanıt vektörü Y , $n \times p$ boyutlu ve p ranklı tasarım matrisi X , $p \times 1$ boyutlu parametre vektörü β ve $n \times 1$ boyutlu 0 ortalamalı ve σ^2 varyanslı hata değişkenlerinin vektörü ε ile gösterilsin. Bu durumda doğrusal regresyon modeli

$$Y = X\beta + \varepsilon \quad (2.1)$$

biçiminde yazılır. Hata değişkenine ilişkin varsayımlar altında, bu modele ilişkin β parametre vektörünün EKK tahmin edicisi

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

ile bulunur. $\hat{\beta}$ vektörü kullanılarak elde edilen tahmin değerleri $\hat{Y}$ vektörüyle gösterilirse bu vektör

$$\hat{Y} = X\hat{\beta} \quad (2.2)$$

ile bulunur. Artık vektörü olan e ise

$$e = Y - \hat{Y}$$

ile verilir. Hata değişkenlerinin varyansı olan σ^2 için yansız bir tahmin edici aşağıdaki gibidir.

$$\hat{\sigma}^2 = e^T e / (n - p)$$

Regresyon analizine başlamadan önce etkili gözlem(ler) olup olmadığının belirlenmesi gerekir. Çünkü önceki bölümde bahsedildiği gibi bu gözlem(ler) regresyon sonuçlarını değiştirme eğilimindedir. Etkili gözlemlerin saptanması için kullanılan tanı istatistiklerine ilişkin temel kavramlardan bazıları kısaca incelenecektir.

2.2. Hat Matrisi

EKK yöntemiyle oluşturulan bir regresyon modelinde her bir gözlemin parametre tahminleri üzerinde ve buna bağlı olarak $\hat{y}_i$ tahmin değerleri üzerinde ne kadar bir aykırılığa veya etkililiğe sahip olduğunun belirlenmesi oldukça önemlidir. Projeksiyon matrisi olarak da bilinen Hat matrisi bu bilgiyi verir ve Studentleştirilmiş Artıklarla beraber verinin geneline uymayan noktaları bulmayı sağlar.

Eş. (2.2),

$$\begin{aligned}\hat{Y} &= X\hat{\beta} \\ &= X(X^T X)^{-1} X^T Y \\ &= HY\end{aligned}$$

biçiminde tekrar yazılırsa, buradaki $n \times n$ boyutlu olan H matrisi Hat Matrisi olarak adlandırılır [7]. Bu durumda H matrisi aşağıdaki formülle elde edilir.

$$H = X(X^T X)^{-1} X^T$$

Simetrik ve idempotent bir matris olan H matrisinin i . köşegen elemanına karşılık gelen değer h_{ii} ile gösterilir ve kaldırıcı değeri olarak adlandırılır. Hoaglin ve Welsch (1978) i . gözleme karşılık gelen kaldırıcı değerinin $2p/n$ ' den büyük olması durumunda, bu gözlemi yüksek dereceli kaldırıcı noktası olarak tanımlamışlardır. Bu nokta bağımsız değişken değeri bakımından verinin geneline uymayan bir gözlemdir. Bundan dolayı yüksek dereceli kaldırıcı noktaları aykırı veya etkili gözlem olabilir [7]. Hoaglin ve Welsch (1978) i . gözleme karşılık gelen kaldırıcı değerinin hesaplanması için alternatif bir formülü

$$h_{ii} = x_i (X^T X)^{-1} x_i^T, \quad (i = 1, 2, \dots, n)$$

olarak tanımlamışlardır [7].

H matrisinin önemli bir özelliği de köşegen elemanlarının yani h_{ii} değerlerinin toplamının tahmin edilecek parametre sayısı p ' ye eşit olması ve $0 \leq h_{ii} \leq 1$ olmasıdır. Ayrıca H matrisinin rankı tasarım matrisi X ' in rankına eşittir [7].

2.3. Etki Eğrisi ve Etki Eğrisinin Örnek Uyarlamaları

2.3.1. Etki eğrisi

Etki eğrisi ilk kez Von Misses (1947) tarafından önerilmiştir. Hampel (1968) ve Andrews ve ark. (1972) etki eğrisini tahmin edicileri karşılaştırmak ve mevcut tahmin tekniklerini geliştirmek amacıyla kullanmışlardır. Özellikle, Hampel (1974) etki eğrisinin sağlam (robust) tahmin edicilerdeki rolü üzerine çalışmıştır[10,18].

Etki fonksiyonu olarak da bilinen etki eğrisinin genel bir tanımı, sonsuz boyutlu bir uzayda tanımlı ve belli bir dağılıma sahip bir tahmin edicinin fonksiyonel bir gösterilişle ilk türetilişi biçiminde yapılabilir [18].

Birikimli dağılım fonksiyonu F olan ve R^m ’ de tanımlı bir yığından çekilen bağımsız ve aynı dağılımlı örnek $z_1, z_2, \dots, z_n$ olsun ve bu örneğe dayalı k boyutlu vektörel değerli istatistik T_n ile gösterilsin. T_n istatistiğine ilişkin parametre θ ile gösterilirse bu parametre fonksiyonel olarak $\theta = T(F)$ biçiminde yazılabilir. Ayrıca deneysel birikimli dağılım fonksiyonu $\hat{F}$ ile gösterilirse T_n istatistiği fonksiyonel olarak $T(\hat{F})$ biçiminde gösterilebilir. Bunu bir örnekle göstermek gerekirse $m = k = 1$ iken örnek ortalaması

$$T_n = \frac{1}{n} \sum z_i = \bar{z}$$

olsun. Bu durumda $T(\hat{F})$ fonksiyoneli

$$T(\hat{F}) = \int z d\hat{F} = \bar{z}$$

ile verilir [10].

Herhangi bir $z_0 \in R^m$ noktasında belirlenen birikimli dağılım fonksiyonu δ_{z_0} olsun ve

$$\delta_{z_0} = F(z; z_0) = \begin{cases} 1, & z \geq z_0 \\ 0, & z < z_0 \end{cases}$$

ile gösterilsin. τ , bozulma katsayısı olmak üzere, F ve $(1-\tau)F + \tau\delta_{z_0}$ birikimli dağılım fonksiyonları için T_n istatistiğine karşılık gelen fonksiyoneller sırasıyla $T\{F\}$ ve $T\{(1-\tau)F + \tau\delta_{z_0}\}$ olsun. Bu durumda F birikimli dağılım fonksiyonu üzerindeki T istatistiğinin etki eğrisi

$$IC(z) = \lim_{\tau \rightarrow 0^+} \frac{T\{(1-\tau)F + \tau\delta_{z_0}\} - T\{F\}}{\tau}, \quad (0 < \tau < 1) \quad (2.3)$$

olarak tanımlanır. Burada limitin her $z \in R^m$ noktası için tanımlı olması zorunluluğu vardır [10,18].

Etki eğrisi, asimptotik varyansın bulunmasında kullanılmasının yanı sıra sezgisel değerlendirmelerde kullanılan birçok sağlam istatistiğin türetilmesinde de kullanılmaktadır. Bununla birlikte bir istatistiğin her bir gözleme karşı duyarlılığını etki eğrisinin sonlu örnek uyarlamalarıyla incelemek daha kullanışlıdır.

2.3.2. Etki eğrisinin örnek uyarlamaları

Hampel (1974) Etki Eğrisi' nin tanı yöntemleri için kullanışlı olduğunu öne sürmüştür. Fakat Mallows (1975) bunun F birikimli dağılım fonksiyonuna göre yapılan bir tahmin tekniği olduğunu ve etki eğrisinin sonlu bir örnek uyarlamasına ilişkin bir yöntemin, F birikimli dağılım fonksiyonunu kullanmaya bağımlı olmayı gidererek daha fazla bilgi sağlayacağını belirtmiştir. Böylece Mallows (1975), regresyon parametrelerinin EKK tahminlerini etki eğrisine uygulayarak, bir örneğe ilişkin gözlem değerlerine dayalı etki eğrisinin sonlu örnek uyarlamaları olan deneysel etki eğrisi (empirical influence curve, EIC) ve örnek etki eğrisi (sample influence curve, SIC) yöntemlerini bulmuştur. Bu fikir geneldir ve diğer durumlara uygulanabilir. Bundan dolayı gözlem silme yaklaşımında, etki eğrisinin sonlu örnek uyarlamaları dikkate alınmaktadır. Çünkü bu yaklaşımla her bir gözlem değerini değerlendirerek etkili gözlemleri saptamak, sonlu örnek uyarlamalarıyla daha kullanışlı olur [10].

i . gözleme ilişkin bağımsız değişken değerleri x_i vektörüyle ve bağımlı değişken değeri de y_i ile gösterilsin ve $z_i = (x_i^T, y_i)^T$ vektörü ($i=1,2,\dots,n$) ilgili birime ilişkin gözlem değerlerinin bir vektörü olsun. (2.1) modeli dikkate alınır, $\hat{\beta}$ tahmin

edicişi için örnek birikimli dağılım fonksiyonu $\hat{F}$, Eş. (2.3)' te yerine konursa deneysel etki eğrisi (EIC)

$$EIC(z_i) = n(X^T X)^{-1} x_i e_i$$

biçiminde yazılır [10]. i . birimin silinmesiyle elde edilen örnek birikimli dağılım fonksiyonu $\hat{F}_{(i)}$, Eş. (2.3)' te $\hat{\beta}$ tahmin edicişi için yerine konursa, gözlem silme tekniğine dayalı olan yeni deneysel etki eğrisi

$$EIC_{(i)}(z_i) = (n-1)(X_{(i)}^T X_{(i)})^{-1} x_i (y_i - x_i^T \hat{\beta}_{(i)})$$

ile elde edilir [10]. Burada $(X_{(i)}^T X_{(i)})^{-1} = (X^T X - x_i x_i^T)^{-1}$ olur. $\hat{\beta}_{(i)}$ ise i . birimin silinmesiyle elde edilen β parametre vektörünün EKK tahmin edicisidir ve

$$\hat{\beta}_{(i)} = (X_{(i)}^T X_{(i)})^{-1} X_{(i)}^T Y_{(i)}$$

ile verilir.

Etki eğrisinin örnek uyarlamasının daha arzu edilen bir biçimi, Eş. (2.3)' te $\hat{\beta}$ tahmin edicişi için F yerine $\hat{F}$ ve τ yerine $-1/(n-1)$ alınarak elde edilen örnek etki eğrisi (SIC)' dir ve

$$\begin{aligned} SIC(z_i) &= -(n-1)(\hat{\beta}(\hat{F}_{(i)}) - \hat{\beta}(\hat{F})) \\ &= (n-1)(\hat{\beta} - \hat{\beta}_{(i)}) \\ &= \frac{(n-1)(X^T X)^{-1} x_i e_i}{1 - h_{ii}} \end{aligned} \tag{2.4}$$

olarak elde edilir. Burada h_{ii} terimi H matrisinin i . köşegen elemanıdır.

Etkili gözlemlerin saptanması için uygulamada sıkça kullanılan Cook uzaklığı SIC kullanılarak türetilir.

Etki eğrisinin sonlu örnek uyarlamaları olan her bir gözleme ilişkin $EIC(z_i)$, $EIC_{(i)}(z_i)$ ve $SIC(z_i)$, p boyutlu birer vektördür. Bu vektörlerin normları ilgili gözlemin etkili gözlem olup olmadığına ilişkin bir kanıya varmak için kullanılabilir.

2.4 Studentleştirilmiş Artıklar

Bilinen artıkların varyansı hem σ^2 parametresinin hem de h_{ii} istatistiğinin bir fonksiyonu olduğundan bu artıklar ölçeğe dayalı bir dağılıma sahiptir. Birçok tanı yöntemi için bu niceliklerin ikisine de bağlı olmayan, artıkların studentleştirilmiş bir biçimini kullanmak daha kullanışlıdır. Margolin (1977) Studentleştirme terimini, ölçeğe dayalı bir istatistik olan U istatistiğinin, bir T tahminine bölümü şeklinde tanımlamıştır. Bu bölümle elde edilen $S = U/T$ oranı, ilgilenilmeyen parametrelerden (nuisance parameters) bağımsız bir dağılıma sahip olacaktır. Ayrıca Studentleştirilmiş Artıklar U ve T istatistiklerinin genellikle aynı veriden türetildiği ve bağımlı olduğu Dahili Studentleştirme ve U ve T istatistiklerinin bağımsız olduğu Harici Studentleştirme kavramları altında ikiye ayrılır [10,19,20].

2.4.1. Dâhili studentleştirilmiş artıklar

EKK regresyonunda Standartlaştırılmış Artıklar olarak da isimlendirilen Dahili Studentleştirilmiş Artıklar aşağıdaki formülle elde edilir.

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

Burada $\hat{\sigma}^2 = \sum_{i=1}^n e_i^2 / (n-p)$ olarak tanımlanır ve artık kareler ortalaması olarak adlandırılır [10].

Doğrusal bir modelde, Dahili Studentleştirilmiş Artıklar ortalaması 0 ve varyansı 1 olan normal dağılıma sahiptir. Birçok durumda, özellikle büyük çaptaki veri setlerinde, artıkların varyansı dengede tutulur. Bu durumda, Dahili Studentleştirilmiş Artıklar ve Harici Studentleştirilmiş Artıklar arasında önemsiz derecede küçük farklılıklar mevcut olabilir. Böylece bu iki istatistik çoğu kez denk bilgileri verir. Bununla birlikte büyük Dahili Studentleştirilmiş Artık değerine ve büyük h_{ii} değerine sahip herhangi bir nokta potansiyel olarak regresyon tahminlerini değiştirme eğilimine sahiptir [1,10].

2.4.2. Harici studentleştirilmiş artıklar (RSUDENT)

Harici Studentleştirilmiş Artıklar için bağımsız e_i değerlerinden hesaplanan σ^2 parametresinin bir tahmin edicisine gerek vardır. Hataların normalliği varsayımı altında i . gözlemin silinmesiyle hesaplanan varyansın tahmin edicisi olan $\hat{\sigma}_{(i)}^2$ istatistiği aşağıdaki şekilde tanımlanır.

$$\begin{aligned} \hat{\sigma}_{(i)}^2 &= \frac{(n-p)\hat{\sigma}^2 - e_i^2 / (1-h_{ii})}{n-p-1} \\ &= \hat{\sigma}^2 \left(\frac{n-p-r_i^2}{n-p-1} \right), \quad (i=1,2,\dots,n) \end{aligned} \quad (2.5)$$

Eş. (2.5)' ten yararlanarak Harici Studentleştirilmiş Artıklar aşağıdaki gibi tanımlanır.

$$t_i = \frac{e_i}{\hat{\sigma}_{(i)} \sqrt{(1-h_{ii})}}, \quad (i=1,2,\dots,n) \quad (2.6)$$

Burada $\hat{\sigma}_{(i)}^2$ ve e_i bağımsızdır [10]. Bu durumda $\hat{\sigma}_{(i)}$ terimi, Eş. (2.6)' da yerine konursa harici studentleştirilmiş artıkların bir başka biçimi

$$t_i = r_i \left(\frac{n-p-1}{n-p-r_i^2} \right)^{1/2}, \quad (i=1,2,\dots,n)$$

ile verilir [10].

t_i istatistiği $n-p$ serbestlik dereceli student-t dağılır. Her bir t_i değeri ilgili kritik değerle karşılaştırılarak aykırı gözlemler belirlenir[10,20]. Buna bağlı olarak mutlak değerce oldukça büyük R-SUDENT değerine sahip gözlemler etkili gözlem olabilir.

3. ETKİLİ GÖZLEMLERİN SAPTANMASI İÇİN TANI İSTATİSTİKLERİ

Bir verideki gözlemlerden biri veya birkaçı verinin genel yapısına uymuyor ve bunlar regresyon modeli parametrelerinin EKK tahminlerinde dikkate değer farklılaşmalara neden oluyorsa bu tür gözlem veya gözlem kümeleri etkili gözlemler olarak tanımlanır. Bundan dolayı etkili gözlem veya etkili gözlem kümelerinin saptanması, regresyon modeli oluşturma sürecinde önemli bir adımdır.

Etkili gözlemler verinin genel yapısına uymayan gözlemler oldukları için, literatürde aykırı gözlemlerin belirlenmesi için kullanılan istatistiklerin bazıları etkili gözlemlerin tespiti için de kullanılabilir. Bunlardan ikisi, kaldıraç değeri olarak da bilinen H matrisinin köşegen elemanları olan h_{ii} ve Dahili Studentleştirilmiş Artıklar olan r_i istatistikleridir. Oldukça büyük h_{ii} ve r_i değerleri ilgili gözlemin etkili gözlem olduğunun belirtisidir. Bu yöntemler dışında, etkili gözlemleri saptamanın daha iyi bir yolu i . gözlemi veya verinin alt kümesi olan bir gözlem grubunu veriden çıkartarak regresyon sonuçlarının nasıl değişeceğinin incelenmesidir. Gözlem silme yöntemi olarak adlandırılan bu yöneme dayalı tanı istatistikleri etkili gözlemlerin saptanmasında oldukça kullanışlıdır. Bu tanı istatistikleri iki başlık altında toplanabilir. Bunlar içinde bir etkili gözlemi saptamaya dayalı tanı istatistiklerine *DFBETAS*, *DFFITS* ve Pena' nın S_i istatistiği örnek gösterilebilir. Hem bir etkili gözlemi hem de etkili gözlem gruplarını saptamaya dayalı tanı istatistiklerine de Cook Uzaklığı ve *COVRATIO* istatistiği örnek verilebilir.

Bazı durumlarda iki veya daha fazla gözlem tek başlarına regresyon katsayıları üzerinde etkili gözükmüyorken bu gözlemlerin tahminler üzerinde ortak etkisi mevcut olabilir. Buna bağlı olarak bu iki gözlem bir etkili gözlemi saptamak için kullanılan ve gözlem silme tekniğine dayalı olan tanı istatistikleriyle tespit edilemez. Bu durumda maskeleye (masking) söz konusudur. Maskeleye sorununun tam zıttı ise etkili bir gözlemin etkili olmayan bir gözlemle beraber regresyon parametrelerinin tahminlerini değiştirme eğiliminde olduğu durumdur. Böylece etkili

olmayan bu gözlem, saptanan etkili gözlem grubunun elemanı olarak belirlenir. Bu durum ise yanılığa düşürme (swamping) olarak bilinir[17].

Bu bölümde Cook uzaklığı, *DFBETAS*, *DFFITS*, *COVRATIO* ve Pena' nın S_i istatistiği tanıtıldıktan sonra maskeleye ve yanılığa-düşürme sorunları hakkında bilgi verilecektir.

3.1. Bir Etkili Gözlemi Saptamaya Dayalı Tanı İstatistikleri

Bir gözlemin tek başına etkili bir gözlem olup olmadığı, bu gözlemin veriden atılmasıyla geri kalan veriden hesaplanan tahminlerin nasıl değiştiğini gözleyerek saptanabilir. Eğer elde edilen tahminlerde dikkate değer bir değişme varsa bu gözlem etkili bir gözlemdir. Bu yöntemeye dayalı tanı istatistikleri, bir etkili gözlemi saptamaya dayalı tanı istatistikleri başlığı altında toplanır.

Belsley ve ark. (1980) tarafından önerilen, bir gözlem silme tekniğine dayalı iki kullanışlı tanı istatistiği *DFBETAS* ve *DFFITS* istatistikleridir. Ayrıca Pena (2005) tarafından önerilen S_i tanı istatistiği regresyon katsayıları üzerinde tek başına etkili olan gözlemin belirlenmesinde kullanışlıdır. Bu bölümde tek etkili gözlem saptamaya dayalı bu üç tanı istatistiği tanıtılacaktır.

3.1.1. DFBETAS tanı istatistiği

Belsley ve ark. (1980) tarafından önerilen $DFBETAS_{ij}$ istatistiği, i . gözlemin veriden silinmesiyle buna karşılık gelen standart sapmaya bağlı olarak j . regresyon katsayısı $\hat{\beta}_j$ tahmininin ne kadar değişeceğini gösteren bir istatistiktir.

i . gözlemin veriden silinmesiyle elde edilen β parametre vektörünün EKK tahmin edicisi $\hat{\beta}_{(i)}$ vektörüyle gösterilsin. Orijinal veriden elde edilen β parametre

vektörünün EKK tahmin edicisi $\hat{\beta}$ ile $\hat{\beta}_{(i)}$ arasındaki fark $DFBETA_i$ ile gösterilirse bu fark istatistiği

$$DFBETA_i \equiv \hat{\beta} - \hat{\beta}_{(i)} = \frac{(X^T X)^{-1} x_i^T e_i}{1 - h_{ii}} \quad , \quad (i = 1, 2, \dots, n) \quad (3.1)$$

ile verilir [9]. Fakat $DFBETA_i$ sadece vektörel değerli bir sonuç vermektedir ve skaler değerli sonuç veren bir istatistik ile etkili gözlemleri saptamak daha kolaydır. Bu durumda Eş. (3.1)' in geliştirilmiş bir biçimi olan $DFBETAS_{ij}$ istatistiği kullanılır[9].

$\hat{\beta}$ vektörünün j . elemanı $\hat{\beta}_j$ ve $\hat{\beta}_{(i)}$ vektörünün j . elemanı da $\hat{\beta}_{j(i)}$ olsun. $C = (X^T X)^{-1} X^T$ matrisi dikkate alınırsa $\hat{\beta}_j$ tahmini ile $\hat{\beta}_{j(i)}$ tahmini arasındaki fark

$$\hat{\beta}_j - \hat{\beta}_{j(i)} = \frac{c_{ji} e_i}{1 - h_{ii}} \quad , \quad (i = 1, 2, \dots, n; j = 1, 2, \dots, p)$$

biçiminde yazılır [9]. Burada c_{ij} , C matrisinin j . satır i . sütun elemanıdır. Veriden bir gözlemin silinmesiyle elde edilen sonuçlarda çoğu kez $\hat{\beta}_j$ istatistiğinin varyansı da kullanılır. Bu durumda $\hat{\beta}_j$ istatistiğinin varyansı

$$\begin{aligned} Var(\hat{\beta}_j) &= \sigma^2 (X^T X)^{-1}_{jj} \\ &= \sigma^2 \sum_{k=1}^n c_{jk}^2 \end{aligned}$$

ile verilir. Buna göre $\hat{\beta}_j$ tahminindeki değişimin ölçüsü

$$\begin{aligned}
DFBETAS_{ij} &\equiv \frac{\hat{\beta}_j - \hat{\beta}_{j(i)}}{\hat{\sigma}_{(i)} \sqrt{(X^T X)^{-1}_{jj}}} \\
&= \frac{c_{ji}}{\sqrt{\sum_{k=1}^n c_{jk}^2}} \cdot \frac{e_i}{\hat{\sigma}_{(i)} (1 - h_{ii})}
\end{aligned}$$

biçiminde elde edilir [9]. Burada $\hat{\sigma}_{(i)}$ terimi i . gözlemin silinmesiyle elde edilen standart sapmadır ve

$$\hat{\sigma}_{(i)} = \sqrt{\frac{1}{n-p} \sum_k (y_k - x_k \hat{\beta}_{(i)})^2}$$

ile verilir [9]. C matrisinin j . satırı c_j ve Harici Studentleştirilmiş Artıklar t_i olmak üzere $DFBETAS_{ij}$ istatistiğinin alternatif bir formülü

$$DFBETAS_{ij} = \frac{c_{ji}}{\sqrt{c_j c_j^T}} \cdot \frac{t_i}{\sqrt{1 - h_{ii}}} \quad , \quad (i = 1, 2, \dots, n; j = 1, 2, \dots, p)$$

olarak bilinir. Belsley ve ark. (1980) tarafından önerilen kritere göre, eğer $|DFBETAS_{ij}| > 2/\sqrt{n}$ ise i . gözlem etkili bir gözlemdir [9].

3.1.2. DFFITS tanı istatistiği

Tek başına etkili olan gözlemi saptamak için kullanılan bir diğer tanı istatistiği Belsley ve ark. (1980) tarafından önerilen $DFFITS$ istatistiğidir. Gözlem silme tekniğine dayalı olan bu istatistik i . gözlemin tahmin değerindeki değişimine dayanır.

$DFFIT_s$ istatistiği tüm birimlerden elde edilen tahmin değeri ile i . gözlemin veriden silinmesiyle elde edilen yeni tahmin değeri arasındaki farkın büyüklüğüne dayalı olarak bu gözlemin etkili bir gözlem olup olmadığını belirlemek için kullanılır [9].

Bu iki tahmin değeri arasındaki fark $DFFIT_i$ ile gösterilirse bu istatistik

$$DFFIT_i \equiv \hat{y}_i - \hat{y}_{i(i)} = x_i (\hat{\beta} - \hat{\beta}_{(i)}) = \frac{h_{ii} e_i}{1 - h_{ii}} \quad , \quad (i = 1, 2, \dots, n) \quad (3.2)$$

ile verilir. Bu tanı ölçüsünün avantajı regresyon modelinde kullanılan özel koordinat sistemlerine bağlı olmamasıdır [9]. Ölçeklendirmeden kurtulmak amacıyla Eş. (3.2)'deki farkı $\hat{\sigma}_{(i)} \sqrt{h_{ii}}$ terimine bölmek gerekir. Bu durumda bu farkın standartlaştırılmış biçimi $DFFITS_i$ ile gösterilir ve

$$\begin{aligned} DFFITS_i &= \frac{\hat{y}_i - \hat{y}_{i(i)}}{\hat{\sigma}_{(i)} \sqrt{h_{ii}}} \\ &= \left(\frac{h_{ii}}{1 - h_{ii}} \right)^{1/2} \frac{e_i}{\hat{\sigma}_{(i)} \sqrt{1 - h_{ii}}} \quad , \quad (i = 1, 2, \dots, n) \end{aligned} \quad (3.3)$$

biçiminde elde edilir [9]. Eş. (3.3)'teki formülde sol taraf, yani $[h_{ii}/(1 - h_{ii})]^{1/2}$ fonksiyonu, i . gözlemin kaldıraç gücüne karşılık gelmektedir. Sağ taraftaki oran ise i . gözleme ilişkin Harici Studentleştirilmiş Artık değeridir. Böylece bu iki kısmın çarpımı olan $DFFITS_i$ istatistiği aşağıdaki gibi yeniden yazılabilir [9].

$$DFFITS_i = \left(\frac{h_{ii}}{1 - h_{ii}} \right)^{1/2} t_i \quad , \quad (i = 1, 2, \dots, n) \quad (h_{ii} \neq 1)$$

Burada büyük h_{ii} değerine ve büyük Harici Studentleştirilmiş Artığa sahip gözlemlerin etkili gözlem olacağı açıkça görülmektedir. Buna dayanarak Belsley ve

ark. (1980) $DFFITS_i$ istatistiğiyle etkili gözlemleri belirlemek için bir karar kuralı önermişlerdir. Buna göre $|DFFITS_i| > 2\sqrt{p/n}$ ise i . gözlem etkili gözlemdir [9].

3.1.3. Pena' nın $S_{(i)}$ tanı istatistiği

Gözlem silme tekniğine dayalı olarak tahmin değerlerindeki farklılaşmanın incelendiği diğer bir istatistik Pena (2005) tarafından önerilen $S_{(i)}$ istatistiğidir. Bu istatistik i . gözlemin tahmininin her bir gözlemin tek tek silinmesiyle nasıl değişeceğini ölçen alternatif bir yöntemdir. Böylece etkili gözlemler verideki diğer gözlemlerin yardımıyla belirlenir.

(2.1) modeli dikkate alındığında tüm veriden elde edilen $\hat{\beta}_j$ istatistikleriyle hesaplanan i . gözlemin tahmin değeri $\hat{y}_i$ ile veriden bir gözlemin çıkartılmasıyla elde edilen $\hat{\beta}_{j(k)}$ istatistikleriyle hesaplanan i . gözlemin tahmin değeri $\hat{y}_{i(k)}$ arasındaki farka ilişkin vektör

$$s_i = (\hat{y}_i - \hat{y}_{i(1)}, \dots, \hat{y}_i - \hat{y}_{i(n)})^T$$

biçiminde tanımlansın [15]. Bu i . gözlemin tahmin değerinin verideki her bir gözlemin silinmesine karşı duyarlılığını göstermektedir. Böylece $S_{(i)}$ istatistiği, s_i vektörünün standartlaştırılmış karesel normu olacak biçimde

$$S_{(i)} = \frac{s_i^T s_i}{p \cdot \hat{\sigma}_{(\hat{y}_i)}} \quad , \quad (i = 1, \dots, n) \quad (3.4)$$

ile verilir [15]. Burada $\hat{y}_i - \hat{y}_{i(k)} = h_{ki} e_k / (1 - h_{kk})$ ve $\hat{\sigma}_{(\hat{y}_i)} = s^2 h_{ii}$ alınırsa Eş. (3.4)' te tanımlanan istatistik

$$S_{(i)} = \frac{1}{ps^2 h_{ii}} \sum_{k=1}^n \frac{h_{ki}^2 e_k^2}{(1-h_{kk})^2} \quad , \quad (i, k = 1, \dots, n)$$

biçiminde yeniden yazılır [15]. $S_{(i)}$ istatistiği için alternatif bir formül Cook Uzaklığı istatistiğinin de kullanıldığı

$$S_{(i)} = \sum_{k=1}^n \rho_{ki}^2 D_i \quad , \quad (i, k = 1, \dots, n)$$

formülü ile verilir [15]. Burada $\rho_{ik} = (h_{ik}^2 / h_{ii} h_{kk})^{1/2} \leq 1$ olur ve bu $\hat{y}_i$ ile $\hat{y}_k$ arasındaki korelasyona karşılık gelir. $\hat{\beta}_{(k)}$ tahmin vektörü kullanılarak elde edilen k . gözleme ilişkin artık değeri

$$\begin{aligned} e_{k(k)} &= y_k - \hat{\beta}_{(k)} x_k \\ &= e_k / (1 - h_{kk}) \end{aligned}$$

ile verilsin [15]. Bu artık değeri kullanarak $S_{(i)}$ istatistiği

$$S_{(i)} = \frac{1}{ps^2} \sum_{k=1}^n w_{ki} e_{k(k)}^2 \quad , \quad (i, k = 1, \dots, n)$$

ile gözlem silme tekniğine dayalı olarak elde edilen tahminlere ilişkin artıkların ağırlıklı kombinasyonu biçiminde yeniden yazılabilir. Buradaki w_{ki} terimi

$$w_{ki} = \frac{h_{ik}^2}{ph_{ii}(1-h_{kk})^2}$$

ile verilir [15].

Hiçbir aykırı gözlem olmaması ve h_{ii} değerlerinin tümünün küçük olması durumunda $S_{(i)}$ istatistiğinin beklenen değeri yaklaşık olarak $1/p$ olur. Diğer bir deyişle yüksek dereceli kaldırıcı noktalarının olmadığı bir veride gözlemlerin tümünün aynı duyarlılığa sahip olması beklenir. Bu beklenen değer kaldırıcı noktalara oldukça bağlı olan Cook uzaklığına göre önemli bir avantajdır [15].

Pena (2005) etkili gözlemlerin saptanması için önerdiği S_i istatistiğine ilişkin

$$\left| S_{(i)} - \text{med}(S) \right| \geq 4.5 \times \text{MAD}(S_{(i)}) \quad (3.5)$$

karar kuralını önermiştir. Eğer bu eşitsizlik sağlanırsa i . gözlem etkili gözlemdir. (3.5) eşitsizliğinde $\text{med}(S)$ değeri $S_{(i)}$ değerlerinin medyanı, $\text{MAD}(S_{(i)})$ ise $S_{(i)}$ değerlerinin medyandan sapmalarının mutlak değerlerinin medyanıdır [15].

3.2. Hem Bir Etkili Gözlemi Hem de Etkili Gözlem Gruplarını Saptamaya Dayalı Tanı İstatistikleri

Herhangi bir gözlem veriden atıldığında, regresyon tahminleri üzerinde önemli derecede bir değişime neden oluyorsa bu gözlem tek başına (bireysel) etkili gözlemdir. Bununla birlikte bazı veriler için birden fazla gözlemin yani bir gözlem grubunun regresyon tahminleri üzerinde etkisi söz konusu olabilir. Etkili bir gözlem grubu, etkili bir gözlemin doğal bir genelleştirilmiş halidir. Bu durumda bu grubu oluşturan ve regresyon tahminleri üzerinde ortak etkisi olan her bir gözlemin bireysel olarak da etkisi vardır. Bu tip durumlarda çoklu gözlem silme tekniği, bir gözlem silme tekniğinden daha güvenilir sonuçlar verir [21]. Fakat gözlemlerin ortak etkilerini incelerken maskeleye (masking) ve yanılgıya düşürme (swamping) problemlerine dikkat etmek gerekir. Bu bölümde uygulamada çokça kullanılan Cook uzaklığı ve *COVRATIO* tanı istatistikleri tanıtılacaktır.

3.2.1. Cook uzaklığı

Cook (1977) tarafından önerilen tanı istatistiği Cook Uzaklığı D_i ile gösterilir. Bu istatistik tüm veriye dayalı EKK tahmin vektörü $\hat{\beta}$ ile verinin bir alt kümesinin silinmesiyle hesaplanan EKK tahmin vektörü $\hat{\beta}_{(i)}$ arasındaki karesel uzaklığın bir ölçüsüdür.

Cook Uzaklığı istatistiğinin türetilmesinde kullanılan ve Eş. (2.4)' te verilen β parametre vektörünün tahmin edicisi için örnek etki eğrisi

$$SIC(z_i) = (n-1) \left(\hat{\beta} - \hat{\beta}_{(i)} \right)^2, \quad (i=1, 2, \dots, n) \quad (3.6)$$

ile verilir. Örnek etki eğrisi $p \times 1$ boyutlu bir vektör olduğundan dolayı etkili gözlemlerin belirlenmesi için kullanışlı değildir. Bununla birlikte skaler sonuç veren bir istatistik etkili gözlemlerin saptanmasında daha kullanışlıdır. Bundan dolayı $\hat{\beta}$ ile $\hat{\beta}_{(i)}$ arasındaki karesel uzaklık etkili gözlemlerin saptanmasında kullanılabilir. Eş. (3.6), $\hat{\beta}$ ile $\hat{\beta}_{(i)}$ arasındaki fark biçiminde yazılırsa

$$(n-1)^{-1} SIC(z_i) = \hat{\beta} - \hat{\beta}_{(i)}, \quad (i=1, 2, \dots, n)$$

ile verilir [10]. Burada eşitliğin sağ tarafı $(n-1)$ sabitinden kurtulmuş olur. Böylece simetrik ve pozitif tanımlı (veya pozitif yarı tanımlı) $p \times p$ boyutlu bir M matrisine ve pozitif değerli bir c faktörüne dayalı olarak $\hat{\beta}$ ile $\hat{\beta}_{(i)}$ arasındaki karesel uzaklık $D_i(M, c)$ normları ile hesaplanır. Bu durumda $D_i(M, c)$ normu

$$\begin{aligned}
D_i(M, c) &= (n-1)^{-2} \frac{(SIC(z_i))^T M (SIC(z_i))}{c} \\
&= \frac{(\hat{\beta}_{(i)} - \hat{\beta})^T M (\hat{\beta}_{(i)} - \hat{\beta})}{c}, \quad (i = 1, 2, \dots, n)
\end{aligned} \tag{3.7}$$

ile elde edilir [10]. Cook (1977) tarafından önerilen prosedüre göre $M = X^T X$ ve $c = p\hat{\sigma}^2$ alınırsa ve Eş. (3.7)' de yerine koyulursa Cook Uzaklığı

$$D_i(X^T X, p\hat{\sigma}^2) = \frac{(\hat{\beta}_{(i)} - \hat{\beta})^T X^T X (\hat{\beta}_{(i)} - \hat{\beta})}{p\hat{\sigma}^2}, \quad (i = 1, 2, \dots, n)$$

ile verilir [10]. Bu tanı istatistiği, $X^T X$ matrisine bağlı olarak $\hat{\beta}$ tahmininden $\hat{\beta}_{(i)}$ tahminine olan karesel uzaklığı vermektedir. Ayrıca bu istatistiğe karşılık gelen güven aralıkları, merkezi $\hat{\beta}$ olan ve $X^T X$ matrisinin özdeğerleri ve özvektörleriyle belirlenen elipslere karşılık gelmektedir.

Cook Uzaklığı, Dahili Studentleştirilmiş Artıkların ve Hat matrisinin köşegen elemanlarının yer aldığı

$$\begin{aligned}
D_i &= \frac{r_i^2 \text{Var}(\hat{y}_i)}{p \text{Var}(e_i)} \\
&= \frac{r_i^2}{p} \frac{h_{ii}}{1 - h_{ii}}, \quad i = 1, 2, \dots, n
\end{aligned}$$

formülüyle de hesaplanabilir [10]. Bu oran x_i vektörünün geri kalan verinin kitle merkezine olan uzaklığını ifade etmektedir. Ayrıca büyük artıklara sahip gözlemler de etkili gözlem olarak yorumlanabileceği için, Dahili Studentleştirilmiş Artıklarla hesaplanan D_i değerlerinden oldukça büyük olanlar etkili gözlemlerin göstergesidir.

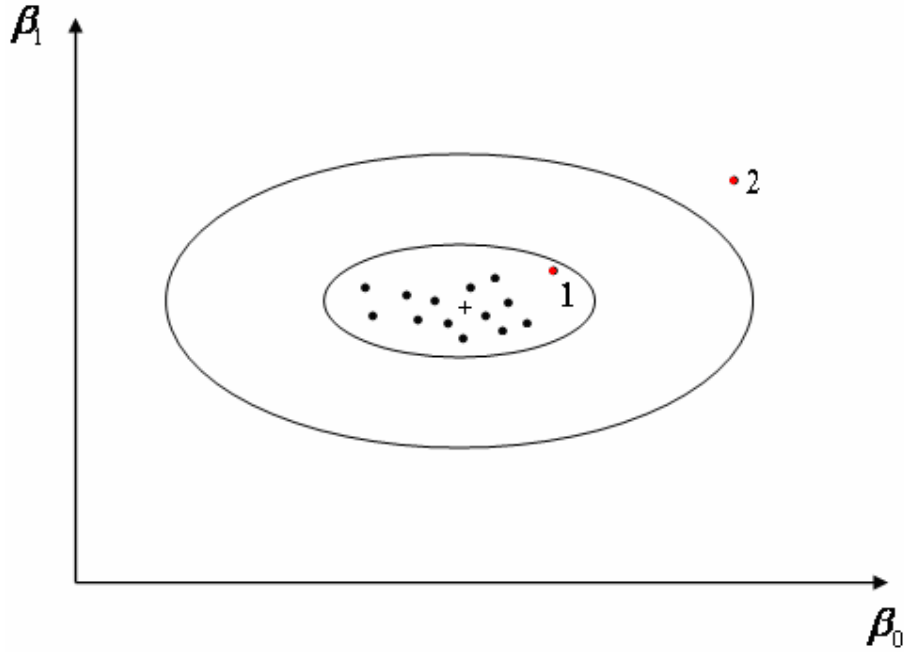
Cook Uzaklığı için alternatif bir formül gözlem silme yaklaşımıyla hesaplanan tahmin değeriyle verinin tümüne dayalı tahmin değeri arasındaki farka dayalı olan

$$D_i = \frac{(\hat{y}_{(i)} - \hat{y}_i)^T (\hat{y}_{(i)} - \hat{y}_i)}{p\hat{\sigma}^2} \quad , \quad (i = 1, 2, \dots, n)$$

ile verilir [10].

Cook Uzaklığı $\hat{\beta}_{(i)}$ ile $\hat{\beta}$ arasındaki karesel uzaklığa dayalı bir istatistiktir. Bu durumda bu istatistik için kritik değer $\hat{\beta}_{(i)}$ ile $\hat{\beta}$ arasındaki karesel uzaklığı temel olarak çıkartılır.

Verinin tümünden elde edilen β parametre vektörünün EKK tahmin edicisi $\hat{\beta}$ vektöründen oluşturulan bir elipsoid klasik anlamdaki bir güven elipsoidine karşılık gelmektedir. Eğer i . gözlemin veriden çıkartılmasıyla elde edilen β parametre vektörünün EKK tahmin edicisi $\hat{\beta}_{(i)}$ vektörü, $\hat{\beta}$ vektörü kullanılarak oluşturulan %50' lik güven elipsoidinin sınırında ya da bu sınırın dışında ise bu durumda oldukça büyük bir sapma söz konusudur ve bu gözlemin etkili olduğu söylenebilir. Çünkü normal şartlarda etkili olmayan bir gözlem veriden çıkartıldığında verinin geri kalanından elde edilen $\hat{\beta}_{(i)}$ vektörü %10 ile %20 civarında oluşturulan bir güven elipsoidi içinde olmalıdır. Buna bir örnek vermek gerekirse $\beta = [\beta_0, \beta_1]$ parametre vektörü dikkate alınsın. Bu parametreler için oluşturulan %10' luk ve %50' lik güven elipsoidleri Şekil3.1' deki gibi olsun.



Şekil 3.1. İki regresyon katsayısının bulunduğu bir model için $\hat{\beta}_{0(i)}$ ve $\hat{\beta}_{1(i)}$ tahminlerine ait grafik

Şekil 3.1’ de tüm veriden elde edilen regresyon katsayıları merkezi oluşturmak üzere her bir gözleme ilişkin gözlem silme tekniğine dayalı olarak hesaplanan tahmin değerlerinin grafiği verilmiştir. Burada içteki elipsoid %10’ luk bir güven elipsoidine, dıştaki elipsoid %50’ lik bir güven elipsoidine karşılık gelmektedir. 2 ile gösterilen gözlem veriden silindikten sonra verinin geri kalanından elde edilen $\hat{\beta}_{(2)}$ tahmini %50’ lik güven elipsoidi dışındadır. Dolayısıyla bu gözlemin regresyon tahminlerini olumsuz yönde etkileyeceği söylenebilir. Bir diğer gözlem olan 1. gözlem ise %10’ luk güven elipsoidi içinde kaldığı için regresyon tahminlerini dikkate değer bir şekilde değiştirme eğiliminde değildir. Bir başka deyişle bu gözlem verinin geneline uyduğu için etkili gözlem değildir.

$\hat{\beta}_{(i)}$ vektörünün güven elipsoidinin dışında olduğu, grafikte bu vektörün elemanlarıyla belirlenen noktanın güven elipsoidinin merkezine olan uzaklığı ile belirlenir. Bu uzaklık i . gözlem için hesaplanan D_i değerine karşılık gelmektedir.

Bir başka deyişle bu uzaklık Cook Uzaklığı istatistiğiyle hesaplanır. Eğer i . gözlemin veriden silinmesiyle elde edilen $\hat{\beta}_{(i)}$ vektörünün grafikteki koordinatı %50'lik güven elipsoidinin sınırında ise $D_i \cong F_{0.50,p,n-p}$ olacaktır. Bu durumda $D_i > F_{0.50,p,n-p}$ ise grafikte $\hat{\beta}_{(i)}$ vektörüne ilişkin nokta güven elipsoidinin dışında olacaktır. Bundan dolayı verideki herhangi bir gözlem için hesaplanan Cook Uzaklığı $F_{0.50,p,n-p}$ kritik değerinden büyükse bu gözlem etkili bir gözlemdir. Burada dikkat edilmesi gereken nokta Cook Uzaklığı istatistiğinin F dağılımına yakınsadığıdır [10].

$F_{0.5,p,n-p} \cong 1$ olduğundan dolayı Cook Uzaklığı için kritik değerin 1 alınması etkili gözlemlerin incelenmesinde pratiklik sağlar. Bundan dolayı uygulamada Cook Uzaklığı istatistiği için kritik değer 1 alınmaktadır [10].

Daha önce belirtildiği gibi Cook Uzaklığı'nın bir başka özelliği etkili gözlem gruplarının saptanması için de kullanılabilmesidir. Bu durumda veriden i . gözlemin silinmesi yerine, hacmi örnek hacminden küçük olmak üzere, verinin m büyüklüğünde ve I ile gösterilen bir alt kümesi silinir. Bu gözlem gruplarına ilişkin Cook Uzaklığı değerleri hesaplandıktan sonra bu değerler $F_{0.50,p,n-p}$ ile karşılaştırılarak ilgili gözlem kümesinin regresyon sonuçlarını değiştirme eğiliminde olup olmadıkları incelenir.

3.2.2. Kovaryans oranı (COVRATIO)

Cook uzaklığı, $DFBETAS$ ve $DFITS$ tanı istatistikleri sadece tahmin edilen regresyon katsayıları ve tahmin değerleri üzerinde etkili olan gözlemleri saptamada kullanılır. Fakat bu istatistikler tam olarak tahminin kesinliğine (precision of estimation) ilişkin herhangi bir bilgi sağlamaz. Varyans-kovaryans matrisinin determinantı, kesinliğin uygun bir sayısal ölçüsü olarak oldukça yaygın kullanılmasından dolayı, genelleştirilmiş varyans olarak tanımlanmıştır [1,9]. Bu durumda $\hat{\beta}$ tahmininin genelleştirilmiş varyansı aşağıdaki gibidir.

$$\begin{aligned} G\text{Var}(\hat{\beta}) &= \det\{\text{Var}(\hat{\beta})\} \\ &= \det\{\hat{\sigma}^2(X^T X)^{-1}\} \end{aligned}$$

Veriden i . gözlemin silinmesiyle elde edilen $\hat{\beta}_{(i)}$ tahmininin kovaryans matrisi $\hat{\sigma}_{(i)}^2(X_{(i)}^T X_{(i)})^{-1}$ ile hesaplanır. $\hat{\beta}_{(i)}$ tahmininin genelleştirilmiş varyansının $\hat{\beta}$ tahmininin genelleştirilmiş varyansına oranı i . gözlemin etkili bir gözlem olup olmadığına ilişkin bir sonuca varmak için kullanılabilir. Gözlem silme tekniğiyle elde edilen bu oran, kovaryans oranı ($COVRATIO_i$) olarak tanımlanır ve

$$COVRATIO_i = \frac{\det\{\hat{\sigma}_{(i)}^2(X_{(i)}^T X_{(i)})^{-1}\}}{\det\{\hat{\sigma}^2(X^T X)^{-1}\}}, \quad (i = 1, 2, \dots, n) \quad (3.8)$$

ile verilir [9]. Eş. (3.8)' de $\det(X_{(i)}^T X_{(i)})^{-1} = (1 - h_{ii}) \det(X^T X)^{-1}$ olduğundan dolayı bu oran için alternatif bir formül

$$\begin{aligned} COVRATIO_i &= \left(\frac{\hat{\sigma}_{(i)}^2}{\hat{\sigma}^2} \right)^p \left(\frac{1}{1 - h_{ii}} \right) \\ &= \frac{1}{\left(\frac{n - p - 1}{n - p} + \frac{t_i^2}{n - p} \right)^p (1 - h_{ii})} \end{aligned}$$

ile verilir. Burada t_i terimi, i . gözleme karşılık gelen Harici Studentleştirilmiş Artık terimidir [9].

i . birime ilişkin $COVRATIO$ değerinin 1' den büyük olması durumunda bu gözlemin tahminin kesinliğini arttıracak ve $COVRATIO$ değerinin 1' den küçük olması durumunda da i . gözlemin tahminin kesinliğini bozacağı açıktır. Belsley ve ark.

(1980), $|COVRATIO_i - 1| > 3p/n$ olması durumunda i . gözlemin etkili gözlem olabileceğini öne sürmüşlerdir [1,9].

Kovaryans oranı, Cook uzaklığı gibi etkili gözlem gruplarını saptamak için de kullanışlı bir ölçüdür. Eş. (3.8)' deki formül genelleştirilirse verinin m büyüklüğünde ($m < n$) bir alt kümesi olan ve I ile gösterilen alt kümesinin silinmesiyle elde edilen $COVRATIO_I$ değeri,

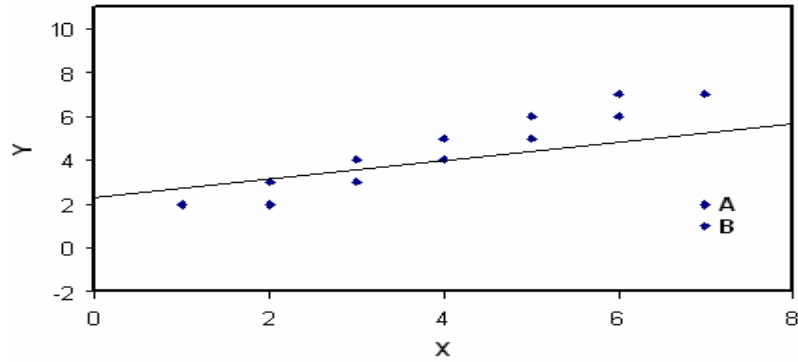
$$COVRATIO_I = \frac{\det\left\{\hat{\sigma}_{(I)}^2 \left(X_{(I)}^T X_{(I)}\right)^{-1}\right\}}{\det\left\{\hat{\sigma}^2 \left(X^T X\right)^{-1}\right\}}$$

ile verilir[9]. Tek gözlem için $COVRATIO$ istatistiğinde kullanılan karar kuralı, gözlem grupları $COVRATIO$ ile incelenirken de kullanılmaktadır.

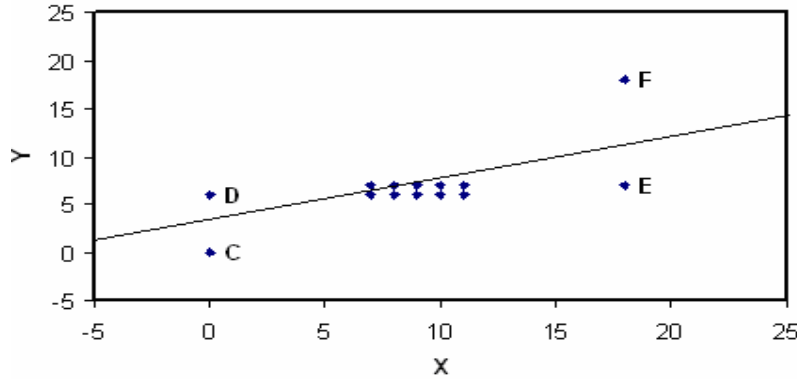
3.3. Maskeleme ve Yanılgıya-Düşürme

Bazı verilerde etkili bir gözlem diğer gözlem veya gözlemler veriden atılmadan saptanamaz. Bu durumda tek etkili gözlemi saptamaya dayalı tanı istatistikleri başarısız olur. Bu tip veriler için maskeleme söz konusudur.

Literatürde maskelemeye ilişkin farklı tanımlar vardır. Chatterjee ve Hadi (1986), iki veya daha fazla gözlem tek başlarına regresyon katsayıları üzerinde etkili gözükmüyorken bu gözlemlerin tahminler üzerinde ortak olarak etkisi olması durumunu maskeleme olarak tanımlamışlardır [4]. Rousseeuw ve Leroy (1987) da maskelemeyi, bir etkili gözlemin veriden atılmasından sonra ilk başta etkili görünmeyen bir gözlemin regresyon parametre tahminleri üzerinde etkili bir gözlem olarak ortaya çıkması şeklinde tanımlamışlardır [20]. Bu iki tanımlama aynıdır ve koşullu etkililikle ilgilidir. Aynı tanımlamalar verideki ikiden fazla etkili gözlem için de yapılabilir. Maskelemeye ilişkin iki örnek durum Şekil 3.2' de verilmiştir.



(a)



(b)

Şekil 3.2. Maskelemeye ilişkin iki örnek durum

(a) Maskelemeye ilişkin birinci örnek durum

(b) Maskelemeye ilişkin ikinci örnek durum

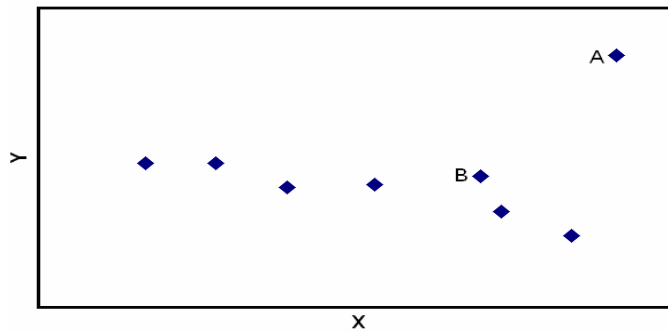
Şekil 3.2(a)'daki örnekte koşullu etkililik söz konusudur. Bu durumda bireysel olarak etkili olan A ve B noktaları ayrı ayrı incelendiğinde tek etkili gözlemi saptamaya dayalı tanı istatistikleri bu gözlemleri etkili gözlem olarak saptayamaz. Çünkü bunlar birbirlerini gizlemektedir. Fakat bu iki gözlemin ortak etkileri incelenirse regresyon tahminlerini dikkate değer bir biçimde değiştirebileceği gözlenir. Bu durumda A ve B noktaları birbirlerini maskelemektedir.

Şekil 3.2(b) incelenirse $\{C,D\}$ ve $\{E,F\}$ nokta çiftlerinin regresyon sonuçları üzerinde ortak olarak etkili olmadığı görülmektedir. Bununla birlikte C ve F noktaları bireysel olarak etkili gözlemlerdir. Bu durumda tek başına etkili olan F

noktası C noktasını gizlemektedir. Aynı şekilde tek başına etkili olan C noktası da F noktasını gizlemektedir. Bu durumda bu iki nokta birbirini maskeleymektedir. Bu verideki gözlemlerin ikili alt kümeleri etkili gözlem kümelerini saptamak için kullanılan tanı istatistikleriyle incelenirse $\{C,F\}$ nokta çifti etkili gözlem grubu olarak saptanır.

Gray ve Ling (1984) maskeleymeyi belirlemek için Hat Matrisi' nin köşegen olmayan elemanlarından ve alt matrislerinden yararlanabileceğini öne sürmüşlerdir [21]. Ayrıca maskeleyenin belirlenmesi için Atkinson (1986) en küçük medyan kareler yöntemine dayalı Unmasked yöntemini bulmuştur [22].

Etkili gözlem grupları incelenirken karşılaşılabilecek diğer bir problem maskeleyenin karşıtı olarak bilinen yanılgıya-düşürme (swamping) durumudur. Etkili gözlem gruplarını saptamak için kullanılan tanı istatistikleri çoğu zaman etkili olan bir gözlemle etkili olmayan bir gözlemi etkili gözlem grubu olarak saptayabilir. Bu durumda etkili olmayan gözlemin etkili gözlem olduğu yorumuna varılır. Bu durum yanılgıya düşürme olarak bilinir [17]. Yanılgıya-düşürmeye ilişkin bir örnek durum Şekil 3.3' te verilmiştir.



Şekil 3.3. Yanılgıya-düşürmeye ilişkin bir örnek durum

Yanılgıya-düşürmeye ilişkin yukarıdaki örnek durum incelendiğinde, tek başına oldukça etkili olan A noktasıyla gözlemlerin genel eğilimine uyan B noktasının regresyon tahminleri üzerinde ortak bir etkiye sahip olup olmadığı etkili gözlem gruplarını belirlemek için kullanılan tanı istatistikleriyle incelenirse, bu iki gözlem

etkili gözlem grubu olarak saptanır. Bu durumda etkili olmayan B noktası etkiliymiş gibi değerlendirilir.

Hem maskeleye hem de yanılgıya-düşürme, etkili gözlem gruplarının saptanması için kullanılan çoğu tanı istatistiğinde yanlış kararlar almaya sebep olabilir. Bu da araştırmacının yapacağı regresyon analizini olumsuz bir biçimde etkiler. Bu durumda etkili gözlemler incelenirken maskeleye ve yanılgıya-düşürme sorunlarının mevcut olup olmadığı da incelenmelidir.

4. ÇOKLU REGRESYONDA ETKİLİ GÖZLEMLERİN SAPTANMASI İÇİN GRAFİKSEL YÖNTEM

Grafiksel yöntemler veri analizinde özellikle de doğrusal modellerin oluşturulmasında önemli rol oynarlar. Çünkü grafiksel yöntemler veriye ilişkin bazı ön bilgileri ve modelleme yapıldıktan sonra modelin yapısına ilişkin bilgileri sunması nedeniyle çok kullanışlıdır. Bu yöntemlerin kullanılma nedenleri

- Verideki hataları saptamak (örneğin aykırı gözlemler tipografik hatanın bir sonucu olabilir),
 - Verinin biçimini (desenini) tanımak (verideki gözlem gruplarını, aykırı gözlemleri, etkili gözlemleri görmek),
 - Değişkenler arasındaki ilişkileri araştırmak,
 - Olağanüstü durumları keşfetmek,
 - Varsayımların sağlandığını veya reddedildiğini görmek,
 - Oluşturulan modelin uygunluğunu değerlendirmek,
 - Gerekliyse veriyi iyileştirici eylemleri önermek (gerekli dönüşümlerin yapılması, deneyin tekrarlanması, daha fazla veri toplanması gibi),
 - Yapılan analizin sonuçlarını genelleştirmek,
- olarak sınıflandırılabilir [14].

Grafiksel yöntemler iki grupta sınıflandırılır. Bunlardan birincisini, verideki hataları düzeltmek ve oluşturulan modelin uygunluğunu değerlendirmek amacıyla yapılan *modelleme öncesi grafikler* oluşturur. Bunlar, tek boyutlu grafikler (histogram, dal-yaprak grafiği, dot grafiği, box-plot), iki boyutlu grafikler, rotating (üç boyutlu serpmeye) grafikleri ve dinamik grafikler başlıkları altında toplanabilir. İkincisini ise varsayımların sağlanıp sağlanmadığının incelenmesi gibi amaçlarla kullanılan *modelleme sonrası grafikler* oluşturur. Bu tip grafikler, doğrusallık ve normallik ile ilgili varsayımların incelenmesi için kullanılan grafikler, aykırı gözlemlerin ve etkili gözlemlerin saptanması için kullanılan tanı grafikleri (diagnostic plots) ve

değişkenlerin etkililiklerinin incelenmesi için kullanılan tanı grafikleri başlıkları altında toplanabilir [14].

Basit doğrusal regresyonda aykırı ve etkili gözlemlerin görsel olarak tespit edilmesi için kullanılan grafiksel yöntem, bağımlı değişkene karşı bağımsız değişken için oluşturulan serpm diyagramının incelenmesine dayanır. Bağımlı değişken tahmin değerlerine karşı artıklar için oluşturulan serpm diyagramı incelenerek de verinin geneline uymayan gözlemler hakkında bir kanıya ulaşılabilir. Fakat bağımsız değişken sayısı 2 iken oluşan üç boyutlu serpm diyagram ile verinin geneline uymayan gözlemleri saptamak görsel yolla mümkün olmaz. Bağımsız değişken sayısının ikiden büyük olduğu durumlarda ise bağımlı değişkene karşı bağımsız değişkenlerin serpm diyagramı boyut sorunu yaratır. Bu sorunu gidermek için, Li ve ark. (2001) çoklu regresyonda etkili gözlemleri saptamak için kullanılan bir grafiksel yöntem önermişlerdir. Tanı grafiklerine (diagnostic plots) dayalı olan bu grafiksel yöntem hem daha kolay yorumlayabilmeyi sağlamak hem de hesaplamalarla diğer yöntemlere göre daha az uğraşmak amacıyla geliştirilmiştir [16]. Bu yöntem adımsal bir yöntemdir ve her bir öz değere karşılık gelen tanı grafiklerinin belli bir algoritmaya göre oluşturulup, ayrı ayrı incelenmesine dayanır. Bu bölümde Li ve ark. (2001) tarafından önerilen bu grafiksel yöntem tanıtılacaktır.

İkinci bölümde de bahsedildiği gibi i . gözleme ilişkin deneysel etki eğrisi olarak ifade edilen $EIC(z_i)$ bir vektöre karşılık gelmektedir. Bu vektörün normu ilgili gözlemin etkili olup olmadığı konusunda bilgi verir. a bir vektör ise, bu vektörün normu $\|a\|$ ile gösterilir. Buna göre etkili gözlemi belirlemek için

$$\max_i \|EIC(z_i)\| \quad (4.1)$$

bir kriter olarak kullanılabilir.

Tüm örnek noktaları için bir S uzayı

$$S = \left\{ z \mid z = Z^T w = \sum_{i=1}^n w_i z_i; \quad w = [w_1, \dots, w_n]^T \right\}$$

olarak tanımlansın. Burada p tahmin edilecek parametre sayısı olmak üzere

$z_i = [x_1, x_2, \dots, x_{p-1}, y]$ ile verilir. Bu durumda (4.1) ifadesi

$$\max_{\|w\|_2=1} \|EIC[z(w)]\|_2 \quad (4.2)$$

biçiminde tekrar yazılır [16]. (4.2) ile verilen maksimum probleminin optimum çözümü olan $n \times 1$ boyutlu w^* vektörünün elemanları, birimlerin regresyon parametrelerinin EKK tahminleri üzerinde etkilerine ilişkin en büyük oranları verir [16]. (4.2) maksimum problemi, gerekli matematiksel işlemler yapıldıktan sonra

$$\text{Max}_{w \neq 0} \left\{ (w^T E w) (w^T U w) / (w^T w)^2 \right\}^{1/2} \quad (4.3)$$

biçiminde yeniden yazılabilir. Burada $E = ee^T$, $e = Y - X\hat{\beta} = (I - X(X^T X)^{-1} X^T)Y$ ve $U = X(X^T X)^{-2} X^T$ ile elde edilir [16].

Li ve ark. (2001), Eş. (4.3)' teki maksimum problemi için w^* vektörünün hesaplanışına yardımcı olan Teorem 1' i öne sürmüşlerdir. Ayrıca bu teoremden grafik yönteminde kullanılan $n \times 2$ boyutlu $w^{(j)}$ matrisinin, w^* vektöründen yararlanarak elde edilişi de verilmektedir. Bu teoremden kullanılmak üzere, Li ve ark. (2001) Lemma 1 ve Lemma 2' yi de önermişlerdir. Lemma 1, E matrisi ve U matrisinin tekil değer ayrıştırmasına (singular value decomposition, SVD) ilişkindir. Lemma 2 de w^* vektörünün bulunmasında kullanılan T^* vektörünün optimum değerlerinin çıkartılışını göstermektedir.

Lemma 1: $Y \in R^n$, $X \in R^{n \times p}$ ve $\text{rank}(X) = p$ için $E = ee^T$, $U = X(X^T X)^{-2} X^T$ ve $e = (I - X(X^T X)^{-1} X^T)Y$ olmak üzere, P bir dik matris olsun.

$$P^T EP = \text{diag}\{O_{p \times p}, O_{(n-p-1) \times (n-p-1)}, e^T e\}$$

$$P^T UP = \text{diag}\{A_{p \times p}, O_{(n-p-1) \times (n-p-1)}, 0\}$$

Burada $A = \text{diag}\{a_1, \dots, a_p\}$ ile gösterilir ve $a_1 \geq \dots \geq a_p > 0$ değerleri $X^T X$ matrisinin özdeğerlerine karşılık gelmektedir[16].

Lemma 2: $T = [t_1, \dots, t_n]^T$ için $a_1 \geq \dots \geq a_p > 0$ olsun ve birbirinden farklı J sayıda a_j özdeğerleri için s_j çokluğu verilsin ($\sum_{j=1}^J s_j = p < n$). Böylece

$$\text{Max}_{T \neq 0} \left[t_n^2 \sum_{j=1}^p a_j t_j^2 \right] / \left[\sum_{j=1}^n t_j^2 \right]^2 = a_1 / 4$$

maksimum problemi tanımlansın. Elemanları sıfırdan farklı olmak üzere herhangi bir keyfi olarak belirlenen $s_1 \times 1$ boyutlu vektör r ile gösterilirse, bu maksimum probleminin en uygun çözümleri $T^* = 2^{-1/2} \left[r^T / (r^T r)^{1/2}, 0, 1 \right]^T$ vektörüne eşittir. Pratikte optimum çözüm vektörleri T_1^* ve T_2^* ile gösterilir. Eğer $s_1 = 1$ ise $T_1^* = 2^{-1/2} [1, 0, \dots, 0, 1]^T$ ve $T_2^* = 2^{-1/2} [-1, 0, \dots, 0, 1]^T$ olur [16].

Lemma 1 ve *Lemma 2*'nin ispatları için [16]'e bakınız.

Teorem : $Y \in R^n$, $X \in R^{n \times p}$ ve $\text{rank}(X) = p$ için $E = ee^T$, $U = X(X^T X)^{-2} X^T$ ve $e = (I - X(X^T X)^{-1} X^T)Y$ verilsin. $a_1 \geq \dots \geq a_p > 0$ olmak üzere $A = \text{diag}(a_1, \dots, a_p)$ matrisi ve Lemma 1' deki P matrisi verilmiş olsun. J sayıda birbirinden farklı a_j özdeğeri için

i) Eş. (4.3) ile verilen maksimum probleminin sonucu

$$\text{Max}_{w \neq 0} \left\{ (w^T E w) (w^T U w) / (w^T w)^2 \right\}^{1/2} = a_1 / 4$$

olur. Bu durumda normalleştirilmiş en uygun çözümler $w^{(1)} = PT^{(1)}$ olarak verilir. Eğer $s_1 = 1$ ise, $w_1^{(1)}$ ve $w_2^{(1)}$ sırasıyla $PT_1^{(1)}$ ve $PT_2^{(1)}$ ile elde edilir. Burada $T^{(1)} = T^*$, $T_1^{(1)} = T_1^*$ ve $T_2^{(2)} = T_2^*$, Lemma 2' de verilmiştir.

ii) B matrisi, $n \times j$ boyutlu ve P matrisinin ilk j sütununu kapsayan bir matris olsun. Bu durumda Eş. (4.3) ile verilen maksimum problemi

$$\text{Max}_{w \neq 0, B^T w = 0} \left\{ (w^T E w) (w^T U w) / (w^T w)^2 \right\}^{1/2} = a_{j+1} / 4, \quad (j = 1, \dots, p-1)$$

olur. Keyfi olarak belirlenen elemanları sıfırdan farklı $s_{j+1} \times 1$ boyutlu bir r vektörü

$$\text{için } T^{(j+1)} = 2^{-1/2} \left[O_{L \times 1}, r^T / (r^T r)^{1/2}, O_{m \times 1}, 1 \right]_{n \times 1}^T \quad \text{ile} \quad \text{gösterilir} \left(L = \sum_j^i s_j \quad \text{ve} \right.$$

$m = n - L - 1 - s_{j+1}$). Böylece normalleştirilmiş en uygun çözüm $w^{(j+1)} = PT^{(j+1)}$ ile elde edilir. Eğer $s_{j+1} = 1$ ise $w_1^{(j+1)}$ ve $w_2^{(j+1)}$ vektörleri sırasıyla $PT_1^{(j+1)}$ ve $PT_2^{(j+1)}$ ile elde edilir ve burada $T_1^{(j+1)}$ ve $T_2^{(j+1)}$ değerleri sırasıyla $2^{-1/2} [O_{L \times 1}, 1, O_{m \times 1}, 1]^T$ ve $2^{-1/2} [O_{L \times 1}, -1, O_{m \times 1}, 1]^T$ ile verilir[16].

Bu teoremden belirtilen ve saptama matrisi olarak da tanımlanan $w^{(j)}$ matrisi, Eş. (4.3) ile verilen maksimum probleminin optimum sonucu olan w^* vektörünün dönüştürülmüş bir biçimidir. Bu matris $[w_1^{(j)}, w_2^{(j)}]$ olmak üzere iki vektörün birleştirilmiş biçimidir ve gözlemlerin en küçük kareler tahmini $\hat{\beta}$ vektörünü değiştirme eğilimlerine ilişkin oranları verir[16]. Li ve ark. (2001), $w_l^{(j)}$ vektörlerini ($l=1,2$) hesaplamak amacıyla bu teoremi bir algoritmaya çevirmişlerdir. Bu algoritma EK-1’de verilmiştir.

$X^T X$ matrisinin özdeğerleri sırasıyla $a_1 \geq \dots \geq a_p > 0$ olmak üzere bu teoremin geometrik açıklaması şu şekilde yapılır. $n \times 2$ boyutlu olan saptama matrisi $w^{(j)}$, artık vektörüyle birleşik bir matristir ve açıklayıcı değişken değerlerinin süper-elipsoidinin j . ana ekseninin izdüşümüne karşılık gelir. $w^{(j)}$ matrisinin en duyarlı yanı aykırı gözlemleri, oluşturulan süper elipsoidin sınırları içinde tutmasıdır.

$w_l^{(j)}$ vektörünün ($j=1,2,\dots,p$; $l=1,2$) özellikleri

- $w_l^{(j)}$ ’lerin örnek ortalamaları sıfır,
- $w_l^{(j)}$ ’lerin örnek varyansları $1/(n-1)$,
- $w_1^{(j)}$ ve $w_2^{(j)}$ için örnek kovaryanslar sıfır

olarak sıralanabilir[16].

j . tanı grafiği $w_1^{(j)}$ ve $w_2^{(j)}$ vektörlerinin elemanlarıyla oluşturulur ve bu grafik özdeğerler $a_1 \geq \dots \geq a_p > 0$ olmak üzere j . özdeğere karşılık gelir. Bu durumda özdeğer sayısı kadar mevcut olan tanı grafiğinin her birinin ayrı ayrı incelenmesi yoluyla verinin yapısı hakkında bilgi elde edilir. Fakat burada bir sorun söz konusu olmaktadır. Herhangi bir tanı grafiğinde etkili görünmeyen bir gözlem diğer grafikte etkili gözüküyorsa bu durum ilgili gözlemin etkili gözlem olup olmaması konusunda bir kararsızlığa neden olabilir. Bu sorunu engellemek için Li ve ark. (2001) her bir tanı grafiğine ilişkin iki tane gösterge değişkeni geliştirmiştir. Gösterge değişkeni

bakımından büyük değerli olan grafikler verinin yapısına uymayan gözlemleri açık bir şekilde sunar. $\|A\|_F$ değeri A matrisinin Frobenius normu olmak üzere gösterge değişkenleri

$$d^{(j)} = \|M^{(j)} - M^{(j)}(L_2)\|_F / \|M^{(j)}(L_2)\|_F, \quad (j=1,2,\dots,p) \quad (4.4)$$

$$\delta^{(j)} = \|w_0^{(j)}\|, \quad (j=1,2,\dots,p) \quad (4.5)$$

ile verilir[16]. Eş. (4.4)' te belirtilen gösterge dikkate alınırsa $M^{(j)}$ ve $M^{(j)}(L_2)$ matrisleri sırasıyla kovaryans matrisinin sağlam tahminini ve en çok olabilirlik, EÇOB, tahminini ifade etmektedir. Eş. (4.5)' te belirtilen gösterge değişkeni için $w_0^{(j)} = [w_{01}^{(j)}, w_{02}^{(j)}]^T$ vektörü, $[w_1^{(j)}, w_2^{(j)}]^T$ matrisinin sütun vektörlerinin konum parametrelerinin sağlam tahminleridir [16].

Li ve ark. (2001) tarafından önerilen, j . tanı grafiğinin etkili gözlemleri içerip içermediğinin anlaşılması amacıyla kullanılan diğer bir gösterge değişkeni göreceli duyarlılık faktörü (relative sensitivity factor, RSF) olarak tanımlanır. j . özdeğer a_j olmak üzere

$$\lambda^{(j)} = \frac{a_j^{1/2}}{\sum_{j=1}^p a_j^{1/2}}, \quad (j=1,2,\dots,p) \quad (4.6)$$

ile verilir [16]. Bu faktör j . özdeğere karşılık gelen tanı grafiği için hesaplanır. Göreceli duyarlılık faktörü değeri en büyük olan tanı grafiği, etkili gözlem veya gözlem gruplarının en açık şekilde saptanabileceği grafiklerdir [16].

Li ve ark. (2001) tanı grafiklerinde verinin geneline uymayan gözlemlerin tespit edilebilmesi için ρ yarıçaplı deneysel güven elipslerinin oluşturulabileceğini belirtmişlerdir. Deneysel güven elipsleri

$$(w - w_0^{(j)})^T [M^{(i)}]^{-1} (w - w_0^{(j)}) = \rho$$

ile elde edilir [16]. İlgili tanı grafiği için güven elipsi oluşturulduktan sonra her bir gözleme karşılık gelen noktalar görsel olarak incelenir. Güven elipsinin dışında olan noktalar etkili gözlem olarak saptanır.

Güven elipslerinin olumsuz bir özelliği, bu elipslerin oluşturulmasında kullanılan sağlam tahminlerin adımsal metodlarla hesaplanışının zaman almasıdır. Ayrıca adımsal olmayan metodlar da aykırı ve etkili gözlemlere karşı duyarlıdır. Böylece etkililik potansiyeli olan herhangi bir gözlem güven elipsinin içine girebilir ve buna bağlı olarak da bu gözlem grafik yöntemi ile saptanamayabilir. Bundan dolayı aykırı ve etkili gözlemlerin çok olduğu veri türleri için güven elipslerinin kullanılması uygun değildir.

Etkili gözlemlerin saptanması için önerilen bu grafiksel yöntem küçük bir uygulama üzerinden gösterilmek istenirse, literatürde etkili gözlemlerin saptanması için kullanılan yöntemlerin değerlendirilmesi için yaygın olarak kullanılan ve etkili gözlem içeren Fare Verisi (Rat Data) ele alınabilir [10]. Veri seti, farelerin karaciğerine uygulanan bir ilacın miktarının incelenmesi suretiyle yapılan bir deneyden elde edilmiştir. 19 fare rastgele seçilmiştir ve yanıt değişkeni karaciğerdeki dozun oranı (y) ve üç bağımsız değişkenler vücut ağırlığı (x_1), karaciğer ağırlığı (x_2) ve vücuda verilen doz oranı (x_3) olarak belirlenmiştir. Veri seti Çizelge 4.1' de sunulmuştur.

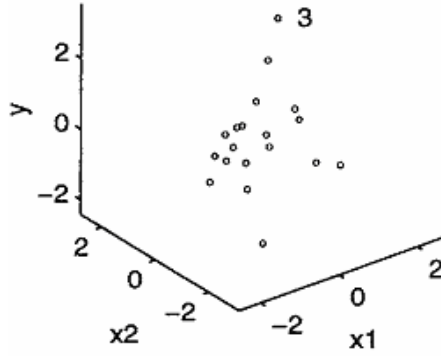
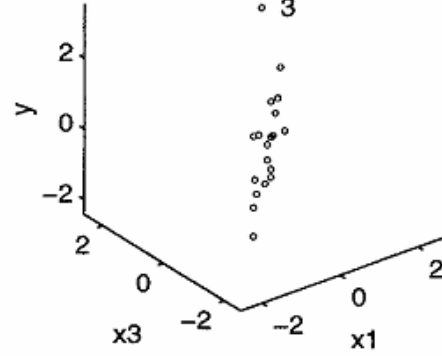
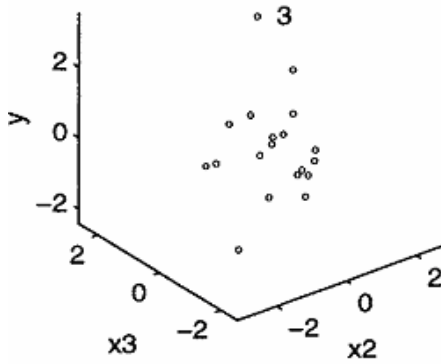
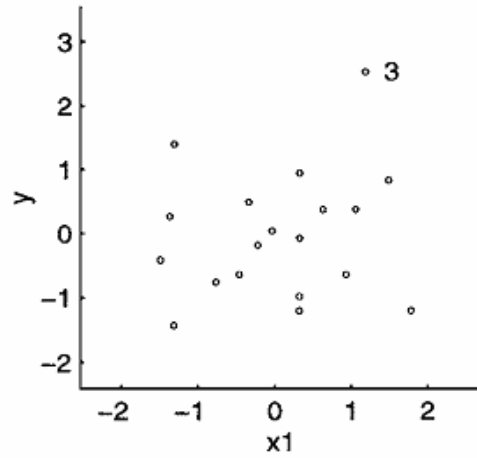
Çizelge 4.1. Fare verisi (Rat data)

Gözlem no	- vücut Ağırlığı(gr)	- karaciğer ağırlığı(gr)	- doz oranı	y - karaciğerdeki dozun oranı
1	176	6,50	0,88	0,42
2	176	9,50	0,88	0,25
3	190	9,00	1,00	0,56
4	176	8,90	0,88	0,23
5	200	7,20	1,00	0,23
6	167	8,90	0,83	0,32
7	188	8,00	0,94	0,37
8	195	10,00	0,98	0,41
9	176	8,00	0,88	0,33
10	165	7,90	0,84	0,38
11	158	6,90	0,80	0,27
12	148	7,30	0,74	0,36
13	149	5,20	0,75	0,21
14	163	8,40	0,81	0,28
15	170	7,20	0,85	0,34
16	186	6,80	0,94	0,28
17	146	7,30	0,73	0,30
18	181	9,00	0,90	0,37
19	149	6,40	0,75	0,46

Cook ve Weisberg (1982) bu veri için aşağıdaki doğrusal modeli oluşturmuşlardır.

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \varepsilon_i, \quad i = 1, \dots, 19$$

3. gözlemi etkili olan bu veri seti için serpm diyagramları Şekil 4.1' de verilmiştir. İlk üç diyagram, bağımsız değişkenlerin ikili kombinasyonları ile yanıt değişkeni dikkate alınarak oluşturulan üç boyutlu diyagramlardır. Bu diyagramların üçünde de üçüncü gözlem etkili gözlem olmasından ziyade aykırı gözlem olarak gözükmektedir. İki boyutlu serpm diyagramında ise gözlemlere ilişkin yorumlama zorluğu doğmaktadır. Çünkü çoklu regresyonda, özellikle de yüksek boyutlu problemlerde, bağımlı değişkene karşı bağımsız değişken için oluşturulan iki boyutlu serpm diyagramlarının tek tek incelenmesiyle verinin yapısı hakkında açık bir bilgi elde etmek zordur.

(a) x_1, x_2 ve y için serpmeye diyagramı(b) x_1, x_3 ve y için serpmeye diyagramı(c) x_2, x_3 ve y için serpmeye diyagramı(d) x_1 ve y için serpmeye diyagramı

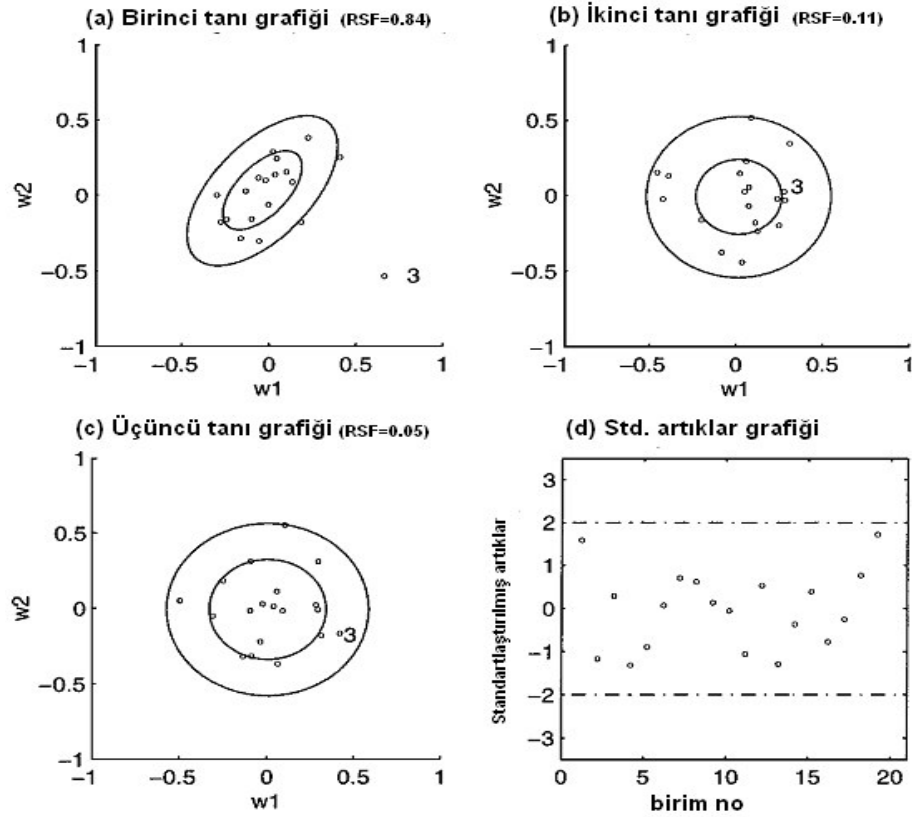
Şekil 4.1. Fare verisi için serpmeye diyagramları.

Tanı grafikleri oluşturulmadan önce ham veri standartlaştırılmıştır. Sonra her bir özdeğere karşılık gelen grafiklere ilişkin gösterge değişken değerleri hesaplanıp Çizelge 4.2' de verilmiştir.

Çizelge 4.2. Tüm tanı grafikleri için gösterge değişkenlerinin değerleri

Grafik no		
1	0,517	0,059
2	0,034	0,018
3	0,045	0,008

Çizelge 4.2’ de verilen $d^{(j)}$ ve $\delta^{(j)}$ gösterge değişkeni değerleri incelendiğinde $d^{(1)}$ ve $\delta^{(1)}$ değerlerinin diğerlerine göre oldukça büyük olması 1. tanı grafiğinin etkili gözlemleri açık bir biçimde sunabileceğini ortaya koymaktadır. Bununla birlikte verinin yapısını ayrıntılı bir şekilde görmek istenirse diğer tanı grafikleri de incelenebilir.



Şekil 4.2. Fare verisi için tanı grafikleri

Li ve ark. (2001) tarafından geliştirilen algoritmaya göre oluşturulan grafikler Şekil4.2’ de verilmiştir. Birinci grafik incelendiğinde $w_1^{(1)}$ ve $w_2^{(1)}$ arasında doğrusal ilişki olduğu açık bir şekilde görülebilir. Ayrıca bu grafikteki noktalar incelendiğinde 3. gözlem verinin geneline uymadığı gözlenmektedir. Li ve ark. (2001) tarafından önerilen kritere göre, bu nokta oluşturulan deneysel güven elipsinin dışında olduğu için regresyon tahminlerini değiştirme potansiyeline sahiptir [16]. Diğer grafiklerde ise gözlemler düzgün bir yayılım göstermektedir. Yani diğer grafiklerde 3. gözlem dışında verinin geneline uymayan başka gözlem yoktur. Bu durumda 3. gözlemin verideki tek etkili gözlem olduğu sonucuna varılır.

5. SİMÜLASYON ÇALIŞMASI

Bu bölümde önce etkili gözlemlerin saptanması için kullanılan *DFBETAS* , *DFFITS* , Cook Uzaklığı, *COVRATIO* , $S_{(i)}$ istatistikleri ve grafik yöntemi, örnek hacminin ve bağımsız değişken sayısının farklı durumları için simülasyon kullanılarak etkili gözlem içeren verideki etkili gözlemi saptama oranı bakımından karşılaştırılmıştır. Sonra örnek hacmi ve bağımsız değişken sayısı sabit iken, etkili gözlem ve etkili gözlem gruplarının hata terimleri mutlak değer olarak daha da büyütülerek bunların verinin merkezinden uzaklaştırıldığı durumda, bu yöntemler etkili gözlem içeren verideki etkili gözlemi saptama oranı bakımından karşılaştırılmıştır. Bu iki simülasyon çalışmasında hem bir etkili gözlemi saptamak için kullanılan tanı yöntemleri hem de etkili gözlem gruplarını saptamak için kullanılan tanı yöntemleri ele alınmıştır.

DFBETAS , *DFFITS* ve *COVRATIO* tanı istatistiklerine ilişkin karar kuralında yer alan kritik değerler tahmin edilecek parametre sayısı olan p ve örnek hacmi olan n değerlerinin bir fonksiyonu olup örnek hacmi arttıkça bu kritik değerler küçülür. Böylece ilgili tanı istatistiği, verinin geneline uymamasıyla birlikte EKK tahminlerini değiştirme eğilimi düşük olan bir gözlemi bile etkili gözlem olarak saptayabilir. Bu sorunu dikkate alan Belsley ve ark. (1980) *DFBETAS* , *DFFITS* ve *COVRATIO* tanı istatistiklerinin hacmi 100' den büyük örnekler için kullanışlı olmadığını belirtmişlerdir[9]. Bu nedenle simülasyon çalışmasında örnek hacimleri 20, 30 ve 40 olarak alınmıştır. Bağımsız değişken sayısı da 2, 3 ve 4 olarak belirlenmiştir.

Bu bölümde yapılan simülasyon çalışmalarında veri üretmek için Hadi ve Simonoff (1993) tarafından yapılan çalışmadaki veri üretme yönteminden yararlanılmıştır[23]. Hadi ve Simonoff (1993) tarafından yapılan çalışmada veri aykırı gözlem içermek amacıyla üretildiği için bu çalışmada aynı yöntemle veri etkili gözlem içerecek biçimde MATLAB2008a programı kullanılarak üretilmiştir.

Birden fazla etkili gözlemin bulunduğu bir veride etkili gözlem veya gözlem grupları tanı istatistikleri kullanılarak saptanırken karşılaşılabilecek sorunlardan ikisi maskeleye ve yanılgıya düşürmedir. Bu çalışmada, simülasyon maskeleye ve yanılgıya düşürmeyle karşılaşılmayacak biçimde tasarlanmıştır.

Etkili bir gözlemin kaldıraç değeri etkili olmayan gözlemlerinkine göre daha büyüktür. Bu tanımdan yararlanarak 1 indisi ile gösterilecek olan etkili gözleme ilişkin kaldıraç değerini büyütmek amacıyla bu gözlemin $p-1$ sayıda bağımsız değişkenlerin değerleri, verinin geneline uyan gözlemlere ilişkin bağımsız değişkenlerin değerlerinin türetildiği $[0,15]$ aralığının en uç değeri olan 15 olarak belirlenmiştir. Bu gözlemin hata değişken değeri ise $\varepsilon_1 = -5$ olarak alınmıştır. Buradaki amaç bu gözlemin hata değişken değerini mutlak değerce arttırarak, bu gözlemin artık değerini mutlak değerce büyülmektir. Sonra parametrelerinin değerleri 1 olan

$$Y_1 = 1 + X_{11} + \dots + X_{1,p-1} + \varepsilon_1$$

modeline göre etkili gözlemin bağımlı değişken değeri türetilmiştir. Simülasyon boyunca etkili gözleme ilişkin bağımsız değişkenlerin değerleri ve bağımlı değişken değerleri sabit kalmıştır.

Verinin geneline uyan $n-1$ sayıda gözlem için $p-1$ sayıda bağımsız değişken değerleri $[0,15]$ aralığında tekdüze dağılımdan türetilmiştir. Sonra simülasyon aşağıdaki adımlar doğrultusunda yapılmıştır.

Adım 1: Bağımsız değişken sayısı $p-1$ olmak üzere, $p \times 1$ boyutlu β parametre vektörünün tüm elemanlarına 1 değeri atanır.

Adım 2: Hata değişkenlerinin değerleri, $\varepsilon_i \sim N(0,1)$ dağılımından üretilir.

Adım 3: $p-1$ sayıda bağımsız değişkenlerin değerleri ve hata değişkenlerinin değerleri kullanılarak ilgili modele yani

$$Y_i = \beta_0 + \beta_1 X_{i1} + \dots + \beta_{p-1} X_{i,p-1} + \varepsilon_i \quad (i = 2, 3, \dots, n)$$

modeline göre bağımlı değişken değerleri türetilir.

Adım 4: Türetilen verideki her bir gözlem için $DFBETAS$, $DFFITS$, $COVRATIO$, D_i ve S_i istatistiklerinin değerleri hesaplanır ve her bir tanı istatistiğine ilişkin karar kuralına göre 1 indisi ile gösterilen gözlemin etkili bir gözlem olup olmadığına karar verilir.

Adım 5: Aynı veri için grafik yöntemine göre tanı grafikleri oluşturulur. Eş. (4.6)' da verilen ve göreceli duyarlılık faktörü olarak tanımlanan karar değişkenine göre etkili gözlemi belirleme gücü en yüksek olan grafik alınır. Bu grafikte etkili gözlem haricindeki diğer gözlemlere ilişkin noktaların merkeze uzaklıkları

$$r = (w - w_0^{(j)})^T [M^{(j)}]^{-1} (w - w_0^{(j)})$$

ile hesaplanıp en büyük uzaklık deneysel güven elipsinin yarıçap uzunluğu olan ρ olarak alınır. Eğer 1 indisi ile gösterilen gözleme ilişkin noktanın elipsin merkezine uzaklığı, ρ değerinden büyük ise bu nokta deneysel güven elipsinin dışındadır ve grafik yöntemine göre bu gözlem etkili bir gözlemdir.

Her 1 000 tekrarda verinin geneline uyan gözlemler için bağımsız değişkenlerin değerleri yeniden üretilmek üzere bu deneme 100 000 kez tekrarlanmıştır. Sonra etkili gözlem içeren veride, bir etkili gözlemi saptamak için kullanılan bu tanı yöntemlerinin etkili bir gözlemi tespit etme oranları hesaplanmıştır.

5.1. Bağımsız Değişken Sayısı ve Örnek Hacminin Farklı Değerleri için Tanı Yöntemlerinin Karşılaştırılması

Bağımsız değişken sayısı ve örnek hacminin farklı değerleri için tanı yöntemlerinin karşılaştırılmasında önce etkili gözlem içeren bir veride bir etkili gözlemi saptamaya dayalı tanı yöntemleri, etkili gözlemi saptama oranı bakımından karşılaştırılmıştır. Sonra aynı bağımsız değişken sayıları ve örnek hacimleri için etkili bir gözlem grubu içeren veride etkili gözlem grubunu saptamaya dayalı tanı yöntemleri, bu etkili gözlem grubunu saptama oranı bakımından karşılaştırılmıştır.

5.1.1. Bir etkili gözlem için tanı yöntemlerinin karşılaştırılması

Bağımsız değişken sayısının ve örnek hacminin farklı değerleri için ilk olarak bir etkili gözlem için tanı yöntemleri ele alınmıştır. İki bağımsız değişkenin olduğu bir model, yani $p = 3$ durumu, için bölümün başında belirtilen adımlar doğrultusunda yapılan simülasyonla elde edilen sonuçlar Çizelge 5.1’ de verilmiştir.

Çizelge 5.1. $p = 3$ iken bir etkili gözlem için tanı yöntemlerinin etkili bir gözlem içeren veride etkili gözlemi saptama oranları

$p = 3$			
	$n = 20$	$n = 30$	$n = 40$
$DFITS_{(1)}$	0,9995	0,9997	0,9997
$DFBETAS_{1(1)}$	0,9994	0,9997	0,9998
$DFBETAS_{2(1)}$	0,9947	0,9994	0,9998
$DFBETAS_{3(1)}$	0,9948	0,9996	0,9998
$D_{(1)}$	0,8759	0,7744	0,6027
$COVRATIO_{(1)}$	0,9641	0,9955	0,9996
$S_{(1)}$	0,9614	0,9820	0,9870
Grafik Yöntemi	0,9440	0,9852	0,9949

Çizelge 5.1’ deki sonuçlar incelendiğinde, farklı örnek hacimleri altında $DFFITS_{(i)}$ ve $DFBETAS_{j(i)}$ istatistiklerinin etkili gözlemi saptama oranlarının büyük olduğu görülmektedir. Bununla birlikte, Cook Uzaklığı istatistiğinin etkili gözlemi saptama oranı örnek hacmi arttıkça düşmektedir. $COVRATIO_{(i)}$, $S_{(i)}$ istatistiklerinin ve grafik yönteminin etkili gözlemi saptama oranları ise örnek hacmi arttıkça büyümektedir.

Modelde üç bağımsız değişkenin olduğu durum yani $p = 4$ iken bir etkili gözlem için tanı istatistikleri ve grafik yöntemini karşılaştırmak amacıyla yapılan simülasyon bölümün başında belirtilen adımlar doğrultusunda oluşturulmuştur. Elde edilen sonuçlar Çizelge 5.2’ de verilmiştir.

Çizelge 5.2. $p = 4$ iken bir etkili gözlem için tanı yöntemlerinin etkili bir gözlem içeren veride etkili gözlemi saptama oranları

$p = 4$			
	$n = 20$	$n = 30$	$n = 40$
$DFFITS_{(1)}$	0,9993	0,9995	0,9998
$DFBETAS_{1(1)}$	0,9994	0,9995	0,9995
$DFBETAS_{2(1)}$	0,9623	0,9922	0,9978
$DFBETAS_{3(1)}$	0,9647	0,9921	0,9979
$DFBETAS_{4(1)}$	0,9645	0,9926	0,9984
$D_{(1)}$	0,8431	0,7228	0,5634
$COVRATIO_{(1)}$	0,8978	0,9839	0,9967
$S_{(1)}$	0,9935	0,9988	0,9993
Grafik Yöntemi	0,8372	0,9486	0,9809

Yukarıdaki çizelgede verilen sonuçlara göre örnek hacminin tüm durumları için $DFFITS_{(i)}$, $DFBETAS_{j(i)}$ ve $S_{(i)}$ istatistiklerinin etkili gözlemi saptama oranları yüksektir. Fakat örnek hacmi 20 iken $DFBETAS_{j(i)}$ istatistiğinin etkili gözlemi saptama oranı $DFFITS_{(i)}$ ve $S_{(i)}$ istatistiklerine ilişkin oranlara göre daha düşüktür.

Ayrıca bu sonuçlara göre örnek hacmi arttıkça etkili gözlemi saptama oranı artan yöntemler $COVRATIO_{(i)}$ istatistiği ve grafik yöntemidir. Cook Uzaklığı istatistiği için sonuçlar incelendiğinde örnek hacmi arttıkça bu istatistiğe ilişkin etkili gözlemi saptama oranının düştüğü görülmektedir.

Bir etkili gözlemi saptamak için kullanılan tanı yöntemlerine ilişkin son simülasyon bağımsız değişken sayısı 4 iken, bölümün başında belirtilen adımlar doğrultusunda yapılmıştır. Simülasyon sonuçları Çizelge 5.3' te verilmiştir.

Çizelge 5.3. $p = 5$ iken bir etkili gözlem için tanı yöntemlerinin etkili bir gözlem içeren veride etkili gözlemi saptama oranları

$p = 5$			
	$n = 20$	$n = 30$	$n = 40$
$DFFITS_{(1)}$	0,9994	0,9996	0,9998
$DFBETAS_{1(1)}$	0,9998	0,9995	0,9998
$DFBETAS_{2(1)}$	0,9221	0,9709	0,9885
$DFBETAS_{3(1)}$	0,9171	0,9678	0,9889
$DFBETAS_{4(1)}$	0,9206	0,9702	0,9895
$DFBETAS_{5(1)}$	0,9168	0,9714	0,9895
$D_{(1)}$	0,7831	0,6843	0,5234
$COVRATIO_{(1)}$	0,7926	0,9593	0,9895
$S_{(1)}$	0,9967	0,9998	0,9995
Grafik Yöntemi	0,7139	0,8851	0,9500

Çizelge 5.3' te verilen sonuçlar incelendiğinde, $p = 4$ durumunda olduğu gibi $DFFITS_{(i)}$, $DFBETAS_{j(i)}$ ve $S_{(i)}$ istatistiklerinin etkili olarak üretilen gözlemi saptama oranlarının ele alınan tüm örnek hacimlerinde büyük olduğu görülmektedir. Fakat örnek hacmi 20 iken $DFBETAS_{j(i)}$ istatistiğinin bu etkili gözlemi saptama oranı, $DFFITS_{(i)}$ ve $S_{(i)}$ istatistiği için elde edilen oranlara göre daha düşüktür. $COVRATIO_{(i)}$ istatistiği ve grafik yöntemine ilişkin etkili gözlemi saptama oranları

incelendiğinde $n = 20$ iken bu yöntemlere ilişkin etkili gözlemi saptama oranlarının küçük olmasıyla birlikte örnek hacmi arttıkça bu yöntemlere ilişkin oranların arttığı görülmektedir. Son olarak örnek hacmi arttıkça Cook Uzaklığı istatistiğinin etkili gözlem olan ve 1 indisi ile gösterilen gözlemi saptama oranının düştüğü gözlenmektedir.

Örnek hacmi artarken $DFFITs_{(i)}$, $DFBETAS_{j(i)}$ ve $COVRATIO_{(i)}$ istatistiklerinin etkili bir gözlemi saptama oranlarının artmasının nedeni bu istatistiklere ilişkin karar kuralında yer alan kritik değerlerin örnek hacminin azalan bir fonksiyonu olmasıdır. Özellikle bu artış $COVRATIO_{(i)}$ istatistiği için daha açık bir şekilde görülmektedir.

Bağımsız değişken sayısının tüm durumlarında örnek hacmi arttıkça Cook Uzaklığı istatistiğinin etkili gözlemi saptama oranı düşmektedir. Çünkü gözlem silme tekniğine dayalı olarak etkili bir gözlem için elde edilen $\hat{\beta}_{(i)}$ tahmini, örnek hacmi arttıkça $\hat{\beta}_{(i)}$ tahminleri için oluşturulan ve bir elipsoide karşılık gelen güven bölgesine yaklaşır. Bu durumda etkili gözlem için elde edilen $\hat{\beta}_{(i)}$ tahmininin güven elipsoidinin merkezine uzaklığının karesi olan bu gözleme ilişkin Cook Uzaklığı istatistiği değeri küçülür. Böylece Cook Uzaklığı yöntemiyle etkili gözlemin saptanma oranı düşer.

Pena' nın $S_{(i)}$ istatistiğine ilişkin sonuçlar incelendiğinde bağımsız değişken sayısının ve örnek hacminin tüm durumlarında bu istatistiğe ilişkin saptama oranlarının büyük olduğu gözlenmektedir. Bunun nedeni $S_{(i)}$ istatistiğine ilişkin karar kuralından kaynaklanmaktadır. Bu istatistiğe ilişkin karar kuralı

$$\left| S_{(i)} - \text{med}(S) \right| \geq 4.5 \times \text{MAD}(S_{(i)})$$

ile verilir. Burada $\text{MAD}(S_{(i)})$ ifadesi $S_{(i)}$ değerlerinin medyanadan sapmalarının mutlak değerlerinin medyanıdır. Yukarıdaki karar kuralına bağlı olarak etkili gözlem

için hesaplanan $S_{(i)}$ değerinin medyandan mutlak değer bakımından sapması, verinin geneline uyan gözlemlerinkine göre büyüktür. Bağımsız değişken sayısının tüm durumlarında örnek hacmi artarken $\text{med}(S)$ ve $\text{MAD}(S_{(i)})$ değerlerinde büyük bir değişim olmamaktadır. Böylece örnek hacmi arttıkça etkili gözlem için hesaplanan $S_{(i)}$ değerinin medyandan mutlak değerce sapması büyük bir değişim göstermediği için bu istatistiğin etkili gözlemi saptama oranı yüksek olacaktır.

Grafik yöntemine ilişkin sonuçlar incelendiğinde bağımsız değişken sayısının tüm durumlarında örnek hacmi arttıkça bu yönteme ilişkin etkili gözlemi saptama oranının arttığı görülmektedir. Eğer veride etkili gözlem yoksa tüm grafiklerde gözlemlere ilişkin noktalar geniş bir yayılım gösterecektir. Fakat veride en az bir etkili gözlem varsa görelî duyarlılık faktörüne bağılı olarak belirlenen grafikte verinin geneline uyan noktalar grafiğin merkezi olan $(0,0)$ noktası etrafında kümelenecektir ve etkili gözlem ilişkin nokta da verinin geneline uyan noktaların oluşturduğu bu kümeden uzak bir konumda bulunacaktır. Veride en az bir etkili gözlem varken bağımsız değişken sayısının tüm durumları için örnek hacmi arttıkça verinin geneline uyan gözlemler grafiğin merkezine daha da yaklaşmakta ve buna bağılı olarak güven elipsleriyle belirlenen güven bölgesi daha da daralmaktadır. Böylece grafik yönteminin bir etkili gözlemi saptama oranı örnek hacmi arttıkça artmaktadır.

Örnek hacmi sabit tutulduğunda ele alınan tanı yöntemlerinin Çizelge 5.1, Çizelge 5.2 ve Çizelge 5.3' teki sonuçlara göre incelenmesi bu tanı yöntemlerinin bağımsız değişken sayısındaki artışa göre değerlendirilmesi açısından önemlidir. Örnek hacmi sabit iken $DFFITs_{(i)}$ ve $S_{(i)}$ tanı istatistiklerinin etkili gözlemi saptama oranlarının büyük değişimler göstermediği açıkça görülmektedir. Bununla birlikte bağımsız değişken sayısı arttıkça $DFBETAS_{j(i)}$, $COVRATIO_{(i)}$, Cook Uzaklığı istatistikleri ve grafik yöntemine ilişkin etkili gözlemi saptama oranları azalmaktadır. Özellikle bu azalma grafik yöntemi için diğer yöntemlerinkine nazaran büyüktür. Bu da regresyon probleminin boyutu arttıkça grafik yönteminin etkili gözlemi saptama duyarlılığının azaldığını göstermektedir.

5.1.2. Etkili gözlem grupları için tanı yöntemlerinin karşılaştırılması

Etkili gözlem gruplarını saptamak için kullanılan tanı yöntemlerinden $COVRATIO_{(I)}$, Cook Uzaklığı istatistikleri ve grafik yöntemini karşılaştırmak amacıyla yapılan simülasyon bir etkili gözlemi saptamak için kullanılan tanı yöntemlerini karşılaştırmak amacıyla yapılan simülasyona benzer biçimde yapılmıştır. Fakat bu simülasyonda iki etkili gözlemden oluşan etkili bir gözlem grubunun bulunduğu veriler üretilmiştir. Bu amaç doğrultusunda, üretilen verilerde etkili gözlem grubunun elemanları 1 ve 2 indisleri ile gösterilirse bu gözlemlerin bağımsız değişkenlerinin değerleri 15 ve hata değişkenlerinin değerleri $\varepsilon_1 = -5$ ve $\varepsilon_2 = -5,5$ olarak belirlenmiştir. Sonra

$$Y_i = \beta_0 + \beta_1 X_{i1} + \dots + \beta_{p-1} X_{i,p-1} + \varepsilon_i \quad (i = 3, 4, \dots, n)$$

modeli dikkate alınarak bölümün başında belirtilen adımlar doğrultusunda simülasyon yapılmıştır.

Bağımsız değişken sayısının iki olduğu durum yani $p = 3$ durumu için etkili gözlem grubu içeren bir veride etkili gözlem gruplarını saptamak için kullanılan üç tanı yöntemine ilişkin simülasyon sonuçları Çizelge 5.4' te verilmiştir.

Çizelge 5.4. $p = 3$ iken etkili gözlem grupları için tanı yöntemlerinin iki gözlemden oluşan etkili gözlem grubu içeren veride etkili gözlem grubunu saptama oranları

$p = 3$			
	$n = 20$	$n = 30$	$n = 40$
$D_{\{1,2\}}$	0,9989	0,9997	0,9998
$COVRATIO_{\{1,2\}}$	0,9790	0,9991	0,9995
Grafik Yöntemi	0,7450	0,8913	0,9253

Çizelge 5.4’ teki sonuçlar incelendiğinde Cook Uzaklığı ve $COVRATIO_{\{t\}}$ istatistiklerinin bu etkili gözlem grubunu saptama oranlarının ele alınan tüm örnek hacimlerinde oldukça büyük olduğu görülmektedir. Bununla birlikte tüm örnek hacimleri için diğer yöntemlere nazaran grafik yöntemin etkili gözlem grubunu saptama oranının düşük olduğu gözlenmektedir. Ayrıca örnek hacmi arttıkça grafik yönteminin bu etkili gözlem gurubunu saptama oranı da artmaktadır.

Bağımsız değişken sayısının üç olduğu durum yani $p = 4$ durumu için etkili gözlem gruplarını saptamaya dayalı tanı yöntemlerine ilişkin simülasyon sonuçları Çizelge 5.5’ te verilmiştir.

Çizelge 5.5. $p = 4$ iken etkili gözlem grupları için tanı yöntemlerinin iki gözlemden oluşan etkili gözlem grubu içeren veride etkili gözlem grubunu saptama oranları

$p = 4$			
	$n = 20$	$n = 30$	$n = 40$
$D_{\{1,2\}}$	0,9992	0,9996	0,9998
$COVRATIO_{\{1,2\}}$	0,9389	0,9937	0,9993
Grafik Yöntemi	0,3130	0,5917	0,7232

Çizelge 5.5’ teki sonuçlar incelendiğinde Cook Uzaklığı istatistiğine ilişkin etkili gözlem grubunu saptama oranının örnek hacminin tüm durumları için yüksek olduğu açıkça görülmektedir. $COVRATIO_{\{t\}}$ istatistiği ve grafik yöntemine ilişkin etkili gözlem grubunu saptama oranları ise örnek hacmi arttıkça artmaktadır. Buna ek olarak grafik yöntemi $n = 20$ iken etkili gözlem grubunu oldukça küçük bir oranla saptamaktadır.

Bağımsız değişken sayısının dört olduğu durum yani $p = 5$ durumu için etkili gözlem grupları için tanı yöntemlerine ilişkin simülasyon sonuçları Çizelge 5.6’ da verilmiştir.

Çizelge 5.6. $p = 5$ iken etkili gözlem grupları için tanı yöntemlerinin iki gözlemden oluşan etkili gözlem grubu içeren veride etkili gözlem grubunu saptama oranları

$p = 5$			
	$n = 20$	$n = 30$	$n = 40$
$D_{\{1,2\}}$	0,9986	0,9991	0,9996
$COVRATIO_{\{1,2\}}$	0,8733	0,9855	0,9980
Grafik Yöntemi	0,1826	0,3340	0,5654

Çizelge 5.6' daki sonuçlar incelendiğinde, farklı örnek hacimleri altında Cook Uzaklığı istatistiğinin etkili gözlem grubunu saptama oranının dikkate değer biçimde büyük olduğu görülmektedir. Bununla birlikte $COVRATIO_{\{1\}}$ istatistiği ve grafik yöntemine ilişkin oranlar farklı örnek hacimleri altında Cook Uzaklığı istatistiği için elde edilen oranlara göre daha küçüktür. Ayrıca bu iki yöntem için elde edilen etkili gözlem grubunu saptama oranları örnek hacmi arttıkça sayısal olarak büyümektedir.

Örnek hacmi $n = 20$ iken etkili gözlem gruplarını saptamak için kullanılan tanı yöntemlerine ilişkin Çizelge 5.4, Çizelge 5.5 ve Çizelge 5.6' daki sonuçlar incelendiğinde bağımsız değişken sayısı arttıkça $COVRATIO_{\{1\}}$ istatistiği ve grafik yöntemi için bulunan etkili bir gözlem grubunu saptama oranlarının küçüldüğü gözlenmektedir. Özellikle bu küçülme grafik yöntemi için daha da açık bir şekilde görülmektedir. Bununla birlikte ele alınan tüm örnek hacimleri ve bağımsız değişken sayıları için Cook Uzaklığı istatistiğine ilişkin etkili gözlem grubunu saptama oranları diğer yöntemler göre büyüktür.

Cook Uzaklığı istatistiğinin etkili bir gözlem grubunu saptama oranının ele alınan tüm örnek hacimleri ve bağımsız değişken sayıları için yüksek olması, etkili bir gözlem grubunun etkili bir gözleme göre EKK tahminlerini değiştirme eğiliminin daha yüksek olmasından dolayıdır. Bir başka ifadeyle m hacimli etkili bir gözlem

grubunun bulunduğu veride bu etkili gözlem grubunun silinmesiyle geri kalan gözlemlerden elde edilen EKK tahmini $\hat{\beta}_{(i)}$ ile tüm gözlemlere dayalı olarak elde edilen $\hat{\beta}$ arasındaki karesel uzaklık, bir etkili gözlemin bulunduğu veride bu etkili gözlemin silinmesiyle diğer gözlemlerden elde edilen $\hat{\beta}_{(i)}$ tahmini ile $\hat{\beta}$ tahmini arasındaki karesel uzaklıktan büyük olacaktır. Bundan dolayı etkili gözlem gruplarını saptamaya dayalı Cook Uzaklığı istatistiği, bir etkili gözlemi saptamaya dayalı olarak kullanılan Cook Uzaklığı istatistiğinden daha yüksek bir saptama oranına sahiptir.

Bağımsız değişken sayısı sabitken, örnek hacmi arttıkça $COVRATIO_{\{i\}}$ istatistiği için bulunan etkili gözlem grubunu saptama oranlarının artmasının nedeni ise bu istatistiğe ilişkin karar kuralında yer alan kritik değerin örnek hacminin azalan bir fonksiyonu olmasıdır.

Grafik yöntemi, çok boyutlu bir regresyon problemini iki boyutlu bir probleme dönüştürmek suretiyle oluşturulan p sayıda tanı grafikleri setinin, her bir grafikteki noktalar için oluşturulan güven bölgesi kriterine bağlı olarak incelemesine dayalıdır. Veride bir etkili gözlem grubu bulunduğu anda bu etkili gözlem grubunu saptama oranı bakımından grafik yönteminin incelenmesi, bu yöntemin regresyon problemini nasıl dönüştürdüğünü değerlendirmek açısından önemlidir. Bağımsız değişken sayısı sabitken grafik yöntemiyle etkili gözlem grubunu saptama oranlarına ilişkin sonuçlar incelendiğinde örnek hacminin artmasıyla bu yöntemin etkili gözlem grubunu daha iyi saptadığı açıkça görülmektedir. Fakat örnek hacmi sabit iken bağımsız değişken sayısı arttıkça grafik yönteminin etkili gözlem grubunu saptama oranı düşmektedir. Buna bağlı olarak regresyon probleminin boyutu arttıkça grafik yönteminin regresyon problemini dönüştürme eğiliminin azaldığı sonucuna varılır.

5.2. Bağımsız Değişken Sayısı ve Örnek Hacmi Sabitken Tanı Yöntemlerinin Karşılaştırılması

Bu simülasyonda, bağımsız değişken sayısı ve örnek hacmi sabit iken etkili gözlemin ve etkili gözlem grubunu oluşturan gözlemlerin hata değişken değerleri mutlak değerce artırılarak, bunların verinin merkezinden uzaklaştırıldığı durumlarda, tanı yöntemlerinin etkili gözlem içeren veride etkili gözlemi ve etkili gözlem grubunu saptama oranları elde edilmiştir. Buradaki amaç etkili gözlem veya etkili gözlem grubundaki gözlemlerin hata değişken değerleri daha da arttığı durumlarda tanı yöntemlerini karşılaştırmaktır.

İlk olarak bir etkili gözlemi saptamak için kullanılan tanı yöntemleri örnek hacmi 20 ve bağımsız değişken sayısı 2 iken bölümün başında belirtilen adımlar doğrultusunda simülasyonla incelenmiştir. Simülasyonda önce etkili gözlem olarak belirlenen ve 1 indisi ile gösterilen gözleminin hata değişken değeri $\varepsilon_1 = -5$ alınmıştır. Döngü

$$Y_i = \beta_0 + \beta_1 X_{i1} + \dots + \beta_{p-1} X_{i,p-1} + \varepsilon_i \quad (i = 2, 3, \dots, n)$$

modeli dikkate alınarak bölümün başında verilen adımlar doğrultusunda, her 1 000 tekrara da bir verinin geneline uyan gözlemlerin bağımsız değişkenlerinin değerleri yeniden üretilmek üzere 100 000 kez tekrar edilmiştir ve tanı yöntemlerinin etkili olan bu gözlemi saptama oranları elde edilmiştir. Sonra etkili gözlemin hata değişken değeri önce $\varepsilon_1 = -7,5$ sonra da $\varepsilon_1 = -10$ alınıp bu değerler için aynı simülasyon tekrarlanmıştır. Sonuçlar Çizelge 5.7' de verilmiştir.

Çizelge 5.7. $n = 20$ ve $p = 3$ iken bir etkili gözlem için tanı yöntemlerinin farklı hata değişken değerlerine göre üretilen etkili gözlemi saptama oranları

$n = 20, p = 3$			
	$\varepsilon_1 = -5$	$\varepsilon_1 = -7,5$	$\varepsilon_1 = -10$
$DFFITs_{(1)}$	0,9995	1	1
$DFBETAS_{1(1)}$	0,9994	1	1
$DFBETAS_{2(1)}$	0,9947	0,9985	0,9989
$DFBETAS_{3(1)}$	0,9948	0,9985	0,9990
$D_{(1)}$	0,8759	0,9958	0,9997
$COVRATIO_{(1)}$	0,9641	0,9996	1
$S_{(1)}$	0,9614	0,9886	0,9932
Grafik Yöntemi	0,9440	0,9943	0,9989

Bağımsız değişken sayısı ve örnek hacmi sabitken, veride kaldıraç değeri diğer gözlemlerinkine göre büyük olan herhangi bir gözlemin hata değişken değeri arttıkça, bu gözlemin regresyon tahminlerini değiştirme eğiliminin artması beklenir. Çizelge 5.7’ deki sonuçlar da bunu doğrulamaktadır. Çünkü etkili gözlemin hata değişkeni değeri arttıkça etkili bir gözlemin saptanması için kullanılan bu altı tanı yönteminin verideki etkili gözlemi saptama oranları artmaktadır. Özellikle bu artış Cook Uzaklığı istatistiği için daha açık bir şekilde görülmektedir.

$n = 20$ ve $p = 3$ iken etkili gözlem gruplarını saptamak için kullanılan tanı yöntemlerine ilişkin simülasyon bir öncekine benzer biçimde bölümün başında belirtilen adımlar doğrultusunda yapılmıştır. Simülasyon yapılırken etkili gözlem grubunu oluşturan iki gözlemin hata değişkeni değerleri sırasıyla $\varepsilon_1 = -5$ ve $\varepsilon_2 = -5,5$ alınıp bu gözlemlerin bağımsız değişken değerleri de kullanılarak bağımlı değişken değerleri elde edilmiştir. Sonra bölümün başında verilen adımlar doğrultusunda, her 1 000 tekrara da bir verinin geneline uyan gözlemlerin bağımsız değişkenlerinin değerleri yeniden üretilmek üzere

$$Y_i = \beta_0 + \beta_1 X_{i1} + \dots + \beta_{p-1} X_{i,p-1} + \varepsilon_i \quad (i = 3, 4, \dots, n)$$

modeli dikkate alınıp, bu deneme 100 000 kez tekrar edilmiştir ve etkili bir gözlem grubu içeren veride etkili gözlem gruplarını saptamak için kullanılan tanı yöntemlerinin bu etkili gözlem grubunu saptama oranları elde edilmiştir. Aynı simülasyon bu iki gözlemin hata değişkeni değerleri önce $\varepsilon_1 = -7$, $\varepsilon_2 = -7,5$ için sonra $\varepsilon_1 = -10$, $\varepsilon_2 = -10,5$ için yapıp sonuçlar Çizelge 5.8’ de verilmiştir.

Çizelge 5.8. $n = 20$ ve $p = 3$ iken etkili gözlem grupları için tanı yöntemlerinin farklı hata değişken değerlerine göre üretilen iki etkili gözlemin oluşturduğu etkili gözlem grubunu saptama oranları

$n = 20, p = 3$			
	$\varepsilon_1 = -5$ $\varepsilon_2 = -5,5$	$\varepsilon_1 = -7$ $\varepsilon_2 = -7,5$	$\varepsilon_1 = -10$ $\varepsilon_2 = -10,5$
$D_{\{1,2\}}$	0,9989	1	1
$COVRATIO_{\{1,2\}}$	0,9790	0,9987	1
Grafik Yöntemi	0,7450	0,8829	0,9094

Çizelge 5.8’ deki sonuçlar incelendiğinde verideki etkili gözlem grubunu oluşturan gözlemlerin hata değişkeni değerlerinin artmasıyla $COVRATIO_{(I)}$, Cook Uzaklığı istatistikleri ve grafik yönteminin bu gözlem grubunu saptama oranlarının arttığı gözlenmektedir. $COVRATIO_{(I)}$ ve Cook Uzaklığı istatistiklerinin etkili gözlem grubunu saptama oranlarının ele alınan tüm örnek hacimlerinde yüksek olmasıyla birlikte grafik yöntemi için deneysel olarak elde edilen oran bu iki istatistiğe göre daha küçüktür.

6. SONUÇ VE ÖNERİLER

Bu çalışmanın birinci bölümünde aykırı ve etkili gözlemler tanıtılıp, veride etkili gözlem veya etkili gözlem grupları bulunduğunda regresyon sonuçlarının nasıl bir değişim göstereceği hakkında bilgi verilmiştir. Ayrıca literatürdeki bu konu ile ilgili bazı çalışmalara da yer verilmiştir.

İkinci bölümde regresyon analizi ve tanı yöntemlerine ilişkin temel kavramlar verilmiştir. Etkili gözlemlerin mutlak değerce artıkları ve kaldıraç değerleri büyük olduğu için doğal olarak tanı yöntemleri bunlara ilişkin değişkenlere bağlıdır. Bundan dolayı tanı yöntemlerinin temelini oluşturan bu değişkenler etkili gözlemlerin saptanması için yapılan analizlerde tanı yöntemlerine yardımcı bir rol üstlenir.

Üçüncü bölümde bir etkili gözlemi ve etkili gözlem gruplarını saptamak için uygulamada sıkça kullanılan tanı istatistiklerine yer verilmiştir. Ayrıca bu istatistiklere ilişkin karar kuralları da belirtilmiştir. Bunlara ek olarak tanı istatistikleri kullanılırken karşılaşılabilecek iki sorun olan maskeleye ve yanılgıya düşürme hakkında bilgi verilmiştir.

Dördüncü bölümde çoklu regresyonda etkili gözlem veya gözlem gruplarının saptanması için kullanılan yeni bir yöntem olan grafik yöntemine yer verilmiştir. Grafik yöntemi yüksek boyutlu bir regresyon problemini iki boyutlu tanı grafiklerinin bir setine indirgeyerek, bu grafiklerin görsel olarak incelenmesine dayalı bir tekniktir. Bu dönüştürme işlemi hesaplama bakımından benzer grafiksel yöntemlere nazaran daha kolaydır.

Bu çalışmanın beşinci bölümünde ise Cook uzaklığı, $DFBETAS$, $DFFITs$, $COVRATIO$, S_i istatistikleri ve grafik yöntemi simülasyon çalışmasıyla, etkili gözlem veya gözlem grubu içeren veride bu gözlem veya gözlem grubunu saptama oranı bakımından karşılaştırılmıştır. Elde edilen sonuçlara göre bağımsız değişken

sayısı ve örnek hacminin tüm durumları için bir etkili gözlemi saptamaya dayalı tanı istatistiklerinden $DFFITS_{(i)}$ ve $DFBETAS_{(i)}$ istatistikleri etkili gözlemi yüksek bir oranla saptamaktadır. Fakat $p=3$ durumu için örnek hacmi küçük iken $DFFITS_{(i)}$ istatistiğinin $DFBETAS_{(i)}$ istatistiğine göre etkili gözlemi daha yüksek bir oranla saptadığı gözlenmektedir. Diğer bir tanı istatistiği olan $COVRATIO_{(i)}$ istatistiği, gerek bir etkili gözlemi gerekse etkili bir gözlem gurubunu saptamak için kullanılsın, örnek hacmi arttıkça bu istatistiğine ilişkin saptama oranının arttığı açıkça görülmektedir. Bu çalışmada incelenen etkili bir gözlemi saptamaya dayalı tanı yöntemlerinden üçü olan $DFFITS_{(i)}$, $DFBETAS_{(i)}$ ve $COVRATIO_{(i)}$ istatistiklerinin örnek hacmi arttıkça etkili gözlemi saptama oranlarının artmasının nedeni bu istatistiklere ilişkin kritik değerlerin örnek hacminin azalan bir fonksiyonu olmasıdır. Çünkü n teriminin değerinin artmasıyla bu istatistiklere ilişkin kritik değerler sayısal olarak düşmektedir. Buna bağlı olarak da bu üç tanı istatistiği etkili gözlemi daha düşük bir oranla saptamaktadır.

Bir etkili gözlemi saptamak için kullanılan tanı yöntemlerinden Pena' nın $S_{(i)}$ istatistiği ise etkili gözlemi saptama oranı yüksek olan diğer bir istatistiktir. Bunun nedeni $S_{(i)}$ istatistiğine ilişkin karar kuralından kaynaklanmaktadır. Örnek hacmi kaç olursa olsun $S_{(i)}$ değerlerinin medyanı sayısal olarak küçük değişimler gösterir. i . gözlemin etkili olduğuna, bu gözlem için elde edilen $S_{(i)}$ değerinin, medyandan sapmasına dayanarak karar verilir. Etkili gözleme ilişkin $S_{(i)}$ değeri, değer aralığındaki en uç değer olduğundan dolayı örnek hacmi kaç olursa olsun bu istatistik ilgili gözlemi etkili olarak saptayacaktır.

Verinin tümünden hesaplanan EKK tahmin vektörü $\hat{\beta}$ ile verinin bir alt kümesinin silinmesiyle hesaplanan EKK tahmin vektörü $\hat{\beta}_{(i)}$ arasındaki karesel uzaklığa dayanan Cook Uzaklığı istatistiğinin bir etkili gözlemi saptama oranı aynı örnek hacmi ve aynı bağımsız değişken sayısı altında diğer yöntemlerin saptama oranlarına

göre daha düşüktür. Ayrıca örnek hacmi arttıkça Cook Uzaklığı istatistiğinin etkili gözlemi saptama oranı düşmektedir. Çünkü etkili bir gözlem için elde edilen $\hat{\beta}_{(i)}$ tahmini örnek hacmi arttıkça $\hat{\beta}_{(i)}$ tahminleri için oluşturulan ve bir elipsoide karşılık gelen güven bölgesine yaklaşmaktadır. Fakat veride etkili bir gözlem gurubu mevcutsa, bu etkili gözlem gurubunun etkili bir gözleme nazaran regresyon tahminlerini değiştirme eğilimi daha fazla olduğu için bu grubun veriden silinmesiyle geri kalan gözlemlerden elde edilen $\hat{\beta}_{(i)}$ tahmini $\hat{\beta}_{(i)}$ tahminleri için oluşturulan güven bölgesine uzak olacaktır. Böylece Cook Uzaklığı istatistiğiyle bu etkili gözlem grubu daha kolay saptanmaktadır. Simülasyon sonuçları da bunu desteklemektedir.

Bu çalışmada incelenen son yöntem olan grafik yöntemi için elde edilen simülasyon sonuçlarına göre, bağımsız değişken sayısının tüm durumlarında örnek hacmi arttıkça bu yöntem için elde edilen, bir etkili gözlemi saptama oranları da artmaktadır. Bunun nedeni grafik tekniği yöntemi gereğince, yüksek boyutlu bir regresyon probleminin iki boyutlu bir probleme indirgenmesi sonucu elde edilen $w_l^{(j)}$ vektörlerinin ($l = 1, 2; j = 1, 2, \dots, p$) değerlerine bağlı olarak oluşturulan grafiklerde, örnek hacmi arttıkça, verinin geneline uyan gözlemlere ilişkin noktaların, $w_l^{(j)}$ vektörünün konum parametreleriyle belirlenen güven bölgesinin merkezinde kümelenmesi ve etkili gözleme ilişkin noktanın bu kümeden konumca uzak kalmasıdır.

Örnek hacmi ve bağımsız değişken sayısı sabit tutulup etkili gözlemin mutlak değerce hata değişken değerini arttırarak bu gözlemin verinin merkezinden uzaklaştırıldığı durumlar için yapılan simülasyon çalışmasının sonuçları incelendiğinde bir etkili gözlemi saptamaya dayalı tanı yöntemlerinin, etkili gözleme ilişkin hata değişkeni değerine bağlı olarak bu gözlemin mutlak değerce artık değerinin artmasıyla, bu gözlemi saptama oranının daha büyük olduğu gözlenmektedir. Aynı durum etkili gözlem grubunu saptamaya dayalı yöntemler için de geçerlidir. Etkili bir gözlem, tanımı gereği kaldıraç değeri ve mutlak değer bakımından artık değeri diğer gözlemlerinkine nazaran büyük olan gözlemdir.

Kaldıraç değeri büyük olan bir gözlemin hata değişken değerine bağlı olarak mutlak değerce arttığını büyütmek bu gözlemin regresyon sonuçlarını değiştirme eğilimini güçlendirmektedir. Buna bağlı olarak da ilgili tanı yöntemine ilişkin etkili gözlemi saptanma oranı büyük olacaktır. Aynı durum etkili gözlem gruplarının saptanması için kullanılan tanı yöntemleri için de geçerlidir.

Bu çalışmada incelenen tanı yöntemlerinden $DFFITS_{(i)}$, $DFBETAS_{(i)}$ ve $COVRATIO_{(i)}$ istatistiklerine ilişkin kritik değerlerin örnek hacmine bağlı olarak azalan bir fonksiyon olduğu önceki bölümde belirtilmişti. Örnek hacmi büyük olan bir veride, kaldıraç değeri ve mutlak değerce artışı büyük olan bir gözlem regresyon sonuçlarını değiştirme eğiliminde olmasa bile bu istatistiklere ilişkin karar kurallarında yer alan kritik değerlerin küçük olmasından dolayı bu istatistiklere göre etkili gözlem olarak saptanabilir. Bundan dolayı $DFFITS_{(i)}$, $DFBETAS_{(i)}$ ve $COVRATIO_{(i)}$ istatistikleriyle verideki gözlemlerin etkili olup olmadığı incelenirken aynı zamanda bu gözlemlerin kaldıraç değerlerinin ve studentleştirilmiş artık değerlerinin de incelenmesi etkili gözlemin daha iyi saptanması açısından kolaylık sağlar.

Sonuç olarak bir etkili gözlemin saptanmak için kullanılan yöntemlerden verinin geneline uymayan bu gözlemi saptama bakımından en duyarlı tanı yöntemleri $DFFITS_{(i)}$ ve Pena' nın $S_{(i)}$ istatistikleridir. Bu iki yöntemden sonra $DFBETAS_{(i)}$ ve $COVRATIO_{(i)}$ yöntemlerinin de etkili gözlemi saptama bakımından duyarlılığı yüksektir. Cook Uzaklığı istatistiği ise etkili gözlem gruplarını saptamak için kullanılan yöntemler içinde bu gözlem gruplarını saptama bakımından en iyi yöntemdir. Son olarak yüksek boyutlu bir regresyon probleminin iki boyutlu bir probleme indirgenmesi amacıyla veriye bir dönüşüm uygulandığı grafik yöntemi ise regresyon probleminin boyutu arttıkça etkili gözlem ve gözlem gruplarını saptama bakımından duyarlılığı azalmaktadır.

KAYNAKLAR

1. Montgomery, D.C., Peck, E.A., Vining, G.G., "Introduction to Linear Regression Analysis", *Willey Series in Probability and Mathematical Statistics*, New York, 207-219 (2001).
2. Cook, R.D., "Detection of Influential Observations in Linear Regression", *Technometrics*, 19 (1): 15-18 (1977a).
3. Cook, R.D., "Influential Observations in Linear Regression", *Journal of The American Statistical Association*, 74 (365): 169-174 (1979).
4. Chatterjee, S., Hadi, A.S., "Influential Observations, High Leverage Points and Outliers in Linear Regression", *Statistical Science*, 1(3): 379-416 (1986).
5. Margolin, B.H., "The Distribution of Internally Studentized Statistics via Laplace Transform Inversion", *Biometrika*, 64: 573-582 (1977).
6. David, H.A., "Order Statistics, 2nd Edn.", *Willey*, New York, 110-150 (1981).
7. Hoaglin, D.C., Welsch, R.E., "The Hat Matrix in Regression and ANOVA", *The American Statistician*, 32 (1): 17-22 (1978).
8. Andrews, D.F., Pregibon, D., "Finding Outliers That Matter", *J. Roy. Statist. Soc., Ser. B.* 40: 85-93 (1976).
9. Belsley, D.A., Kuh, E., Welsch, R.E., "Regression Diagnostics: Identifying Influential Data and Sources of Collinearity", *Willey Series in Probability and Mathematical Statistics*, New York, 6-84 (1980).
10. Cook, R.D., Weisberg, S., "Residuals and Influence in Regression", *Chapman and Hall*, New York, 10-20, 101-156 (1982)
11. Velleman, P.F., Welsch, R.E., "Efficient Computing of Regression Diagnostics", *The American Statistician*, 35(4): 234-242 (1981).
12. Welsch, R.E., "Influence Functions and Regression Diagnostics. In Modern Data Analysis", *Academic*, New York (1982).
13. Hadi A.S., "A New Measure of Overall Potential Influence in Linear Regression", *Computational Statistics and Data Analysis*, 14: 1-27 (1992).
14. Chatterjee S., Hadi A.S., "Regression Analysis by Example", *Willey Series in Probability and Mathematical Statistics*, New Jersey, 85-115 (2006).

15. Pena, D., "A New Statistics for Influence in Linear Regression", *American Statistical Association and the American Society for Quality*, 47(1): 1-12 (2005)
16. Li, B., Martin, E.B., Morris, A.J., "A Graphical Technique for Detecting Influential Cases in Regression Analysis", *Communications in Statistics – Theory and Methods*, 30(3): 463-483 (2001).
17. Lawrance, A.J., "Deletion Influence and Masking in Regression", *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1): 181-189 (1995).
18. Hampel F.R., "The Influence Curve and Its Role in Robust Estimation, *Journal of The American Statistical Association*, 69(346): 383-393 (1974).
19. David, H.A., Nagaraja, H.N., "Order Statistics", *Wiley Series in Probability and Statistics*, New Jersey, 121-148 (2003).
20. Rousseeuw, P.J., Leroy, A.M., "Robust Regression and Outlier Detection", Wiley, New York 216-244 (2003).
21. Gray, J.B., Ling, R.F., "K-Clustering as a Detection Tool for Influential Subsets in Regression", *Technometrics*, 26(4): 305-318 (1984).
22. Atkinson, A.C., "Masking Unmasked", *Biometrika*, 73(3): 533-541 (1986).
23. Hadi, A.S., Simonoff, J.S., "Procedures for the Identification of Multiple Outliers in Linear Models", *Journal of the American Statistical Association*, 88(424): 1264-1272 (1993).

EKLER

EK-1. Grafik yönteminde $w_l^{(j)}$ vektörünün hesaplanması için algoritma

$(l=1, 2; j=1, 2, \dots, p)$

Adım 1: X matrisinin faktöriyel QR ayrıştırması, $n \times n$ boyutlu $Q = [Q_1, Q_2]$ matrisi için $X = QR$ biçiminde elde edilir. Burada tüm elemanları sıfır olan matris O olmak üzere $R = \begin{bmatrix} R_1^T, O_{(n-p) \times p}^T \end{bmatrix}^T$ matrisi $n \times n$ boyutlu bir üst üçgen matrisi, R_1 tekil olmayan $p \times p$ boyutlu bir üst üçgen matrisi ve Q_1 matrisi $n \times p$ boyutlu bir dik matristir.

Adım 2: $p \times p$ boyutlu R_1 matrisinin tekil değer ayrıştırması

$$R_1^T = P_1 \text{diag} \{a_1^{-1/2}, \dots, a_p^{-1/2}\} P_2^T$$

ile hesaplanır. Burada $p \times p$ boyutlu P_1 ve P_2 matrisleri dik matrislerdir.

Adım 3: $\phi = Q \text{diag} \{O_{p \times p}, I_{(n-p) \times (n-p)}\} Q^T Y$ ve $\phi_0 = \phi / (\phi^T \phi)^{1/2}$ hesaplanır.

Adım 4: Keyfi olarak belirlenen $s_j \times 1$ boyutlu r_j vektörü için, $u_j = Q_1 \begin{bmatrix} G_j^T, O^R \end{bmatrix}^T r_j$ hesaplanır. Burada G_j vektörü P_1 matrisinin j . sütunudur.

Adım 5: $w_l^{(j)} = (\phi_0 + u_j) / 2^{1/2}$ hesaplanır.

ÖZGEÇMİŞ

Kişisel Bilgiler

Soyadı, adı : ACARLAR, Irmak
Uyruğu : T.C.
Doğum tarihi ve yeri : 18.04.1984 Ankara
Medeni hali : Bekar
Telefon : 0 (312) 476 19 23
e-mail : irmakacarlar@gazi.edu.tr

Eğitim

Derece	Eğitim Birimi	Mezuniyet tarihi
Yüksek Lisans	Gazi Üniversitesi/ İstatistik A.B.D.	2009
Lisans	Gazi Üniversitesi/ İstatistik Bölümü	2006
Lise	Sokullu Mehmet Paşa Lisesi	2002

İş Deneyimi

Yıl	Yer	Görev
2007-	Gazi Üniversitesi	Araştırma Görevlisi

Yabancı Dil

İngilizce

Hobiler

Bilgisayar teknolojileri, Futbol, Sinema