

**ÇOKLU REGRESYONDA ETKİLİ GÖZLEMLERİN BELİRLENMESİNDE
İKİ YENİ YAKLAŞIM**

Irmak ACARLAR

**DOKTORA TEZİ
İSTATİSTİK ANABİLİM DALI**

**GAZİ ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ**

MAYIS 2015

İrmak ACARLAR tarafından hazırlanan “ÇOKLU REGRESYONDA ETKİLİ GÖZLEMLERİN BELİRLENMESİNDE İKİ YENİ YAKLAŞIM” adlı tez çalışması aşağıdaki jüri tarafından OY BİRLİĞİ ile Gazi Üniversitesi İstatistik Anabilim Dalında DOKTORA TEZİ olarak kabul edilmiştir.

Danışman: Prof. Dr. Hamza GAMGAM
İstatistik Anabilim Dalı, Gazi Üniversitesi

Bu tezin, kapsam ve kalite olarak Doktora Tezi olduğunu onaylıyorum

.....

Başkan : Prof. Dr. Meral ÇETİN
İstatistik Anabilim Dalı, Hacettepe Üniversitesi

Bu tezin, kapsam ve kalite olarak Doktora Tezi olduğunu onaylıyorum

.....

Üye : Prof. Dr. Mehmet Akif BAKIR
İstatistik Anabilim Dalı, Gazi Üniversitesi

Bu tezin, kapsam ve kalite olarak Doktora Tezi olduğunu onaylıyorum

.....

Üye : Doç. Dr. Cemal ATAKAN
İstatistik Anabilim Dalı, Ankara Üniversitesi

Bu tezin, kapsam ve kalite olarak Doktora Tezi olduğunu onaylıyorum

.....

Üye : Doç. Dr. Fikri GÖKPINAR
İstatistik Anabilim Dalı, Gazi Üniversitesi

Bu tezin, kapsam ve kalite olarak Doktora Tezi olduğunu onaylıyorum

.....

Tez Savunma Tarihi: 14/05/2015

Jüri tarafından kabul edilen bu tezin Doktora Tezi olması için gerekli şartları yerine getirdiğini onaylıyorum.

.....
Prof. Dr. Şeref SAĞIROĞLU
Fen Bilimleri Enstitüsü Müdürü

ETİK BEYAN

Gazi Üniversitesi Fen Bilimleri Enstitüsü Tez Yazım Kurallarına uygun olarak hazırladığım bu tez çalışmada;

- Tez içinde sunduğum verileri, bilgileri ve dokümanları akademik ve etik kurallar çerçevesinde elde ettiğimi,
- Tüm bilgi, belge, değerlendirme ve sonuçları bilimsel etik ve ahlak kurallarına uygun olarak sunduğumu,
- Tez çalışmada yararlandığım eserlerin tümüne uygun atıfta bulunarak kaynak gösterdiğimi,
- Kullanılan verilerde herhangi bir değişiklik yapmadığımı,
- Bu tezde sunduğum çalışmanın özgün olduğunu,

bildirir, aksi bir durumda aleyhime doğabilecek tüm hak kayıplarını kabullendiğimi beyan ederim.

Irmak ACARLAR

14/05/2015

ÇOKLU REGRESYONDA ETKİLİ GÖZLEMLERİN BELİRLENMESİNDE İKİ YENİ
YAKLAŞIM
(Doktora Tezi)

İrmak ACARLAR

GAZİ ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ
Mayıs 2015

ÖZET

Çoklu Doğrusal Regresyonda veri kümesinde yer alan etkili gözlem(ler), model parametrelerinin En Küçük Kareler tahmin değerlerinde önemli derecede farklılaşmalara neden olur. Ayrıca bu gözlemler, tahmin edicilerin varyanslarının tahmin değerinin büyümesine de neden olabilirler. Buna dayalı olarak istatistiksel çıkarsamaların sonuçları anlamsız olabilir. Bu dikkate değer farklılaşmalar modelin açıklanabilirliğini azalttığı için veri kümesindeki etkili gözlem veya etkili gözlem gruplarının saptanması regresyon analizinin verimliliği açısından önemlidir. Bu çalışmanın birinci bölümünde çoklu doğrusal regresyon analizinde etkili gözlemleri saptamanın önemi ve literatürde bu konuyla ilgili yapılan çalışmalardan bahsedilmiştir. Çalışmanın ikinci bölümünde, çoklu doğrusal regresyon modeli, kaldırma değeri, artıkların türleri, etki eğrisi ve etki eğrisinin örnek uyarlamaları gibi temel kavramlar verilmiştir. Sonra üçüncü bölümde etkili gözlemlerin saptanması için literatürde yer alan bazı tanı yöntemleri tanıtılmıştır. Dördüncü ve beşinci bölümlerde sırasıyla etkili gözlemlerin saptanması için örnek etki eğrisine dayalı bir grafiksel yöntem ve Biplot yöntemine dayalı bir grafiksel yöntem önerisi verilmiştir. Klasik tanı yöntemleri ile önerilen iki grafik yöntemine ilişkin bir uygulama ise altıncı bölümde yer almaktadır. Yedinci bölümde klasik tanı yöntemleri ve önerilen iki grafik yöntemini karşılaştırmak için simülasyon çalışması verilmiştir. Çalışmanın son bölümünde ise sonuç ve öneriler sunulmuştur.

Bilim Kodu : 205.1.066
Anahtar Kelimeler : Aykırı Gözlemler, Biplot, Çoklu Doğrusal Regresyon Analizi, Etkili Gözlemler, Kaldırma Değeri, Tekil Değer Ayırıştırması
Sayfa Adedi : 139
Danışman : Prof. Dr. Hamza GAMGAM

TWO NEW APPROACHES FOR DETECTION OF INFLUENTIAL OBSERVATIONS IN MULTIPLE REGRESSION

(Ph. D. Thesis)

Irmak ACARLAR

GAZİ UNIVERSITY

GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

May 2015

ABSTRACT

Influential observations in data cause some significant differentiation on Least Square Estimations of model parameters in multiple linear regression. These observations would also cause increasing in the variances of estimations. Therefore, results of statistical inferences would be insignificant. Since these notable differentiations decrease the explicable of the model, the detection of the influential observation or the influential groups in data is important for the efficiency of regression analysis. The importance of detecting influential observations and some studies about these topics in literature are referred in the first section of this study. Some basic concepts as multiple linear regression model, leverage value, some types of residuals, influence curve and its sample versions are given in the second section of the study. Then some diagnostic methods in order to detect influential observations in literature are demonstrated in the third section. In the fourth and fifth sections a graphical method based on sample influence curve and a graphical method based on Biplot method are proposed to detect influential observations respectively. Two applications of classical methods and the two graphical methods proposed in this study are included in the sixth section. In the seventh section a simulation study is given in order to compare the classical diagnostics and the two graphical methods in this study. At the end of this study, conclusion and recommendations are presented.

Science Code : 205.1.066
Key Words : Outliers, Biplot, Multiple Linear Regression Analysis, Influential Observations, Leverage Value, Singular Value Decomposition
Page Number : 139
Supervisor : Prof. Dr. Hamza GAMGAM

TEŞEKKÜR

Çalışmalarım boyunca değerli yardım ve katkılarıyla beni yönlendiren hocam Prof. Dr. Hamza GAMGAM'a; tez izleme toplantılarında bu çalışmayla ilgili değerli görüş ve önerilerini paylaşan hocam Prof. Dr. Mehmet Akif BAKIR ve hocam Doç. Dr. Cemal ATAKAN'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.....	x
ŞEKİLLERİN LİSTESİ.....	xii
SİMGELER VE KISALTMALAR.....	xiv
1. GİRİŞ.....	1
2. TEMEL KAVRAMLAR	7
2.1. Çoklu Doğrusal Regresyon Modeli.....	7
2.2. Kaldıraç Değeri (Leverage Value)	8
2.3. Artıklar ve Artıkların Standart Biçimleri	10
2.3.1. Sıradan artıklar	10
2.3.2. Student türü artıklar (Studentized residuals)	11
2.3.3. Sıkıştırma artıkları (PRESS residuals)	13
2.3.4. Tekrarlanan artıklar (Recursive residuals)	15
2.4. Etki Eğrisi ve Sonlu Örnek Uyarlamaları	16
2.4.1. Etki eğrisi (Influence curve).....	16
2.4.2. Etki eğrisinin sonlu örnek uyarlamaları	17
2.5. Maskeleye ve Yanılgıya Düşürme	19
2.5.1. Maskeleye sorunu	19
2.5.2. Yanılgıya düşürme sorunu	20

3. ETKİLİ GÖZLEMLERİN SAPTANMASI İÇİN KULLANILAN TANI YÖNTEMLERİ	23
3.1. Etkili Gözlemlerin Saptanması için Tanı İstatistikleri	25
3.1.1. Cook uzaklığı istatistiği	25
3.1.2. DFBETAS istatistiği	29
3.1.3. DFFITS istatistiği.....	31
3.1.4. Kovaryans oranı (COVRATIO).....	32
3.1.5. Andrews-Pregibon AP istatistiği.....	34
3.1.6. Cook ve Weisberg olabilirlik istatistiği LD	35
3.1.7. Hadi'nin K istatistiği	37
3.1.8. Pena'nın S istatistiği.....	41
3.2. Etkili Gözlemlerin Saptanması için Grafiksel Yöntemler	44
3.2.1. Eklenmiş değişken grafikleri (Added variable plots).....	45
3.2.2. Kısmi artık grafikleri (Partial residual plots)	48
4. ÖRNEK ETKİ EĞRİSİNE DAYALI BİR GRAFİKSEL YÖNTEM ÖNERİSİ	51
5. BİPLOT YAKLAŞIMINA DAYALI BİR GRAFİKSEL YÖNTEM ÖNERİSİ	57
5.1. Biplot Tekniğine İlişkin Temel Kavramlar	59
5.2. Biplot Yardımıyla Etkili Gözlemlerin Saptanması için Bir Grafiksel Yöntem Önerisi	63
6. UYGULAMA.....	69
6.1. Fare Verisi (Rat Data) Uygulaması.....	69
6.2. New York'taki Nehirler Verisi (New York Rivers Data) Uygulaması.....	77
7. SİMÜLASYON ÇALIŞMASI.....	87

Sayfa

7.1. Bir Gözlem Silme Yaklaşımına Dayalı Tanı İstatistikleri ile Önerilen İki Grafik Yönteminin Karşılaştırılması.....	89
7.2. Birden Fazla Gözlem Silme Yaklaşımına Dayalı Tanı İstatistikleri ile Önerilen İki Grafik Yönteminin Karşılaştırılması	106
7.3. Parametre Sayısı ve Örnek Hacmi Sabitken Önerilen İki Yöntemin Klasik Tanı Yöntemleriyle Kaldıraç Değerindeki Değişim Bakımından Karşılaştırılması	118
7.4. Parametre Sayısı ve Örnek Hacmi Sabitken Önerilen İki Yöntemin Klasik Tanı Yöntemleriyle Hata Değişkeni Değerindeki Değişim Bakımından Karşılaştırılması	121
8. SONUÇ VE ÖNERİLER	125
KAYNAKLAR	131
EKLER.....	135
EK-1. Fare Verisi	136
EK-2. New York'taki Nehirler Verisi.....	137
ÖZ GEÇMİŞ	139

ÇİZELGELERİN LİSTESİ

Çizelge	Sayfa
Çizelge 6.1. Fare verisi için tanı istatistiklerinin değerleri ve kritik değerler	72
Çizelge 6.2. Fare verisi için biplotta noktalara ilişkin Mahalanobis Uzaklıklarının karesi ve kritik değer	76
Çizelge 6.3. New York'taki Nehirler Verisi için tanı istatistiklerinin değerleri ve kritik değerler	80
Çizelge 6.4. New York'taki Nehirler Verisinde bulunan 4, 5, 7 ve 19 numaralı gözlemlerin farklı kombinasyonları için bir gözlem ve birden fazla gözlem silme yaklaşımına dayalı Cook Uzaklığı, COVRATIO ve K istatistiklerinin değerleri ve bunlara karşılık gelen kritik değerler	82
Çizelge 6.5. New York'taki Nehirler Verisi için biplotta noktalara ilişkin Mahalanobis Uzaklıklarının karesi ve kritik değer	86
Çizelge 7.1. Bir gözlem silmeye dayalı tanı istatistikleri ve kritik değerleri	88
Çizelge 7.2. Bir gözlem silme yaklaşımına dayalı klasik tanı istatistikleri ile BYDGY ve SICDGY için etkili gözlemi saptama oranları	91
Çizelge 7.3. Bir etkili gözlemin bulunduğu veri kümesinde etkili gözlemi saptama oranları bakımından iki yönlü varyans analiziyle elde edilen p-değerleri	94
Çizelge 7.4. Bir gözlem silme yaklaşımına dayalı tanı istatistikleri ile BYDGY ve SICDGY için etkili olmayan gözlemlerden en az birini etkili olarak saptama oranları	99
Çizelge 7.5. Bir etkili gözlemin bulunduğu veri kümesinde etkili olmayan gözlemlerden en az birini saptama oranları bakımından iki yönlü varyans analiziyle elde edilen p-değerleri	102
Çizelge 7.6. Birden fazla gözlem silme yaklaşımına dayalı tanı istatistikleri ile BYDGY ve SICDGY için etkili gözlem grubunu saptama oranları	108
Çizelge 7.7. İki etkili gözlemin bulunduğu veri kümesinde etkili gözlem grubunu saptama oranları bakımından iki yönlü varyans analiziyle elde edilen p-değerleri	110
Çizelge 7.8. Birden fazla gözlem silme yaklaşımına dayalı tanı istatistikleri ile BYDGY ve SICDGY için etkili olmayan gözlem çiftlerinin en az birini etkili olarak saptama oranları	113

Çizelge	Sayfa
Çizelge 7.9. İki etkili gözlemin bulunduğu veri kümesinde verinin geneline uyan gözlemlerin ikili gruplarından en az birini etkili gözlem grubu olarak saptama oranları bakımından iki yönlü varyans analiziyle elde edilen p-değerleri	114
Çizelge 7.10. $n = 20$ ve $p = 3$ iken önerilen iki yöntemle bir gözlem silme yaklaşımına dayalı tanı yöntemlerinin, farklı bağımsız değişken değerlerine göre üretilen etkili gözlemi saptama oranları.....	119
Çizelge 7.11. $n = 20$ ve $p = 3$ iken önerilen iki yöntemle birden fazla gözlem silme yaklaşımına dayalı tanı yöntemlerinin, farklı bağımsız değişken değerlerine göre üretilen iki etkili gözlemin oluşturduğu etkili gözlem grubunu saptama oranları.....	120
Çizelge 7.12. $n = 20$ ve $p = 3$ iken önerilen iki yöntemle bir gözlem silme yaklaşımına dayalı tanı yöntemlerinin, farklı hata değişken değerlerine göre üretilen etkili gözlemi saptama oranları.....	122
Çizelge 7.13. $n = 20$ ve $p = 3$ iken önerilen iki yöntemle birden fazla gözlem silme yaklaşımına dayalı tanı yöntemlerinin, farklı hata değişken değerlerine göre üretilen iki etkili gözlemin oluşturduğu etkili gözlem grubunu saptama oranları	123

ŞEKİLLERİN LİSTESİ

Şekil	Sayfa
Şekil 3.1. Basit Doğrusal Regresyon Modeli için $\hat{\beta}_{0(i)}$ ve $\hat{\beta}_{1(i)}$ tahminlerine ait grafik.....	28
Şekil 5.1. İki boyutlu serpm diyagramı	58
Şekil 5.2. Biplot gösterimi	58
Şekil 5.3. Modelde iki bağımsız değişken varken $\hat{\beta}_{1(i)}$ ve $\hat{\beta}_{2(i)}$ tahminleri için bir serpm diyagramı örneği	64
Şekil 6.1. Fare verisi için harici student tipi artıklar (t_i ; $i = 1,2,...,19$) ile gözlem numarası için oluşturulan serpm diyagramı ve sınırlar	70
Şekil 6.2. Fare verisi için kaldırıcı değerleri (h_{ii} ; $i = 1,2,...,19$) ile gözlem numarası için oluşturulan serpm diyagramı ve sınır	70
Şekil 6.3. Fare verisi için SICDGY ile elde edilen w değişkeni ile gözlem numarası için oluşturulan serpm diyagramı ve sınırlar	74
Şekil 6.4. Fare verisi için $\hat{\beta}_{j(i)}$ tahminlerine dayalı biplot.....	75
Şekil 6.5. Fare verisi için $\hat{\beta}_{j(i)} - \hat{\beta}_j$ tahminlerine dayalı biplot	76
Şekil 6.6. New York'taki Nehirler Verisi için harici student tipi artıklar (t_i ; $i = 1,2,...,20$) ile gözlem numarası için oluşturulan serpm diyagramı ve sınırlar	78
Şekil 6.7. New York'taki Nehirler Verisi için kaldırıcı değerleri (h_{ii} ; $i = 1,2,...,20$) ile gözlem numarası için oluşturulan serpm diyagramı ve sınır	78
Şekil 6.8. New York'taki Nehirler Verisi için SICDGY ile elde edilen w_i ($i = 1,2,...,20$) ile gözlem numarası için oluşturulan serpm diyagramı ve sınırlar	84
Şekil 6.9. New York'taki Nehirler Verisi için $\hat{\beta}_{j(i)} - \hat{\beta}_j$ tahminlerine dayalı biplot	85
Şekil 7.1. Bir etkili gözlemin bulunduğu veri kümesinde etkili gözlemi saptama oranları bakımından iki yönlü varyans analizinden sonra oluşturulan ana etki grafikleri	96

Şekil	Sayfa
Şekil 7.2. Bir etkili gözlemin bulunduğu veri kümesinde etkili gözlemi saptama oranları bakımından iki yönlü varyans analizinden sonra oluşturulan etkileşim etkisi grafikleri	98
Şekil 7.3. Bir etkili gözlemin bulunduğu veri kümesinde verinin geneline uyan gözlemlerden en az birini etkili olarak saptama oranları bakımından iki yönlü varyans analizinden sonra oluşturulan ana etki grafikleri	103
Şekil 7.4. Bir etkili gözlemin bulunduğu veri kümesinde verinin geneline uyan gözlemlerden en az birini etkili olarak saptama oranları bakımından iki yönlü varyans analizinden sonra oluşturulan etkileşim etkisi grafikleri	105
Şekil 7.5. İki etkili gözlemin bulunduğu veri kümesinde etkili gözlem grubunu saptama oranları bakımından iki yönlü varyans analizinden sonra oluşturulan ana etki grafikleri	111
Şekil 7.6. İki etkili gözlemin bulunduğu veri kümesinde etkili gözlem grubunu saptama oranları bakımından iki yönlü varyans analizinden sonra oluşturulan etkileşim etkisi grafikleri	112
Şekil 7.7. İki etkili gözlemin bulunduğu veri kümesinde verinin geneline uyan gözlemlerin ikili gruplarından en az birini etkili gözlem grubu olarak saptama oranları bakımından iki yönlü varyans analizinden sonra oluşturulan ana etki grafikleri	115
Şekil 7.8. İki etkili gözlemin bulunduğu veri kümesinde verinin geneline uyan gözlemlerin ikili gruplarından en az birini etkili gözlem grubu olarak saptama oranları bakımından iki yönlü varyans analizinden sonra oluşturulan etkileşim etkisi grafikleri	116

SİMGELER VE KISALTMALAR

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

Simgeler	Açıklamalar
n	Örnek hacmi
p	Model parametrelerinin sayısı
k	Örnek hacminin en az iki elemanlı bir alt kümesinin eleman sayısı
Y	Bağımlı değişken değerlerinin vektörü
$Y_{(i)}$	i . gözlem veriden çıkartıldıktan sonra geri kalan gözlemler için Y vektörü
$Y_{(I)}$	I kümesinin veriden çıkartılmasıyla geri kalan gözlemler için Y vektörü
y_i	i . gözlemin bağımlı değişken değeri
X	Tasarım matrisi
x_i	Tasarım matrisinin i . satırı
$X_{(i)}$	i . gözlem veriden çıkartıldıktan sonra geri kalan gözlemler için X matrisi
$X_{(I)}$	I kümesinin veriden çıkartılmasıyla geri kalan gözlemler için X matrisi
Z	X matrisi ve Y vektörünü içeren matris
z_i	Z matrisinin i . satırı
β	Model parametrelerinin vektörü
ε	Hata terimlerinin vektörü
I	Örneğin k hacimli bir alt kümesi
H	Şapka matrisi
$h_{ii'}$	Şapka matrisinin i . satır, i' . sütun elemanı
$\hat{\beta}$	En Küçük Kareler tahmin vektörü
$\hat{\beta}_{(i)}$	Bir gözlem silme yaklaşımına dayalı EKK tahmin vektörü
$\hat{\beta}_{(I)}$	Birden fazla gözlem silme yaklaşımına dayalı EKK tahmin vektörü
$\hat{Y}$	Bağımlı değişken değerlerinin tahminlerinin vektörü
$\hat{y}_{(i)}$	Bir gözlem silme yaklaşımına dayalı olarak elde edilen tahmin değeri
e	Artıkların vektörü

Simgeler	Açıklamalar
e_i	i . gözleme ilişkin artık değeri
r_i	i . gözleme ilişkin dahili student türü artık değeri
t_i	i . gözleme ilişkin harici student türü artık değeri
e_I	I kümesindeki gözlemlere ilişkin artık değerlerinin vektörü
r_I	I kümesindeki gözlemlere ilişkin r_i değerleri vektörü
t_I	I kümesindeki gözlemlere ilişkin t_i değerleri vektörü
σ^2	Varyans
$\hat{\sigma}^2$	Artık kareler ortalaması (Varyansın tahmini)
$\hat{\sigma}_{(i)}^2$	i . gözleme ilişkin artığın dahil edilmediği artık kareler ortalaması
$\hat{\sigma}_{(I)}^2$	I daki gözlemlerin artıklarının dahil edilmediği artık kareler ortalaması
$e_{S,i}$	i . gözlem için sıkıştırma artığı
$e_{R,i}$	i . gözlem için tekrarlanan artık değeri
e_K	Kısmi artık vektörü
$D_{(i)}$	Bir gözlem silme yaklaşımına dayalı Cook Uzaklığı İstatistiği
$D_{(I)}$	Birden fazla gözlem silme yaklaşımına dayalı Cook Uzaklığı İstatistiği
$DFBETAS_{j(i)}$	i . gözlem veriden silindiğinde j . parametre için DFBETAS İstatistiği
$DFFITS_{(i)}$	i . gözlem için DFFITS İstatistiği
$COVRATIO_{(i)}$	i . gözlem için COVRATIO İstatistiği
$COVRATIO_{(I)}$	I gözlem kümesi için COVRATIO İstatistiği
$AP_{(i)}$	i . gözlem için AP İstatistiği
$LD_{(i)}$	i . gözlem için LD İstatistiği
$K_{(i)}$	i . gözlem için K İstatistiği
$K_{(I)}$	I gözlem kümesi için K İstatistiği
w_i	i . gözlemin, EKK tahminlerini değiştirme eğilimine ilişkin değişken
w	w_i değişkenlerinin vektörü
$\tilde{\sigma}_w$	w_i değişkeninin dağılış parametresinin sağlam tahmin edicisi
Λ	Hedef matris
L	Sol matris
L_i	Sol matrisin i . satır vektörü
R	Sağ matris

Simgeler	Açıklamalar
R_j	Sağ matrisin j . sütun vektörü
Ψ	$\hat{\beta}_{(i)}^T$ vektörlerinin oluşturduğu matris
Δ	$(\hat{\beta}_{(i)} - \hat{\beta})^T$ vektörlerinin oluşturduğu matris
ξ_i^2	i . gözlem için BYDGY'ne dayalı Mahalanobis Uzaklığı

Kısaltmalar	Açıklamalar
EKK	En Küçük Kareler
EÇOB	En Çok Olabilirlik
IC	Etki Eğrisi (Influence Curve)
SIC	Örnek Etki Eğrisi (Sample Influence Curve)
EIC	Deneysel Etki Eğrisi (Empirical Influence Curve)
MAD	Medyandan Mutlak Sapmalar (Median Absolute Deviation)
TDA	Tekil Değer Ayırıştırması
SICDGY	SIC'ye Dayalı Grafik Yöntemi
BYDGY	Biplot Yaklaşımına Dayalı Grafik Yöntemi

1. GİRİŞ

Örnek ortalaması ve örnek varyansı gibi birçok istatistiğin değeri hesaplanırken, örneği oluşturan gözlemlerin her biri aynı ağırlığa sahiptir. Bununla birlikte veri kümesinin geneline uymayan gözlemler örnekten elde edilen bu istatistiklerin değerlerini olumsuz şekilde etkiler. Regresyon Analizinde de gözlemlerin geneline uymayan bir veya daha fazla gözlemi içeren veri kümeleriyle karşılaşılabilir. Veri kümesinde gözlemlerin geneline uymayan, artık değeri diğer gözlemlerin artık değerlerine göre mutlak değerce anormal şekilde büyük olan bir gözlem, aykırı gözlem olarak tanımlanır. Gözlemlerin geneline uymayıp, regresyon parametrelerinin en küçük kareler (EKK) tahminlerinde önemli derecede farklılaşmalara neden olan bir gözlem ise etkili gözlem olarak tanımlanır. Etkili bir gözlemin artışı genellikle mutlak değerce büyük olur. Fakat bazı veri kümeleri için etkili gözlemin artığının mutlak değerce küçük olduğu da gözlenebilir. Eğer etkili gözlemin artışı mutlak değerce büyük ise bu gözlem aynı zamanda aykırı bir gözlemdir.

Bir veya daha fazla etkili gözlem içeren bir veri kümesi için verinin geneline uymayan bu gözlemlere ilişkin artıklar, genellikle mutlak değerce büyük olduklarından, artık kareler ortalaması olarak da bilinen hata değişkeninin varyansının yansız tahmini bu gözlemlerden etkilenecek büyük değerler alır. Ayrıca veri kümesinde bulunan etkili gözlemler hata terimlerine ilişkin normallik varsayımının sağlanmamasına da neden olabilir. Bu durum istatistiksel çıkarımların yapılması bakımından sorun yaratır. Her ne kadar büyük hacimli örnekler için hata değişkenlerine ilişkin normallik varsayımı sağlansa da, varyansın tahmininin büyük olması, regresyon modelindeki parametrelerin EKK tahminlerine ilişkin standart hataların tahminlerinin büyük olmasına ve buna bağlı olarak regresyon analizine ilişkin çıkarımların istatistiksel olarak anlamsız olarak sonuçlanmasına neden olabilir.

Etkili gözlem tanımı ve bu gözlemin saptanmasının önemi ilk kez Cook (1977) tarafından ortaya atılmasıyla günümüze kadar literatürde bu konu ile ilgili çok sayıda çalışma yapılmıştır. Etkili gözlemlerin saptanması için önerilen yöntemler tanı istatistikleri ve grafiksel yöntemler olarak iki başlık altında toplanabilir. Genellikle bu yöntemlerin çoğu gözlem silme yaklaşımına dayalıdır. Gözlem silme yaklaşımında veri kümesinden bir gözlem çıkarıldığında parametre tahminlerinin nasıl değişeceği araştırılır. Bundan dolayı bu yaklaşım veri kümesindeki her bir gözlemin regresyon sonuçlarını değiştirme

eğilimlerini incelemek amacıyla tanı yöntemleri için bir araç olmuştur. Gözlem silme yaklaşımı veri kümesinden gözlemlerin birer birer silinmesiyle regresyon tahminlerinin incelenmesi olarak düşünülebileceği gibi, silinecek eleman sayısı toplam gözlem sayısının yarısından küçük olmak üzere, iki veya daha fazla gözlemi içeren gözlem gruplarının silinmesiyle regresyon tahminlerinin incelenmesi şeklinde de düşünülebilir. Bu durumda veri kümesinin, eleman sayısı birden farklı alt kümelerinin regresyon sonuçları üzerine etkisi incelenir. Bununla birlikte bu alt kümelerdeki eleman sayısı arttıkça, veri kümesinin büyüklüğüne dayalı olarak regresyon sonuçlarına etkisi incelenecek gözlem gruplarının sayısı da artmaktadır. Bu durumda gözlemlerin toplam sayısı n ve regresyon sonuçlarına etkisi incelenecek gözlem grubundaki eleman sayısı k , ($1 < k < n/2$), olmak üzere $\binom{n}{k}$

sayıda mümkün gözlem grubu söz konusu olmaktadır. Bundan dolayı k 'nın küçük olduğu durumlar için gözlem gruplarının regresyon sonuçları üzerine etkisinin incelenmesi uygulamada tercih edilmektedir. Bir gözlem silme yaklaşımına dayalı tanı istatistikleri regresyon tahminleri üzerinde bireysel (tek başına) etkisi olan gözlemleri saptamak için kullanılmaktadır. k büyüklüğünde bir gözlem kümesi silme yaklaşımına dayalı tanı istatistikleri ise regresyon tahminleri üzerinde ortak etkisi olan gözlem gruplarını saptamak için kullanılır.

Literatürde etkili gözlemlerin saptanması için birçok tanı istatistiği önerilmiştir. Etkili gözlemlerin saptanması amacıyla bir tanı istatistiği Cook (1977) tarafından geliştirilen ve uygulamada yaygın olarak kullanılan karesel bir form olan Cook Uzaklığı İstatistiğidir. Gözlem silme yaklaşımına dayalı olan Cook Uzaklığı İstatistiği hem gözlemlerin regresyon tahminleri üzerine birer birer etkisinin incelenmesi amacıyla kullanılabilir, hem de eleman sayısı birden farklı iken veri kümesinin k hacimli alt kümelerinin regresyon tahminleri üzerine etkisinin incelenmesi amacıyla da kullanılabilir. Etkili gözlemlerin saptanması için kullanılan diğer bazı tanı istatistikleri Welsch ve Kuh (1977); Andrews ve Pregibon (1978); Belsley, Kuh ve Welsch (1980); Atkinson (1981); Cook ve Weisberg (1982); Welsch (1982); Hadi (1992) ve Pena (2005) tarafından çalışılmıştır. Bu tanı istatistiklerinin bazı dezavantajları aşağıdaki gibidir;

- k büyüklüğünde gözlem grupları tanı istatistikleriyle incelenirken k 'nın değeri $n/2$ ' ye yaklaştıkça, $\binom{n}{k}$ kadar olan mümkün gözlem grubu sayısı artar. Böylece çok sayıda gözlem grubunun incelenmesi durumu ortaya çıkar.
- Bazı durumlarda iki veya daha fazla gözlem regresyon tahminleri üzerine bireysel olarak etkili gözüküyorken, bu gözlemlerin tahminler üzerinde ortak etkisi mevcut olabilir. Buna bağlı olarak regresyon tahminleri üzerine ortak etkisi olan bu gözlemler, bir etkili gözlemi saptamak için kullanılan tanı istatistikleriyle tespit edilemez. Bu sorun maskeleye (masking) olarak bilinir(Lawrance, 1995).
- Bazı durumlarda etkili gözlem gruplarının saptanması için kullanılan tanı istatistikleri, etkili bir gözlem ile gerçekte etkili olmayan ve verinin geneline uyan bir gözlemi, etkili bir gözlem grubu olarak tespit edilmesine neden olabilir. Bu durumda etkili olmayan gözlem de etkili gözlem gibi yorumlanır. Bu sorun yanılgıya düşürme (swamping) olarak bilinir(Lawrance, 1995).

Tanı istatistikleri kullanılarak etkili gözlemlerin belirlenmesi amacıyla her bir tanı istatistiği için bir karar kuralı mevcuttur. Veri kümesinde bir gözlemin etkili gözlem olarak saptanması için bu gözlemin ilgili tanı istatistiği değerinin veya bunun mutlak değerinin karar kuralında yer alan kritik değerden (cutoff value) büyük olması koşulu söz konusudur. Nitekim bazı tanı istatistiklerinin karar kurallarında yer alan kritik değer örnek hacmine karşı duyarlıdır. Belsley ve diğerleri (1980) tarafından önerilen DFBETAS, DFFITS ve COVRATIO ele alınırsa bu istatistiklere karşılık gelen kritik değerler örnek hacminin bir fonksiyonu olduğundan ve örnek hacmi arttıkça ilgili kritik değerler küçüldüğünden, bu istatistiklerle büyük hacimli örneklerde etkili gözlemler kolay bir biçimde belirlenebilir. Bunun yanında bu kritik değerlerin küçülmesi verinin geneline uyan ve etkili olmayan gözlemlerden bazılarının da etkili olarak saptanmasına neden olabilir. Bu durum araştırmacıyı yanlış karar vermeye yönlendirir.

Hadi (1992) tarafından önerilen K istatistiği ele alınırsa, her bir gözlem için hesaplanan $K_{(i)}$ değerleri yeni bir veri kümesi gibi düşünüldüğünde, karar kuralında yer alan kritik değer bu değerlerin ortalaması ve varyansının tahminlerine dayalıdır. Veri kümesinde bir etkili gözlem varken etkili gözlemin $K_{(i)}$ değerinin, bu kritik değeri büyütme eğilimi

olmayabilir. Bununla birlikte veri kümesinde etkili gözlem sayısı arttıkça, kritik değer etkili gözlemlerin $K_{(i)}$ değerlerinden olumsuz bir biçimde etkilenebilir.

Pena (2005) tarafından önerilen $S_{(i)}$ tanı istatistiği için karar kuralında yer alan kritik değer dikkate alınır, her bir gözlemin $S_{(i)}$ tanı istatistiği değerleri yeni bir veri gibi düşünüldüğünde, ilgili kritik değer bu verinin ortalama ve varyansının sağlam tahminlerine dayalıdır. Dolayısıyla bu kritik değer, etkili gözlemin neden olduğu büyük $S_{(i)}$ değerlerine karşı duyarlılığı azdır. Bununla birlikte sağlam tahminlere dayalı bu kritik değer, örnek hacmi arttıkça küçülme eğiliminde olduğundan etkili olmayan gözlemlerin de etkili olarak saptanmasına neden olabilir. Bu durum araştırmacının yanlış karar almasına yol açar.

Basit Doğrusal Regresyonda modelin uygunluğu görsel olarak incelenmek istenirse, ilk başvurulacak yol bağımlı değişkenle bağımsız değişken arasında oluşturulan serpm diyagramının incelenmesidir. Bu serpm diyagramı model seçimi doğru yapılmışken verinin geneline uymayan gözlem veya gözlem kümelerinin de görsel saptanmasında yardımcı olur. Özellikle etkili gözlem(ler)in bulunduğu bir veri kümesi için doğrusal bir regresyon modeli uygun iken, tahmin edilen regresyon doğrusunun da yer aldığı, bağımlı değişkenle bağımsız değişken arasında oluşturulan serpm diyagramı incelendiğinde regresyon doğrusunun y eksenini kestiği noktanın veya regresyon doğrusunun eğiminin beklenenden farklı olduğu gözlenebilir. Benzer bir durum modelde bir bağımsız değişken varken eğrisel bir model için de geçerli olabilir. Bu durumda etkili gözlem(ler)in görsel olarak saptanması kolaydır. Bununla birlikte iki bağımsız değişkenin olduğu bir doğrusal regresyon problemi için bağımlı değişkenle bağımsız değişkenler arasında oluşturulan serpm diyagramı üç boyutludur ve bir regresyon düzlemi söz konusudur. Bu durumda bu serpm diyagramında bir hacim söz konusu olduğu için etkili gözlemlerin görsel olarak saptanması zorlaşır. Tahmin edilecek parametre sayısı p , bağımsız değişken sayısı $p - 1$ olmak üzere $(p - 1) > 2$ iken bağımlı değişkenle bağımsız değişkenler arasında oluşturulan p boyutlu serpm diyagramının çizimi mümkün değildir. Böyle bir durum için kullanılan bir grafiksel yöntem serpm diyagramı matrisidir. Serpm diyagramı matrisi bağımlı değişken ve bağımsız değişkenlerin ikili serpm diyagramlarının birlikte gösterilmesiyle oluşturulur. Ayrıca bu grafik değişkenler arasındaki sıfıncı dereceden korelasyonun incelenmesi için kullanılabilir. Ancak çoğu durumda serpm diyagramı matrisi bağımsız değişkenlerle bağımlı değişken arasında oluşturulan modele ilişkin kullanışlı olabilecek

ortak bir bilgiyi açıkça sağlayamadığından, etkili gözlemlerin saptanması açısından kullanışlı olmayabilir(Cook, 1998: 49-51).

Projeksiyon matrisi olarak da bilinen Hat matrisinin köşegen elemanları etkili gözlemlerin belirlenmesinde kullanılan araçlardan biridir. Hoaglin ve Welsch (1978) kaldırıcı değeri olarak da bilinen, Hat matrisinin köşegen elemanlarından $2p/n$ değerinden büyük olan gözlemlerin her birini yüksek dereceli kaldırıcı noktası olarak tanımlamıştır. Genellikle etkili gözlemin artığı mutlak değerce büyük olduğundan ve bu gözlem aynı zamanda yüksek dereceli kaldırıcı noktası da olduğundan, etkili gözlemin saptanması amacıyla artıklar ve kaldırıcı değerleri ortak incelenebilir. Bu amaçla uygulamada etkili gözlemlerin incelenmesi için bir ön çalışma, artıkların diyagramı ile kaldırıcı değerlerinin diyagramının ortak incelenmesi biçiminde yapılabilir. *i.* gözleme ilişkin artığın varyansı bu gözleme ilişkin kaldırıcı değerinin bir fonksiyonu olduğu için artıkların dağılımı dağılış parametresi bakımından bağımsız değildir. Bundan dolayı uygulamada genellikle artıklar yerine student türü artıklardan yararlanılmaktadır(Cook ve Weisberg, 1982: 118).

Etkili gözlemlerin saptanması için görsel bir başka yol da Cook Uzaklığı İstatistiği değerlerinin diyagramının incelenmesidir. Cook Uzaklığı için kritik değer $F_{0.50,p,n-p}$ olarak bilinir. Bu kritik değer 1'e yakın bir değer olduğu için uygulamada pratiklik sağlamak amacıyla Cook Uzaklığı değerleri 1'e göre değerlendirilir. Buradan hareketle Cook Uzaklığı İstatistiği değerlerinin diyagramında 1 değerini aşan noktalara karşılık gelen gözlemler etkili gözlem olarak saptanır(Cook ve Weisberg, 1982: 118).

Çoklu regresyonda serpm diyagramını oluşturma sorunu, gözlemlerin görsel olarak incelenmesine ve bu gözlemlerden verinin geneline uymayan gözlemlerin görsel olarak saptanmasına imkan vermemektedir. Bu sorunu gidermek amacıyla Li, Martin ve Morris (2001) çoklu regresyonda etkili gözlemlerin saptanması için bir grafiksel yöntem geliştirmişlerdir. Bu yöntemdeki ana fikir yüksek boyutlu bir çoklu regresyon problemini iki boyutlu p tane tanı grafiğine indirgeyerek, bu iki boyutlu grafiklerin görsel olarak incelenmesine dayanır. Bu yöntem diğer serpm diyagramlarına göre daha kolay bir yorumlamayı sağlamaktadır. Ancak etkili gözlemlerin saptanması amacıyla kullanılan diğer yöntemlerde de olduğu gibi, bu yöntemde de bir hesaplama zamanı söz konusu olduğundan pratik bir yöntem değildir. Buna ek olarak bağımsız değişken sayısı arttıkça, iki boyutlu tanı grafiklerinin sayısı da artmakta ve tüm bu grafiklerin birlikte incelenmesi

pratik olmamaktadır. Li, Morris ve Martin (2003) aynı yaklaşımı kullanarak temel bileşenler analizinde etkili gözlemlerin saptanması için bir grafiksel yöntem daha önermişlerdir. Her iki yöntem de Hampel (1974) tarafından geniş olarak çalışılan ve sağlam (robust) istatistiksel yöntemlerin bazılarının temelini oluşturan etki eğrisi (influence curve/function) kavramının sonlu örnek uyarlamalarından deneysel etki eğrisi (empirical influence curve/function) kavramına dayalıdır.

Bu çalışmada çoklu doğrusal regresyonda etkili gözlemlerin saptanması için SIC fonksiyonuna dayalı ve Biplot yaklaşımına dayalı iki grafik yöntemi önerilmiştir. Önerilen yöntemlerden birincisi, tasarım matrisinin her bir gözleme karşılık gelen satır vektörlerinin ağırlıklı toplamı ve gözlemlerin bağımlı değişken değerlerinin ağırlıklı toplamının yer aldığı SIC fonksiyonunun normuna dayalı bir amaç fonksiyonunu maksimum yapan w_i ağırlıklarını elde etmeye dayalıdır. Bu ağırlıklar gözlemlerin model parametre tahminlerini değiştirme eğilimlerine ilişkin katsayılardır. Dolayısıyla w_i ve gözlem numarası için oluşturulan serpmeye diyagramı görsel olarak incelenerek etkili gözlemler tespit edilmektedir. Ayrıca etkili gözlemlerin kolay bir biçimde saptanması amacıyla w_i değişkeninin değerleri için sınırlar oluşturulmuştur.

Önerilen ikinci yöntem olan Biplot yaklaşımına dayalı grafik yöntemi, parametre sayısı $p \geq 3$ iken, veri kümesindeki her bir gözlemin model parametrelerinin tahminleri üzerinde değişim miktarlarına dayalı olarak oluşturulan ve değişkenlerle gözlemlerin aynı grafikte gösterilebildiği bir yöntemdir. Ayrıca her bir parametre tahminindeki değişim birer değişken gibi düşünüldüğünde, bu grafik yöntemiyle, değişim bakımından parametre tahminlerinin ikili doğrusal ilişkileri de incelenebilmektedir. Buna ek olarak Biplotta etkili gözlemlerin görsel olarak saptanmasını kolaylaştırmak amacıyla Mahalanobis Uzaklığına dayalı bir elips sınır olarak belirlenmiştir.

Bu çalışmada önerilen iki yöntemin avantajlarından biri diğer grafiksel yöntemlerin aksine, tahmin edilecek parametre sayısının $p \geq 3$ olduğu durumda etkili gözlemlerin tek bir grafik üzerinden görsel olarak saptanabilmesidir. Diğer bir avantajı ise veri kümesindeki etkili gözlem gruplarının saptanması için kullanılan klasik tanı istatistiklerinde karşılaşılan,

$k > 2$ olmak üzere $\binom{n}{k}$ kadar olan EKK tahmini yapma gereksinimine ihtiyaç

duyulmamasıdır.

2. TEMEL KAVRAMLAR

Bu bölümde çoklu doğrusal regresyon modeli, etkili gözlemlerin saptanması için tanı yöntemlerinin temelini oluşturan bazı istatistikler ve kavramlar tanıtılmıştır. Daha sonra maskeleme ve yanılığa düşürme sorunlarından bahsedilmiştir.

2.1. Çoklu Doğrusal Regresyon Modeli

$n \times 1$ boyutlu gözlemlerin bağımlı değişken değerlerinin vektörü Y , $n \times p$ boyutlu ve p ranklı ($p < n$) tasarım matrisi X , $p \times 1$ boyutlu parametre vektörü β , 0 ortalamalı ve $\sigma^2 I$ varyanslı hata değişkenlerinin $n \times 1$ boyutlu vektörü ε ile gösterilmek üzere çoklu doğrusal regresyon modeli,

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

ile verilir. X matrisinde i . gözleme karşılık gelen satırın değerleri sütun vektörü olacak biçimde $x_i = [1 \ x_{i1} \ x_{i2} \ \dots \ x_{ip-1}]^T$ ile gösterilsin. i . gözleme ilişkin x_i bilgisinin ve bu gözleme ilişkin bağımlı değişken değerinin yer aldığı bir z_i vektörü $z_i = (x_i^T, y_i)^T$, ($i=1, 2, \dots, n$), ile tanımlansın. Bu durumda bu vektörlerin oluşturduğu matris $Z = [z_1, z_2, \dots, z_n]^T = (X, Y)$ ile verilir.

Hata değişkenlerine ilişkin varsayımlar altında Eş. 2.1 ile verilen modelde yer alan β parametre vektörünün EKK tahmin edicisi,

$$\hat{\beta} = (X^T X)^{-1} X^T Y \quad (2.2)$$

ile verilir. $\hat{\beta}$ vektörü kullanılarak elde edilen tahmin vektörü $\hat{Y}$ ile gösterilirse bu vektör,

$$\hat{Y} = X\hat{\beta} \quad (2.3)$$

ile bulunur. Artık vektörü e ise

$$e = Y - \hat{Y} \quad (2.4)$$

ile verilir. i . hata değişkeninin varyansı olan σ^2 için yansız bir tahmin edici,

$$\hat{\sigma}^2 = e^T e / (n - p) \quad (2.5)$$

ile verilir.

Etkili gözlemlerin saptanması için önerilen tanı yöntemlerinin çoğu için temel araçlar kaldıraç değeri (leverage value), artıklar, etki eğrisi ve etki eğrisinin sonlu örnek uyarlamalarıdır. Sonraki kısımlarda bunlar kısaca tanıtılmıştır.

2.2. Kaldıraç Değeri (Leverage Value)

EKK yöntemiyle tahmin edilen bir regresyon modelinde gözlemlerin her birinin regresyon modeli parametrelerinin tahminleri ve bağımlı değişken değerlerinin tahminleri üzerinde bir etkililiğe sahip olup olmadığının belirlenmesi oldukça önemlidir. Projeksiyon matrisi olarak da bilinen Hat matrisinin köşegen elemanları bu bilgiyi verir ve Student Törü Artıklarla beraber verinin geneline uymayan noktaların tespitinde bir ön inceleme amacıyla kullanılabilir.

Eş. 2.3 genişletilirse,

$$\begin{aligned} \hat{Y} &= X \hat{\beta} \\ &= X (X^T X)^{-1} X^T Y \\ &= HY \end{aligned} \quad (2.6)$$

elde edilir. Burada $n \times n$ boyutlu H matrisi şapka (hat) matrisi veya izdüşüm (projeksiyon) matrisi olarak bilinir. I_n matrisi $n \times n$ boyutlu birim matris olmak üzere, Eş. 2.4'te verilen artıkların vektörü olan e , izdüşüm matrisi ile

$$\begin{aligned} e &= Y - \hat{Y} \\ &= (I_n - H)Y \end{aligned} \quad (2.7)$$

biçiminde verilir. Bu durumda H matrisi

$$H = X(X^T X)^{-1} X^T \quad (2.8)$$

olarak yazılır. Bu matris simetrik ve eşgüçlü olduğundan $(I_n - H)$ matrisi de aynı özelliğe sahiptir.

H matrisinin i . köşegen elemanına karşılık gelen değer h_{ii} ile gösterilir ve kaldıraç değeri olarak adlandırılır. i . gözleme karşılık gelen kaldıraç değeri,

$$h_{ii} = x_i^T (X^T X)^{-1} x_i, \quad (i = 1, 2, \dots, n) \quad (2.9)$$

ile verilir.

Hoaglin ve Welsch (1978) i . gözleme karşılık gelen kaldıraç değerinin $2p/n$ ' den büyük olması durumunda, bu gözlemi *yüksek dereceli kaldıraç noktası* (high leverage point) olarak tanımlamışlardır. Ayrıca yüksek dereceli kaldıraç noktalarının verinin geneline uymayan gözlemler olabileceklerini de belirtmişlerdir. Bundan dolayı kaldıraç değerleri artıklarla beraber incelenmesi etkili gözlemlerin saptanması için kullanılabilir.

Veri kümesinin k elemanlı ($1 < k < n/2$) ve I ile gösterilen bir alt kümesi dikkate alınırsa Eş. 2.9'un genişletilmiş bir biçimi

$$H_I = X_I^T (X^T X)^{-1} X_I \quad (2.10)$$

ile verilir. Burada $k \times p$ boyutlu X_I^T matrisi, $n \times p$ boyutlu X tasarım matrisinde I gözlem kümesindeki gözlemlere karşılık gelen satırların oluşturduğu matristir. $k \times k$ boyutlu H_I matrisi etkili gözlem gruplarının saptanması için önerilen tanı istatistiklerinde kullanılan araçlardan biridir.

2.3. Artıklar ve Artıkların Standart Biçimleri

2.3.1. Sıradan artıklar

Regresyonda i . hata terimine ilişkin varsayım $\varepsilon_i \sim (0, \sigma^2)$ biçiminde verilir. Tüm hata terimlerinin bilinmeyen varyanslarının σ^2 olması regresyon varsayımlarından bir diğeri olan homojen varyans varsayımı olarak bilinir. Fakat artıklar için varyansların eşitliği söylenemez. i . artığa ilişkin beklenen değer 0 olmasına karşın, buna ilişkin varyans,

$$\text{var}(e_i) = (1 - h_{ii})\sigma^2 \quad (2.11)$$

olduğundan artıkların varyansları farklıdır. Artıkların vektörü e olmak üzere Eş. 2.11'in genelleştirilmiş bir biçimi,

$$\text{var}(e) = (\mathbf{I}_n - H)\sigma^2 \quad (2.12)$$

ile verilir. Eş. 2. 7'de verilen artık vektörü tekrar ele alınırsa

$$\begin{aligned} e &= (\mathbf{I}_n - H)Y \\ &= (\mathbf{I}_n - H)(X\beta + \varepsilon) \\ &= \left[\mathbf{I}_n - X(X^T X)^{-1} X^T \right] (X\beta + \varepsilon) \\ &= X\beta + \varepsilon - X\beta - X(X^T X)^{-1} X^T \varepsilon \\ &= (\mathbf{I}_n - H)\varepsilon \end{aligned} \quad (2.13)$$

elde edilir. Bu eşitlikten yararlanarak i . gözleme ilişkin artık

$$\begin{aligned} e_i &= (1 - h_{ii})y_i - \sum_{i' \neq i}^n h_{ii'} y_{i'} \\ &= (1 - h_{ii})\varepsilon_i - \sum_{i' \neq i}^n h_{ii'} \varepsilon_{i'} \end{aligned} \quad (2.14)$$

ile verilir. Regresyonda bağımsız değişken değerleri sabit olduğundan, bağımsız değişkene dayalı olan h_{ii} ve $h_{ii'}$ de sabittir. Bu durumda yukarıdaki eşitlikten de görüldüğü üzere artık terimlerinin dağılımı hata teriminin dağılımına ve örnek hacmine dayalıdır. Hata teriminin

dağılımı normal dağılım olduğu varsayımı altında artık terimi de normal dağılır. Hata terimlerinin dağılımına ilişkin bir bilgi yoksa, örnek hacmi arttıkça artık teriminin dağılımı normal dağılıma yakınsar.

2.3.2. Student türü artıklar (Studentized residuals)

Tanı yöntemleri için artıkların standart bir formunu elde etmek amacıyla birçok dönüşüm önerilmiştir. Bunlardan en basiti i . artığın, artık kareler ortalaması olarak bilinen $\hat{\sigma}^2$ tahminine bölünmesidir. Bu durumda artık kareler ortalaması her bir artık için birer yaklaşık varyans gibi alınır. Bununla birlikte artıkların varyansları eşit olmadığından daha iyi bir standart biçim, artıklara özgü varyansları kullanan Student Türü Artıklar (Studentized Residuals) olarak bilinir. Eş. 2.11'den de görüldüğü üzere bilinen e_i artıklarının 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) *Student Türü* terimini, ölçeğe dayalı bir istatistiğin, bu istatistiğin varyansının bir tahminine bölümü şeklinde tanımlamıştır. Bu bölümle elde edilen oran, ilgilenilmeyen parametrelerden (nuisance parameters) bağımsız bir dağılıma sahip olacaktır. Ayrıca Student Türü Artıklar, bu oranın pay ve paydasındaki istatistiklerin aynı veriden hesaplandığı Dahili Student Türü Artıklar (Internal Studentized Residuals) ve bu istatistiklerinin bağımsız olduğu Harici Student Türü Artıklar (External Studentized Residuals) olarak ikiye ayrılır.

i . gözlem için artığın varyansı Eş. 2.11'de verilmişti. Bu varyansın tahminini kullanan, i . gözleme ilişkin Dahili Student Türü Artık,

$$r_i = \frac{e_i}{\hat{\sigma}\sqrt{(1-h_{ii})}}, \quad (i = 1, 2, \dots, n) \quad (2.15)$$

ile verilir (Cook ve Weisberg, 1982: 18). Bu artıkların beklenen değeri 0'a eşittir. Varyansları ise e_i artıklarının varyanslarının aksine sabittir ve 1'e eşittir. Dahili Student Türü Artıklar dağılımı Student-t dağılımının özelliğini göstermesine rağmen Eş. 2.15'te verilen formülde pay ve payda bağımsız olmadığından dağılımları Student-t değildir.

Dahili Student Türü Artıkların veri kümesinin k elemanlı ($1 < k < n/2$) ve I ile gösterilen bir alt kümesi için genişletilmiş hali r_I^2 ile gösterilir ve

$$r_I^2 = \frac{e_I^T (\mathbf{I}_k - H_I)^{-1} e_I}{\hat{\sigma}^2} \quad (2.16)$$

ile verilir. Burada $\mathbf{I}_k$ matrisi $k \times k$ boyutlu birim matristir. e_I vektörü ise e_i artığının genişletilmiş bir formu olan ve e vektöründeki artıklardan, I gözlem kümesine ait elemanlara karşılık gelen artıkların oluşturduğu vektördür.

Eş. 2.11'de verilen varyans formülünde varyansın tahmin edicisi için kullanılan $\hat{\sigma}^2$ hesaplanırken tüm gözlemlere ilişkin artıklardan yararlanılmaktadır. Dolayısıyla i . gözleme ilişkin artığının varyansının tahmininde bu gözleme ilişkin artığın bilgisi de kullanılmaktadır. Harici Student Türü Artıklarda ise i . gözleme ilişkin artığın varyansının hesaplanması için kullanılan artık kareler ortalamasında bu gözleme ilişkin artık bilgisinden yararlanılmaz. i . gözleme ilişkin artığın dahil edilmediği artık kareler ortalaması $\hat{\sigma}_{(i)}^2$ ile gösterilirse bu,

$$\begin{aligned} \hat{\sigma}_{(i)}^2 &= \frac{(n-p)\hat{\sigma}^2 - e_i^2/(1-h_{ii})}{n-p-1} \\ &= \hat{\sigma}^2 \left(\frac{n-p-r_i^2}{n-p-1} \right), \quad (i = 1, 2, \dots, n) \end{aligned} \quad (2.17)$$

ile verilir (Cook ve Weisberg, 1982: 20). Hataların normal dağıldığı varsayımı altında, $\hat{\sigma}_{(i)}^2$ hesaplanırken i . gözleme ilişkin artığın dahil edilmemesi $\hat{\sigma}_{(i)}^2$ ile e_i 'nin bağımsız olmasını sağlar. Eş. 2.17'deki tahmin ediciden yararlanarak i . gözleme ilişkin Harici Student Türü Artık,

$$t_i = \frac{e_i}{\hat{\sigma}_{(i)} \sqrt{1-h_{ii}}}, \quad (i = 1, 2, \dots, n) \quad (2.18)$$

ile verilir (Cook ve Weisberg, 1982: 20). t_i istatistiğinin dağılımı $n-p-1$ serbestlik dereceli student-t dağılımıdır. Bazı kaynaklarda Harici Student Türü Artıklar, Jackknife

Artıklar olarak isimlendirilir. t_i istatistiği için bir diğer formül Dahili Student Türü Artık cinsinden

$$t_i = r_i \left(\frac{n-p-1}{n-p-r_i^2} \right)^{1/2}, \quad (i=1,2,\dots,n) \quad (2.19)$$

biçimindedir (Cook ve Weisberg, 1982: 20).

Dahili Student Türü Atıkların veri kümesinin k elemanlı ($1 < k < n/2$) ve I ile gösterilen bir alt kümesi için genişletilmiş hali t_I^2 ile gösterilir ve

$$t_I^2 = \frac{e_I^T (\mathbf{I}_k - H_I)^{-1} e_I}{\hat{\sigma}_{(I)}^2} \quad (2.20)$$

ile verilir. Burada $\hat{\sigma}_{(I)}^2$ tahmini Eş. 2.17’de verilen $\hat{\sigma}_{(i)}^2$ tahmininin I gözlem kümesi bakımından genişletilmiş halidir ve I gözlem kümesine ait artıkların bilgisini içermez.

Birçok durumda, özellikle büyük çaptaki veri setlerinde, artıkların varyansı dengede tutulur. Bu durumda, Dahili Student Türü Artıklar ve Harici Student Türü Artıklar arasında önemsiz derecede küçük farklılıklar mevcut olabilir.

2.3.3. Sıkıştırma artıkları (PRESS residuals)

Aykırı gözlemlerin ve etkili gözlemlerin saptanmasında kullanılan artıklara ilişkin bir diğer yaklaşım, $\hat{y}_{(i)}$ terimi, i . gözlem veri kümesinden çıkarıldıktan sonra, verinin geri kalanından tahmin edilen regresyon modeli kullanılarak elde edilen i . gözlemin bağımlı değişken değerinin tahmini olmak üzere, $y_i - \hat{y}_{(i)}$ farkına dayalı olan Sıkıştırma Artıkları (PRESS Residuals) olarak bilinir. Bunun temel mantığını açıklamak amacıyla veri kümesinde i ile gösterilen herhangi bir gözlemin, bağımlı değişken değerinin diğer gözlemlerin bağımlı değişken değerlerine göre olağan dışı olduğu dikkate alınsın. Bu durumda bu gözlem regresyon tahminleri üzerinde bir etkisi söz konusu olabilir. Buna bağlı olarak bu nokta regresyon doğrusunu veya düzlemini, kendi artığını mutlak değerce küçültecek şekilde değiştirme eğiliminde olduğundan, regresyon tahminlerini etkileme

eğilimi olan bu gözlemin bağımlı değişkeninin tahmini $\hat{y}_i$ değeri, bu gözlemin y_i bağımlı değişken değerine yakın bir değer alacaktır. Bundan dolayı i . gözlem veri kümesinden çıkarıldıktan sonra verinin geri kalanından tahmin edilen regresyon modeli bu gözlemden olumsuz bir biçimde etkilenmediğinden, bu gözleme ilişkin $y_i - \hat{y}_{(i)}$ farkının mutlak değerce büyük olması beklenir.

i . gözlem için Sıkıştırma Artığı $e_{S,i}$ ile gösterilir ve

$$e_{S,i} = \frac{y_i - \hat{y}_{(i)}}{1 - h_{ii}} \quad (2.21)$$

ile verilir(Montgomery, Peck ve Vining, 2001: 151). Sıkıştırma Artıkları, Harici Student Türü Artıklar gibi gözlem silme yaklaşımına dayalı olduğundan, bazı kaynaklarda bu artıklar Silinmiş Artıklar (Deleted Residuals) olarak da isimlendirilir. Eş. 2.21'den de görüldüğü üzere eğer i . gözlem regresyon sonuçlarını değiştirme eğilimindeyse, hem pay kısmı büyük olur hem de bu gözlem aynı zamanda yüksek dereceli kaldıraç noktası olduğundan h_{ii} değeri de büyük olur ve payda kısmı küçülme eğilimi gösterir. Bu durumda büyük değerli $e_{S,i}$, ilgili gözlemin etkili gözlem olabileceğine bir kanıt ifade eder(Montgomery ve diğerleri, 2001: 151).

Sıkıştırma Artıklarının beklenen değeri sıfırdır. Varyansları ise

$$\text{var}(e_{S,i}) = \text{var}\left(\frac{e_i}{1 - h_{ii}}\right) = \frac{1}{(1 - h_{ii})^2} [\sigma^2 (1 - h_{ii})] = \frac{\sigma^2}{1 - h_{ii}} \quad (2.22)$$

ile verilir. Bu durumda $e_{S,i}$ artığının standartlaştırılmış hali

$$\frac{e_{S,i}}{\sqrt{\text{var}(e_{S,i})}} = \frac{\frac{e_i}{1 - h_{ii}}}{\sqrt{\frac{\sigma^2}{1 - h_{ii}}}} = \frac{e_i}{\sqrt{\sigma^2 (1 - h_{ii})}} \quad (2.23)$$

olarak elde edilir. Yukarıdaki eşitlikte σ^2 parametresinin tahmin edicisi $\hat{\sigma}^2$ kullanılırsa $e_{s,i}$ artığının standartlaştırılmış hali i . gözlem için dahili student türü artığa eşit olur.

Aykırı gözlemlerin belirlenmesinde Standartlaştırılmış Artıklar, Student Türü Artıklar ve Sıkıştırma Artıkları oldukça kullanışlıdır. Bununla birlikte etkili gözlemlerin saptanmasında bu artıklar tek başlarına yeterli olmayabilir. Bu artıkların Kaldıraç Değerleriyle birlikte incelenmesi veri kümesindeki etkili gözlemlerin belirlenmesinde bir ön bilgi sunar. Ayrıca bu artıklar ve Kaldıraç Değeri etkili gözlemlerin saptanması için önerilen tanı yöntemlerinde kullanışlı araçlardır.

2.3.4. Tekrarlanan artıklar (Recursive residuals)

Artıkların diğer bir biçimi de Tekrarlanan Artıklar (Recursive Residuals) olarak bilinir. Bu artıklar hesaplanırken ilk p tane gözleme ilişkin Tekrarlanan Artık sıfır alınır. Sonra bu ilk p tane gözlem ele alınarak her bir adımda yeni oluşturulan bu p hacimli veri kümesine geri kalan $n-p$ gözlem sırasıyla birer birer eklenir ve $n-p$ tane Tekrarlanan Artık hesaplanır. i . gözleme ilişkin tekrarlanan artık

$$e_{R,i} = \frac{y_i - x_i^T \hat{\beta}_{i-1}}{\sqrt{1 + x_i^T (X_{i-1}^T X_{i-1})^{-1} x_i}}, \quad i = p+1, \dots, n \quad (2.24)$$

ile verilir (Galpin ve Hawkins, 1984). Burada $\hat{\beta}_{i-1}$ vektörü ilk $i-1$ gözlemden elde edilen β parametre vektörünün EKK tahmini, x_i vektörü elemanları i . gözlemin bağımsız değişken değerleri olan sütun vektörü ve X_{i-1} matrisi X tasarım matrisinin ilk $i-1$ satırının oluşturduğu matristir.

Tekrarlanan Artıklar, Sıradan Artıkların aksine bağımsız ve aynı dağılımlıdır. Hata terimine ilişkin $\varepsilon_i \sim N(0, \sigma^2)$ varsayım altında i . gözlem için Tekrarlanan Artık $e_{R,i} \sim N(0, \sigma^2)$ dağılır.

2.4. Etki Eğrisi ve Sonlu Örnek Uyarlamaları

2.4.1. Etki eğrisi (Influence curve)

Etki eğrisi ilk kez Von Mises (1947) tarafından önerilmiştir. Hampel (1974) ve Andrews, Bickel, Hampel, Huber, Rogers ve Tukey (1972) etki eğrisini, tahmin edicileri karşılaştırmak ve mevcut tahmin yöntemleri geliştirmek amacıyla kullanmışlardır. Özellikle, Hampel (1974) etki eğrisinin sağlam tahmin edicilerdeki rolü üzerine çalışmıştır.

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(Hampel, 1974).

Birikimli dağılım fonksiyonu F olan ve R^m ’ de tanımlı bir yığından çekilen bağımsız ve aynı dağılımlı örnek $t_1, t_2, \dots, t_n$, olsun ve bu örneğe dayalı p boyutlu vektörel değerli istatistik T_n ile gösterilsin. T_n vektörel değerli istatistiği, θ parametresinin tahmin edicisi olmak üzere 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 $p=1$ iken örnek ortalaması

$$T_n = \frac{1}{n} \sum_{i=1}^n t_i = \bar{t} \quad (2.25)$$

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

$$T(\hat{F}) = \int t d\hat{F} = \bar{t} \quad (2.26)$$

ile verilir(Hampel, 1974).

Herhangi bir $t_0 \in R^m$ noktasında belirlenen birikimli olasılıkların fonksiyonu ϕ_{t_0} olsun. τ , bozulma katsayısı olmak üzere, F ve $(1-\tau)F + \tau\phi_{t_0}$ birikimli dağılım fonksiyonları için T_n

istatistiğine karşılık gelen fonksiyoneller sırasıyla $T\{F\}$ ve $T\{(1-\tau)F + \tau\phi_{t_0}\}$ olsun. Bu durumda F birikimli dağılım fonksiyonu üzerindeki T istatistiğinin etki eğrisi

$$IC(t) = \lim_{\tau \rightarrow 0^+} \frac{T\{(1-\tau)F + \tau\phi_{t_0}\} - T\{F\}}{\tau}, \quad (0 < \tau < 1) \quad (2.27)$$

olarak tanımlanır. Burada limitin her $t \in R^m$ noktası için tanımlı olması zorunluluğu vardır (Hampel, 1974).

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.

2.4.2. Etki eğrisinin sonlu 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 yöntemi olduğunu ve etki eğrisinin sonlu bir örnek uyarlamasının 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*) kavramlarını bulmuştur. Bu fikir geneldir ve diğer durumlara uygulanabilir. Bundan dolayı gözlem silme yaklaşımında, her bir gözlem değerini değerlendirerek etkili gözlemleri saptamak, sonlu örnek uyarlamalarıyla daha kullanışlı olduğundan etki eğrisinin sonlu örnek uyarlamaları dikkate alınmaktadır. (Cook ve Weisberg, 1982: 108).

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 ($i=1,2,...,n$) ilgili birime ilişkin gözlem değerlerinin bir vektörü olsun. Bu durumda bağımsız değişkenlerin ve bağımlı değişkenin değerlerinin yer aldığı bir matris

$$Z = [X_{n \times p} : Y_{n \times 1}]_{n \times (p+1)} \quad (2.28)$$

olarak tanımlansın. Eş. 2.1'deki model dikkate alınırsa, $\hat{\beta}$ tahmin edicisi için örnek birikimli dağılım fonksiyonu $\hat{F}$, Eş. 2.27'de yerine konursa deneysel etki eğrisi (*EIC*)

$$EIC(z_i) = n(X^T X)^{-1} x_i e_i \quad (2.29)$$

biçiminde yazılır (Cook ve Weisberg, 1982: 108). i . birimin silinmesiyle elde edilen örnek birikimli dağılım fonksiyonu $\hat{F}_{(i)}$, Eş. 2.27'de $\hat{\beta}$ tahmin edicisi için yerine konursa, gözlem silme yaklaşımına dayalı olan yeni deneysel etki eğrisi

$$EIC_{(i)}(z_i) = (n-1)(X_{(i)}^T X_{(i)})^{-1} x_i (y_i - x_i^T \hat{\beta}_{(i)}) \quad (2.30)$$

ile elde edilir (Cook ve Weisberg, 1982: 109). Burada $X_{(i)}$ matrisi, tasarım matrisinden i . satır çıkarıldıktan sonra elde edilen matristir. Dolayısıyla $(X_{(i)}^T X_{(i)})^{-1} = (X^T X - x_i x_i^T)^{-1}$ olur. $\hat{\beta}_{(i)}$ ise i . birimin silinmesiyle elde edilen β parametre vektörünün EKK tahmin edicisidir ve

$$\hat{\beta}_{(i)} = (X_{(i)}^T X_{(i)})^{-1} X_{(i)}^T Y_{(i)} \quad (2.31)$$

ile verilir. Burada $Y_{(i)}$ vektörü, bağımlı değişken değerlerine ilişkin vektörde i . gözlemin bağımlı değişken değerinin çıkarılmasıyla elde edilen vektördür. Veri kümesinin k elemanlı ($1 < k < n/2$) ve I ile gösterilen bir alt kümesi veri kümesinden çıkartıldıktan sonra geri kalan $n-k$ gözlemin bilgisinden elde edilen β parametre vektörünün EKK tahmini $\hat{\beta}_{(I)}$ ile gösterilirse bu tahmin vektörü

$$\hat{\beta}_{(I)} = (X_{(I)}^T X_{(I)})^{-1} X_{(I)}^T Y_{(I)} \quad (2.32)$$

ile hesaplanır. Burada $X_{(I)}$ matrisi, X tasarım matrisinden I alt kümesini oluşturan gözlemlerin numaralarına karşılık gelen satırların çıkartılmasıyla; $Y_{(I)}$ vektörü ise Y vektöründen I alt kümesini oluşturan gözlemlerin numaralarına karşılık gelen elemanların çıkartılmasıyla elde edilir. Etki eğrisinin örnek uyarlamasının daha arzu edilen bir biçimi, Eş. 2.27’de $\hat{\beta}$ tahmin edicisi için F yerine $\hat{F}$ ve τ yerine $-1/(n-1)$ alınarak elde edilen örnek etki eğrisi (SIC)’dir ve

$$\begin{aligned}
 SIC(z_i) &= -(n-1) \left(\hat{\beta}(\hat{F}_{(i)}) - \hat{\beta}(\hat{F}) \right) \\
 &= (n-1) \left(\hat{\beta} - \hat{\beta}_{(i)} \right) \\
 &= \frac{(n-1)(X^T X)^{-1} x_i e_i}{1 - h_{ii}} \\
 &= \frac{(n-1)(X^T X)^{-1} x_i e_i}{1 - x_i^T (X^T X)^{-1} x_i}
 \end{aligned} \tag{2.33}$$

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

Etkili gözlemlerin saptanması için uygulamada sıkça kullanılan Cook uzaklığı, SIC kullanılarak türetilir (Cook ve Weisberg, 1982: 114).

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.

2.5. Maskeleye ve Yanılgıya Düşürme Sorunları

2.5.1. Maskeleye sorunu

Birden fazla etkili gözlemin bulunduğu bazı veri kümelerinde bir gözlemi silmeye dayalı tanı yöntemleri regresyon tahminlerini olumsuz yönde değiştirme eğilimi olan gözlemleri teker teker saptayamayabilir. Bu durumda veri kümesindeki etkili gözlemleri saptamak için veri kümesinin alt kümeleri olan k hacimli ($1 < k < n/2$) gözlem grupları silinerek, tanı yöntemleriyle etkili gözlemler saptanmaya çalışılır. Eğer bu k hacimli gözlem grubu veri

kümesindeki tüm etkili gözlemleri içeriyorsa, bu etkili gözlem grubunun veri kümesinden tamamen çıkartılması veya bunlara uygun ağırlıkların atanması ile bunların regresyon tahminlerini değiştirme eğilimlerinin azaltılması tahmin edilen regresyon modelinin, veri kümesi için büyük ölçüde uygun olmasını sağlar.

Birden fazla etkili gözlemin olduğu veri kümeleri için bir gözlemi silmeye dayalı tanı yöntemlerinin bu etkili gözlemleri teker teker saptayamaması durumu bir başarısızlıktır. Bu sorun *Maskeleme Sorunu* (Masking) olarak bilinir.

Literatürde maskelemeye ilişkin farklı tanımlar vardır. Chatterjee ve Hadi (1986), iki veya daha fazla gözlem tek başlarına regresyon katsayıları üzerinde etkili gözüküyorken, bu gözlemlerin tahminler üzerinde ortak etkisi olması durumunu maskeleme olarak tanımlamışlardır. Rousseeuw ve Leroy (2003) da maskelemeyi, bir etkili gözlemin veriden atılmasından sonra ilk başta etkili görünmeyen bir gözlemin regresyon parametrelerinin tahminleri üzerinde etkili bir gözlem olarak ortaya çıkması şeklinde tanımlamışlardır. Bu iki tanımlama aynıdır ve koşullu etkililikle ilgilidir. Aynı tanımlamalar veri kümesindeki ikiden fazla etkili gözlem için de yapılabilir.

Gray ve Ling (1984) maskelemeyi belirlemek için Hat Matrisi'nin köşegen olmayan elemanlarından ve alt matrislerinden yararlanabileceğini öne sürmüşlerdir. Ayrıca maskelemenin belirlenmesi için Atkinson (1986) en küçük medyan kareler yöntemine dayalı Maske Düşürme (Unmasked) yöntemini bulmuştur.

2.5.2. Yanılgıya düşürme sorunu

Maskelemenin karşıtı olarak bilinen *Yanılgıya Düşürme Sorunu* (swamping) etkili gözlem grupları incelenirken karşılaşılabilecek diğer bir problemdir. 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 (Atkinson, 1986; Lawrance, 1995).

Hem maskeleme hem de yanılgıya-düşürme, etkili gözlem gruplarının saptanması için kullanılan çoğu tanı istatistiğinde yanlış kararlar almaya sebep olabilir. Bu da

arařtırmacının yapacađı regresyon analizini olumsuz bir biçimde etkiler. Bu durumda etkili gözlemler incelenirken maskeleme ve yanılgıya-düşürme sorunlarının mevcut olup olmadığı da incelenmelidir.

3. ETKİLİ GÖZLEMLERİN SAPTANMASI İÇİN KULLANILAN TANI YÖNTEMLERİ

Regresyonda verinin geneline uymayan ve artık değeri mutlak değerce büyük olan gözlemler aykırı gözlem olarak tanımlanır. Bazı durumlarda veri kümesinde yer alan gözlemlerden birinin veya bir kaçının artıkları mutlak değerce büyük olup bu gözlem veya gözlemlere ilişkin kaldıraç değerleri de verinin geneline uyan diğer gözlemlere ilişkin kaldıraç değerlerinden oldukça büyük olabilir. Bu durumda hem aykırı hem de yüksek dereceli kaldıraç noktası olan bu gözlem veya gözlemler regresyon sonuçlarını olumsuz şekilde değiştirme eğiliminde olabilir. Bu tip gözlemler etkili gözlem olarak tanımlanır ve aykırı gözlemlerden farklı bir şekilde incelenir.

Etkili gözlemlerin saptanmasının çeşitli amaçları vardır. Kimi zaman etkili bir gözlemin saptanması ilgili bilim dalında çalışan araştırmacının hedefi olabilir. İlaç geliştirme üzerine deneysel olarak çalışan bir eczacı, deney sonucu elde ettiği veriyi regresyon analizi kullanarak analiz ederken saptadığı etkili gözlemin, üzerinde çalıştığı ilaca yan etkisi olan bir hastaya ilişkin olduğunu tespit edip, bu hastanın ilaca tepkisinin nedenlerini araştırabilir. Nitekim bu araştırma ile ilacın yan etkilerini belirleyebilir. Bununla birlikte genellikle regresyon analizinin amacı bir istatistiksel model oluşturmaktır. Bu durumda regresyon modelinin parametrelerinin tahminlerini olumsuz bir biçimde değiştirme eğiliminde olan ve buna bağlı olarak regresyon sonuçlarını beklenmedik bir şekilde değiştiren gözlemler olan etkili gözlemlerin saptanması regresyon modelini oluşturma sürecinde önemli bir aşamadır.

Literatürde etkili gözlemlerin saptanması için önerilen yöntemlerin sayısı oldukça fazladır. Bu yöntemlerin temel araçları artıkların özel biçimleri olan Student Türü Artıklar, Sıkıştırma Artıkları ve Kaldıraç Değeri olarak bilinen h_{ii} kavramlarına dayalıdır. Bunun yanı sıra Etki Eğrisinin Örnek Uyarlamaları olan Deneysel Etki Eğrisi (EIC) ve Örnek Etki Eğrisi (SIC) kavramları da etkili gözlemlerin saptanması için önerilen yöntemlerde kullanılmaktadır.

Etkili gözlemlerin saptanması için kullanılan tanı yöntemlerinin çoğunda gözlem silme yaklaşımından yararlanılır. Bu yaklaşımla veri kümesinden bir gözlemin veya veri kümesinin bir alt kümesini oluşturan ve eleman sayısı birden büyük olan bir gözlem

kümesinin veriden çıkartılmasıyla, geri kalan gözlemlerden elde edilen regresyon tahminlerinin ve sonuçlarının nasıl değiştiği incelenir. Gözlem silme yaklaşımının kullanıldığı tanı istatistiklerinden en bilineni Cook Uzaklığı İstatistiği'dir. Cook (1977) tarafından önerilen bu istatistik regresyon tahminleri üzerinde hem bireysel olarak etkili olan bir gözlemi saptamak için, hem de ortak etkiye sahip bir etkili gözlem grubunu saptamak için kullanılabilir. Bunun haricinde benzer istatistikler DFBETAS, DFFITS, COVRATIO, Pena'nın S İstatistiği, Andrews-Pregibon' un AP İstatistiği, Cook-Weisberg Olabilirlik İstatistiği LD, Hadi'nin K İstatistiği etkili gözlemlerin saptanmasında kullanılmaktadır.

Gözlem silme yaklaşımı tanı yöntemlerinde oldukça yaygın kullanılmasına rağmen, örnek hacminin büyük durumları için pratiklik sağlamaz. Ayrıca bu yaklaşıma dayalı yöntemlerde tanı yöntemleri için iki önemli sorun olan maskeleye ve yanılgıya düşürme problemlerine de neden olur.

Basit doğrusal regresyonda, etkili gözlem veya gözlem gruplarının görsel olarak saptanması için bir yöntem de bağımlı değişkenle bağımsız değişken için serpmeye diyagramının incelenmesidir. Ancak bağımsız değişken sayısı arttıkça serpmeye diyagramlarının oluşturulmasında boyut sorunu ortaya çıkar. Bu amaçla etkili gözlemlerin saptanması için boyut indirgemeye dayalı bir grafiksel yöntem Li ve diğerleri (2001) tarafından önerilmiştir.

Etkili gözlemlerin görsel olarak saptanması amacıyla bir ön çalışma olarak Sıradan Artık Değerleri, Student Türü Artık Değerleri, Sıkıştırma Artık Değerleri ve Kaldıraç Değerlerine ilişkin grafikler incelenebilir. Ayrıca her bir gözlem için elde edilen tanı istatistiklerinin grafikleri de etkili gözlemleri görsel olarak saptamak için kullanılabilir. Bu grafikler incelenirken ilgili gözlemin etkili gözlem olup olmadığının karar verilmesinde ilgili istatistiğin karar kuralından yararlanılır. Bunların haricinde Eklenmiş Değişken Grafikleri (Added Variable Plots), Kısmi Artık Grafikleri (Partial Residual Plots) etkili gözlemlerin saptanması için artık değerlerine dayalı diğer grafiksel yöntemlerden bazılarıdır.

Bu bölümde etkili gözlemlerin saptanması için kullanılan tanı istatistikleri ve grafiksel yöntemler verilmiştir.

3.1. Etkili Gözlemlerin Saptanması için Tanı İstatistikleri

3.1.1. Cook uzaklığı istatistiği

Etkili gözlemlerin saptanmasında sıkça kullanılan bir istatistik olan Cook Uzaklığı İstatistiği, veri kümesinin tümüne dayalı olarak hesaplanan EKK tahmin vektörü $\hat{\beta}$ ile veri kümesinden bir gözlemin çıkartılmasıyla hesaplanan EKK tahmin vektörü $\hat{\beta}_{(i)}$ arasındaki karesel uzaklığın bir ölçüsüdür. Cook Uzaklığı İstatistiği $D_{(i)}$ ile gösterilir.

Cook Uzaklığı istatistiğinin türetilmesinde kullanılan ve Eş. 2.33'de verilen, β parametre vektörünün tahmin edicisi için örnek etki eğrisi

$$SIC(z_i) = (n-1) \left(\hat{\beta} - \hat{\beta}_{(i)} \right)^2, \quad (i = 1, 2, \dots, n) \quad (3.1)$$

ile verilir. Örnek etki eğrisi $p \times 1$ boyutlu bir vektör olduğundan 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. 3.1, $\hat{\beta}$ ile $\hat{\beta}_{(i)}$ arasındaki fark biçiminde yazılırsa

$$(n-1)^{-1} SIC(z_i) = \hat{\beta} - \hat{\beta}_{(i)}, \quad (i = 1, 2, \dots, n) \quad (3.2)$$

elde edilir (Cook ve Weisberg, 1982: 110). 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 ζ faktörüne dayalı olarak $\hat{\beta}$ ile $\hat{\beta}_{(i)}$ arasındaki karesel uzaklık $D_{(i)}(M, \zeta)$ normları ile hesaplanır. Bu durumda $D_{(i)}(M, \zeta)$ normu

$$D_{(i)}(M, \zeta) = (n-1)^{-2} \frac{[SIC(z_i)]^T M [SIC(z_i)]}{\zeta} \quad (3.3)$$

$$= \frac{(\hat{\beta}_{(i)} - \hat{\beta})^T M (\hat{\beta}_{(i)} - \hat{\beta})}{\zeta}, \quad (i = 1, 2, \dots, n)$$

ile elde edilir (Cook ve Weisberg, 1982: 114). Cook (1977) tarafından önerilen bir yaklaşıma göre $M = X^T X$ ve $\zeta = p\hat{\sigma}^2$ alınıp, Eş. 3.3'te yerine koyulursa Cook Uzaklığı

$$D_{(i)}(X^T X, p\hat{\sigma}^2) = \frac{(\hat{\beta}_{(i)} - \hat{\beta})^T X^T X (\hat{\beta}_{(i)} - \hat{\beta})}{p\hat{\sigma}^2}, \quad (i = 1, 2, \dots, n) \quad (3.4)$$

ile verilir (Cook ve Weisberg, 1982: 115). 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. Eş. 3.4'te verilen biçimiyle Cook Uzaklığı, Mahanobis Uzaklığı'na benzer biçimdedir.

Örnek hacmi büyük iken Eş. 3.4'te verilen formülüyle Cook Uzaklığı İstatistiği, her bir gözlem için $\hat{\beta}_{(i)}$ tahmin vektörleri ayrı ayrı hesaplanacağından, pratik değildir. Bu durumda Cook Uzaklığı İstatistiği'nin hesaplama bakımından daha pratik bir biçimi ikinci bölümde verilen Dahili Student Türü Artıkların ve Hat matrisinin köşegen elemanlarının yer aldığı

$$D_{(i)} = \frac{r_i^2 \text{Var}(\hat{y}_i)}{p \text{Var}(e_i)} \quad (3.5)$$

$$= \frac{r_i^2 h_{ii}}{p(1 - h_{ii})}, \quad i = 1, 2, \dots, n$$

formülüyle verilir (Cook ve Weisberg, 1982: 117). Eş. 3.5'te sağ taraftaki $h_{ii}/(1 - h_{ii})$ oranı, x_i vektörünün geri kalan verinin kitle merkezine olan uzaklığını ifade etmektedir. Ayrıca büyük artıklara sahip gözlemler de etkili gözlem olarak yorumlanabileceği için, Dahili

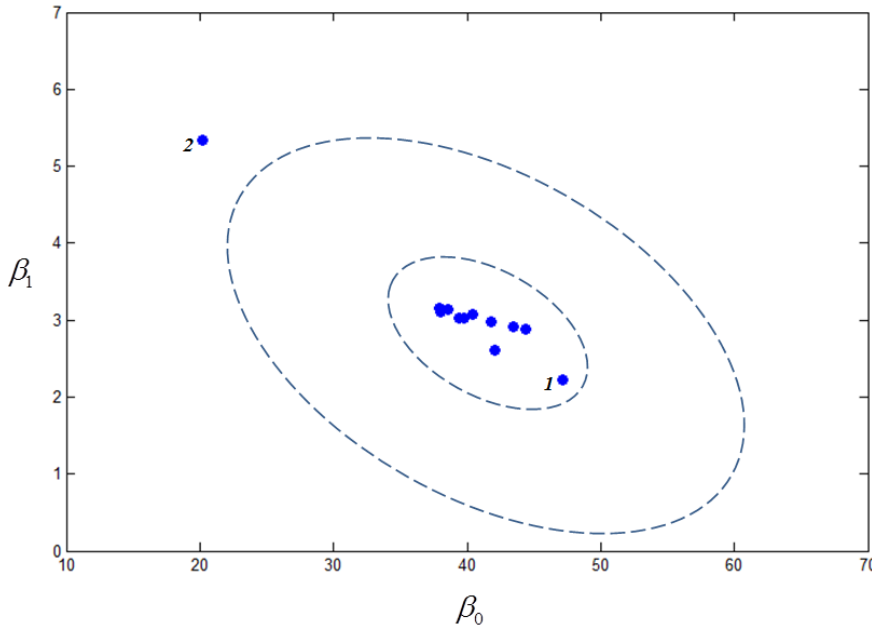
Student Türü Artıklarla hesaplanan $D_{(i)}$ değerlerinden oldukça büyük olanlar etkili gözlemlerin göstergesidir.

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

$$D_{(i)} = \frac{(\hat{Y}_{(i)} - \hat{Y})^T (\hat{Y}_{(i)} - \hat{Y})}{p\hat{\sigma}^2}, \quad (i = 1, 2, \dots, n) \quad (3.6)$$

ile verilir (Cook ve Weisberg, 1982: 116). Eş. 3.4 ve Eş. 3.6'dan görüldüğü üzere gözlem silme yaklaşımına dayalı olarak $\hat{\beta}$ tahmin vektöründeki değişim üzerinden Cook Uzaklığı İstatistiği hesaplandığı gibi $\hat{y}$ tahmin vektöründeki değişim üzerinden de bu istatistik hesaplanabilir.

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 alarak çıkartılır. Veri kümesinin 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 veri kümesinden 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. 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 Şekil 3.1'deki gibi olsun.



Şekil 3.1. Basit Doğrusal Regresyon Modeli için $\hat{\beta}_{0(i)}$ ve $\hat{\beta}_{1(i)}$ tahminlerine ait grafik

Şekil 3.1’de basit doğrusal regresyon modeli için tüm veriden elde edilen regresyon katsayıları merkezi oluşturmak üzere, her bir gözleme ilişkin gözlem silme yaklaşımına 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 $\hat{\beta}_{(2)} = (\hat{\beta}_{0(2)}, \hat{\beta}_{1(2)})$ noktası %50’lik güven elipsoidi dışındadır. Dolayısıyla bu gözlemin regresyon tahminlerini olumsuz yönde etkilediği 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 İstatistiğiyle 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 (Cook ve Weisberg, 1982: 116).

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 k 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.

$F_{0.50,p,n-p} \cong 1$ olduğundan, Cook Uzaklığı için kritik değerin 1 alınması etkili gözlemlerin incelenmesinde pratiklik sağlar. Bu nedenle uygulamada Cook Uzaklığı istatistiği için kritik değer 1 alınmaktadır (Cook ve Weisberg, 1982: 118). Buna ek olarak Kim ve Storer (1996) Monte-Carlo simülasyon yöntemiyle örnek hacminin ve bağımsız değişken sayısının bazı durumlarında Cook Uzaklığı İstatistiği için kritik değerleri elde etmiştir.

3.1.2. DFBETAS istatistiği

Belsley ve diğerleri (1980) tarafından önerilen $DFBETAS_{j(i)}$ istatistiği, veri kümesinden i . gözlemin silinmesiyle elde edilen 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. Verinin tümünden elde edilen β parametre vektörünün EKK tahmin edicisi $\hat{\beta}$ ile $\hat{\beta}_{(i)}$ arasındaki fark $DFBETA_i$ ile gösterilirse bu fark istatistiği

$$DFBETA_i \equiv \hat{\beta} - \hat{\beta}_{(i)} = \frac{(X^T X)^{-1} x_i e_i}{1 - h_{ii}} \quad , \quad (i = 1, 2, \dots, n) \quad (3.7)$$

ile verilir (Belsley ve diğerleri, 1980: 13). Ancak $DFBETA_i$ sadece vektörel değerli bir sonuç vermektedir ve skaler değerli sonuç veren bir istatistik ile etkili gözlemleri saptamak

daha kolaydır. Bu durumda Eş. 3.7'nin geliştirilmiş bir biçimi olan $DFBETAS_{j(i)}$ istatistiği kullanılır (Belsley ve diğerleri, 1980: 13).

$\hat{\beta}$ 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

$$\hat{\beta}_j - \hat{\beta}_{j(i)} = \frac{c_{ji} e_i}{1 - h_{ii}}, \quad (i = 1, 2, \dots, n; j = 1, 2, \dots, p) \quad (3.8)$$

biçiminde yazılır (Belsley ve diğerleri, 1980: 13). Burada c_{ji} terimi, 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

$$\begin{aligned} \text{Var}(\hat{\beta}_j) &= \sigma^2 (X^T X)^{-1}_{jj} \\ &= \sigma^2 \sum_{s=1}^n c_{js}^2 \end{aligned} \quad (3.9)$$

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

$$\begin{aligned} DFBETAS_{j(i)} &= \frac{\hat{\beta}_j - \hat{\beta}_{j(i)}}{\hat{\sigma}_{(i)} \sqrt{(X^T X)^{-1}_{jj}}} \\ &= \frac{c_{ji}}{\sqrt{\sum_{s=1}^n c_{js}^2}} \frac{e_i}{\hat{\sigma}_{(i)} (1 - h_{ii})} \end{aligned} \quad (3.10)$$

biçiminde elde edilir (Belsley ve diğerleri, 1980: 13). Burada $\hat{\sigma}_{(i)}$ terimi i . gözlemin silinmesiyle elde edilen standart sapmadır ve

$$\hat{\sigma}_{(i)} = \sqrt{\frac{1}{n-p} \sum_{s=1}^n (y_s - x_s^T \hat{\beta}_{(i)})^2} \quad (3.11)$$

ile verilir (Belsley ve diğerleri, 1980: 14). C matrisinin j . satırı c_j ve Harici Student Türü Artıklar t_i olmak üzere $DFBETAS_{j(i)}$ istatistiğinin alternatif bir formülü

$$DFBETAS_{j(i)} = \frac{c_{ji}}{\sqrt{c_j c_j^T}} \frac{t_i}{\sqrt{1-h_{ii}}} \quad , \quad i = 1, 2, \dots, n; j = 1, 2, \dots, p \quad (3.12)$$

olarak bilinir. Belsley ve diğerleri (1980) tarafından önerilen kritere göre, eğer $|DFBETAS_{j(i)}| > 2/\sqrt{n}$ ise i . gözlem etkili bir gözlemdir. Cook Uzaklığı İstatistiği $D_{(i)}$, $\hat{\beta}$ tahmin vektöründe yer alan regresyon katsayılarının tümünü ortak olarak olumsuz bir biçimde etkileyen gözlemleri saptamak için kullanılmasına karşın, $DFBETAS_{j(i)}$ istatistiği her bir $\hat{\beta}_j$ tahminini ayrı ayrı ele alarak bunları olumsuz bir biçimde etkileyen gözlemleri saptamak amacıyla kullanılır. Bununla birlikte kullandığı tahminler bakımından bu iki tanı istatistiği benzerdir.

3.1.3. DFFITS İstatistiği

Bireysel etkili olan gözlemi saptamak için kullanılan bir diğer tanı istatistiği Belsley ve diğerleri (1980) tarafından önerilen $DFFITS_{(i)}$ istatistiğidir. Gözlem silme yaklaşımına dayalı olan bu istatistik, i . gözlemin tahmin değerindeki değişimine dayanır.

$DFFITS_{(i)}$ 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 (Belsley ve diğerleri, 1980: 15).

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

$$DFFIT_i \equiv \hat{y}_i - \hat{y}_{i(i)} = x_i^T (\hat{\beta} - \hat{\beta}_{(i)}) = \frac{h_{ii} e_i}{1 - h_{ii}} \quad , \quad (i = 1, 2, \dots, n) \quad (3.13)$$

ile verilir. Bu tanı ölçüsünün avantajı regresyon modelinde kullanılan özel koordinat sistemlerine bağlı olmamasıdır (Belsley ve diğerleri, 1980: 15). Ölçeklendirmeden

kurtulmak amacıyla Eş. 3.13'teki farkı $\hat{\sigma}_{(i)}\sqrt{h_{ii}}$ terimine bölmek gerekir. Bu durumda elde edilen bu farkın standartlaştırılmış biçimi $DFFITS_{(i)}$ ile gösterilir ve

$$\begin{aligned} DFFITS_{(i)} &= \frac{\hat{y}_i - \hat{y}_{i(i)}}{\hat{\sigma}_{(i)}\sqrt{h_{ii}}} \\ &= \left(\frac{h_{ii}}{1-h_{ii}} \right)^{1/2} \frac{e_i}{\hat{\sigma}_{(i)}\sqrt{1-h_{ii}}}, \quad i=1,2,\dots,n \end{aligned} \quad (3.14)$$

biçimindedir (Belsley ve diğerleri, 1980: 15). Eş. 3.14'teki formülde sol taraf, yani $\left[h_{ii}/(1-h_{ii}) \right]^{1/2}$ fonksiyonu, i . gözlemin kaldıraç gücüne karşılık gelmektedir. Sağ taraftaki oran ise i . gözleme ilişkin Harici Student Türü Artık değeridir. Böylece bu iki kısmın çarpımı olan $DFFITS_{(i)}$ istatistiği

$$DFFITS_{(i)} = \left(\frac{h_{ii}}{1-h_{ii}} \right)^{1/2} t_i, \quad i=1,2,\dots,n; h_{ii} \neq 1 \quad (3.15)$$

biçiminde yeniden yazılabilir (Belsley ve diğerleri, 1980: 15). Burada büyük h_{ii} değerine ve büyük Harici Student Türü Artığa sahip gözlemlerin etkili gözlem olacağı açıkça görülmektedir. Buna dayanarak Belsley ve diğerleri (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 (Belsley ve diğerleri, 1980: 28).

3.1.4. Kovaryans oranı (COVRATIO)

Cook uzaklığı, $DFBETAS_{(i)}$ ve $DFFITS_{(i)}$ tanı istatistikleri sadece tahmin edilen regresyon katsayıları ve tahmin değerleri üzerinde etkili olan gözlemleri saptamada kullanılır. Fakat bu istatistikler tam olarak tahminin duyarlılığına (precision of estimation) ilişkin herhangi bir bilgi sağlamaz. Varyans-kovaryans matrisinin determinantı, duyarlılığın uygun bir sayısal ölçüsü olarak oldukça yaygın kullanılmasından dolayı, genelleştirilmiş varyans olarak tanımlanmıştır (Belsley ve diğerleri, 1980: 22). Bu durumda $\hat{\beta}$ tahmininin genelleştirilmiş varyansı

$$\begin{aligned} G\text{Var}(\hat{\beta}) &= \det\{\text{Var}(\hat{\beta})\} \\ &= \det\{\hat{\sigma}^2(X^T X)^{-1}\} \end{aligned} \quad (3.16)$$

biçimindedir (Belsley ve diğerleri, 1980: 22). Veri kümesinden 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 yaklaşımıyla elde edilen bu oran, kovaryans oranı (COVRATIO) olarak tanımlanır ve

$$COVRATIO_{(i)} = \frac{\det[\hat{\sigma}_{(i)}^2(X_{(i)}^T X_{(i)})^{-1}]}{\det[\hat{\sigma}^2(X^T X)^{-1}]}, \quad i = 1, 2, \dots, n \quad (3.17)$$

ile verilir (Belsley ve diğerleri, 1980: 22). Eş. 3.17’de sağ tarafta pay kısmında, $\det(X_{(i)}^T X_{(i)})^{-1} = (1 - h_{ii})^{-1} \det(X^T X)^{-1}$ olduğundan bu oran için alternatif bir formül

$$\begin{aligned} COVRATIO_{(i)} &= \left(\frac{\hat{\sigma}_{(i)}^2}{\hat{\sigma}^2} \right)^p \left(\frac{1}{1 - h_{ii}} \right) \\ &= \frac{1}{\left(\frac{n - p - 1}{n - p} + \frac{t_i^2}{n - p} \right)^p (1 - h_{ii})}, \quad i = 1, 2, \dots, n \end{aligned} \quad (3.18)$$

ile verilir. Burada t_i terimi, i . gözleme karşılık gelen Harici Student Türü Artık terimidir (Belsley ve diğerleri, 1980: 22).

i . gözleme ilişkin $COVRATIO_{(i)}$ değerinin 1’ den büyük olması durumunda, bu gözlemin tahminin duyarlılığını arttıracak ve $COVRATIO_{(i)}$ değerinin 1’ den küçük olması durumunda da i . gözlemin tahminin duyarlılığını bozacağı açıktır. Dolayısıyla karar kuralı olarak $|COVRATIO_{(i)} - 1| > 3p/n$ olması durumunda i . gözlemin etkili gözlem olabileceği önerilmiştir (Belsley ve diğerleri, 1980: 28).

Kovaryans oranı, Cook uzaklığı gibi etkili gözlem gruplarını saptamak için de kullanışlı bir ölçüdür. Eş. 3.17'deki formül genelleştirilirse verinin k büyüklüğünde ($k < n$) bir alt kümesi olan ve I ile gösterilen alt kümesinin silinmesiyle elde edilen $COVRATIO_{(I)}$ değeri,

$$COVRATIO_{(I)} = \frac{\det[\hat{\sigma}_{(I)}^2 (X_{(I)}^T X_{(I)})^{-1}]}{\det[\hat{\sigma}^2 (X^T X)^{-1}]} \quad (3.19)$$

ile verilir(Belsley ve diğerleri, 1980: 33). Tek gözlem için $COVRATIO_{(I)}$ istatistiğinde kullanılan karar kuralı, gözlem grupları $COVRATIO_{(I)}$ ile incelenirken de kullanılmaktadır.

3.1.5. Andrews-Pregibon AP istatistiği

Gözlem silme yaklaşımına dayalı diğer bir tanı istatistiği, Andrews ve Pregibon (1978) tarafından önerilmiştir. $AP_{(i)}$ ile gösterilen bu istatistik için öncelikle tüm gözlemlerin bağımsız değişkenlerinin değerlerini ve bağımlı değişkeninin değerlerini içeren bir $Z = [X_{n \times p} : Y_{n \times 1}]_{n \times (p+1)}$ matrisi tanımlanır. $Z_{(i)}$ matrisi, i . gözlemin veri kümesinden çıkarıldığı durumda geri kalan gözlemlerin bağımlı ve bağımsız değişkenlerinin değerlerinin oluşturduğu matris olarak tanımlanırsa, bu matris kısaca Z matrisinden i . satırın çıkarılmış biçimi olarak ifade edilebilir. Bu durumda Andrews ve Pregibon'un $AP_{(i)}$ İstatistiği

$$AP_{(i)} = \frac{|Z_{(i)}^T Z_{(i)}|}{|Z^T Z|}, \quad i = 1, 2, \dots, n \quad (3.20)$$

ile verilir(Andrews ve Pregibon, 1978). Eş. 3.20'de eşitliğin sağ tarafındaki oranın payda kısmı $|Z^T Z| = e^T e |X^T X|$ olarak da ifade edilebilir. Hesaplama kolaylığı amacıyla $AP_{(i)}$ istatistiği için pratik bir formül bilinen artıkların ve kaldıraç değerinin kullanıldığı

$$AP_{(i)} = 1 - h_{ii} - \frac{e_i^2}{e^T e} = \left(1 - \frac{r_i^2}{n - p}\right)(1 - h_{ii}), \quad i = 1, 2, \dots, n \quad (3.21)$$

ile verilir(Andrews ve Pregibon, 1978; Meloun ve Militky, 2001). Verinin k hacimli bir alt kümesinin ($1 < k < n/2$) veri kümesinden çıkartılmasıyla etkili gözlem gruplarını saptamak amacıyla $AP_{(i)}$ istatistiğinin genişletilmiş hali de Andrews ve Pregibon (1978) tarafından önerilmiştir. Veri kümesinin k hacimli alt kümesi I ile gösterilmek üzere $AP_{(I)}$ İstatistiği

$$AP_{(I)} = \frac{|Z_{(I)}^T Z_{(I)}|}{|Z^T Z|}$$

ile verilir(Andrews ve Pregibon, 1978). Burada $Z_{(I)}$ matrisi, veri kümesinden I alt kümesi çıkarıldıktan sonra geri kalan gözlemlerin bağımlı ve bağımsız değişken değerlerinin oluşturduğu matristir.

$AP_{(i)}$ istatistiği için karar kuralı Meloun ve Militky (2001) tarafından verilmiştir. Bu karar kuralına göre $(1 - AP_{(i)}) > 2(p+1)/n$ ise i . gözlem etkili gözlemdir.

3.1.6. Cook ve Weisberg olabilirlik istatistiği LD

Cook ve Weisberg (1982) tarafından önerilen diğer bir tanı istatistiği β parametre vektörünün ve σ^2 parametresinin En Çok Olabilirlik (EÇOB) Tahminlerine dayalı olan Cook ve Weisberg'in Olabilirlik İstatistiği'dir. Bu istatistik, log-olabilirlik fonksiyonunun tüm gözlemler için elde edilen maksimum değerinden, log-olabilirlik fonksiyonunun i . gözlem veri kümesinden çıkartıldıktan sonra geri kalan gözlemlerden elde edilen maksimum değeri arasındaki farka dayalıdır.

Bir θ parametre vektörü için log-olabilirlik fonksiyonu $L(\theta)$ ile tanımlansın. Bu durumda θ parametre vektörünün EÇOB Tahmini $\tilde{\theta}$ vektörü olmak üzere, bu tahmin ile hesaplanan log-olabilirlik fonksiyonu $L(\tilde{\theta})$ ile gösterilsin. Gözlem silme yaklaşımıyla veri kümesinden i . gözlem çıkarıldığında geri kalan gözlemlerden elde edilen θ parametre vektörü için EÇOB Tahmini $\tilde{\theta}_{(i)}$ vektörü olmak üzere bu tahmin ile hesaplanan log-olabilirlik fonksiyonu ise $L(\tilde{\theta}_{(i)})$ ile gösterilsin. Bu durumda Cook ve Weisberg Olabilirlik İstatistiği $LD_{(i)}$ ile gösterilmek üzere bu istatistik

$$LD_{(i)} = 2[L(\tilde{\theta}) - L(\tilde{\theta}_{(i)})] \quad (3.22)$$

ile verilir(Cook ve Weisberg, 1982: 183). Burada θ parametre vektörü, β parametre vektörü olarak seçilirse bu durumda β parametre vektörünün EÇOB tahmi edicisi olan $\tilde{\theta} = \tilde{\beta}$ vektörü söz konusudur. θ parametre vektörü skaler bir değer olan σ^2 parametresi olarak seçilirse bu durumda σ^2 parametresinin EÇOB tahmi edicisi olan $\tilde{\theta} = \tilde{\sigma}^2$ söz konusu olur. Eğer θ parametre vektörü $\theta = [\theta_1 : \theta_2]$ olmak üzere $\theta_1 = \beta$ ve $\theta_2 = \sigma^2$ olarak seçilirse bu durumda her iki parametrenin de EÇOB tahmi edicisi söz konusudur. Buradaki temel fikir gözlem silme yaklaşımıyla regresyon parametrelerinin EÇOB tahminlerindeki dikkate değer farklılaşmaya neden olan gözlemi log-olabilirlik fonksiyonu ile saptamaktır. Bu gözlem etkili gözlem olarak yorumlanır.

σ^2 parametresinin EÇOB tahmin edicisi $\tilde{\sigma}^2$ için gözlem silme yaklaşımı uygulanmayıp, β parametre vektörünün EÇOB tahmin edicisi için gözlem silme yaklaşımı uygulandığında, etkili gözlemlerin saptanması amacıyla $\tilde{\theta} = \tilde{\beta}$ ve $\tilde{\theta}_{(i)} = \tilde{\beta}_{(i)}$ alınarak Eş. 3.22'de yerine koyulursa $LD_i(\tilde{\beta})$ istatistiği

$$LD_{(i)}(\tilde{\beta}) = n \times \ln \left[\frac{d_i \times h_{ii}}{1 - h_{ii}} + 1 \right] \quad (3.23)$$

ile verilir(Meloun ve Militky, 2001). Burada $d_i = r_i / (n - p)$ ile elde edilir.

β parametre vektörünün EÇOB tahmin edicisi olan $\tilde{\beta}$ için gözlem silme yaklaşımı uygulanmadığında, σ^2 parametresinin EÇOB tahmin edicisi için gözlem silme yaklaşımı kullanılarak etkili gözlemi saptamak amacıyla $\tilde{\theta} = \tilde{\sigma}^2$ ve $\tilde{\theta}_{(i)} = \tilde{\sigma}_{(i)}^2$ alınarak Eş. 3.22'de yerine koyulursa $LD_{(i)}(\tilde{\sigma}^2)$ istatistiği

$$LD_{(i)}(\tilde{\sigma}^2) = n \times \ln \left[\frac{n}{n-1} \right] + n \times \ln(1 - d_i) + \frac{d_i(n-1)}{1 - d_i} - 1 \quad (3.24)$$

olarak verilir(Meloun ve Militky, 2001).

$\tilde{\theta} = [\tilde{\theta}_1 : \tilde{\theta}_2]$ olmak üzere $\tilde{\theta}_1 = \tilde{\beta}$ ve $\tilde{\theta}_2 = \tilde{\sigma}^2$ alınarak bu iki tahmin edici için gözlem silme yaklaşımı uygulanırsa Eş. 3.22’de yerine koyulursa $LD_{(i)}(\tilde{\beta}, \tilde{\sigma}^2)$ istatistiği

$$LD_i(\tilde{\beta}, \tilde{\sigma}^2) = n \times \ln \left[\frac{n}{n-1} \right] + n \times \ln(1-d_i) + \frac{d_i(n-1)}{(1-d_i)(1-h_{ii})} - 1 \quad (3.25)$$

ile verilir (Meloun ve Militky, 2001).

Eşitlik (3.25)’te verilen Cook ve Weisberg Olabilirlik İstatistiği olan $LD_{(i)}$ istatistiğinin dağılımı v serbestlik dereceli ki-kare dağılımına yakınsar. Buradaki v serbestlik derecesi θ parametre vektörünün eleman sayısıdır. Bu durumda eğer i . gözlem için Cook ve Weisberg Olabilirlik İstatistiği $LD_{(i)} > \chi_{v,\alpha}^2$ ise bu gözlem etkili gözlemdir.

3.1.7. Hadi’nin K istatistiği

Hadi (1992) tarafından önerilen $K_{(i)}$ tanı istatistiği, x -uzayında, y -uzayında veya xy -uzayında potansiyel olarak etkili olan gözlemlerin aykırı değer olup olmadığını saptamaya dayalı bir istatistiktir.

Hadi (1992) tarafından önerilen istatistikle regresyon tahminlerinde bozulmaya neden olan gözlemleri saptamak amacıyla, hem veri kümesindeki gözlemler birer birer incelenebilir, hem de veri kümesindeki k adet gözlemin oluşturduğu ($1 < k < n/2$) ve I ile gösterilen gözlem gruplarının kombinasyonları incelenebilir. Buradan hareketle Hadi (1992) gözlemlerin birer birer incelendiği $K_{(i)}$ istatistiğinden önce gözlemlerin gruplar halinde incelendiği ve $K_{(i)}$ istatistiğinin I gözlem grubu bakımından genelleştirilmiş hali olan $K_{(I)}$ istatistiğini önermiştir.

$K_{(I)}$ İstatistiğinin türetilmesi için tasarım matrisi X , bağımlı değişken değerlerinin vektörü Y ve artık değerlerin vektörü e

$$X = \begin{bmatrix} X_{(I)} \\ X_I^T \end{bmatrix} : \begin{matrix} (n-k) \times p \\ k \times p \end{matrix} \quad (3.26)$$

$$Y = \begin{bmatrix} Y_{(I)} \\ Y_I \end{bmatrix} \quad \begin{array}{l} : (n-k) \times 1 \\ : k \times 1 \end{array} \quad (3.27)$$

$$e = \begin{bmatrix} e_{(I)} \\ e_I \end{bmatrix} \quad \begin{array}{l} : (n-k) \times 1 \\ : k \times 1 \end{array} \quad (3.28)$$

biçiminde parçalansın. Burada X_I matrisi, I gözlem kümesini oluşturan gözlemlerin tasarım matrisi bilgisini içeren $p \times k$ boyutlu bir matris; Y_I vektörü I gözlem kümesini oluşturan gözlemlerin bağımlı değişken değerlerinden oluşturulmuş $k \times 1$ boyutlu bir vektör; e_I vektörü I gözlem kümesini oluşturan gözlemlerin artık değerlerinden oluşturulmuş $k \times 1$ boyutlu bir vektördür. Hadi (1992) tarafından önerilen $K_{(I)}$ İstatistiğini elde etmek amacıyla ilk aşamada x -uzayına göre gözlem gruplarının aykırı gözlem grubu olup olmadığı dikkate alınır. Gözlemlerin x -uzayında aykırı değer olup olmadığı

$$\det[X_I^T (X_{(I)}^T X_{(I)})^{-1} X_I] \quad (3.29)$$

ile saptanabilir(Hadi, 1992). Burada X_I matrisi, X tasarım matrisinde I gözlem kümesindeki gözlemlere karşılık gelen satırların oluşturduğu matristir. $X_{(I)}$ matrisi ise X matrisinden k büyüklüğünde gözlem grubuna ilişkin satırlar çıkarıldıktan sonra geri kalan satırların oluşturduğu matristir. Eş. 3.29'da yer alan $X_{(I)}^T X_{(I)}$ çarpımı

$$X_{(I)}^T X_{(I)} = X^T X - X_I X_I^T \quad (3.30)$$

ile verilir. $X_{(I)}^T X_{(I)}$ çarpımının tersi

$$\begin{aligned} (X_{(I)}^T X_{(I)})^{-1} &= (X^T X - X_I X_I^T)^{-1} \\ &= (X^T X)^{-1} + (X^T X)^{-1} X_I [\mathbf{I}_k - X_I^T (X^T X)^{-1} X_I]^{-1} X_I^T (X^T X)^{-1} \end{aligned} \quad (3.31)$$

olarak elde edilir. Burada $\mathbf{I}_k$ matrisi $k \times k$ boyutlu birim matristir. $k \times k$ boyutlu H_I matrisi

$$H_I = X_I^T (X^T X)^{-1} X_I \quad (3.32)$$

ile tanımlansın. Bu matris, bir gözlem için hesaplanan kaldıraç değeri olan h_{ii} değerinin birden fazla gözlem için genelleştirilmiş biçimidir. Bu durumda Eş. 3.29'da determinantın içindeki kısım

$$\begin{aligned} H_{I(I)} &= X_I^T (X_{(I)}^T X_{(I)})^{-1} X_I \\ &= H_I + H_I (\mathbf{I}_k - H_I)^{-1} H_I \\ &= H_I [\mathbf{I}_k + (\mathbf{I}_k - H_I)^{-1} H_I] \\ &= H_I (\mathbf{I}_k - H_I)^{-1} \end{aligned} \quad (3.33)$$

olarak elde edilir.

k elemanı olan ve I ile gösterilen gözlem grubunun, veri kümesinde geri kalan $n - k$ adet gözleme göre bağımsız değişken değerleri bakımından, diğer bir tabirle x -uzayında, aykırı gözlem grubu olup olmadığını daha iyi belirlemek amacıyla Hadi (1992) önerdiği istatistikte Eş. 3.33'te elde edilen sonucun izinin daha kullanışlı olacağını belirtmiştir. Bu durumda $K_{(I)}$ istatistiğinde kullanılmak amacıyla

$$\text{iz}[X_I^T (X_{(I)}^T X_{(I)})^{-1} X_I] = \text{iz}[H_I (\mathbf{I}_k - H_I)^{-1}] \quad (3.34)$$

elde edilir (Hadi, 1992).

$K_{(I)}$ istatistiği için ikinci aşamada I gözlem grubunun y -uzayı bakımından aykırı bir gözlem grubu olup olmadığı dikkate alınır. Bunun için en uygun yol Student Türü Artıkların I gözlem grubu için genişletilmiş bir biçimini dikkate almaktır. I gözlem grubunun bilgisi kullanılmadan elde edilen β parametre vektörünün EKK tahmini $\hat{\beta}_{(I)}$ vektörü kullanılarak elde edilen I gözlem grubuna ilişkin $k \times 1$ boyutlu artık vektörü $e_{I(I)}$ ile gösterilsin. Bu vektör

$$e_{I(I)} = Y_I - X_I \hat{\beta}_{(I)} \quad (3.35)$$

ile verilir(Hadi, 1992). $e_{I(I)}$ artık vektörünün varyansı $(\mathbf{I}_k + H_{I(I)})$ ile orantılıdır. Ölçü biriminin etkisini arındırmak amacıyla $e_{I(I)}$ artık vektörü

$$e_{I(I)}^* = (\mathbf{I}_k + H_{I(I)})^{-1/2} (\mathbf{Y}_I - X_I^T \hat{\beta}_{(I)}) \quad (3.36)$$

biçiminde ele alınsın. I gözlem grubu veriden çıkartılınca, $\hat{\beta}_{(I)}$ tahmin vektöründe bir iyileşme söz konusu olduğundan, $e_{I(I)}^*$ artıkları mutlak değerce büyük olacaktır. Bundan dolayı Eş. 3.36'da verilen $e_{I(I)}^*$ vektörünün elemanlarının değerleri mutlak değerce büyük olursa bu durum ele alınan I gözlem grubunun geri kalan $n-k$ gözlemin geneline uymadığının göstergesidir. Bundan yararlanarak Hadi (1992)

$$\frac{e_{I(I)}^T (\mathbf{I}_k + H_{I(I)})^{-1/2} (\mathbf{I}_k + H_{I(I)})^{-1/2} e_{I(I)}}{e_{(I)}^T e_{(I)}} = \frac{e_{I(I)}^T (\mathbf{I}_k + H_{I(I)})^{-1} e_{I(I)}}{e^T e - e_I^T e_I} \quad (3.37)$$

oranını oluşturmuştur. Nitekim $e_{I(I)} = (\mathbf{I}_k + H_{I(I)})e_I$ ve $(\mathbf{I}_k + H_{I(I)}) = (\mathbf{I}_k - H_I)^{-1}$ olduğundan Eş. 3.37

$$\begin{aligned} \frac{e_I^T (\mathbf{I}_k + H_{I(I)}) (\mathbf{I}_k + H_{I(I)})^{-1} (\mathbf{I}_k + H_{I(I)}) e_I}{e^T e - e_I^T e_I} &= \frac{e_I^T (\mathbf{I}_k + H_{I(I)}) e_I}{e^T e - e_I^T e_I} \\ &= \frac{e_I^T (\mathbf{I}_k - H_I)^{-1} e_I}{e^T e - e_I^T e_I} \end{aligned} \quad (3.38)$$

biçiminde yeniden yazılır(Hadi, 1992).

Eş. 3.34 ve Eş. 3.38'de elde edilenlerden yararlanarak Hadi (1992)

$$K_I = \frac{p}{k} \frac{e_I^T (\mathbf{I}_k - H_I)^{-1} e_I}{e^T e - e_I^T e_I} + \frac{1}{k} \text{iz} [H_I (\mathbf{I}_k - H_I)^{-1}] \quad (3.39)$$

istatistiğini önermiştir. Eş. 3.39'da verilen $K_{(I)}$ istatistiğinin bir gözlem silme yaklaşımına göre sadeleştirilmiş biçimi $K_{(i)}$ ile gösterilir ve

$$K_{(i)} = \frac{p}{(1-h_{ii})} \frac{e_i^2}{(e^T e - e_i^2)} + \frac{h_{ii}}{(1-h_{ii})}, \quad i = 1, 2, \dots, n \quad (3.40)$$

verilir(Hadi, 1992). Eş. 3.39’da verilen $K_{(i)}$ istatistiği, Eş. 3.40’da verilen $K_{(i)}$ istatistiği’nin genelleştirilmiş bir biçimi olduğundan, bu iki istatistiğin yorumu benzerdir. Eş. 3.40’da verilen $K_{(i)}$ İstatistiği ele alınırsa eşitliğin sağ tarafındaki ilk oran i . gözlemin y -uzayı bakımından aykırı bir gözlem olup olmadığını ifade etmektedir. Eş. 3.40’ın sağ tarafındaki ikinci oran olan $h_{ii}/(1-h_{ii})$ oranı ise i . gözlemin x -uzayı bakımından aykırı bir gözlem olup olmadığını ifade eder. Bu durumda i . gözlemin $K_{(i)}$ değeri büyük ise, bu gözlem hem x -uzayı hem de y -uzayı bakımından aykırı olduğu için regresyon tahminlerini olumsuz bir biçimde değiştirme eğilimine sahiptir. Bu durumda bu gözlem potansiyel olarak etkili bir gözlemdir.

Hadi (1992), $K_{(i)}$ tanı istatistiği için $\bar{K} + \varsigma \sqrt{\hat{\sigma}_K^2}$ değerini kritik değer olarak önermiştir. Burada $\bar{K}$ terimi $K_{(i)}$ istatistiği için ortalama, $\hat{\sigma}_K^2$ ise bu istatistik için varyansa karşılık gelmektedir. ς sabiti ise güven katsayısıdır ve bu katsayı 2 veya 3 olarak alınabilir. Bu durumda bu kritik değerden büyük olan $K_{(i)}$ değerlerine karşılık gelen gözlem etkili gözlemdir. Ayrıca Hadi (1992) bu kritik değer için ortalamanın ve varyansın sağlam alternatiflerinin kullanılabileceğini de belirtmiştir.

$K_{(i)}$ tanı istatistiği, bilinen artıkların ve kaldıraç değerlerinin bir fonksiyonu olarak düşünüldüğünde diğer tanı istatistiklerine benzer. Bununla birlikte diğerlerinden ayırıcı özelliği $K_{(i)}$ istatistiğinin, artıkları ve kaldıraç değerlerini birleştirme yoludur. Diğer birçok tanı istatistiği artıkların ve kaldıraç değerlerinin çarpımsal bir fonksiyonuyken, $K_{(i)}$ istatistiği bu iki kavramın toplamsal bir fonksiyonudur.

3.1.8. Pena’nın S istatistiği

Gözlem silme yaklaşımına 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 veri kümesindeki diğer gözlemlerin yardımıyla belirlenir.

Eş. 2.1’de verilen model dikkate alındığında, tüm veriden elde edilen $\hat{\beta}$ tahminiyle hesaplanan i . gözlemin tahmin değeri $\hat{y}_i$ ile veri kümesinden her biri i' ile gösterilmek üzere bu gözlemlerin teker teker çıkartılmasıyla elde edilen $\hat{\beta}_{(i')}$ tahminiyle hesaplanan i . gözlemin tahmin değeri $\hat{y}_{i(i')}$ arasındaki farka ilişkin vektör

$$s_i = \left(\hat{y}_i - \hat{y}_{i(1)}, \dots, \hat{y}_i - \hat{y}_{i(n)} \right)^T \quad (3.41)$$

biçiminde tanımlansın (Pena, 2005). Bu i . gözlemin tahmin değerinin veri kümesindeki her bir gözlemin silinmesine karşı duyarlılığını göstermektedir. Böylece $S_{(i)}$ istatistiği, s_i vektörünün standartlaştırılmış karesel normu olacak biçimde

$$S_{(i)} = \frac{s_i^T s_i}{p \cdot \text{var}(\hat{y}_i)} \quad , \quad (i = 1, 2, \dots, n) \quad (3.42)$$

ile verilir (Pena, 2005). Burada $\hat{y}_i - \hat{y}_{i(i')} = h_{ii'} e_{i'} / (1 - h_{ii'})$ ve $\text{var}(\hat{y}_i) = \hat{\sigma}^2 h_{ii}$ alınırsa, Eş. 3.42’de tanımlanan istatistik

$$S_{(i)} = \frac{1}{p \hat{\sigma}^2 h_{ii}} \sum_{i'=1}^n \frac{h_{ii'}^2 e_{i'}^2}{(1 - h_{ii'})^2} \quad , \quad (i, i' = 1, 2, \dots, n) \quad (3.43)$$

biçiminde yeniden yazılır (Pena, 2005). $S_{(i)}$ istatistiği için alternatif bir formül Cook Uzaklığı İstatistiği’nin de kullanıldığı

$$S_{(i)} = \sum_{i'=1}^n \rho_{ii'}^2 D_{(i)} \quad , \quad (i, i' = 1, 2, \dots, n) \quad (3.44)$$

formülü ile verilir (Pena, 2005). Burada $\rho_{ii'} = (h_{ii'}^2 / h_{ii} h_{ii'})^{1/2} \leq 1$ olur ve bu $\hat{y}_i$ ile $\hat{y}_{i'}$ arasındaki korelasyona karşılık gelir. $\hat{\beta}_{(i')}$ tahmin vektörü kullanılarak elde edilen i' . gözleme ilişkin artık değeri

$$\begin{aligned}
e_{i'(i')} &= y_{i'} - x_{i'}^T \hat{\beta}_{(i')} \\
&= \frac{e_{i'}}{1 - h_{i'i'}}
\end{aligned} \tag{3.45}$$

ile verilsin (Pena, 2005). Bu artık değer kullanılarak $S_{(i)}$ istatistiği

$$S_{(i)} = \frac{1}{p\hat{\sigma}^2} \sum_{i'=1}^n \varpi_{i'i} e_{i'(i')}^2, \quad (i, i' = 1, 2, \dots, n) \tag{3.46}$$

ile gözlem silme yaklaşımına dayalı olarak elde edilen tahminlere ilişkin artıkların ağırlıklı kombinasyonu biçiminde yeniden yazılabilir. Buradaki $\varpi_{i'i}$ terimi

$$\varpi_{i'i} = \frac{h_{ii'}^2}{ph_{ii}(1 - h_{i'i'})^2}, \quad (i, i' = 1, 2, \dots, n) \tag{3.47}$$

ile verilir (Pena, 2005).

Veri kümesinde 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 kaldıraç noktalarının olmadığı bir veri kümesinde gözlemlerin tümünün aynı duyarlılığa sahip olması beklenir. Bu beklenen değer kaldıraç noktalarına oldukça bağlı olan Cook uzaklığına göre önemli bir avantajdır (Pena, 2005).

Pena (2005) etkili gözlemlerin saptanması için önerdiği $S_{(i)}$ istatistiğine ilişkin

$$|S_{(i)} - \text{med}(S)| \geq 4,5 \text{MAD}(S) \tag{3.48}$$

karar kuralını önermiştir. Eğer bu eşitsizlik sağlanırsa i . gözlem etkili gözlemdir. 3.48 eşitsizliğinde $\text{med}(S)$ değeri $S_{(i)}$ değerlerinin medyanı, $\text{MAD}(S_i)$ ise $S_{(i)}$ değerlerinin medyandan sapmalarının mutlak değerlerinin medyanıdır.

3.2. Etkili Gözlemlerin Saptanması için Grafiksel Yöntemler

Regresyon analizinde kullanılan grafikler gerek model tahmin aşamasından önce, gerekse model tahmin aşamasından sonra oldukça kullanışlı araçlardır. Özellikle bağımlı değişken, bağımsız değişken ve artıkların ikili kombinasyonları için oluşturulan serpmeye diyagramları gibi iki boyutlu grafiklerin görsel bir bilgi sunması regresyon analizinde tahmin ve çıkarsama aşamasında karşılaşılabilecek ve araştırmacı tarafından fark edilemeyen bazı sorunların tespitinde kullanılabilir. Genel olarak grafiksel yöntemlerin kullanılma nedenleri:

- Veri kümesindeki hataları saptamak (örneğin aykırı gözlemler kayıt hatasının bir sonucu olabilir),
- Veri kümesinin biçimini (desenini) tanımak (veri kümesindeki gözlem gruplarını, aykırı gözlemleri, etkili gözlemleri görmek),
- Değişkenler arasındaki ilişkileri araştırmak,
- Olağanüstü durumları keşfetmek,
- Varsayımların sağlandığını veya reddedildiğini görmek,
- Oluşturulan modelin uygunluğunu değerlendirmek,
- Gerekiyorsa veri kümesini iyileştirici eylemleri önermek (gerekli dönüşümlerin yapılması, deneyin tekrarlanması, daha fazla veri toplanması gibi),
- Yapılan analizin sonuçlarını genelleştirmek,

olarak sıralanabilir (Chatterjee ve Hadi, 2006: 98). Bu amaçlar doğrultusunda regresyon analizinde grafiksel yöntemler iki başlık altında toplanır. Bunlardan birincisini, veri kümesindeki 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 hata terimlerine ilişkin 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.

Bu bölümde etkili gözlemlerin saptanması için kullanılan bazı grafikler tanıtılmıştır.

3.2.1. Eklenmiş değişken grafikleri (Added variable plots)

Mosteller ve Tukey (1977) tarafından önerilen Eklenmiş Değişken Grafiği çoklu regresyonda hem bağımsız değişkenlerin incelenmesi hem de etkili gözlemlerin saptanması amacıyla kullanılan bir grafik yöntemidir. Bu grafiğin kullanım amaçları

- Bağımsız değişkenlerden j . bağımsız değişkenin bağımlı değişkeni açıklama payının önemli olup olmadığını görsel olarak değerlendirmek,
- Her bir gözlemin $\hat{\beta}$ tahmin vektörünün her bir elemanı üzerindeki etkisini görsel olarak değerlendirmek,

olarak sıralanabilir. Birinci amaç doğrultusunda Eklenmiş Değişken Grafiği, j . bağımsız değişken bağımlı değişken gibi ele alınıp bu değişken ile diğer bağımsız değişkenler arasında oluşturulan regresyon modeliyle elde edilen artıkların, bağımlı değişken ile diğer bağımsız değişkenler arasında oluşturulan regresyon modeliyle elde edilen artıklarla karşılaştırılması biçiminde, bu iki artığa ilişkin serpmeye diyagramı oluşturularak elde edilir. Bunun dışında aynı işlemler modele katılmamış başka bir bağımsız değişken bakımından her bir gözleme ilişkin bir bilgi mevcutsa bu bağımsız değişken için de Eklenmiş Değişken Grafiği oluşturulabilir. Bu karşılaştırma da tahmin edilen iki regresyon modeli için elde edilen artık vektörlerinin serpmeye diyagramının oluşturulmasıyla yapılır. Eklenmiş Değişkenler Grafiği ile Kısmi Korelasyon Katsayısı (Partial Correlation Coefficient) için artıkların hesaplanması benzer şekilde yapılmaktadır. Aynı zamanda bu grafik ikinci amaç doğrultusunda her bir gözlemin j . bağımsız değişkene karşılık gelen $\hat{\beta}_j$ tahminini değiştirme eğilimine sahip olup olmadığının belirlenmesi için de kullanılır.

X tasarım matrisi $X = \begin{bmatrix} X^{(-)} & X^{(+)} \end{bmatrix}$ biçiminde parçalansın. Burada $X^{(+)}$ vektörü j .

bağımsız değişkenin değerlerinin vektörü, $n \times (p-1)$ boyutlu $X^{(-)}$ matrisi ise tasarım

matrisinde j . sütunun çıkartılmasıyla geri kalan sütunların oluşturduğu matristir. Bu durumda Eş. 2.1’de verilen regresyon modeli

$$Y = X^{(-)}\gamma_1 + X^{(+)}\gamma_2 + \varepsilon \quad (3.49)$$

biçiminde tekrar yazılabilir (Atkinson ve Riani, 2000: 22). Burada γ_1 parametresi $(p-1) \times 1$ boyutlu bir vektör ve γ_2 parametresi bir skalerdir. Bu parametrelerin bir vektör biçiminde yazılmasıyla tüm bağımsız değişkenleri içeren doğrusal regresyon modeli için β parametre vektörü elde edilir. Dolayısıyla Eş. 3.49’da β parametre vektörü $\beta^T = [\gamma_1^T : \gamma_2]$ şeklinde parçalanmış olarak ifade edilir. γ_1 parametre vektörü ve γ_2 parametresinin EKK tahminleri sırasıyla $\hat{\gamma}_1$ ve $\hat{\gamma}_2$ ile gösterilirse Eş. 3.49’da belirtilen model parametrelerinin tahmini için normal denklemler

$$X^{(-)T} X^{(-)} \hat{\gamma}_1 + X^{(-)T} X^{(+)} \hat{\gamma}_2 = X^{(-)T} Y \quad (3.50)$$

$$X^{(+T} X^{(-)} \hat{\gamma}_1 + X^{(+T} X^{(+)} \hat{\gamma}_2 = X^{(+T} Y \quad (3.51)$$

ile verilir (Atkinson ve Riani, 2000: 22). Eş. 3.50’de eşitliğin her iki tarafı $[X^{(-)T} X^{(-)}]^{-1}$ ile çarpılırsa $\hat{\gamma}_1$ tahmin vektörü

$$\hat{\gamma}_1 = [X^{(-)T} X^{(-)}]^{-1} X^{(-)T} Y - [X^{(-)T} X^{(-)}]^{-1} X^{(-)T} X^{(+)} \hat{\gamma}_2 \quad (3.52)$$

biçiminde elde edilir (Atkinson ve Riani, 2000: 22). Eş. 3.52’de elde edilen $\hat{\gamma}_1$ tahmin vektörü Eş. 3.51’de yerine konursa $\hat{\gamma}_2$ tahmini

$$\hat{\gamma}_2 = \frac{X^{(+T} [\mathbf{I} - H^{(-)}] Y}{X^{(+T} [\mathbf{I} - H^{(-)}] X^{(+)}} \quad (3.53)$$

olarak elde edilir (Atkinson ve Riani, 2000: 22). Burada $n \times n$ boyutlu

$H^{(-)} = X^{(-)} [X^{(-)T} X^{(-)}]^{-1} X^{(-)T}$ matrisi j . bağımsız değişkenin modelden çıkarıldığı

durumda geri kalan bağımsız değişkenlerin bilgisinden elde edilen hat matrisidir. Bu durumda bağımlı değişkenin tahmin vektörü

$$\begin{aligned}
 \hat{Y} &= X^{(-)}\hat{\gamma}_1 + X^{(+)}\hat{\gamma}_2 \\
 &= X^{(-)}\left\{\left[X^{(-)T}X^{(-)}\right]^{-1}X^{(-)T}Y - \left[X^{(-)T}X^{(-)}\right]^{-1}X^{(-)T}X^{(+)}\hat{\gamma}_2\right\} + X^{(+)}\hat{\gamma}_2 \\
 &= H^{(-)}Y - H^{(-)}X^{(+)}\hat{\gamma}_2 + X^{(+)}\hat{\gamma}_2 \\
 &= H^{(-)}Y + (\mathbf{I}_n - H^{(-)})X^{(+)}\hat{\gamma}_2
 \end{aligned} \tag{3.54}$$

ile verilmek üzere artık vektörü

$$\begin{aligned}
 e &= Y - \hat{Y} \\
 &= Y - \left[H^{(-)}Y + (\mathbf{I}_n - H^{(-)})X^{(+)}\hat{\gamma}_2 \right] \\
 &= (\mathbf{I}_n - H^{(-)})Y - (\mathbf{I}_n - H^{(-)})X^{(+)}\hat{\gamma}_2 \\
 &= e_{(Y|X^{(-)})} - e_{(X^{(+)}|X^{(-)})} \times \hat{\gamma}_2
 \end{aligned} \tag{3.55}$$

elde edilir. Eş. 3.55'te elde edilen sonuçta iki tip artık vektörü söz konusudur. Bu artık vektörleri

$$e_{(Y|X^{(-)})} = (\mathbf{I}_n - H^{(-)})Y \tag{3.56}$$

$$e_{(X^{(+)}|X^{(-)})} = (\mathbf{I}_n - H^{(-)})X^{(+)} \tag{3.57}$$

ile verilir (Atkinson ve Riani, 2000: 22). Bu artıklardan birincisi olan $e_{(Y|X^{(-)})}$ vektörü modelden j . bağımsız değişken çıkarıldığında, bağımlı değişken ile geri kalan bağımsız değişkenler arasındaki tahmin edilen regresyon modeliyle elde edilen artık vektörüdür. İkincisi olan $e_{(X^{(+)}|X^{(-)})}$ vektörü ise j . bağımsız değişken, bağımlı değişken gibi düşünülüp bu değişken ile diğer bağımsız değişkenler arasındaki tahmin edilen regresyon modeliyle elde edilen artık vektörüdür. İkinci artık vektörünü kullanmanın temel amacı j . bağımsız değişken üzerinden modelde yer alan diğer bağımsız değişkenlerin etkisini arındırmaktır. Bu iki artık vektörü için hesaplanan Pearson korelasyon katsayısı da kısmi korelasyon katsayısı olarak bilinir. Bu iki artık vektörünün oluşturduğu serpm diyagramı ise j . bağımsız değişken için oluşturulan Eklenmiş Değişken Grafiği olarak adlandırılır.

Eklenmiş Değişken Grafiği'nde yer alan noktalardan tüm noktaların geneline uymayan noktalara karşılık gelen gözlemler, regresyon tahminlerini değiştirme eğilimine sahip gözlemler olarak nitelendirilir.

3.2.2. Kısmi artık grafikleri (Partial residual plots)

Eklenmiş Değişken Grafiği'nde olduğu gibi, Ezekiel ve Fox (1959) tarafından önerilen Kısmi Artık Grafiği'nde de artıkların kullanıldığı diğer basit grafiklere göre daha çok hesaplama söz konusudur. Eş. 3.55 tekrar ele alınırsa

$$\begin{aligned}
 e &= Y - \hat{Y} \\
 &= Y - X^{(-)}\hat{\gamma}_1 - X^{(+)}\hat{\gamma}_2 \\
 &= Y - \left\{ X^{(-)} \left[X^{(-)T} X^{(-)} \right]^{-1} X^{(-)T} Y - X^{(-)} \left[X^{(-)T} X^{(-)} \right]^{-1} X^{(-)T} X^{(+)} \hat{\gamma}_2 \right\} - X^{(+)}\hat{\gamma}_2 \\
 &= Y - \left[H^{(-)} Y - H^{(-)} X^{(+)} \hat{\gamma}_2 \right] - X^{(+)}\hat{\gamma}_2 \\
 &= \left[(\mathbf{I}_n - H^{(-)}) Y + H^{(-)} X^{(+)} \hat{\gamma}_2 \right] - X^{(+)}\hat{\gamma}_2 \\
 &= \left[e_{(Y|X^{(-)})} + H^{(-)} X^{(+)} \hat{\gamma}_2 \right] - X^{(+)}\hat{\gamma}_2
 \end{aligned} \tag{3.58}$$

elde edilir (Atkinson, 2004: 64). Bu durumda kısmi artık vektörü e_K ile gösterilmek üzere Eş. 3.58'den yararlanarak bu vektör

$$\begin{aligned}
 e_K &= e + X^{(+)}\hat{\gamma}_2 \\
 &= \left[e_{(Y|X^{(-)})} + H^{(-)} X^{(+)} \hat{\gamma}_2 \right]
 \end{aligned} \tag{3.59}$$

olarak elde edilir (Atkinson, 2004: 64). Eş. 3.59'da elde edilen kısmi artık değişkeni dikey eksen olmak üzere, bu değişken ile modelden çıkarılan bağımsız değişken için oluşturulan serpmeye diyagramı ilgili bağımsız değişken için Kısmi Artık Grafiği olarak tanımlanır (Atkinson, 2004: 64).

Kısmi Artık Grafiği'nin yatay eksenini X_j bağımsız değişkeni olduğundan, bu grafik genellikle modelin doğrusallığı ile ilgili bilgi de verir. Böylece gerekiyorsa uygun dönüşümler veri kümesine uygulanabilir. Bununla birlikte etkili gözlemlerin bu grafiklerle saptanması kolay değildir. Hem Eklenmiş Değişken Grafiği, hem de Kısmi Artık Grafiği ile veri kümesini oluşturan gözlemler incelenirken, bu gözlemlerden regresyon

tahminlerini deęiřtirme eęilimi olanları saptamak amacıyla her bir baęımsız deęiřken iin oluřturulan tm grafiklerin birlikte deęerlendirilmesi gerekir. Tasarım matrisinin stunları birbirine dik iken bu iki grafik yntemi benzer sonular verir(Chatterjee ve Hadi, 1986).

4. ÖRNEK ETKİ EĞRİSİNE DAYALI BİR GRAFİKSEL YÖNTEM ÖNERİSİ

Yığının herhangi bir parametresinin tahmini için, bu yığından rasgele çekilen örnekteki bilginin kullanıldığı, etki eğrisinin örnek uyarlamaları örnekteki her bir gözlem değerinin ilgili istatistiği değiştirme eğilimine ilişkin değerleri vermektedir. Bundan dolayı etki eğrisinin örnek uyarlamaları regresyonda etkili gözlemlerin saptanması için kullanılan tanı yöntemlerine iyi bir temel oluşturmaktadır. Deneysel etki eğrisi (*EIC*), örnek etki eğrisi (*SIC*) ve bir gözlem silme yaklaşımına dayalı deneysel etki eğrisi (*EIC_(i)*) etki eğrisinin örnek uyarlamalarından bazılarıdır. $p \times 1$ boyutlu vektörel bir tahmin edici ele alınırsa, etki eğrisinin örnek uyarlamaları örnekteki her bir gözlem için ayı ayrı hesaplanan $p \times 1$ boyutlu vektörleri verir. Regresyon analizinde bu vektörler bunlara karşılık gelen gözlemlerin, β parametre vektörünün EKK tahminini değiştirme eğilimlerine ilişkin değerleri vermektedir (Cook ve Weisberg, 1982: 108).

Etki eğrisinin örnek uyarlamaları kullanılarak her bir gözlem için hesaplanan regresyon tahminlerini değiştirme değerlerine ilişkin vektörlerinin normlarının en büyüğü, regresyon tahminlerini değiştirme eğilimi en yüksek olan gözleme karşılık gelir. Bu bilgiden yararlanarak Li ve diğerleri (2001) Eş. 2.29'da verilen *EIC* fonksiyonunun normunun en büyüklenmesine dayalı olarak bir grafiksel tanı yöntemi önermişlerdir. Bununla birlikte Cook Uzaklığı İstatistiği'nin de temelini oluşturan *SIC* fonksiyonu etkili gözlemlerin saptanmasında *EIC*'ye göre daha avantajlıdır (Cook ve Weisberg, 1982: 108). Özellikle her bir gözlem için *SIC* fonksiyonuyla elde edilen regresyon tahminlerini değiştirme eğilimlerine ilişkin değerlerin vektörlerinin normları ile her bir gözlem için *EIC* fonksiyonuyla elde edilen normlar karşılaştırıldığında *SIC* ile etkili gözlemler *EIC*'ye göre daha açık bir şekilde belirlenmektedir.

Bu çalışmada önerilen grafik yöntemi, *SIC* fonksiyonuna dayalıdır. Bu grafik yöntemi, veri kümesinde yer alan her bir gözlemin çoklu doğrusal regresyon modeli parametrelerinin tahminlerini değiştirme eğilimlerine ilişkin katsayıların *SIC* fonksiyonu yardımıyla elde edilmesi, bu katsayılara ilişkin kritik değerlerin (alt ve üst sınırların) hesaplanması ve bu bulgularla iki boyutlu bir grafiğin oluşturulması biçiminde üç aşamadan oluşmaktadır.

SIC fonksiyonuyla elde edilen vektörlerin normlarına ilişkin en büyükleme problemi

$$Max_i \|SIC(z_i)\| \quad (4.1)$$

olarak tanımlansın. Burada $\|\bullet\|$ fonksiyonu, vektör normunu ifade etmektedir. Ω uzayı

$$\Omega = \left\{ z \mid z = Z^T w = \sum_{i=1}^n w_i z_i, \quad w = [w_1, w_2, \dots, w_n]^T \right\} \quad (4.2)$$

biçiminde tanımlansın. Z^T matrisinin sütun vektörlerinin gerdiği bir uzay olan Ω uzayında $SIC(z_i)$ örnek etki eğrisi düşünölsün. Ω uzayının tanımından yararlanarak z doğrusal bileşimi w değışken vektörünün bir fonksiyonu biçiminde $z(w)$ ile gösterilebilir. Bu durumda $z(w)$ için,

$$z(w) = z = \begin{bmatrix} X^T w \\ Y^T w \end{bmatrix} \quad (4.3)$$

olarak yazılabilir. Burada $X^T w = x_0$ ve $Y^T w = y_0$ olarak tanımlansın. Dolayısıyla Eş. 4.3'te verilen $z(w)$ veri kümesindeki tüm noktaların ağırlıklı bir bileşimiyle elde edilen bir gözlem gibi düşünölebilir. Bu durumda Eş. 2.33'te verilen örnek etki eğrisi, w vektör değışkeninin bir fonksiyonu olarak,

$$\begin{aligned} SIC[z(w)] &= (n-1)(X^T X)^{-1} X^T w (w^T Y - w^T X \hat{\beta}) / (1 - w^T X (X^T X)^{-1} X^T w) \\ &= (n-1)(X^T X)^{-1} X^T w w^T e / (1 - w^T H w) \end{aligned} \quad (4.4)$$

şeklinde yazılabilir. Birim vektör elde etmek amacıyla, $\|v\|$ olacak şekilde normlaştırma yapılırsa, $v = w / (w^T w)^{1/2}$ olmak üzere yeni bir vektör tanımlansın. Bu durumda v vektörü, w vektörünün yerini alınmak üzere Eş. 4.4'te yer alan $SIC[z(w)]$

$$SIC[z(v)] = \frac{(n-1)(X^T X)^{-1} X^T v v^T e}{1 - v^T H v} \quad (4.5)$$

gibi yazılabilir. Eş. 4.5'te v yerine buna eşit olan $w/(w^T w)^{1/2}$ yazılırsa

$$\begin{aligned}
 SIC[z(w)] &= \frac{(n-1)(X^T X)^{-1} X^T \left\{ \frac{1}{(w^T w)^{1/2}} w \right\} \left\{ \frac{1}{(w^T w)^{1/2}} w^T \right\} e}{\left\{ 1 - \frac{1}{(w^T w)^{2/2}} w^T H w \right\}} \\
 &= (n-1) \frac{1}{(w^T w)} (X^T X)^{-1} X^T w w^T e \frac{(w^T w)}{w^T w - w^T H w} \\
 &= (n-1) \frac{1}{w^T w - w^T H w} (X^T X)^{-1} X^T w w^T e
 \end{aligned} \tag{4.6}$$

olarak elde edilir. Bu çalışmada önerilen grafik yöntemi için n adet w_i değerini elde etmek amacıyla $SIC[z(w)]$ fonksiyonunun normu bir amaç fonksiyonu olarak tanımlanmıştır.

Bu durumda Eş. 4.1'de verilen en büyükleme problemi,

$$Max_{\|w\|=1} \|SIC[z(w)]\| \tag{4.7}$$

biçiminde yazılsın. Eş. 4.7' de verilen en büyükleme probleminin en büyük değerini veren $n \times 1$ boyutlu w^* vektörünün elemanları, her bir gözlemin β parametre vektörünün EKK tahmini $\hat{\beta}$ vektörünü en çok değiştirme eğilimlerine ilişkin katsayılarıdır.

Eş. 4.6'daki sonuçtan yararlanarak, Eş. 4.7'de verilen en büyükleme problemi tekrar yazılırsa,

$$Max_{w \neq 0} \left\{ \frac{(w^T e)^2 (w^T U w)}{(w^T w - w^T H w)^2} \right\}^{1/2} \tag{4.8}$$

elde edilir. Nitekim Eş. 4.8 bir amaç fonksiyonu olduğu için ve bu amaç fonksiyonunun en büyük değerini veren w^* çözümü ile ilgilenildiği için üslü kısım ihmal edilebilir. Bu durumda Eş. 4.8 tekrar yazılırsa,

$$\underset{w \neq 0}{Max} \frac{(w^T e)^2 (w^T U w)}{(w^T w - w^T H w)^2} \quad (4.9)$$

elde edilir. Burada $H = X (X^T X)^{-1} X^T$ ve $U = X (X^T X)^{-2} X^T$ ile verilir.

w^* vektörü gözlemlerin regresyon modeli parametrelerinin EKK tahminlerini en çok değiştirme eğilimlerine ilişkin katsayılar oldukları için bu katsayıların diyagramı incelenerek etkili gözlemler belirlenebilir. Eş. 4.9' da verilen en büyükleme problemi

$$\underset{w \neq 0}{max} \frac{\left(\sum_{i=1}^n w_i e_i \right)^2 \left(\sum_{i=1}^n \sum_{i'=1}^n w_i w_{i'} u_{ii'} \right)}{\left(\sum_{i=1}^n w_i^2 - \sum_{i=1}^n \sum_{i'=1}^n w_i w_{i'} h_{ii'} \right)^2} \quad (4.10)$$

biçiminde tekrar yazılabilir. Burada $u_{ii'}$ değeri U matrisinin i . satır i' . sütun elemanıdır. $h_{ii'}$ ise H matrisinin i . satır i' . sütun elemanı olarak verilir. Nitekim Eş. 4.9 ve Eş. 4.10'da açıkça görüldüğü üzere bu en büyükleme probleminde amaç fonksiyonu doğrusal değildir. Dolayısıyla en iyi çözümünü veren w^* çözüm vektörünü bulmak amacıyla sezgisel yöntemlere başvurulabilir.

SIC fonksiyonuna dayalı yöntemle etkili gözlemleri belirlemek için bir karar kuralı gereklidir. Bu nedenle bu karar kuralı için de bir kritik nokta veya kesme değeri (cutoff value) gerekir. Bu çalışmada önerilen yöntemle bir gözlemin etkili gözlem olup olmadığını belirlemek için Hadi (1992) tarafından $K_{(i)}$ tanı istatistiği için önerilen $\bar{K} + \varsigma \sqrt{\hat{\sigma}_K^2}$ kritik değer bulma yöntemi bir karar kuralı oluşturmak için yol gösterici olmuştur. $K_{(i)}$ istatistiği $[0, +\infty)$ aralığında değer almaktadır. Fakat her bir w_i , $(-1, +1)$ aralığında değer aldığı için w_i değişkeni için alt ve üst sınırlar söz konusu olur. Dolayısıyla w_i için kritik değerler alt ve üst sınır olarak ifade edilirse bunlar sırasıyla

$$\bar{w} - \varsigma \sqrt{\hat{\sigma}_w^2} \quad (4.11)$$

ve

$$\bar{w} + \varsigma \sqrt{\hat{\sigma}_w^2} \quad (4.12)$$

biçiminde tanımlanır. Buradaki ς sabiti güven katsayısıdır ve bu katsayı 2 veya 3 olarak alınabilir. Bu durumda bu kritik değerden büyük olan w_i değerlerine karşılık gelen gözlem etkili gözlemdir. Bununla birlikte Hadi (1992) ortalama veya varyansın sağlam tahminlerinin bu karar kuralı için kullanılabileceğini de belirtmiştir. Aynı yaklaşımı Pena (2005), S istatistiği için de kullanmıştır. Bu durumda alt ve üst sınırlar sırasıyla

$$\text{medyan}(w_i) - \varsigma \tilde{\sigma}_w \quad (4.13)$$

ve

$$\text{medyan}(w_i) + \varsigma \tilde{\sigma}_w \quad (4.14)$$

ile verilsin.

Literatürde standart sapmanın sağlam tahmin edicisi olan $\tilde{\sigma}$ için birçok tahmin edici önerilmiştir. Bu çalışmada w_i değişkeninin dağılış parametresinin sağlam tahmin edicisi $\tilde{\sigma}_w$ için Huber ve Ronchetti (2009) tarafından verilen ve medyandan mutlak sapmalara dayalı olan bir tahmin edici olan

$$\tilde{\sigma}_w = 1,4826 \text{MAD}(w) \quad (4.15)$$

tahmin edicisi kullanılmıştır. Burada MAD fonksiyonu medyandan mutlak sapmaları (median absolute deviation) ifade eder ve

$$\text{MAD}(w) = \text{medyan}_i \left\{ \left| w_i - \text{medyan}_j(w_j) \right| \right\} \quad (4.16)$$

olarak verilir (Huber ve Ronchetti, 2009: 106).

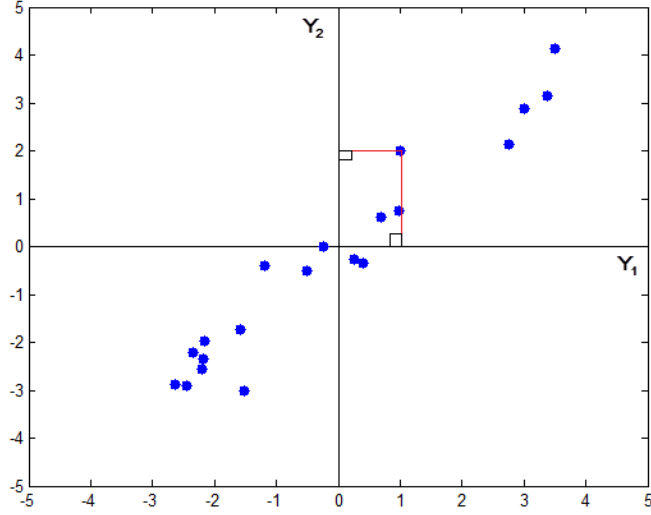
5. BİPLOT YAKLAŞIMINA DAYALI BİR GRAFİKSEL YÖNTEM ÖNERİSİ

Bazı kaynaklarda dikdörtgensel veri kümesi (rectangular data set) olarak ifade edilen, satırları birimlere ve sütunları değişkenlere karşılık gelen $n \times p$ boyutlu bir veri matrisi ele alınarak oluşturulan biplot yöntemi, çok değişkenli bir veri kümesini detaylı bir biçimde betimlemesinin yanı sıra, değişkenler arasındaki doğrusal ilişkileri belirleme ve gözlemleri sınıflandırmada oldukça kullanışlı bir yöntemdir (Greenacre, 2010: 9-14; Gower, Lubbe ve Roux, 2011: 1-7).

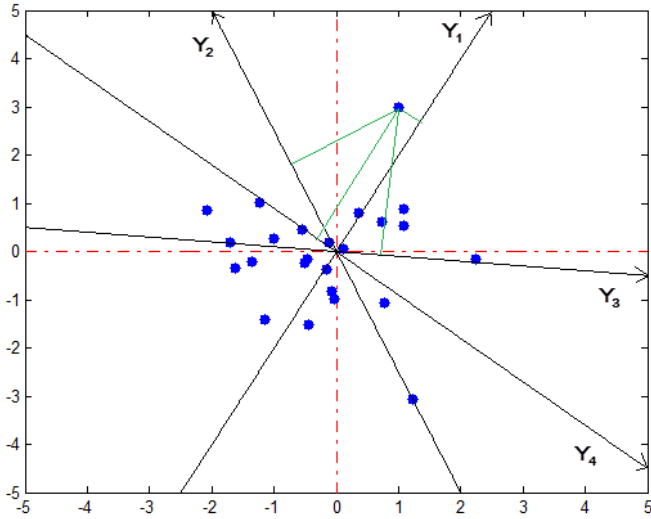
Tekil değer ayrıştırmasına dayalı olan biplot yöntemi ilk olarak çok değişkenli bir veri kümesi için serpmeye diyagramı oluşturmak amacıyla Gabriel (1971) tarafından önerilmiştir. Sonra Gabriel'in biplot yaklaşımı birçok araştırmacı tarafından diğer çok değişkenli analizlere uyarlanmıştır. Biplot yönteminin kullanıldığı bazı çalışmalar iki yönlü çapraz tabloların analizinde Gabriel (1995) tarafından, uygunluk analizinde Greenacre (1984) tarafından ve ayırma analizinde Gardner (2001) tarafından yapılmıştır. Gabriel'in biplot yaklaşımını Gower ve Hand (1996) geliştirmiştir. Ayrıca Caussinus, Fekri, Hakam ve Gazen (2003) temel bileşenler analizinde aykırı gözlemlerin biplot ile saptanması için kullanılan bir yöntem önermişlerdir. Biplot ifadesindeki “bi” öneki birimlerin ve değişkenlerin bir grafikte gösterildiğini ifade etmektedir.

Şekil 5.1 ve Şekil 5.2’de sırasıyla iki boyutlu serpmeye diyagramı ve biplota birer örnek durum verilmiştir. Şekil 5.1’de verilen iki boyutlu serpmeye diyagramından bir noktanın bir eksene iz düşümü o noktanın ilgili eksene karşılık gelen değişkendeki değerini vermektedir. Benzer durum biplot için de geçerlidir. Şekil 5.2’de verilen biplotta bir noktanın herhangi bir değişkene karşılık gelen eksene (veya vektöre) iz düşümü o noktanın ilgili değişkendeki değerine yaklaşık bir değer vermektedir. Böylece biplotun bir avantajının çok boyutlu bir sistemde bulunan gözlem noktalarının iki boyutlu bir uzayda gösterilmesi olduğu açıkça görülmektedir. Şekil 5.1’de verilen iki boyutlu serpmeye diyagramında eksenlerin doğrultusu iki değişken arasındaki ilişkiyi vermez. Bu ilişkiyi görsel olarak yorumlamak için gözlemlere karşılık gelen noktaların yayılımı incelenir. Ancak Şekil 5.2’de verilen biplotta herhangi iki değişkene karşılık gelen vektörler arasındaki açı bu iki değişken arasındaki ilişkiyi değerlendirmek için bir bilgi sağlar. Hatta bu açının kosinüsü yaklaşık olarak bu iki değişkene ilişkin Pearson ilişki katsayısına eşittir.

Bu durumda biplotun iki boyutlu serpmeye diyagramına göre ikinci avantajının da değişkenler arasındaki ilişkinin incelenebilmesi olduğu söylenebilir.



Şekil 5.1. İki boyutlu serpmeye diyagramı



Şekil 5.2. Biplot gösterimi

Bu bölümde biplot yöntemine ilişkin temel kavramlar verilerek, çoklu doğrusal regresyonda biplot yöntemiyle etkili gözlemlerin saptanması için bir grafik yöntemi önerilmiştir.

5.1. Biplot Yöntemine İlişkin Temel Kavramlar

Herhangi $n \times p$ boyutlu bir veri matrisi ele alınsın. Bu matris aynı zamanda dikdörtgensel veri matrisi olarak tanımlanıp genellikle satırları gözlemlere, sütunları da değişkenlere karşılık gelmektedir. Noktaların ilgili değişkenlere göre serpm diyagramında yayılımını incelemek amacıyla çizilen biplot için ilk aşamada aşağıdaki gibi bir eşitlik tanımlanır.

$$\text{hedef matris} = \text{sol matris} \times \text{sağ matris} \quad (5.1)$$

Burada *hedef matrisi* $n \times p$ boyutlu bir matris ($n \geq p$), *sol matris* $n \times m$ boyutlu ve satırları m boyutlu bir uzayda her bir noktanın koordinatlarını veren bir matris, *sağ matris* ise $m \times p$ boyutlu ve sütun vektörleri ilgili değişkenler için m boyutlu uzayda eksenleri veren bir matristir. Bu durumda *hedef matris* Λ , *sol matris* $\mathbf{L}$ ve *sağ matris* $\mathbf{R}$ olmak üzere Eş. 5.1 tekrar yazılırsa

$$\begin{aligned} \Lambda &= \mathbf{L} \mathbf{R}^T \\ &= \begin{bmatrix} L_1^T \\ L_2^T \\ \vdots \\ L_n^T \end{bmatrix} \begin{bmatrix} R_1 & R_2 & \cdots & R_p \end{bmatrix} \\ \Lambda &= \begin{bmatrix} L_1^T R_1 & L_1^T R_2 & \cdots & L_1^T R_p \\ L_2^T R_1 & L_2^T R_2 & \cdots & L_2^T R_p \\ \vdots & \vdots & \ddots & \vdots \\ L_n^T R_1 & L_n^T R_2 & \cdots & L_n^T R_p \end{bmatrix} \end{aligned} \quad (5.2)$$

ile gösterilir (Greenacre, 2010: 19). Burada $i = 1, 2, \dots, n$ ve $j = 1, 2, \dots, p$ olmak üzere, $L_i = [L_{i1}, L_{i2}, \dots, L_{im}]^T$ ve $R_j = [R_{1j}, R_{2j}, \dots, R_{mj}]^T$ ile verilir. Ayrıca yukarıdaki eşitliğin başında verilen $\mathbf{L}$ ve $\mathbf{R}$ matrisleri için $\text{rank}(\mathbf{L}) = m$ ve $\text{rank}(\mathbf{R}) = m$ olarak verilir. Bu durumda bir serpm diyagramı oluşturulmak istenirse m boyutlu bir Öklid uzayı söz konusu olacaktır. Bununla birlikte biplot için 2 boyutlu bir uzay söz konusu olduğundan biplot oluşturmak amacı ile boyut indirgeme yapılır.

Cebirsel olarak boyut indirgeme kavramı ilgili matrisin rankıyla ilgilidir. Dolayısıyla *boyutu-indirgenmiş biplot*, bazı kaynaklarda *rankı-indirgenmiş biplot* olarak da bilinir. Eğer ilgilenilen temel matris olan Λ matrisi $n \times n$ boyutlu simetrik bir matris olsaydı, boyut indirgeme amacıyla tayf ayrıştırması (spectral decomposition) kullanılabilirdi. Fakat uygulamada veri matrisi genellikle $n \times p$ boyutlu veya simetrik olmayan bir matristir. Bu amaçla biplot tekniği prosedüründe, tayf ayrıştırmasının alternatifi olan Tekil Değer Ayrıştırması (TDA) boyut indirgeme için kullanılmaktadır. TDA, $n \times p$ boyutlu veya simetrik olmayan bir matrisin en iyi indirgenmiş-rank yaklaşımının çözümünü vermesinin yanı sıra, ilgili biplotta yer alan gözlem noktaların koordinatlarını ve değişken vektörlerini de vermektedir (Greenacre, 2010: 53). Λ matrisinin tekil değer ayrıştırması,

$$\Lambda = U^* \Sigma^* (V^*)^T \quad (5.3)$$

ile verilir. Burada U^* sütunları $n \times n$ boyutlu $\Lambda \Lambda^T$ matrisinin normlaştırılmış öz vektörleri (sol tekil vektörler) olan $n \times n$ boyutlu bir dik matris, V^* sütunları $p \times p$ boyutlu $\Lambda^T \Lambda$ matrisinin normlaştırılmış öz vektörleri (sağ tekil vektörler) olan $p \times p$ boyutlu bir dik matristir (Golub ve Van Loan, 1996: 70). Λ matrisi tam ranklı bir matris iken $n \times p$ boyutlu Σ^* matrisi ise

$$\Sigma^*_{n \times p} = \begin{bmatrix} \Sigma_{p \times p} \\ \mathbf{O}_{(n-p) \times p} \end{bmatrix} \quad (5.4)$$

ile verilir (Golub ve Van Loan, 1996: 70). Burada $p \times p$ boyutlu Σ matrisi köşgensel bir matristir ve köşegen elemanları Λ matrisinin sıfırdan farklı tekil olmayan değerleridir. $(n-p) \times p$ boyutlu $\mathbf{O}$ matrisinin elemanları sıfırdır. $n \geq p$ iken $\Lambda^T \Lambda$ matrisinin j . özdeğeri a_j olmak üzere Σ matrisi

$$\Sigma = \text{diag}\{a_j\} \quad (5.5)$$

ile verilir (Akdeniz ve Öztürk, 1996: 34). Burada $a_1 \geq a_2 \geq \dots \geq a_k$ olarak bilinir.

U^* matrisinin ilk p sütununun oluşturduğu matris U ve V^* matrisinin ilk p sütununun oluşturduğu matris V olarak alınırsa Eş. 5.3

$$\Lambda = U\Sigma V^T \quad (5.6)$$

biçiminde yeniden yazılabilir(Akdeniz ve Öztürk,1996: 34).

$p > k$ iken Σ matrisinin köşegen elemanlarının k tanesi aynı, geri kalan $(p - k)$ tanesi sıfır olarak alınırsa elde edilen yeni matris $\Sigma_{[k]}$ olmak üzere

$$\hat{\Lambda}_{[k]} = U\Sigma_{[k]}V^T \quad (5.7)$$

matrisi elde edilir(Gower ve diğerleri, 2011: 17). Burada $\hat{\Lambda}_{[k]}$ matrisi $n \times p$ boyutludur ve k adet bileşenin bilgisini içermektedir. Λ matrisini $\hat{\Lambda}_{[k]}$ matrisi biçiminde dönüştürme işlemi Λ 'nın k -boyutlu yaklaşımı olarak tanımlanır. Λ matrisini en iyi temsil eden $\hat{\Lambda}_{[k]}$ matrisini elde etmek amacıyla EKK yaklaşımı kullanılabilir. Bu durumda

$$\|\Lambda - \hat{\Lambda}_{[k]}\|^2 = \text{tr} \left\{ (\Lambda - \hat{\Lambda}_{[k]}) (\Lambda - \hat{\Lambda}_{[k]})^T \right\} \quad (5.8)$$

eşitliği yazılırsa bu eşitlikte sağ tarafı en küçükleyen ve rankı k olan $\hat{\Lambda}_{[k]}$ matrisi bulunmaya çalışılır(Gower ve diğerleri, 2011: 17). $p \times p$ boyutlu bir J matrisi

$$J_{p \times p} = \begin{bmatrix} \mathbf{I}_k & 0 \\ 0 & 0 \end{bmatrix} \quad (5.9)$$

olarak tanımlanırsa Eş. 5.7'deki $\hat{\Lambda}_{[k]}$ matrisinin bir başka biçimi

$$\hat{\Lambda}_{[k]} = U\Sigma JV^T = UJ\Sigma V^T = UJ\Sigma JV^T \quad (5.10)$$

ile verilir(Gower ve diğerleri, 2011: 17). Bu durumda $\hat{\Lambda}_{[k]}$ matrisi

$$\begin{aligned}
\hat{\Lambda}_{[k]} &= (UJ\Sigma)(VJ)^T \\
&= (UJ\Sigma\Theta)(VJ\Theta)^T \\
&= \hat{\mathbf{L}}_{[k]}(\hat{\mathbf{R}}_{[k]})^T
\end{aligned} \tag{5.11}$$

ile verilir. Burada Θ matrisi $p \times p$ boyutlu bir dik matristir ve yukarıda verilen içsel çarpımın temsil ettiği değişkenlere ilişki vektörlerin ve gözlemlere ilişkin noktaların yapısını bozmadan dikey ve yatay eksenleri döndürme işlemi için kullanılmaktadır (Gower ve diğerleri, 2011: 20).

$\hat{\Lambda}_{[k]} = \hat{\mathbf{L}}_{[k]} \hat{\mathbf{R}}_{[k]}^T$ içsel çarpımının elemanları $\hat{\Lambda}_{ij} = \hat{L}_i^T \hat{R}_j$ olmak üzere, $\hat{L}_i$ vektörünün sıfırdan farklı ilk k elemanı, i . gözlem için k -boyutlu koordinat eksenindeki noktayı ifade eder. $\hat{R}_j$ vektörünün sıfırdan farklı ilk k elemanı ise j . değişken için k boyutlu Öklid uzayında bir vektörü ifade eder ve bu vektör j . değişken için bir eksendir. Fakat bu durumda k -boyutlu bir grafik söz konusudur. $k = 2$ alınırsa ilgili p adet değişken için eksenlerin de yer aldığı iki boyutlu bir serpm diyagramı söz konusu olur. Bu serpm diyagramı biplot olarak adlandırılır. Dolayısıyla $k = 2$ için Eş. 5.11

$$\hat{\Lambda}_{[2]} = \hat{\mathbf{L}}_{[2]} \hat{\mathbf{R}}_{[2]}^T \tag{5.12}$$

olarak yazılırsa $n \times 2$ boyutlu $\hat{\mathbf{L}}_{[2]}$ matrisinin satır vektörleri bunlara karşılık gelen gözlemler için noktaları, $p \times 2$ boyutlu $\hat{\mathbf{R}}_{[2]}$ matrisinin satır vektörlerinin doğrultuları ise değişkenlere karşılık gelen eksenleri ifade etmektedir. Böylece oluşturulan iki boyutlu biplotta, birinci boyut birinci bileşene, ikinci boyut ikinci bileşene karşılık gelir.

Biplot için $\hat{\mathbf{L}}_{[2]}$ ve $\hat{\mathbf{R}}_{[2]}$ matrisleri hesaplanmadan önce, noktaların merkezinin orijin olması amacıyla genellikle veri kümesine merkezileştirme işlemi yapılır. Bu matrisler elde edildikten sonra grafiksel gösterimin daha iyi olması amacıyla lamda ölçeklemesi uygulanır. Bu ölçekleme için sabit bir lamda değeri

$$\lambda = \sqrt[4]{\frac{n \|\hat{\mathbf{R}}_{[2]}\|^2}{p \|\hat{\mathbf{L}}_{[2]}\|^2}} \tag{5.13}$$

olarak verilir(Gower ve diğerleri, 2011: 24). Burada $\|\bullet\|$ fonksiyonu bir matris normunu göstermektedir. Bu durumda Eş. 5.12

$$\hat{\Lambda}_{[2]} = (\lambda \hat{\mathbf{L}}_{[2]})(\hat{\mathbf{R}}_{[2]}/\lambda)^T \quad (5.14)$$

biçiminde tekrar yazılabilir(Gower ve diğerleri, 2011: 24). Dolayısıyla $\hat{\mathbf{L}}^* = \lambda \hat{\mathbf{L}}_{[2]}$ ve $\hat{\mathbf{R}}^* = \hat{\mathbf{R}}_{[2]}/\lambda$ olarak tanımlanırsa bu matrisler kullanılarak biplot oluşturulur.

5.2. Biplot Yardımıyla Etkili Gözlemlerin Saptanması için Bir Grafiksel Yöntem Önerisi

Eş. 2.1'deki çoklu regresyon modeli ele alınırsa, bu modelde i . gözleme ilişkin x_i bilgisinin ve bu gözleme ilişkin bağımlı değişken değerinin yer aldığı bir z_i değişkeni $z_i = (x_i^T, y_i)^T$, ($i=1,2,...,n$), ile tanımlansın. Bu durumda, bu vektörlerin oluşturduğu matris $Z = [z_1, z_2, ..., z_n]^T = (X, Y)$ ile verilir.

Hata değişkenlerine ilişkin varsayımlar altında Eş. 2.1 ile verilen modelde yer alan β parametre vektörünün EKK tahmin edicisi,

$$\hat{\beta} = (X^T X)^{-1} X^T Y \quad (5.15)$$

ile verilir. $\hat{\beta}_{(i)}$ ise veri kümesinden i . gözlemin silinmesiyle elde edilen β parametre vektörünün EKK tahmin edicisidir ve

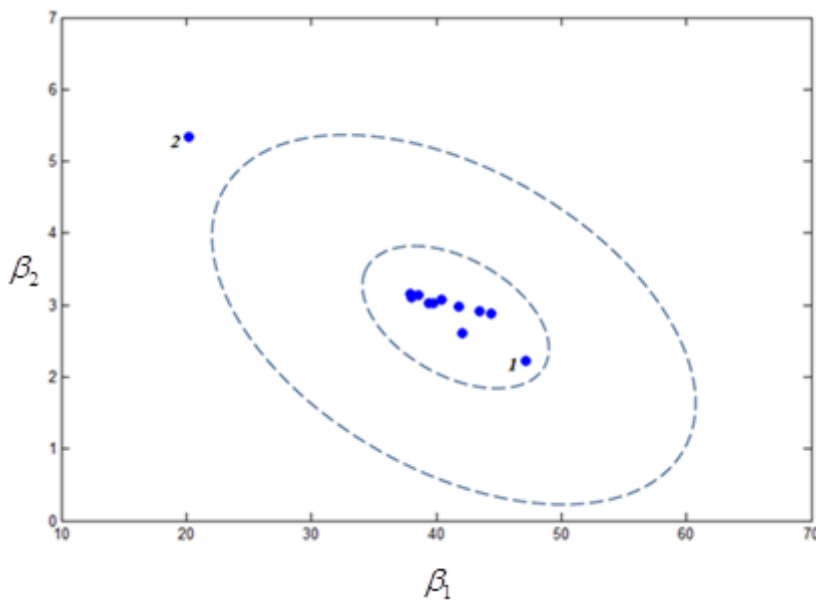
$$\hat{\beta}_{(i)} = (X_{(i)}^T X_{(i)})^{-1} X_{(i)}^T Y_{(i)} \quad (5.16)$$

ile verilir.

Bağımlı değişken ile bağımsız değişkenler arasındaki doğrusal ilişki regresyon modelinde bu bağımsız değişkenlere karşılık gelen parametrelerin tahminleriyle ilgilidir. Bundan dolayı regresyonda modelin anlamlılığı için sabit parametrenin tahmini olan $\hat{\beta}_0$

tahmininden çok, bağımsız değişkenlere ilişkin parametrelerin tahminleri etkide bulunmaktadır. Ayrıca veride bulunan etkili gözlem veya etkili gözlem grubu $\hat{\beta}_0$ tahmininden çok, eğim parametrelerinin tahminlerini olumsuz bir biçimde etkilemektedir. Bu nedenle etkili gözlemlerin saptanması için tanı yöntemleri kullanılmadan önce $\hat{\beta}_0$ tahmini, tahmin edilen modelden çıkarmak amacıyla bağımlı değişken ve bağımsız değişkenlerin değerlerine genellikle merkezileştirme işlemi uygulanır. Nitekim etkili gözlemin saptanmasıyla bu gözlemin veriden çıkarılması veya bu gözlemin ağırlığının düşürülmesi işleminden sonra merkezileştirme yapılmasına gerek duymaksızın regresyon analizine devam edilebilir.

Veri kümesine merkezileştirme yapılmışken iki bağımsız değişkenin olduğu doğrusal regresyon modeli dikkate alındığında $\hat{\beta}_{j(i)}$ ($j=1,2$) tahminleri için oluşturulan serpm diyagramı örneği Şekil 5.3'teki gibidir. Eğer modelde üç tane bağımsız değişken olursa üç boyutlu bir serpm diyagramı oluşturulabilir. Ancak bu diyagramın incelenmesi zordur. Bununla birlikte uygulamada genellikle ikiden fazla bağımsız değişken söz konusu olduğundan, $\hat{\beta}_{j(i)}$ tahminleri için serpm diyagramının oluşturulması mümkün değildir. Bu amaçla bağımsız değişken sayısı ikiden fazla iken $\hat{\beta}_{j(i)}$ 'ler için bir serpm diyagramına ihtiyaç duyulursa biplot tekniğinden yararlanılabilir.



Şekil 5.3. Modelde iki bağımsız değişken varken $\hat{\beta}_{1(i)}$ ve $\hat{\beta}_{2(i)}$ tahminleri için bir serpm diyagramı örneği

Bağımlı değişken ve bağımsız değişken değerlerinin merkezileştirildiği durum altında, tüm gözlemler için $\hat{\beta}_{(i)}^T$ vektörlerinin oluşturduğu matris

$$\Psi = \begin{bmatrix} \hat{\beta}_{(1)}^T \\ \hat{\beta}_{(2)}^T \\ \vdots \\ \hat{\beta}_{(n)}^T \end{bmatrix} = \begin{bmatrix} \hat{\beta}_{1(1)} & \hat{\beta}_{2(1)} & \cdots & \hat{\beta}_{p-1,(1)} \\ \hat{\beta}_{1(2)} & \hat{\beta}_{2(2)} & \cdots & \hat{\beta}_{p-1,(2)} \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\beta}_{1(n)} & \hat{\beta}_{2(n)} & \cdots & \hat{\beta}_{p-1,(n)} \end{bmatrix} \quad (5.17)$$

olarak tanımlansın. Eş. 5.17’de verilen matrisin i . satır j . sütun elemanı, i . gözlem veri kümesinden çıkarıldıktan sonra geri kalan gözlemlerin bilgisi kullanılarak elde edilen $\hat{\beta}_j$ tahminidir. Eş. 5.17’de verilen matris

$$\Psi = \mathbf{1} \times \hat{\beta}^T - \text{diag}\{e_1/(1-h_{11}), \dots, e_n/(1-h_{nn})\} X (X^T X)^{-1} \quad (5.18)$$

olarak yeniden yazılabilir. Burada $n \times 1$ boyutlu $\mathbf{1}$ vektörü tüm elemanları 1 olan vektör, e_i değeri i . gözlem için artık, h_{ii} değeri de kaldıraç değerine karşılık gelmektedir.

Bağımlı değişken ve bağımsız değişken değerlerinin merkezileştirildiği durum altında tüm gözlemler için $(\hat{\beta}_{(i)} - \hat{\beta})^T$ tahmin vektörlerinin oluşturduğu matris ise

$$\Delta = \begin{bmatrix} (\hat{\beta}_{(1)} - \hat{\beta})^T \\ (\hat{\beta}_{(2)} - \hat{\beta})^T \\ \vdots \\ (\hat{\beta}_{(n)} - \hat{\beta})^T \end{bmatrix} = \begin{bmatrix} (\hat{\beta}_{1(1)} - \hat{\beta}_1) & (\hat{\beta}_{2(1)} - \hat{\beta}_2) & \cdots & (\hat{\beta}_{p-1,(1)} - \hat{\beta}_{p-1}) \\ (\hat{\beta}_{1(2)} - \hat{\beta}_1) & (\hat{\beta}_{2(2)} - \hat{\beta}_2) & \cdots & (\hat{\beta}_{p-1,(2)} - \hat{\beta}_{p-1}) \\ \vdots & \vdots & \ddots & \vdots \\ (\hat{\beta}_{1(n)} - \hat{\beta}_1) & (\hat{\beta}_{2(n)} - \hat{\beta}_2) & \cdots & (\hat{\beta}_{p-1,(n)} - \hat{\beta}_{p-1}) \end{bmatrix} \quad (5.19)$$

olarak tanımlansın. Eş. 5.19’da verilen matrisin i . satır j . sütun elemanı, i . gözlem veri kümesinden çıkarıldıktan sonra $\hat{\beta}_j$ tahminindeki net değişim miktarıdır. Eğer i . gözlem $\hat{\beta}_j$ tahminini değiştirmiyorsa bu değer sıfırdır. Buna dayalı olarak Δ matrisinde i . gözleme karşılık gelen satırın tüm elemanları sıfır ise bu ilgili gözleme ilişkin noktanın $p-1$ boyutlu regresyon düzlemi üzerinde olduğu anlamına gelir. Eş. 5.19’daki matris

$$\begin{aligned}\Delta &= \Psi - \mathbf{1} \times \hat{\beta}^T \\ &= -\text{diag}\{e_1/(1-h_{11}), \dots, e_n/(1-h_{nn})\} X (X^T X)^{-1}\end{aligned}\quad (5.20)$$

olarak da tanımlanabilir.

Eş. 5.18’de verilen Ψ matrisine merkezileştirme işlemi uygulanırsa elde edilen yeni matris ile Eş. 5.20’de verilen matris birbirine denk olur. Dolayısıyla bu iki matris ele alınarak oluşturulan biplotlar görsel olarak da aynıdır. Sadece bu biplotların eksenlerinin isimleri farklılık gösterir.

Bu çalışmada Eş. 5.18 ve Eş. 5.20’de tanımlanan matrisler için biplot yöntemi uygulanmıştır. Dolayısıyla biplot oluşturmak için Eş. 5.18 ve Eş. 5.20’de tanımlanan matrislerin 2-boyutlu yaklaşımı, Eş. 5.14’ten yararlanarak sırasıyla

$$\hat{\Psi}_{[2]} = \hat{\mathbf{L}}_{\Psi[2]}^* (\hat{\mathbf{R}}_{\Psi[2]}^*)^T \quad (5.21)$$

ve

$$\hat{\Delta}_{[2]} = \hat{\mathbf{L}}_{\Delta[2]}^* (\hat{\mathbf{R}}_{\Delta[2]}^*)^T \quad (5.22)$$

biçimde tanımlansın. Bu durumda $\hat{\mathbf{L}}_{\Psi[2]}^*$ ve $\hat{\mathbf{L}}_{\Delta[2]}^*$ matrislerinin sütun vektörlerine karşılık gelen değişken vektörleri sırasıyla $\ell_{\Psi} = [\ell_{\Psi 1}, \ell_{\Psi 2}]^T$ ve $\ell_{\Delta} = [\ell_{\Delta 1}, \ell_{\Delta 2}]^T$ ile verilsin. Eş. 5.18 ve Eş. 5.20’de tanımlanan matrislerden yararlanarak oluşturulan biplotta, noktalar için elips biçiminde bir sınır oluşturmak etkili gözlemleri belirlemede kolaylık sağlar. Bu sınır klasik tanı yöntemlerindeki karar kurallarında kullanılan kesme noktasıyla (cutoff point) aynı amaç doğrultusunda belirlenmelidir. Bu durumda oluşturulan iki boyutlu biplotta, noktalar için bir elips Mahalanobis Uzaklığı’na dayalı olarak tanımlanabilir.

$\hat{\mathbf{L}}_{\Delta[2]}^*$ matrisinin sütunlarına karşılık gelen değişkenler 2×1 boyutlu $\ell_{\Delta} = [\ell_{\Delta 1}, \ell_{\Delta 2}]^T$ değişken vektörü ile tanımlansın. Mahalanobis Uzaklığı’nın karesi

$$\Gamma_i^2 = (\ell_{\Delta} - \mu_{\ell_{\Delta}})^T \Sigma_{\ell_{\Delta}}^{-1} (\ell_{\Delta} - \mu_{\ell_{\Delta}}) , \quad (i=1,2,\dots,n) \quad (5.23)$$

olarak tanımlanırsa $\ell_{\Delta} \sim MVN_2(\mu_{\ell_{\Delta}}, \Sigma_{\ell_{\Delta}})$ iken $\Gamma_i^2 \sim \chi_2^2$ dağılmaktadır. $\mu_{\ell_{\Delta}}$ ve $\Sigma_{\ell_{\Delta}}$ parametreleri bilinmediğinden bunların yansız tahmin edicileri

$$\bar{\ell}_{\Delta} = \frac{1}{n} \sum_{i=1}^n \ell_{\Delta i} \quad (5.24)$$

ve

$$\mathbf{S}_{\ell_{\Delta}} = \frac{1}{n-1} \sum_{i=1}^n (\ell_{\Delta i} - \bar{\ell}_{\Delta})(\ell_{\Delta i} - \bar{\ell}_{\Delta})^T \quad (5.25)$$

kullanılabilir. Bu durumda Mahalanobis Uzaklığı'nın karesi

$$\xi_i^2 = (\ell_{\Delta i} - \bar{\ell}_{\Delta})^T \mathbf{S}_{\ell_{\Delta}}^{-1} (\ell_{\Delta i} - \bar{\ell}_{\Delta}), \quad (i=1,2,\dots,n) \quad (5.26)$$

olarak tanımlanır. Bu uzaklık $n/(n-1)^2$ ile çarpılırsa elde edilen yeni değişken

$$\xi_i^2 \frac{n}{(n-1)^2} \sim \text{Beta}\left(\frac{\delta}{2}, \frac{n-\delta-1}{2}\right) \quad (5.27)$$

dağılımına sahiptir (Ververidis ve Kotropoulos, 2008). Burada biplot için boyut sayısı δ olmak üzere $\text{Beta}[\delta/2, (n-\delta-1)/2]$ dağılımı $\delta/2$ ve $(n-\delta-1)/2$ parametrelili beta dağılımını ifade etmektedir. Dolayısıyla bir elips denklemi Mahalanobis Uzaklığı'nın karesi biçiminde

$$(\ell_{\Delta} - \bar{\ell}_{\Delta})^T \mathbf{S}_{\ell_{\Delta}}^{-1} (\ell_{\Delta} - \bar{\ell}_{\Delta}) = \frac{(n-1)^2}{n} F_{\frac{\xi_i^2 n}{(n-1)^2}}^{-1} [1-\alpha] \quad (5.28)$$

ile tanımlanırsa bu elipsin dışında kalan noktalar biplotta verinin geneline uymayan noktalardır ve bu noktalar EKK tahminlerini etkileyen gözlemler olan etkili gözlemlere karşılık gelir. Burada $F_{\frac{\xi_i^2 n}{(n-1)^2}}^{-1}[\cdot]$ fonksiyonu beta dağılım fonksiyonunun ters fonksiyonu olmak üzere $F_{\frac{\xi_i^2 n}{(n-1)^2}}^{-1}[1-\alpha]$ değeri $1-\alpha$ olasılığına karşılık gelen kritik değer (kesme değeri)'dir. Aynı durum $\ell_{\Psi} = [\ell_{\Psi_1}, \ell_{\Psi_2}]^T$ değişken vektörü için de söz konusudur.

Eş. 5.28’de verilen elips denklemine dayalı daha basit bir karar kuralı oluşturulabilir. Bu karar kuralı “*i. gözlem için Mahalanobis Uzaklığı’nın hesaplanan değeri ξ_i^2 olmak üzere, $\xi_i^2 > (n-1)^2/n F_{\xi_i^2 n/(n-1)^2}^{-1}[1-\alpha]$ ise bu gözlem etkili bir gözlemdir.*” biçiminde verilir.

6. UYGULAMA

Bu bölümde çoklu doğrusal regresyonda etkili gözlem veya gözlemlerin saptanması için bilinen yöntemlerle, bu çalışmada önerilen iki grafiksel tanı yöntemi, literatürde etkili gözlem içerdiği bilinen ve tanı yöntemlerinin değerlendirilmesi için sıklıkla kullanılan iki veri kümesi üzerinden incelenmiştir. Cook ve Weisberg (1982) tarafından verilen Fare Verisi (Rat Data) ve Chatterjee ve Hadi (2006) tarafından verilen New York'taki Nehirler Verisi (New York Rivers Data) ele alınan bu iki veri kümesidir.

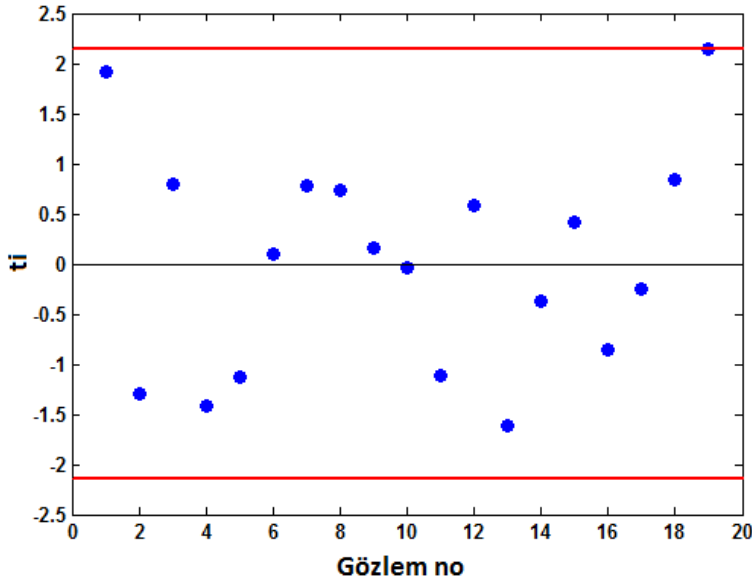
6.1. Fare Verisi (Rat Data) Uygulaması

Cook ve Weisberg (1982) tarafından bahsedilen fare verisi, 3. gözlemi etkili gözlem olduğu bilinen ve literatürde tanı yöntemlerinin incelenmesinde sıklıkla ele alınan popüler bir veri kümesidir. Veri kümesini yapısı ve derlenişi şu şekildedir: Bir ilacın karaciğerdeki etkisi incelenmek amacıyla 19 kobay fare rasgele seçilmiştir. Sonra bu fareler laboratuvar ortamında anestezi edilip ağırlıkları ölçülmüş ve farelere ağız yoluyla ilgili ilaç verilmiştir. Karaciğer ağırlığının vücut ağırlığıyla doğru orantılı olduğu ve verilen ilaç dozunu ağırlıkça büyük karaciğerlerin küçük olan karaciğerlere göre daha fazla absorbe ettiği bilindiğinden, ilacın dozu vücut ağırlığının her bir kilogramı için yaklaşık olarak 40 mg olarak belirlenmiştir. Belli bir süre sonra her bir farenin karaciğeri alınarak bu karaciğerlerin ağırlıkları ölçülmüş ve karaciğerlerinde biriken ilaç doz oranı belirlenmiştir. Bağımlı değişken ve bağımsız değişkenler farenin karaciğerindeki ilaç doz oranı (Y), farenin toplam ağırlığı (X_1), farenin karaciğer ağırlığı (X_2) ve fareye verilen ilaç doz oranı (X_3) olarak tanımlanmıştır. Cook ve Weisberg (1982)'de yer alan bu veri kümesi EK1'de verilmiştir. Yokluk hipotezi "*Farelerin karaciğerlerindeki ilaç doz oranı üzerinde, farelerin vücut ağırlığının, karaciğer ağırlığının ve fareye verilen ilaç doz oranının bir etkisi yoktur.*" biçimindedir. Cook ve Weisberg (1982) bu veri seti için

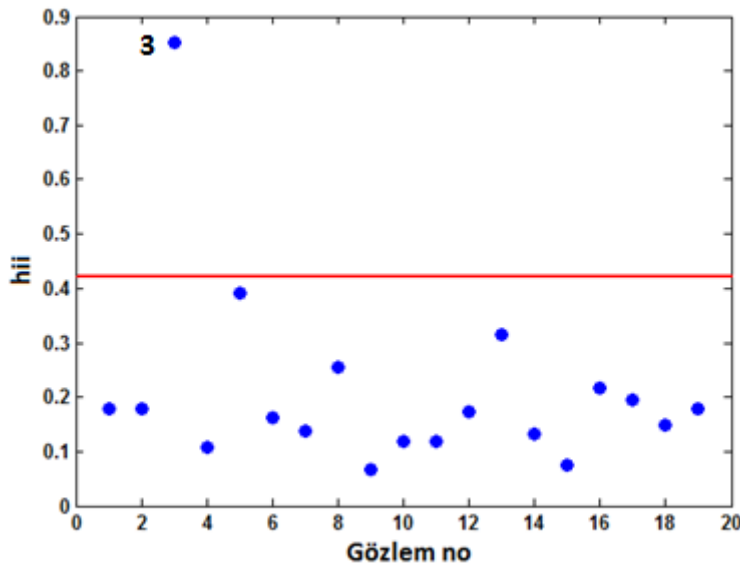
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon \quad (6.1)$$

teorik modelini dikkate almışlardır. Çoklu doğrusal regresyon modelinin tahmini ve istatistiksel çıkarımlardan önce, verinin etkili gözlem veya gözlemler içerip içermediğine bakılır. Bu amaçla tanı yöntemleri uygulanmadan önce artıkların en yaygın kullanılan standart biçimi olan harici student tipi artıklar ve yüksek dereceli kaldıraç noktalarını

belirlemek amacıyla kaldıraç değerleri hesaplanmıştır. Bu çalışmada tüm hesaplamalar ve grafikler MATLAB2011 programı kullanılarak elde edilmiştir. Sırasıyla Şekil 6.1 ve Şekil 6.2’de bu istatistiklere ilişkin grafikler verilmiştir.



Şekil 6.1. Fare verisi için harici student tipi artıklar (t_i ; $i = 1, 2, \dots, 19$) ile gözlem numarası için oluşturulan serpmeye diyagramı ve sınırlar (Alt sınır $-t_{\alpha/2, n-p-1}$ kritik değerine ve üst sınır $+t_{\alpha/2, n-p-1}$ kritik değerine karşılık gelmektedir.)



Şekil 6.2. Fare verisi için kaldıraç değerleri (h_{ii} ; $i = 1, 2, \dots, 19$) ile gözlem numarası için oluşturulan serpmeye diyagramı ve sınır (Bu sınır $2p/n$ kritik değerine karşılık gelmektedir.)

Şekil 6.1’de görüldüğü üzere tüm gözlemlerin harici student tipi artık değeri (t_i) sınırların içinde kalmıştır. Bu da artık değerleri bakımından verinin geneline uymayan hiçbir gözlemin bulunmadığı, bir başka deyişle bu veri kümesinde aykırı bir gözlemin olmadığı anlamına gelmektedir. Bununla birlikte her bir gözleme karşılık gelen kaldırıcı değerleri (h_{ii}) de etkili gözlemlerin saptanmasında önemli bir gösterge olarak kullanılmaktadır. Şekil 6.2 incelendiğinde grafikte sınıra karşılık gelen, kaldırıcı değerleri için kritik değer $2p/n = 0,4210$ iken 3 numaralı gözlemin yüksek dereceli kaldırıcı değeri olduğu açıkça görülmektedir. Bununla birlikte 3. gözleme karşılık gelen harici student türü artık mutlak değerce kritik değer olan $t_{0.05,14}$ değerinden küçüktür. Fakat bu gözlem yüksek dereceli kaldırıcı noktası olduğundan bu gözlemin etkili gözlem olma potansiyeli yüksektir. Bu veri kümesinde etkili gözlemleri saptamak amacıyla tanı istatistikleri ve bunlara karşılık gelen kritik değerler Çizelge 6.1’de verilmiştir.

Çizelge 6.1. Fare verisi için tanı istatistiklerinin değerleri ve kritik değerler

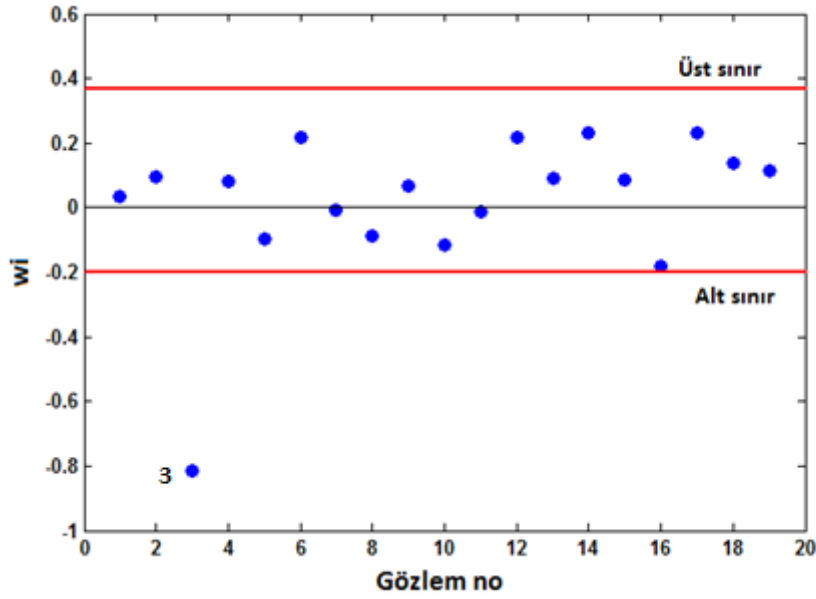
Gözlem no	$D_{(t)}$	$ DFBETAS_{0(t)} $	$ DFBETAS_{1(t)} $	$ DFBETAS_{2(t)} $	$ DFBETAS_{3(t)} $	$ DIFFITS_{(t)} $	$ COVRATIO_{(t)} - 1 $	$1 - AP_{(t)}$	$LD_{(t)}$	$K_{(t)}$	$ S'_{(t)} - med(S) $
1	0,1688	0,0384	0,3149	0,7044	0,2437	0,8920	0,3690	0,3489	1,3466	1,2197	0,1352
2	0,0885	0,1426	0,0977	0,4818	0,1256	0,6088	0,0164	0,2680	0,5117	0,6927	0,1251
3	0,9296	0,2310	1,6677	0,3046	1,7472	1,9048	6,4008	0,8574	4,6656	5,8824	0,5368
4	0,0572	0,1250	0,1269	0,3037	0,1401	0,4944	0,1401	0,2205	0,3785	0,6907	0,1707
5	0,2029	0,5216	0,3963	0,5500	0,2747	0,9095	0,5242	0,4427	1,0744	0,9980	0,0000
6	0,0005	0,0023	0,0136	0,0290	0,0172	0,0427	0,5667	0,1617	0,0289	0,1948	0,1693
7	0,0246	0,1838	0,1504	0,0834	0,1184	0,3096	0,2893	0,1726	0,1246	0,3303	0,0409
8	0,0469	0,2973	0,0594	0,2465	0,0404	0,4262	0,5201	0,2811	0,2361	0,4911	0,0091
9	0,0005	0,0097	0,0179	0,0002	0,0167	0,0427	0,4023	0,0687	0,0278	0,0791	0,1099
10	0,0001	0,0057	0,0099	0,0034	0,0093	0,0140	0,4964	0,1198	0,0274	0,1364	0,2523
11	0,0414	0,2905	0,1938	0,1007	0,1729	0,4104	0,0656	0,1912	0,2257	0,4865	0,1563
12	0,0189	0,2174	0,0253	0,0517	0,0092	0,2689	0,4437	0,1924	0,1012	0,3071	0,0653
13	0,2726	0,7723	0,1439	0,7665	0,1200	1,0989	0,0278	0,4237	1,6878	1,1670	0,1888
14	0,0054	0,0348	0,0462	0,0765	0,0599	0,1423	0,4605	0,1396	0,0446	0,1895	0,1642
15	0,0037	0,0187	0,0406	0,0546	0,0382	0,1188	0,3588	0,0873	0,0346	0,1313	0,0815
16	0,0510	0,1231	0,0056	0,3287	0,0491	0,4475	0,3749	0,2551	0,2575	0,4811	0,0829
17	0,0042	0,1036	0,0145	0,0276	0,0021	0,1262	0,6071	0,1990	0,0433	0,2613	0,0688
18	0,0316	0,1542	0,1904	0,1623	0,1893	0,3522	0,2701	0,1898	0,1597	0,3761	0,1168
19	0,1999	0,8558	0,2501	0,2946	0,1712	0,9952	0,4826	0,3804	1,8102	1,4517	0,1007
Kritik değer	0,8783	0,4588	0,4588	0,4588	0,4588	0,9177	0,6316	0,5263	11,0705	4,6921	0,5257

Çizelge 6.1’de her bir gözlem için Cook Uzaklığı İstatistiğiyle elde edilen değerler incelendiğinde, $D_{(3)} > F_{0.50,4,15}$ ($0,9296 > 0,8783$) olduğundan 3 numaralı gözlemin bu yöntemle göre etkili olarak saptandığı açıktır. Diğer tanı istatistiklerinden *DFBETAS* istatistiğiyle elde edilen sonuçlar incelendiğinde, bu istatistik için kritik değer olan $2/\sqrt{n} = 0,4588$ değerinden mutlak değerce büyük olan *DFBETAS* değerlerinin, $\hat{\beta}_0$ tahmini için 5, 13 ve 19 numaralı gözlemlere; $\hat{\beta}_1$ tahmini için 3 numaralı gözleme; $\hat{\beta}_2$ tahmini için 1, 2, 5 ve 13 numaralı gözlemlere; $\hat{\beta}_3$ tahmini için de 3 numaralı gözleme karşılık geldiği görülmektedir. 3 numaralı gözlemin veri kümesindeki tek etkili gözlem olduğu bilgisi altında *DFBETAS* istatistiği ile bu etkili gözlem net olarak saptanamamaktadır. Benzer bir istatistik olan *DFFITS* istatistiği için elde edilen sonuçlar incelendiğinde ise bu istatistik için hesaplanan kritik değer olan 0,9177 değerinden mutlak değerce büyük *DFFITS*_(i) değerine sahip gözlemlerin 3, 13 ve 19 numaralı gözlemler olduğu görülmektedir. Bu yöntemle gerçekte etkili gözlem olan 3 numaralı gözlem etkili gözlem olarak doğru saptanmış olmasına rağmen, regresyon tahminlerini değiştirme eğilimi olmayan 13 ve 19 numaralı gözlemlerin de etkili olarak saptanması yanlış bir karardır.

COVRATIO istatistiği ve Andrews-Pregibon (AP) istatistiği için sonuçlar incelendiğinde bu iki yöntemle 3 numaralı gözlem etkili gözlem olarak saptanmış, Cook ve Weisberg olabilirlik istatistiği (LD) yöntemiyle hiçbir gözlem etkili olarak saptanamamıştır. Hadi’nin K_i istatistiği ve Pena’nın S_i istatistiği için elde edilen sonuçlara göre de 3 numaralı gözlem etkili gözlem olarak belirlenmiştir.

Bu çalışmada önerilen, SIC fonksiyonunun optimizasyonuna dayalı grafik yöntemi (SICDGY) ile biplot yaklaşımına dayalı grafik yöntemi (BYDGY) incelenmek amacıyla fare verisi ele alınıp uygulama yapılmıştır. İlk olarak SICDGY ele alınsın.

Yüksek dereceli kaldıraç noktaları, potansiyel etkili gözlemler olduğundan SICDGY’nin uygulanması için gerekli koşul veri kümesinde en az bir yüksek dereceli kaldıraç noktasının bulunmasıdır. Fare verisi için Şekil 6.2’den $h_{33} > 2p/n$ olduğu açıkça görülmektedir. Dolayısıyla SICDGY bu veri kümesi için uygulanabilir. SICDGY için grafik Şekil 6.3’teki gibi oluşturulmuştur.

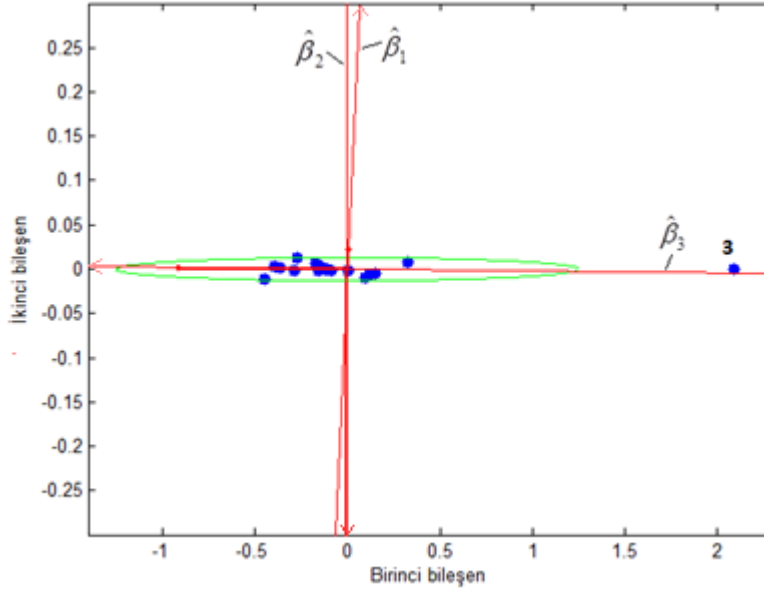


Şekil 6.3. Fare verisi için SICDGY ile elde edilen w değişkeni ile gözlem numarası için oluşturulan serpmeye diyagramı ve sınırlar (alt sınır = -0,1998; üst sınır = 0,3673)

SICDGY yöntemi için amaç fonksiyonunun optimum çözümü olan w_i değerlerini elde etmek amacıyla MATLAB programında *patternsearch* komutundan yararlanılmıştır. $i = 1, 2, \dots, 19$ olmak üzere w_i değerleri veri kümesinde yer alan her bir gözlemin regresyon tahminlerini değiştirme eğilimine ilişkin katsayılar olduğundan, bu değerlerden sıfıra yakın olan değerler regresyon tahminleri üzerinde etkisi olmayan gözlemlere, mutlak değerce büyük olan w_i değerleri ise etkili gözlemlere karşılık gelmektedir. Dolayısıyla Şekil 6.3'ten de görüldüğü üzere 3 numaralı gözlem etkili gözlemdir.

Yüksek dereceli kaldırma noktaları potansiyel etkili gözlemler olduğundan, SIC fonksiyonunun optimizasyonuna dayalı grafiksel yöntemde olduğu gibi biplot yaklaşımına dayalı grafiksel yöntemin (BYDGY) de uygulanması için veri kümesinin en az bir yüksek dereceli kaldırma noktası içermesi gerekmektedir.

$\hat{\beta}_0$ tahminini modelden çıkarmak BYDGY'de daha kolay bir yorumlamayı sağladığından, yöntemin ilk aşamasında, genelliğin kaybı olmaksızın, bağımlı ve bağımsız değişken değerlerine merkezileştirme yapılmıştır. Dolayısıyla modelde sadece bağımsız değişkenlere ilişkin katsayılar kalmıştır. Bu katsayılar tekrar $\hat{\beta}_1$, $\hat{\beta}_2$ ve $\hat{\beta}_3$ ile gösterilsin. Eş. 5.19'da verilen Ψ matrisi ele alındığında, fare verisi için $\hat{\beta}_{j(i)}$ tahminlerine dayalı biplot Şekil 6.4'teki gibi elde edilir.



Şekil 6.4. Fare verisi için $\hat{\beta}_{j(i)}$ tahminlerine dayalı biplot

Bağımlı ve bağımsız değişkenlere merkezileştirme uygulanmışken veri kümesinden i . gözlem çıkartıldığında β_1 , β_2 ve β_3 parametrelerinin tahminleri sırasıyla $\hat{\beta}_{1(i)}$, $\hat{\beta}_{2(i)}$ ve $\hat{\beta}_{3(i)}$ değişkenleri olarak tanımlanırsa, Şekil 6.4'te biplotta yer alan üç adet doğru, bu değişkenlere karşılık gelen eksenleri ifade etmektedir. $\hat{\beta}_{2(i)}$ ve $\hat{\beta}_{3(i)}$ vektörleri arasındaki açı dik açıya yakın olduğundan bu iki değişken için ilişki katsayısı sıfıra yakın bir değerdir. Bununla birlikte $\hat{\beta}_{1(i)}$ ve $\hat{\beta}_{2(i)}$ arasındaki açı geniş ve 180° 'ye yakın olduğundan bu iki değişken için ilişki katsayısı negatif ve mutlak değerce büyük bir değerdir.

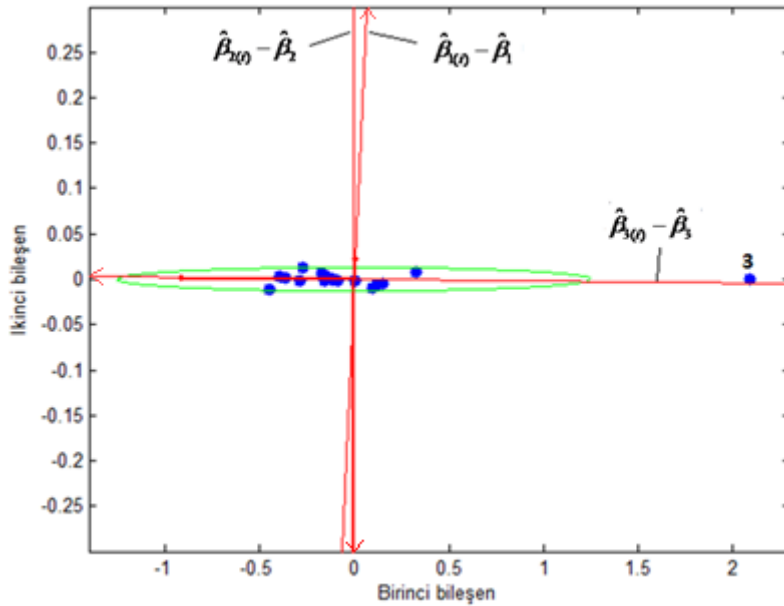
Şekil 6.4'te yer alan biplot gözlemler bakımından incelenirse, noktaların büyük bir çoğunluğunun merkezde kümелendiği gözlenmektedir. Bununla birlikte 3. gözleme ilişkin noktanın, noktaların geneline uymadığı ve bu noktaların koordinatlarının merkezden uzak bir yerde olduğu açıkça görülmektedir. Aynı zamanda 3. gözlem elipsin dışında kaldığından bu gözlemin $\hat{\beta}_j$ tahminleri üzerinde dikkate değer bir farklılaşmaya neden olduğunun göstergesidir. Dolayısıyla 3. gözlem etkili bir gözlemdir. Gözlemlere ilişkin Mahalanobis Uzaklıklarının karesi ve kritik değer Çizelge 6.2'deki gibidir.

Çizelge 6.2. Fare verisi için biplotta noktalara ilişkin Mahalanobis Uzaklıklarının karesi ve kritik değer

Gözlem no	ξ^2	Gözlem no	ξ^2
1	4,6259	11	0,5621
2	2,5443	12	0,0533
3	14,8484	13	5,3023
4	1,1304	14	0,1090
5	2,6125	15	0,1277
6	0,0522	16	1,1099
7	0,3417	17	0,0484
8	0,6355	18	0,8249
9	0,0516	19	0,9763
10	0,0437		
Kritik değer	5,3263		

Çizelgedeki sonuçlardan da görüldüğü üzere sadece ξ_3^2 değeri kritik değerden büyüktür. Dolayısıyla 3. gözlem etkili bir gözlemdir.

Benzer şekilde fare verisi için Δ matrisine dayalı olan biplot Şekil 6.5'teki gibi elde edilmiştir. $j = 1, 2, 3$ olmak üzere tüm gözlemlerin bilgisinden elde edilen EKK tahmini ile j . parametre için, i . gözlem veriden çıkartıldığında elde edilen EKK tahmin değeri arasındaki fark $\hat{\beta}_{j(i)} - \hat{\beta}_j$ değişkeni ile gösterilirse Şekil 6.5'teki biplotta kırmızı ile gösterilen doğrular bu değişkenlere karşılık gelen eksenlerdir.



Şekil 6.5. Fare verisi için $\hat{\beta}_{j(i)} - \hat{\beta}_j$ tahminlerine dayalı biplot

Açıkça görüldüğü üzere Şekil 6.4 ve Şekil 6.5’ te verilen biplotlar görsel olarak denktir. Sadece bu iki biplot için eksenlerin isimleri farklıdır. Şekil 6.5’te verilen biplotta 3 numaralı gözleme karşılık gelen nokta sınırın dışında olduğundan bu gözlem etkili bir gözlemdir. Bu gözlemin sınırın dışında olduğu Mahalanobis Uzaklıkları’na bakılarak da anlaşılabilir. Dolayısıyla bu durum veri kümesinden çıkartılan 3. gözlemin $\hat{\beta}_j$ tahminlerinde dikkate değer bir farklılaşmaya neden olduğunun bir göstergesidir.

6.2. New York’taki Nehirler Verisi (New York Rivers Data) Uygulaması

Bu çalışmada tanı yöntemleri için ele alınan diğer bir veri kümesi Chatterjee ve Hadi (2006) tarafından incelenen New York’taki Nehirler Verisi’dir. 1976’da New York eyaletinde 20 adet nehirden gözlem yoluyla elde edilen veri kümesinin derlenme amacı, ilgili nehrin su kirliliği ile nehrin geçtiği bölgelerde yerleşim yeri, endüstri, orman ve tarım alanı yüzdesi arasındaki ilişkiyi doğrusal regresyon analizi yardımıyla modellemektir. Yapılan bir ön çalışmayla su numunelerinin laboratuvar analizleri sonucunda kirliliğe neden olan maddenin nitrojen olduğu belirlenmiştir. Sonra regresyon analizi için bağımsız değişkenler ve bağımlı değişken

X_1 : Tarım alanı yüzdesi,

X_2 : Ormanlık alan yüzdesi,

X_3 : Yerleşim alanı (meskun alan) yüzdesi,

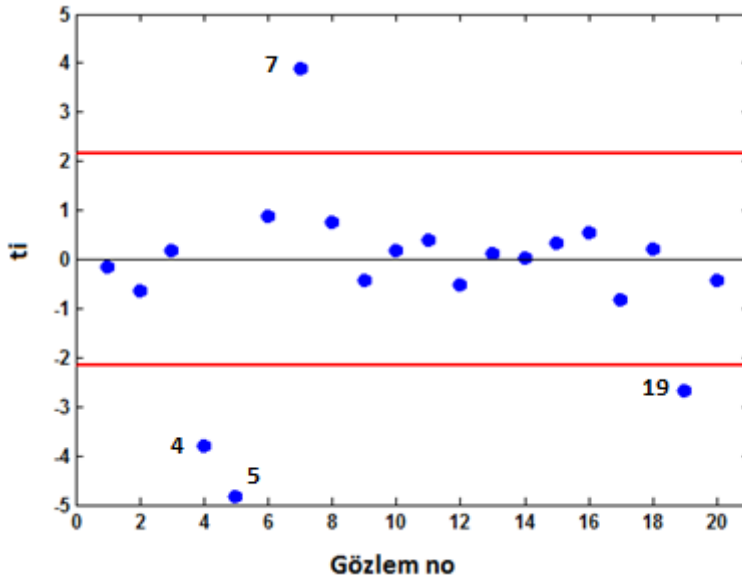
X_4 : Ticari ve endüstriyel alan yüzdesi,

Y : Kış, ilkbahar, yaz ve sonbahar aylarında nitrojen konsantrasyonu oranı (mg/litre)

olarak belirlenmiştir. Veri kümesi EK2’de yer almaktadır. Chatterjee ve Hadi (2006) bu veri kümesi için

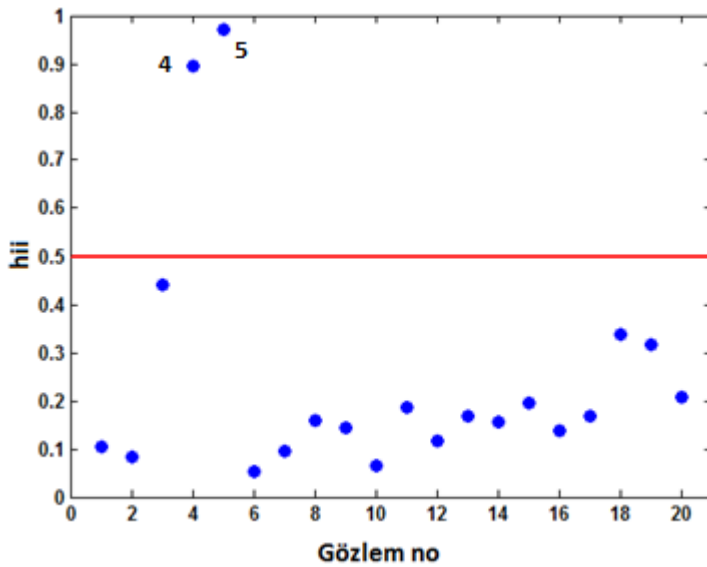
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon \quad (6.2)$$

teorik modelinin uygun olduğunu belirlemişlerdir. Önceki uygulamada olduğu gibi bu uygulamada da tanı yöntemlerinin sonuçları hesaplanmadan önce harici student tipi artıklar ve kaldıraç değerlerinin grafikleri oluşturulup sırasıyla Şekil 6.6 ve Şekil 6.7’de sunulmuştur.



Şekil 6.6. New York'taki Nehirler Verisi için harici student tipi artıklar (t_i ; $i = 1, 2, \dots, 20$) ile gözlem numarası için oluşturulan serpmeye diyagramı ve sınırlar (Alt sınır $-t_{\alpha/2, n-p-1}$ kritik değerine ve üst sınır $+t_{\alpha/2, n-p-1}$ kritik değerine karşılık gelmektedir.)

Şekil 6.6'da belirtilen grafikten de görüldüğü üzere 4, 5, 7 ve 19 numaralı gözlemlere karşılık gelen noktalar sınırların dışında kalmıştır. Bu noktalara karşılık gelen gözlemler artıklar bakımından aykırılık göstermektedir. Bununla birlikte mutlak değerce büyük t_i değerine sahip olan bu dört gözlemin etkili gözlem olduğu net bir şekilde söylenemez.



Şekil 6.7. New York'taki Nehirler Verisi için kaldıraç değerleri (h_{ii} ; $i = 1, 2, \dots, 20$) ile gözlem numarası için oluşturulan serpmeye diyagramı ve sınır (Bu sınır $2p/n$ kritik değerine karşılık gelmektedir.)

Yüksek dereceli kaldıraç noktaları, etkili gözlemlerin göstergesi olduğundan, kaldıraç değerlerini incelemek tanı yöntemleri uygulanmadan önce bir ön bilgi sağlar. Şekil 6.7 incelendiğinde 4 ve 5 numaralı gözlemlere ilişkin noktaların kritik değeri ifade eden sınırın dışında kaldığı açıkça görülmektedir. Dolayısıyla h_{44} ve h_{55} değerleri kritik değer olan $2p/n$ değerinden büyük olduğundan bu değerlere karşılık gelen noktalar yüksek dereceli kaldıraç noktalarıdır. Özellikle bu noktaların artıklar bakımından aykırı birer gözlem olduğu da dikkate alınırsa, bunların etkili gözlem olma potansiyeli diğerlerine göre daha yüksektir. Sonraki aşamada tanı istatistikleri incelenmiştir. New York'taki Nehirler Verisi için tanı istatistiklerinin değerleri hesaplanıp Çizelge 6.3'te verilmiştir.

Çizelge 6.3. New York'taki Nehirler Verisi için tanı istatistiklerinin değerleri ve kritik değerler

Gözlem no	$D_{(t)}$	$ DFBETAS_{0(t)} $	$ DFBETAS_{1(t)} $	$ DFBETAS_{2(t)} $	$ DFBETAS_{3(t)} $	$ DFBETAS_{4(t)} $	$ DIFFITS_{(t)} $	$ COVRATIO_{(t)} - 1 $	$1 - AP_{(t)}$	$LD_{(t)}$	$K_{(t)}$	$ S_{(t)} - med(S) $
1	0,0005	0,0299	0,0332	0,0309	0,0257	0,0038	0,0479	0,5685	0,1072	0,0276	0,1255	0,1255
2	0,0076	0,0010	0,0260	0,0061	0,0282	0,0644	0,1908	0,3374	0,1088	0,0547	0,2337	0,2337
3	0,0056	0,0288	0,0055	0,0392	0,0550	0,0640	0,1623	1,4990	0,4430	0,0593	0,8029	0,8029
4	13,2196	0,6944	0,3961	0,5313	4,7418	10,2818	11,1883	0,6023	0,9491	175,3125	11,3643	11,3643
5	65,4263	0,1232	0,3634	0,0235	8,8010	2,2133	28,5854	0,6316	0,9896	1121,139	38,0066	38,0066
6	0,0091	0,0533	0,0427	0,0434	0,0231	0,0125	0,2118	0,1291	0,1043	0,0611	0,3411	0,3411
7	0,1643	0,1770	0,1386	0,1728	0,3772	0,7379	1,2658	0,9608	0,5669	8,1627	5,0416	5,0416
8	0,0229	0,0488	0,0197	0,0589	0,0466	0,0113	0,3337	0,3654	0,1928	0,1521	0,3980	0,3980
9	0,0066	0,1164	0,1282	0,1206	0,1165	0,0462	0,1765	0,5435	0,1548	0,0566	0,2339	0,2339
10	0,0005	0,0029	0,0087	0,0044	0,0047	0,0075	0,0506	0,4887	0,0666	0,0267	0,0818	0,0818
11	0,0073	0,1517	0,1334	0,1485	0,1153	0,0230	0,1856	0,6435	0,1944	0,0630	0,2820	0,2820
12	0,0071	0,0916	0,0821	0,1057	0,1042	0,0689	0,1842	0,4551	0,1314	0,0567	0,2231	0,2231
13	0,0006	0,0154	0,0092	0,0205	0,0226	0,0231	0,0524	0,6870	0,1679	0,0286	0,2055	0,2055
14	0,0001	0,0072	0,0090	0,0057	0,0051	0,0025	0,0167	0,6710	0,1556	0,0261	0,1846	0,1846
15	0,0054	0,0655	0,0437	0,0792	0,0618	0,0255	0,1587	0,6937	0,2022	0,0532	0,2811	0,2811
16	0,0096	0,0091	0,0159	0,0308	0,0453	0,0902	0,2134	0,4825	0,1555	0,0715	0,2616	0,2616
17	0,0286	0,2879	0,2904	0,3018	0,2629	0,0682	0,3739	0,3346	0,2075	0,1898	0,4471	0,4471
18	0,0044	0,1171	0,1248	0,1079	0,0841	0,0303	0,1442	1,0980	0,3387	0,0512	0,5224	0,5224
19	0,4682	1,6635	1,5656	1,6167	1,2699	0,2501	1,8177	0,7391	0,5473	5,9699	2,6654	2,6654
20	0,0095	0,1091	0,1410	0,1057	0,1008	0,0254	0,2121	0,6741	0,2164	0,0761	0,3221	0,3221
Kritik değer	0,9107	0,4472	0,4472	0,4472	0,4472	0,4472	1,0000	0,7500	0,6000	12,5916	28,9980	1,2402

Çizelge 6.3'teki sonuçlardan ilk önce $D_{(i)}$ değerleri incelendiğinde $D_{(4)} > F_{0.50,p,n-p}$ ($13,2196 > 0,9107$) ve $D_{(5)} > F_{0.50,p,n-p}$ ($65,4263 > 0,9107$) olduğundan 4 ve 5 numaralı gözlemlerin etkili gözlemler olduğu açıkça görülmektedir. Bundan başka her ne kadar $D_{(19)} < F_{0.50,p,n-p}$ ($0,4682 < 0,9107$) olsa da bu değer verinin geneline uyan diğer gözlemlerin Cook Uzaklığı değerlerine göre büyük bir değerdir. $DFBETAS$ istatistiğinin sonuçları incelendiğinde, bu istatistik için kritik değer olan $t_{0,975,14} = 0,4472$ değerinden mutlak değerce büyük olan değerlerin, $\hat{\beta}_0$ tahmini için 4 ve 19 numaralı gözlemlere, $\hat{\beta}_1$ tahmini için 19 numaralı gözleme, $\hat{\beta}_2$ tahmini için 4 ve 19 numaralı gözlemlere, $\hat{\beta}_3$ tahmini için 4, 5 ve 19 numaralı gözlemlere, $\hat{\beta}_4$ tahmini için de 4, 5 ve 7 numaralı gözlemlere karşılık geldiği görülmektedir. Cook Uzaklığı İstatistiği ile etkili gözlem olarak saptanan 4 ve 5 numaralı gözlemler $DFBETAS$ istatistiği ile net bir biçimde etkili gözlem olarak saptanamamıştır. Ayrıca $|DFBETAS_{4(7)}| > 0,4472$ olmasına rağmen, $j = 0,1,2,3$ iken bu gözlem için $DFBETAS$ değerleri kritik değerden mutlak değerce küçük olmasıyla bu gözlemin etkili gözlem olup olmadığı açıkça belirlenememektedir. Bu sonuçlara ek olarak 19. gözlem dikkate alındığında sadece $j = 4$ iken $|DFBETAS_{4(19)}| < 0,4472$ olduğundan bu gözlem de net bir biçimde etkili gözlem olarak saptanamaz.

$DFFITS$ istatistiği için kritik değer $2\sqrt{p/n} = 1$ olarak hesaplanmıştır. Bu kritik değerden mutlak değerce büyük $DFFITS$ değerlerine sahip gözlemler 4, 5, 7 ve 19 numaralı gözlemlerdir. $COVRATIO$ istatistiği ile 3, 7 ve 18 numaralı gözlemler etkili olarak saptanmıştır.

Cook Uzaklığı istatistiği ile etkili olarak saptanan 4 ve 5 numaralı gözlemler, AP istatistiği ve LD istatistiğiyle de etkili olarak saptanmıştır. Hadi'nin K istatistiğiyle 5 numaralı gözlem etkili gözlem olarak belirlenirken, Pena'nın S istatistiğiyle 5 numaralı gözlemin yanı sıra 4, 7, ve 19 numaralı gözlemler de regresyon tahminlerini değiştirme eğiliminde olan gözlemler olarak saptanmıştır. S istatistiğiyle bu dört gözlemin etkili olarak saptanmasının nedeni, bu istatistik için karar kuralında yer alan kritik değerın sağlam tahminlerden hesaplanmasıdır.

New York'taki Nehirler Verisi için bir etkili gözlemi saptamaya dayalı tanı yöntemleri sonuçları ortak incelendiğinde 4, 5 ve 19 numaralı gözlemlerin etkili gözlem olma potansiyeli en yüksek olan gözlemler olduğu kanısına varılabilir. Buna ek olarak DFFITS, COVRATIO ve S istatistiklerine göre 7 numaralı gözlem de etkili olarak saptanmıştır. Bu dört gözlemin farklı kombinasyonları, birden fazla gözlem silme yaklaşımına dayalı olan Cook Uzaklığı, COVRATIO ve K istatistiklerinin genelleştirilmiş formülleriyle incelenebilir. Bu amaçla bu gözlemlerin ikili, üçlü ve dörtlü kümeleri I ile gösterilmek üzere bu I kümesinin veri kümesinden çıkartılmasına (silinmesine) dayalı Cook Uzaklığı, COVRATIO ve K istatistiklerine ilişkin sonuçlar hesaplanmış ve Çizelge 6.4'te kritik değerlerle özetlenmiştir.

Çizelge 6.4. New York'taki Nehirler Verisinde bulunan 4, 5, 7 ve 19 numaralı gözlemlerin farklı kombinasyonları için bir gözlem ve birden fazla gözlem silme yaklaşımına dayalı Cook Uzaklığı, COVRATIO ve K istatistiklerinin değerleri ve bunlara karşılık gelen kritik değerler

	Gözlem no	$D_{(I)}$	$ COVRATIO_{(I)}-1 $	$K_{(I)}$
Bir elemanlı alt kümeler	4	13,2196*	0,6023	11,3643
	5	65,4663*	0,6316	38,0066*
	7	0,1643	0,9608*	5,0416
	19	0,4682	0,7391	2,6654
Kritik değer		0,9107	0,7500	28,9980
İki elemanlı alt kümeler	4, 5	36,1891*	3,9149*	45,1349*
	4, 7	3,7067*	0,6349	12,3362
	4, 19	9,7997*	0,9264*	7,2894
	5, 7	43,3834*	0,9957*	28,2048*
	5, 19	54,9161*	0,9928*	20,8470
	7, 19	0,4277	0,9993*	7,1599
	Kritik değer	0,9107	0,7500	27,7661
Üç elemanlı alt kümeler	4, 5, 7	29,8921*	0,9046*	58,6247
	4, 5, 19	36,2119*	0,9050*	31,3313
	4, 7, 19	1,2307*	0,9914*	12,4251
	5, 7, 19	23,7297*	0,9790*	52,1885
	Kritik değer	0,9107	0,7500	125,6331
Dört elemanlı alt küme	4, 5, 7, 19	18,3035*	0,6042	52,1885
Kritik değer		0,9107	0,7500	208,0439

* Kritik değerden büyük olan değerler.

Çizelge 6.4'te Cook Uzaklığı ($D_{(I)}$) İstatistiği'ne ilişkin sonuçlar incelendiğinde bu dört gözlemin ikili kombinasyonlarından sadece $I = \{7,19\}$ durumu hariç, tüm durumlar için Cook Uzaklığı değerlerinin kritik değerden büyük olduğu görülmektedir. Özellikle $D_{(5,19)} = 54,9161$ değeri en büyük değer olmak üzere $D_{(4,5)}$ ve $D_{(5,7)}$ değerlerinin büyüklüğü dikkate değerdir. Bu durumda 5 numaralı gözlem regresyon tahminlerini

değiştirme eğilimi en yüksek olan gözlem olarak saptanır. $D_{(4,7)}$ ve $D_{(7,19)}$ değerleri ise diğer değerlere göre küçüktür. Bunun yorumu iki hacimli I kümesinde 7 numaralı gözlemin Cook Uzaklığı değerinde bir küçülmeye neden olduğu biçiminde yapılır. Dolayısıyla Cook Uzaklığında küçülmeye neden olan gözlem, verinin geneline uyan gözlem olarak nitelendirilebilir. Bununla birlikte net bir sonuç için diğer kombinasyonlara ilişkin $D_{(i)}$ değerlerinin de incelenmesinde yarar vardır.

Üçlü kombinasyonlar için Cook Uzaklığı sonuçları incelendiğinde, bu dört değer de kritik değerden büyük olduğu açıktır. Fakat üç hacimli I kümesinde 5 numaralı gözlem olduğunda $D_{(4,5,7)}$, $D_{(4,5,19)}$ ve $D_{(5,7,19)}$ Cook Uzaklığı değerlerinin, I kümesinde 5 numaralı gözlem yokken elde edilen $D_{(4,7,19)}$ Cook Uzaklığı değerinden büyük olduğu da dikkate değerdir. Bu durum 5 numaralı gözlemin regresyon tahminlerini etkileme eğilimi en yüksek olan gözlem olduğunun diğer bir göstergesidir. 7 numaralı gözlemin olmadığı $D_{(4,5,19)}$ değeri üçlü kombinasyonlar bakımından en büyük Cook Uzaklığı değerine sahiptir. Bu da 7 numaralı gözlemin etkili gözlem olmadığına diğer bir göstergesidir.

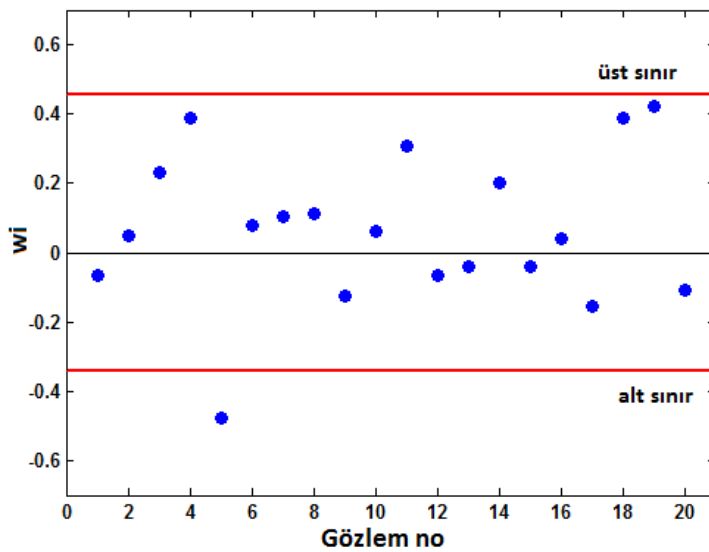
Üçlü kombinasyonlar için hesaplanan Cook Uzaklığı değerlerinde en büyük değerin elde edildiği $I = \{4,5,19\}$ kümesine 7 numaralı gözlemin eklenmesiyle elde edilen yeni Cook Uzaklığı incelendiğinde $D_{(4,5,19)} > D_{(4,5,7,19)}$ olduğu görülmektedir. Bu durumda 7 numaralı gözlemin ilgili I kümesine eklenmesinin Cook Uzaklığı değerlerinde bir düşüşe neden olduğu sonucunu ortaya çıkarmaktadır. Ayrıca $D_{(7)}$ değerinin en küçük değer olması da bu sonucu desteklemektedir. Dolayısıyla 7 numaralı gözlem etkili bir gözlem değildir ve etkili gözlem grubunun $I = \{4,5,19\}$ olduğu söylenebilir. Bu durum, bu tip istatistiklerde karşılaşılan yanılgıya düşürme sorununa bir örnektir. Buna ek olarak bir gözlem silme yaklaşımına dayalı Cook Uzaklığı İstatistiğiyle saptanamayan 19 numaralı gözlem, birden fazla gözlem silme yaklaşımına dayalı Cook Uzaklığı İstatistiğinin genelleştirilmiş formülü kullanılarak etkili gözlem grubu içerisinde etkili olarak saptanmıştır. Bu durum da bu tip istatistiklerde karşılaşılan diğer bir problem olan maskeleye sorununa örnektir.

$COVRATIO_{(i)}$ istatistiği için sonuçlar incelendiğinde $|COVRATIO_{(4,5)} - 1|$ değerinin en büyük olduğu açıkça görülmektedir. Bununla birlikte her ne kadar bir gözlem silme durumunda 7 numaralı gözlem ($I = \{7\}$ iken) etkili olarak saptansa da, $I = \{4,7\}$ ve $I = \{4,5,7,19\}$ iken

elde edilen $|COVRATIO_{(I)}-1|$ değerinde bir azalma gözlenmektedir. Dolayısıyla 7 numaralı gözlemin etkili gözlem olup olmadığına net olarak karar verilemez. Bir Gözlem silme yaklaşımına dayalı $COVRATIO_{(I)}$ İstatistiğiyle etkili gözlem olarak saptanmayan 4, 5 ve 19 numaralı gözlemler birden fazla gözlem silme yaklaşımına dayalı $COVRATIO_{(I)}$ istatistiği kullanılarak $I = \{4,5,19\}$ durumu için etkili gözlem grubu olduğu belirlenmiştir.

I kümesinin farklı hacimleri için K istatistiğine ilişkin kritik değerler de farklı elde edilmektedir. Bunun nedeni I 'nın hacim bakımından her bir kombinasyon için $K_{(I)}$ değerleri yeniden hesaplanacağından, kritik değer formülünde yer alan $\bar{K}$ ve $\hat{\sigma}_K^2$ tahminlerinin de yeniden hesaplanması gerektiğidir. Hadi'nin K istatistiğine ilişkin sonuçlar incelendiğinde, I 'nın farklı kombinasyonları için K istatistiğinin kritik değerlerinin birbirinden farklı olduğu görülmektedir. $K_{(4,5)}$ ve $K_{(5,7)}$ değerleri bunlara karşılık gelen kritik değerden büyük olduğu için $I = \{4,5\}$ ve $I = \{5,7\}$ gözlem grupları etkili gözlem grubu olarak saptanmıştır.

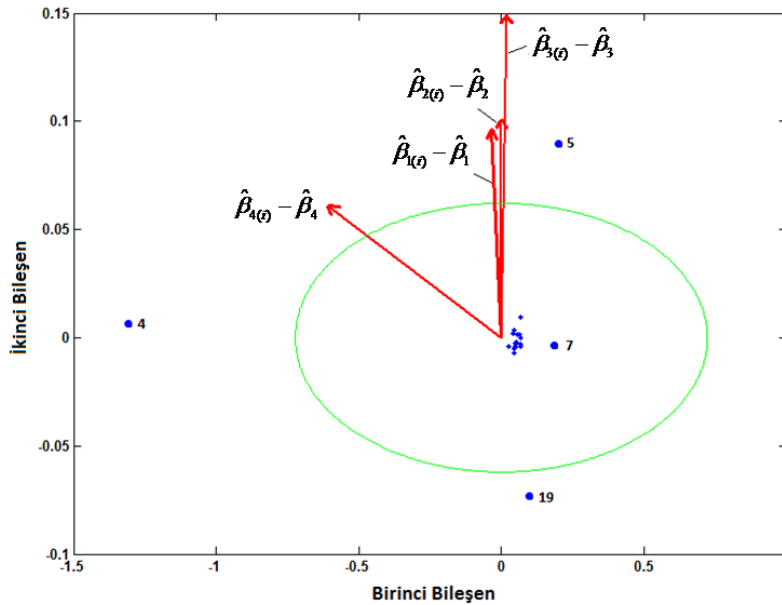
Veri kümesinde en az bir tane yüksek dereceli kaldıraç noktası olduğu için bu çalışmada önerilen ilk yöntem olan SICDGY yöntemi uygulanabilir. i . Gözlemin regresyon tahminlerini değiştirme eğilimine ilişkin katsayı olan w_i değerleri ve kritik değerler elde edilip, SIC'ye dayalı grafik Şekil 6.8'de verilmiştir.



Şekil 6.8. New York'taki Nehirler Verisi için SICDGY ile elde edilen w_i ($i = 1, 2, \dots, 20$) ile gözlem numarası için oluşturulan serpmeye diyagramı ve sınırlar (alt sınır = -0,3412; üst sınır = 0,4551)

Şekil 6.8 incelendiğinde 5 numaralı gözlem sınırların dışında kaldığı için bu etkili gözlem olarak belirlenir. Bununla birlikte etkili gözlem olarak saptanmayan 4 ve 19 numaralı gözlemlerin üst sınıra yakın olması da açıkça görülmektedir.

BYDGY yöntemi New York'taki Nehirler Verisi için uygulanmadan önce h_{ii} değerleri incelenmiş ve bu veri kümesinin en az bir yüksek dereceli kaldırma noktası içerdiği belirlenmiştir. Sonra β_0 parametresinin tahminini, tahmin edilen modelden çıkarmak amacıyla bağımlı değişken ve bağımsız değişkenlerin değerleri standartlaştırılmıştır. Standartlaştırılmış bu veriden elde edilen regresyon modeli tahminler yine $\hat{\beta}_1$, $\hat{\beta}_2$, $\hat{\beta}_3$ ve $\hat{\beta}_4$ ile gösterilsin. Bu durumda i . gözlem veri kümesinden çıkartıldıktan sonra geri kalan gözlemlerden elde edilen regresyon katsayısı $\hat{\beta}_{j(i)}$ olmak üzere, i . gözlemin j . tahmini üzerindeki değişim değeri $\hat{\beta}_{j(i)} - \hat{\beta}_j$ ile gösterilir. Bölüm 5'te tanımlanan Ψ ve Δ matrisleriyle oluşturulan biplotlar aynı bilgiyi sağladıkları için bu matrislerden $\hat{\beta}_{j(i)} - \hat{\beta}_j$ farklarının matrisi olan Δ ile biplot oluşturulsun. New York'taki Nehirler Verisi için $\hat{\beta}_{j(i)} - \hat{\beta}_j$ tahminlerine dayalı biplot Şekil 6.9'da verilmiştir.



Şekil 6.9. New York'taki Nehirler Verisi için $\hat{\beta}_{j(i)} - \hat{\beta}_j$ tahminlerine dayalı biplot

Şekil 6.9’da verilen grafikte dört adet vektör $\hat{\beta}_{j(i)} - \hat{\beta}_j$ değişkenlerine ilişkin eksenlere karşılık gelmektedir. Bu grafik değişkenler bakımından incelenirse $\hat{\beta}_{1(i)} - \hat{\beta}_1$, $\hat{\beta}_{2(i)} - \hat{\beta}_2$ ve $\hat{\beta}_{3(i)} - \hat{\beta}_3$ vektörleri arasındaki açı 0° ’ye yakın olduğundan bu değişkenler arasında ikili ilişkilerin aynı yönlü güçlü olduğu açıkça görülmektedir. Nitekim bu açılar 180° ’ye yakın olsaydı yine güçlü bir ilişkiden söz edilebilirdi. Fakat bu durumda ilişki ters yönlü olurdu. Bununla birlikte $\hat{\beta}_{4(i)} - \hat{\beta}_4$ ile diğer üç değişken arasındaki açı dik açığa yakın olduğu için bu değişkenle diğer üç değişken arasındaki ikili ilişkiler güçsüzdür.

Şekil 6.9’da verilen grafik gözlemler bakımından incelendiğinde 4, 5 ve 19 numaralı gözlemlere ilişkin noktalar, geri kalan tüm gözlemlere ilişkin noktaların geneline uymamaktadır. Dolayısıyla bu noktalar sınırın dışında olduğundan bunlara karşılık gelen gözlemler etkili gözlemler olarak saptanmaktadır. 7 numaralı nokta ise verinin merkezinden uzak olmasına rağmen sınırın içinde kalmaktadır. Dolayısıyla bu gözlem regresyon tahminlerini değiştirme eğilimi yüksek olmayan bir gözlemdir ve etkili gözlem olarak nitelendirilemez. Önerilen bu yöntem değerlendirildiğinde maskeleye ve yanılgıya düşürme sorunlarıyla karşılaşmadığı sonucuna varılır. New York’taki Nehirler Verisi için Biplotta her bir gözleme karşılık gelen noktanın merkeze uzaklıkları Mahalanobis Uzaklığıyla hesaplanmış ve Çizelge 6.5’te verilmiştir.

Çizelge 6.5. New York’taki Nehirler Verisi için biplotta noktalara ilişkin Mahalanobis Uzaklıklarının karesi ve kritik değer

Gözlem no	ξ^2	Gözlem no	ξ^2
1	0,0368	11	0,0368
2	0,0711	12	0,0478
3	0,0569	13	0,0530
4	17,7567	14	0,0451
5	11,5610	15	0,0917
6	0,0302	16	0,0327
7	0,3652	17	0,1795
8	0,0340	18	0,0228
9	0,0435	19	7,4504
10	0,0493	20	0,0355
Kritik değer		5,3614	

Çizelge 6.5’teki sonuçlar incelendiğinde 4, 5 ve 19 numaralı gözlemlere karşılık gelen uzaklıkların kritik değerden büyük olduğu açıktır. Dolayısıyla bu üç gözlemin etkili gözlem olduğu sonucuna tekrar varılır.

7. SİMÜLASYON ÇALIŞMASI

Etkili gözlemlerin saptanması için önerilen tanı yöntemlerinin uygulaması genellikle literatürde etkili gözlem içerdiği bilinen veri kümelerini kullanarak yapılır. Bununla birlikte etkili gözlemlerin saptanması için önerilen tanı yöntemleri, bilinen yöntemlerle simülasyon çalışması ile karşılaştırılır.

Önerilen bir test istatistiği için simülasyon tasarımı, önceden seçilen çeşitli örnek hacmi, parametre sayısı ve parametre değerleri gibi sabitlere dayalı olarak deneysel I. tip hata veya deneysel güç bakımından durağanlık sağlanana kadar gerekli adımların tekrarına dayanır. Bununla birlikte regresyonda etkili gözlemlerin saptanması için önerilen tanı istatistiklerinin bilinen diğer tanı istatistikleri ile karşılaştırılmasındaki simülasyon tasarımı farklıdır, çünkü bu yöntemler n hacimli örnekteki gözlemlerin teker teker regresyon tahminleri üzerindeki etkilerini incelemeye dayalıdır. Dolayısıyla bu çalışmadaki simülasyon tasarımı, klasik simülasyon tasarımlarından farklı olarak belirlenen n ve p değerleri için ele alınan tanı yöntemlerinin etkili gözlemi veya etkili gözlem grubunu saptama oranlarını elde etmeyi amaçlayacak biçimde oluşturulmuştur. Buna ek olarak simülasyon tasarımında veri üretme kısmı Hadi ve Simonoff (1993) tarafından yapılan çalışmadaki yapay veri üretme yaklaşımına dayalı olarak tasarlanmıştır. Bununla birlikte Hadi ve Simonoff (1993) tarafından yapılan çalışmada veri aykırı gözlem içermek amacıyla üretildiği için bu çalışmada aynı yöntemle veri etkili gözlem veya etkili gözlem grubu içerecek biçimde MATLAB2011 programı kullanılarak üretilmiştir.

Bu bölümde önce bir etkili gözlemin saptanması için kullanılan Cook Uzaklığı ($D_{(i)}$), $DFBETAS_{j(i)}$, $DFITS_{(i)}$, $COVRATIO_{(i)}$, $AP_{(i)}$, $LD_{(i)}$, $K_{(i)}$ ve $S_{(i)}$ istatistikleri ile bu çalışmada önerilen SIC'ye dayalı grafik yöntem (SICDGY) ve Biplot yaklaşımına dayalı grafik yöntem (BYDGY) örnek hacmi ve parametre sayısının farklı durumları için simülasyon yöntemiyle karşılaştırılmıştır. Bu karşılaştırma için her bir denemede veri kümesi, bir etkili gözlem içerecek biçimde üretilmiştir. Sonra bir etkili gözlem grubunun saptanması için kullanılan Cook Uzaklığı ($D_{(i)}$), $COVRATIO_{(i)}$ ve $K_{(i)}$ istatistikleri ile SICDGY ve BYDGY; örnek hacmi ve parametre sayısının farklı durumları için simülasyon yöntemiyle karşılaştırılmıştır. Bu karşılaştırmada ise her bir denemede veri kümesi, etkili bir gözlem grubunu temsil eden iki etkili gözlemi içerecek biçimde üretilmiştir.

Bölüm 3’te ele alınan tanı istatistiklerinin formülleri ve karar kuralları Çizelge 7.1’de özetlenmiştir. Bu çizelgede Cook Uzaklığı, COVRATIO ve Hadi’nin K istatistiklerinin genelleştirilmiş biçimleriyle aynı zamanda etkili gözlem grupları da saptanabilmektedir.

Çizelge 7.1. Bir gözlem silmeye dayalı tanı istatistikleri ve kritik değerleri

Tanı istatistiği	Formül	Karar kuralı*
Cook Uzaklığı İstatistiği	$D_{(i)} = \frac{(\hat{\beta}_{(i)} - \hat{\beta})^T X^T X (\hat{\beta}_{(i)} - \hat{\beta})}{p \hat{\sigma}^2}$	$D_{(i)} > F_{0.50, p, n-p}$
DFBETAS İstatistiği	$DFBETAS_{j(i)} = \frac{\hat{\beta}_j - \hat{\beta}_{j(i)}}{\hat{\sigma}_{(i)} \sqrt{(X^T X)^{-1}_{jj}}}$	$ DFBETAS_{j(i)} > \frac{2}{\sqrt{n}}$
DFFITS İstatistiği	$DFFITS_{(i)} = \frac{\hat{y}_i - \hat{y}_{i(i)}}{\hat{\sigma}_{(i)} \sqrt{h_{ii}}}$	$ DFFITS_{(i)} > 2\sqrt{\frac{p}{n}}$
Kovaryans Oranı İstatistiği	$COVRATIO_{(i)} = \frac{\det\{\hat{\sigma}_{(i)}^2 (X_{(i)}^T X_{(i)})^{-1}\}}{\det\{\hat{\sigma}^2 (X^T X)^{-1}\}}$	$ COVRATIO_{(i)} - 1 > \frac{3p}{n}$
Andrews ve Pregibon’un İstatistiği	$AP_{(i)} = \left \frac{Z_{(i)}^T Z_{(i)}}{Z^T Z} \right $	$(1 - AP_{(i)}) > \frac{2(p+1)}{n}$
Cook ve Weisberg’in Olabilirlik İstatistiği	$LD_{(i)} = n \ln \left(\frac{n}{n-1} \frac{n-p-1}{t_i^2 + n-p-1} \right) + \frac{t_i^2 (n-1)}{(1-h_{ii})(n-p-1)} - 1$	$LD_{(i)} > \chi_{v, \alpha}^2$
Hadi’nin İstatistiği	$K_{(i)} = \frac{p}{(1-h_{ii})} \frac{e_i^2}{(e^T e - e_i^2)} + \frac{h_{ii}}{(1-h_{ii})}$	$ K_{(i)} > \bar{K} + 3\sqrt{\hat{\sigma}_K^2}$
Pena’nın İstatistiği	$S_{(i)} = \frac{1}{p \hat{\sigma}^2 h_{ii}} \sum_{i'=1}^n \frac{h_{ii'}^2 e_{i'}^2}{(1-h_{ii'})^2}$	$ S_{(i)} - \text{med}(S) > 4.5 \times \text{MAD}(S)$

* Koşul sağlandığı durumda i . gözlem etkili gözlemdir.

DFBETAS, DFFITS, COVRATIO ve AP istatistiklerinin karar kurallarında yer alan kritik değerler incelendiğinde, bu kritik değerlerin örnek hacminin azalan bir fonksiyonu olduğu açıktır. Bundan dolayı örnek hacmi arttıkça bu kritik değerler küçülecektir. Böylece kritik değerler küçüldüğünden regresyon tahminlerini değiştirme eğilimi olmayan ve verinin geneline uyan gözlemler, etkili gözlem olarak saptanabilir. Bu durumda yanılgıya düşürme durumuyla karşılaşılır. Literatürde bu tip kritik değere sahip tanı istatistiklerinin, hacmi en fazla 100 olan örnekler için uygulanması önerilmektedir (Belsley ve diğerleri, 1980: 28,39). Ancak örnek hacminin büyük olduğu durumlarda gerçekte veri kümesinde bir veya daha

fazla etkili gözlem varsa, verinin geneline uyan diğer gözlemlerin çok sayıda olmasından dolayı etkili gözlemlerin regresyon tahminlerini değiştirme eğilimleri azalır.

Bu kısımda yer alan simülasyon için örnek hacminin 20, 30 ve 40; sabit terim de dahil olmak üzere parametre sayısının 3, 4 ve 5 olduğu durumlar ele alınmıştır. İlk simülasyon sonuçları veri kümesinde bir etkili gözlem olduğu durumda bir gözlem silme yaklaşımına dayalı tanı yöntemleriyle bu etkili gözlemin saptanma oranlarını ve bu yöntemlerle etkili biçimde üretilmeyen gözlemlerin en az bir tanesinin etkili gözlem olarak saptanma oranlarını içermektedir. Buna benzer olarak ikinci simülasyon sonuçları ise veri kümesinde iki etkili gözlem olduğu durumda çoklu gözlem silme yaklaşımına dayalı tanı yöntemleriyle bu etkili gözlem grubunun saptanma oranlarına ve etkili gözlem olarak üretilmeyen gözlemlerin ikili alt kümelerinden en az bir tanesinin etkili gözlem grubu olarak saptanma oranlarına ilişkindir. Üçüncü kısımda örnek hacmi 20 ve parametre sayısı 3 iken veri kümesinde etkili gözlem veya gözlem grubu, bağımsız değişkenlerin değerleri bakımından verinin genelinden daha da uzaklaştırılmak amacıyla üretildiğinde elde edilen saptama oranlarına yer verilmiştir. Son kısımda ise yine örnek hacmi 20 ve parametre sayısı 3 iken veri kümesinde etkili gözlem veya gözlem grubu hata terimi bakımından verinin genelinden daha da uzaklaştırılmak amacıyla üretildiğinde tanı yöntemlerine ilişkin saptama oranları yer almaktadır.

Bilindiği gibi regresyonda bağımsız değişkenler genellikle sabittir. Bununla birlikte tanı yöntemlerinin çoğu tasarım matrisinden elde edilen H matrisinin i . köşegen elemanı olan h_{ii} değişkeninin fonksiyonu olduğundan bağımsız değişken değerleri ilgili tanı istatistiğinin değerinde belirleyici bir role sahiptir. Dolayısıyla bu çalışmadaki simülasyonda her 1000 tekrarda bağımsız değişkenlerin değerleri yeniden üretilmiştir. Ayrıca simülasyona katılan iki koşul ile verinin geneline uyan gözlemlerin yüksek dereceli kaldıraç noktası ve artıklar bakımından aykırı gözlem olması engellenmiştir.

7.1. Bir Gözlem Silme Yaklaşımına Dayalı Tanı İstatistikleri ile Önerilen İki Grafik Yönteminin Karşılaştırılması

Bu kısımda gerçekte veri kümesinde bir etkili gözlem varken, örnek hacmine ve parametre sayısına bağlı olarak bir gözlem silme yaklaşımına dayalı tanı istatistikleri olan Cook Uzaklığı, DFBETAS, DFFITS, COVRATIO, Andrews ve Pregibon'un AP İstatistiği, Cook

ve Weisberg'in Olabilirlik İstatistiği LD, Hadi'nin K İstatistiği, Pena'nın S istatistiği ile bu çalışmada önerilen SIC'ye Dayalı Grafik Yöntemi (SICDGY) ve Biplot Yaklaşımına Dayalı Grafik Yöntemi (BYDGY) için etkili gözlemi saptama oranları simülasyonla elde edilmiştir. Ayrıca bir gözlem silme yaklaşımına dayalı klasik tanı yöntemleriyle önerilen iki yöntem için, etkili gözlem haricinde, regresyon tahminleri üzerinde etkisi olmamak üzere üretilen diğer gözlemlerin en az bir tanesini etkili olarak saptama oranları da simülasyonla elde edilmiştir. Sonra her bir tanı yöntemi için elde edilen simülasyon sonuçları ayrı ayrı ele alınarak iki yönlü ANOVA yöntemi uygulanmış ve sonuçlar yorumlanmıştır. Veri kümesinde bir etkili gözlemin olduğu durum için simülasyon aşağıdaki adımlar doğrultusunda yapılmıştır.

Adım 1: Verinin geneline uyan $n-1$ adet gözlemin $p-1$ adet bağımsız değişkeninin değerleri, bu gözlemler yüksek dereceli kaldıraç noktası olmayacak biçimde, $[0,15]$ aralığında tekdüze dağılımdan üretilir.

Adım 2: Verinin geneline uyan $n-1$ adet gözlemin hata değişken değerleri, bu gözlemler artıklar bakımından aykırı bir değer olmayacak biçimde, $\varepsilon_i \sim N(0,1)$ dağılımından üretilir.

Adım 3: Etkili gözlem i' ile ifade edilmek üzere, bu gözlem için $p-1$ adet bağımsız değişkeninin değerleri $x_{ji'} = 20$, hata terimi $\varepsilon_{i'} = -5$ olarak belirlenir.

Adım 4: $j = 0,1,..., p-1$ iken her $\beta_j = 1$ olmak üzere i' ile ifade edilen etkili gözlemin de bulunduğu n adet gözlem için bağımlı değişken değerleri

$$y_i = \beta_0 + \beta_1 x_{1i} + ... + \beta_{p-1} x_{(p-1)i} + \varepsilon_i \quad (i = 1,2,...,n) \quad (7.1)$$

modeline göre üretilir.

Adım 5: Üretilen veri kullanılarak Çizelge 7.1'de yer alan tanı istatistiklerinin değerleri ve bunlara karşılık gelen kritik değerler hesaplanır. i' ile ifade edilen gözlemin ve diğer gözlemlerin etkili olup olmadığına karar verilir.

Adım 6: SICDGY ve BYDGY ile grafikler oluşturulur. i' ile ifade edilen gözlemin ve diğer gözlemlerin etkili olup olmadığına karar verilir.

Adım 7: Eğer deneme numarası 1000'in katı ise *Adım 1*'e gidilir, değilse *Adım 2*'den devam edilir.

Simülasyon yapılırken etkili gözlemi saptama oranlarının durağanlaştığı deneme sayısının yaklaşık olarak 100 000 olduğu gözlenmiştir. Dolayısıyla deneme sayısı 100 000 olmak üzere, n örnek hacminin ve p parametre sayısının farklı durumları için tanı istatistiklerinin ve önerilen iki grafik yönteminin etkili gözlem olan i' gözlemi saptama oranları MATLAB2011 programı kullanılarak hesaplanmıştır. Elde edilen sonuçlar Çizelge 7.2'de verilmiştir.

Çizelge 7.2. Bir gözlem silme yaklaşımına dayalı klasik tanı istatistikleri ile BYDGY ve SICDGY için etkili gözlemi saptama oranları

Parametre sayısı	Tanı yöntemi	Örnek Hacmi		
		$n = 20$	$n = 30$	$n = 40$
$p = 3$	SICDGY	0,7830	0,9180	0,9480
	BYDGY	0,9794	0,9901	0,9985
	$D_{(i')}$	0,9424	0,9665	0,9866
	$DFBETAS_{j(i')} (j = 0,1,\dots,p)$	0,9586	0,9658	0,9902
	$DFFITs_{(i')}$	0,9696	0,9885	0,9956
	$COVRATIO_{(i')}$	0,6594	0,8220	0,8993
	$AP_{(i')}$	0,9554	0,9698	0,9936
	$LD_{(i')}$	0,9209	0,8926	0,8436
	$K_{(i')}$	0,9742	0,9892	0,9977
	$S_{(i')}$	0,8872	0,8568	0,8366
$p = 4$	SICDGY	0,7590	0,9160	0,9450
	BYDGY	0,9881	0,9990	0,9995
	$D_{(i')}$	0,9295	0,9527	0,9766
	$DFBETAS_{j(i')} (j = 0,1,\dots,p)$	0,8604	0,9763	0,9953
	$DFFITs_{(i')}$	0,9674	0,9797	0,9856
	$COVRATIO_{(i')}$	0,5643	0,7080	0,7914
	$AP_{(i')}$	0,9496	0,9655	0,9929
	$LD_{(i')}$	0,8840	0,8598	0,8118
	$K_{(i')}$	0,9379	0,9871	0,9910
	$S_{(i')}$	0,9627	0,9414	0,9275
$p = 5$	SICDGY	0,7340	0,8800	0,9370
	BYDGY	0,9363	0,9937	0,9993
	$D_{(i')}$	0,8992	0,9471	0,9615
	$DFBETAS_{j(i')} (j = 0,1,\dots,p)$	0,6129	0,8604	0,9596
	$DFFITs_{(i')}$	0,9769	0,9825	0,9913
	$COVRATIO_{(i')}$	0,5064	0,6227	0,7204
	$AP_{(i')}$	0,9561	0,9699	0,9922
	$LD_{(i')}$	0,8483	0,8279	0,7808
	$K_{(i')}$	0,8974	0,9617	0,9875
	$S_{(i')}$	0,9839	0,9625	0,9585

Çizelge 7.2’de verilen simülasyon sonuçları incelendiğinde p ’nin tüm durumları için örnek hacmi arttıkça Cook Uzaklığı İstatistiği’nin etkili gözlemi saptama oranının değer olarak büyüdüğü görülmektedir. Bu durum, p sabitken örnek hacmi arttıkça, Cook Uzaklığı için kritik değer olan $F_{0.50,p,n-p}$ değerinin ikinci serbestlik derecesindeki artışın bu değeri küçülmesi ve ilgili istatistikle etkili gözlemi daha iyi bir şekilde saptanmasının sağlanması biçiminde açıklanabilir. Benzer şekilde örnek hacmi n sabitken parametre sayısı olan p ’nin değeri arttıkça Cook Uzaklığı İstatistiği’nin etkili gözlemi saptama oranının küçüldüğü görülmektedir. Bu küçülme özellikle $n=20$ durumunda belirgin bir biçimdedir. Bu durum, parametre sayısı arttıkça $F_{0.50,p,n-p}$ kritik değerinin büyümesi ve D istatistiğinin etkili gözlemi saptama eğiliminin azalmasıyla açıklanabilir.

DFBETAS, DFFITS, COVRATIO ve AP istatistiklerine ilişkin etkili gözlemi saptama oranları incelendiğinde p sabitken n arttıkça bu oranların değerce büyüdüğü görülmektedir. Bu istatistiklerin kritik değerleri örnek hacminin azalan bir fonksiyonu olduğundan; bu oranlardaki artış, örnek hacmi arttıkça bu istatistiklerin kritik değerlerinin küçülmesi ve etkili gözlemin daha kolay saptanması biçiminde yorumlanabilir.

n sabitken p arttıkça DFBETAS istatistiğine ilişkin oranlarda bir düşme gözlenmektedir. Bunun nedenini açıklamak amacıyla öncelikle bu istatistiğin karar kuralını yorumlamak gereklidir. i . gözlemin bu istatistikle etkili gözlem olarak saptanması için gerekli koşul, bu gözlem için hesaplanan tüm DFBETAS değerlerinin mutlak değerlerinin, kritik değerden büyük olmasıdır. Dolayısıyla p arttıkça i . gözlem için bir DFBETAS değeri daha hesaplanmakta ve i . gözlem için DFBETAS değerlerinin sayısı arttıkça p adet DFBETAS değerlerinin tümüne dayalı olarak etkili gözlemin saptanması daha da zorlaşmaktadır. Nitekim bu durum bir önceki kısımda verilen uygulama sonuçlarında da gözlenmiştir.

n sabitken DFFITS istatistiğine ilişkin oranlar incelendiğinde, bu değerlerdeki değişimin diğer klasik tanı istatistiklerine göre dikkate değer olmadığı görülmektedir. Ayrıca örnek hacmi ve parametre sayısının tüm durumlarında DFFITS için elde edilen oranlar sayısal olarak büyüktür. DFFITS için kritik değer formülünde yer alan $\sqrt{p/n}$ teriminin, kritik değeri çok küçültmesi, bu istatistiğin etkili gözlemi saptama eğiliminin artmasına neden olmaktadır. Fakat bu kritik değerın küçük olması regresyon tahminlerini değiştirme eğilimi olmayan bir gözlemin de etkili gözlem olarak saptanmasına yol açabilir.

Parametre sayısının artmasıyla COVRATIO istatistiğine ilişkin oranlarda azalma gözlenmektedir. Ayrıca n ve p değerlerinin birçok durumunda, COVRATIO istatistiği için elde edilen oranların diğer istatistiklere ilişkin oranlara göre küçük olduğu görülmektedir. Nitekim kovaryans oranlarına dayalı olan bu istatistiğin küçük oranlara sahip olması; çoğu denemede veri kümesinde sadece bir etkili gözlem varken bu gözlemin regresyon parametrelerinin tahminlerine ilişkin kovaryans matrisinde dikkate değer bir değişime neden olmaması biçiminde yorumlanabilir. Bu durumda etkili gözlem için hesaplanan COVRATIO değeri küçülmekte ve bu gözlemin etkili olarak saptanmasını zorlaştırmaktadır.

n sabitken parametre sayısı arttıkça, Andrews ve Pregibon'un AP İstatistiği için elde edilen oranlarda dikkate değer bir değişim gözlenmemiştir. p sabitken örnek hacmi arttıkça, AP istatistiği için elde edilen oranlarda artış gözlenmektedir. Bu artışın nedeni, AP istatistiğinin kritik değerinin, n örnek hacminin azalan bir fonksiyonu olması biçiminde yorumlanabilir. Tüm durumlarda bu istatistiğin etkili gözlemi saptama oranları da klasik tanı istatistiklerinden DFFITS için elde edilen oranlar gibi büyüktür.

p sabitken örnek hacmi arttıkça, Cook ve Weisberg'in Olabilirlik İstatistiği LD için elde edilen etkili gözlemi saptama oranında azalma gözlenmektedir. Benzer olarak n sabitken parametre sayısı arttıkça, bu istatistik için elde edilen etkili gözlemi saptama oranında da azalma görülmektedir. Bu azalışın nedeni, p arttıkça bu istatistik için kritik değer olan $\chi^2_{v,\alpha}$ değerinin serbestlik derecesinin artması, buna bağlı olarak kritik değerin artmasıyla LD istatistiğiyle etkili gözlemin saptama eğiliminin azalması biçiminde yorumlanabilir.

Hadi'nin K İstatistiği için simülasyon sonuçları incelendiğinde p sabitken n arttıkça bu istatistiğin etkili gözlemi saptama oranının da arttığı görülmektedir. Fakat $p=3$ iken örnek hacmini farklı durumlarında bu istatistik için elde edilen oranlar değer olarak birbirlerine yakındır. n sabitken p arttıkça, K İstatistiği'nin etkili gözlemi saptama oranında bir düşüş gözlenmektedir. Özellikle bu düşüş $n=20$ iken oldukça açıktır. p sabitken n arttıkça Pena'nın S İstatistiği için etkili gözlemleri saptama oranında açık bir şekilde düşüş gözlenmektedir. Bununla birlikte n sabitken p arttıkça bu oranlarda bir artış görülmektedir.

Çizelge 7.2'deki sonuçlardan, bu çalışmada önerilen yöntemlerden biri olan SIC'ye dayalı grafiksel yöntem için elde edilen etkili gözlemi saptama oranları incelendiğinde, p sabitken örnek hacmi arttıkça bu yöntem ile etkili gözlemi saptama oranının arttığı görülmektedir.

Özellikle $n=40$ iken SICDGY için oranlar 0,95 civarındadır. $n=20$ iken p arttıkça SICDGY için etkili gözlemi saptama oranında küçülme olmasına rağmen $n=30$ ve $n=40$ iken önerilen bu yöntem için oranlarda dikkate değer bir değişim gözlenmemektedir. Çizelge 7.2'deki sonuçlardan görüldüğü üzere n ve p 'nin farklı durumlarında Biplot Yaklaşımına Dayalı Grafikselsel Yöntem için etkili gözlemi saptama oranlarının genel anlamda diğer yöntemler için elde edilen oranlara göre büyük olduğu açıktır. Bu durum önerilen bu yöntemin veri kümesinde yer alan bir etkili gözlemi saptamada iyi çalıştığının göstergesidir.

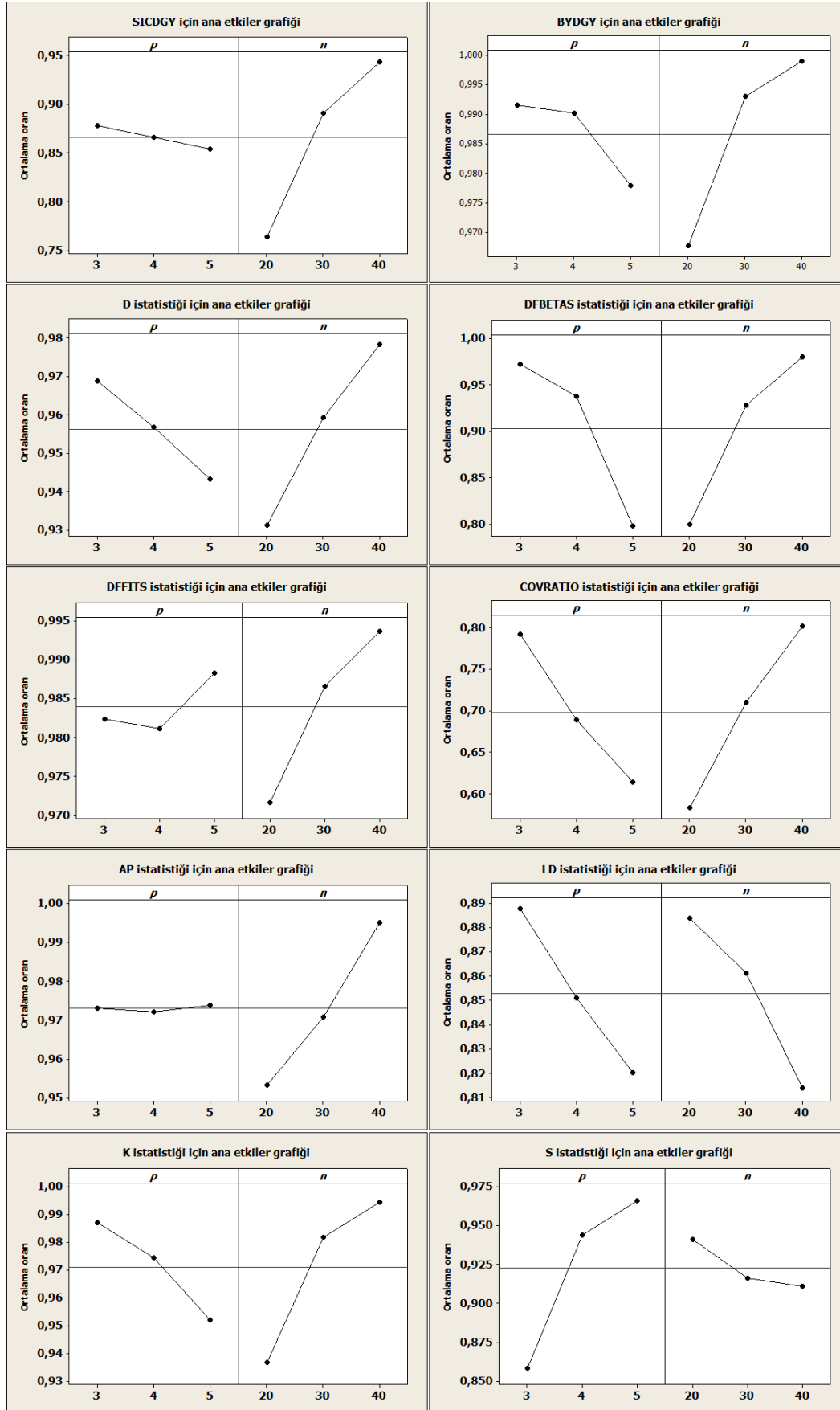
Çizelge 7.2'de verilen simülasyon sonuçları yorumlama kolaylığı sağlamak amacıyla n ve p ana faktörler olarak ele alınıp etkili gözlemi saptama oranı bakımından her yöntem için ayrı ayrı iki yönlü varyans analizi (ANOVA) uygulanmıştır. Bununla birlikte örnek hacmi ve parametre sayısının farklı değerlerinin ikili kombinasyonunu oluşturan 9 farklı durum için simülasyonla elde edilen birer adet oran söz konusudur. Ele alınan iki ana faktörün etkileşim etkisini incelemek amacıyla aynı simülasyon çalışması tekrar yapıp 9 farklı durum için ek olarak birer adet oran daha elde edilmiştir. Böylece ANOVA için bağımlı değişken olarak ele alınan saptama oranına ilişkin varyans analizi tasarımında iki tekrar söz konusu olup, iki ana faktörün etkileşim etkisinin istatistiksel olarak anlamlı olup olmadığının incelenmesi sağlanmıştır. Varyans analizi için deney tasarımı oluşturulduktan sonra n ve p faktörleri ile bu iki faktörün etkileşim etkisinin istatistiksel olarak anlamlılığını test etmek amacıyla İki Yönlü ANOVA yöntemiyle p -değerleri elde edilmiştir. Sonuçlar Çizelge 7.3'teki gibi sunulmuştur.

Çizelge 7.3. Bir etkili gözlemin bulunduğu veri kümesinde etkili gözlemi saptama oranları bakımından iki yönlü varyans analiziyle elde edilen p -değerleri

Tanı yöntemi	Faktörler		Etkileşim
	p	n	
SICDGY	0,038*	0,000*	0,409
BYDGY	0,015*	0,000*	0,004*
$D_{(i')}$	0,004*	0,000*	0,738
$DFBETAS_{j(i')} (j = 0,1 \dots p)$	0,000*	0,000*	0,000*
$DFFITs_{(i')}$	0,148	0,000*	0,421
$COVRATIO_{(i')}$	0,000*	0,000*	0,220
$AP_{(i')}$	0,821	0,000*	0,499
$LD_{(i')}$	0,000*	0,000*	0,259
$K_{(i')}$	0,000*	0,000*	0,000*
$S_{(i')}$	0,000*	0,001*	0,260

* p -değeri $< 0,05$ olduğu sonuçlar

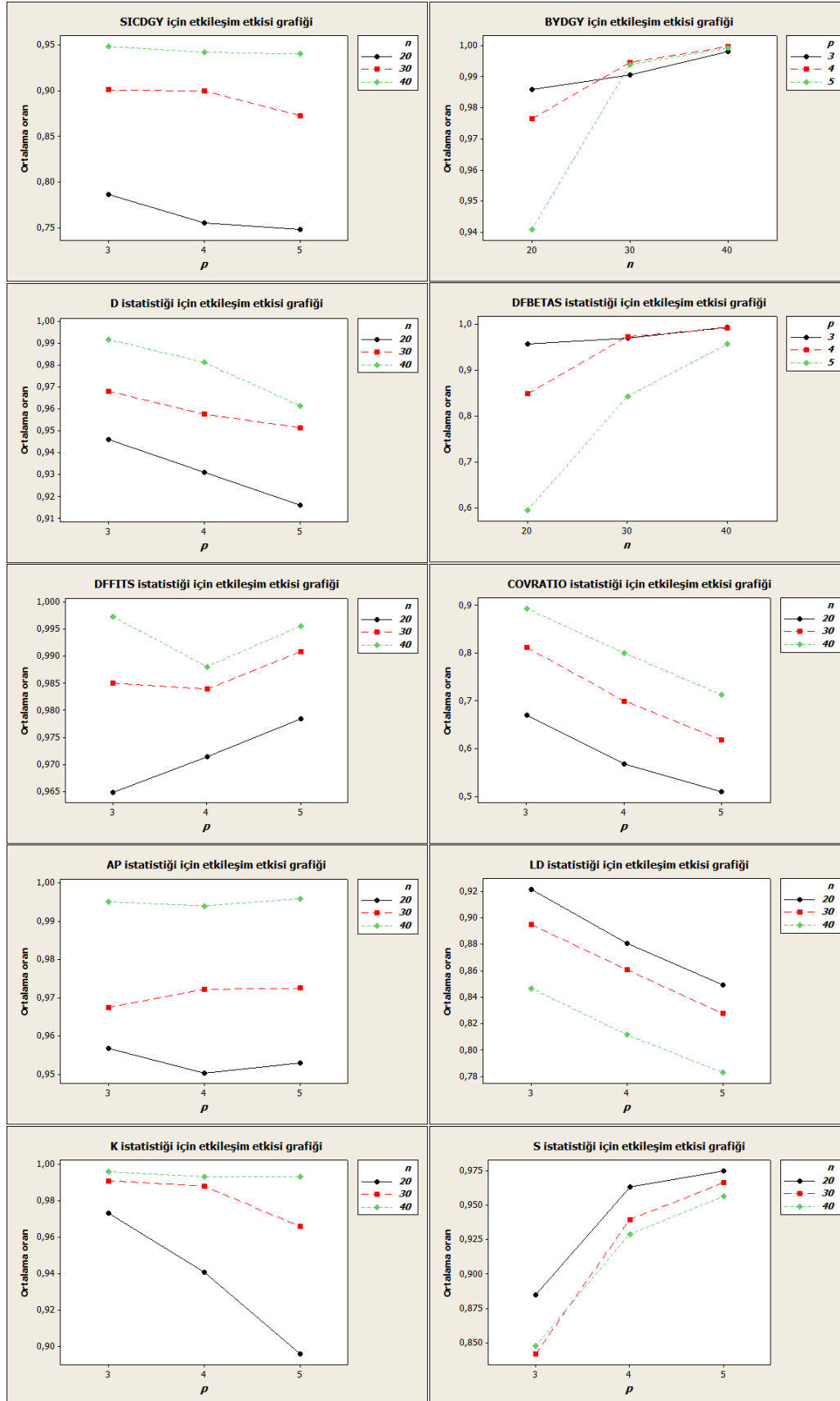
Çizelge 7.3'te İki Yönlü ANOVA yöntemiyle elde edilen p-değeri sonuçları incelendiğinde SIC'ye Dayalı Grafik Yöntemi, Biplot Yaklaşımına Dayalı Grafik Yöntemi, Cook Uzaklığı, DFBETAS, COVRATIO, LD, Hadi'nin K ve Pena'nın S İstatistikleri için elde edilen oranlar üzerinde parametre sayısının etkisinin istatistiksel olarak anlamlı olduğu açıkça görülmektedir. Buna ek olarak ele alınan tüm tanı yöntemleri için elde edilen oranlar üzerinde örnek hacminin etkisinin anlamlı olduğu söylenebilir. Son olarak Çizelge 7.3'te etkileşim etkisi için p-değeri sonuçları incelendiğinde, Biplot Yaklaşımına Dayalı Grafik Yöntemi, DFBETAS ve Hadi'nin K İstatistikleri için n ve p faktörlerinin etkileşim etkisinin istatistiksel olarak anlamlı olduğu sonucuna varılmıştır. Ana etkilere ilişkin anlamlı sonuçlar dikkate alınırsa; örnek hacmi ve parametre sayısındaki artışa dayalı olarak tanı yöntemleriyle etkili gözlemi saptama oranlarındaki değişimi görsel olarak değerlendirmek amacıyla ana etki grafikleri oluşturulmuş ve Şekil 7.1'de verilmiştir.



Şekil 7.1. Bir etkili gözlemin bulunduğu veri kümesinde etkili gözlemi saptama oranları bakımından iki yönlü varyans analizinden sonra oluşturulan ana etki grafikleri

Şekil 7.1'den de görüldüğü üzere n sabitken SICDGY, BYDGY, Cook Uzaklığı, DFBETAS, COVRATIO, LD ve Hadi'nin K İstatistiği yöntemleri için parametre sayısı arttıkça etkili gözlemi saptama oranlarının azaldığı sonucuna varılır. n sabitken parametre sayısı arttıkça Pena'nın S istatistiği için etkili gözlemi saptama oranının arttığı da gözlenmektedir. p sabitken örnek hacmi arttıkça SICDGY, BYDGY ile D, DFBETAS, DFFITS, COVRATIO, AP ve K İstatistikleri için etkili gözlemi saptama oranlarında artış gözleniyorken; LD ve S İstatistikleri için etkili gözlemi saptama oranlarında azalış gözlenmektedir.

n ve p faktörlerinin etkileşim etkisini görsel olarak incelemek amacıyla ele alınan tanı yöntemlerine ilişkin etkileşim etkisi grafikleri oluşturulup; Şekil 7.2'de sunulmuştur.



Şekil 7.2. Bir etkili gözlemin bulunduğu veri kümesinde etkili gözlemi saptama oranları bakımından iki yönlü varyans analizinden sonra oluşturulan etkileşim etkisi grafikleri

Şekil 7.2 incelendiğinde Biplot Yaklaşımına Dayalı Grafik Yöntemi, DFBETAS İstatistiği ve Hadi'nin K İstatistiği için grafiklerde her üç çizginin veya uzantılarının ortak olarak kesiştiği görülmektedir. Dolayısıyla bu istatistikler için elde edilen etkili gözlemi saptama oranları bakımından n ve p faktörlerinin etkileşiminin mevcut olduğu sonucuna varılır.

Bu çalışmada önerilen iki yöntemle bir gözlem silme yaklaşımına dayalı klasik tanı yöntemlerinin etkili olmamak üzere üretilen gözlemlerin en az birini etkili olarak saptama bakımından incelenmesi bu yöntemlerin karşılaştırılması için bir başka bakış açısı sağlayabilir. Bu amaçla bu çalışmada önerilen yöntemlerle klasik tanı yöntemleri için etkili olmayan gözlemlerin en az birini etkili olarak saptama oranları sunulmuştur.

Çizelge 7.4. Bir gözlem silme yaklaşımına dayalı tanı istatistikleri ile BYDGY ve SICDGY için etkili olmayan gözlemlerden en az birini etkili olarak saptama oranları

Parametre sayısı	Tanı yöntemi	Örnek Hacmi		
		$n = 20$	$n = 30$	$n = 40$
$p = 3$	SICDGY	0,2320	0,2550	0,2670
	BYDGY	0,1098	0,1368	0,1583
	$D_{(i')}$	0,1311	0,1098	0,1025
	$DFBETAS_{j(i')} (j = 0, 1, \dots, p)$	0,9642	0,9904	0,9979
	$DFFITs_{(i')}$	0,6771	0,7525	0,8318
	$COVRATIO_{(i')}$	0,9060	0,7982	0,7743
	$AP_{(i')}$	0,1816	0,2938	0,4120
	$LD_{(i')}$	0,0540	0,0488	0,0402
	$K_{(i')}$	0,1270	0,1428	0,1656
	$S_{(i')}$	0,6224	0,6403	0,6507
$p = 4$	SICDGY	0,2770	0,3040	0,3460
	BYDGY	0,1295	0,1560	0,1786
	$D_{(i')}$	0,1788	0,1303	0,1254
	$DFBETAS_{j(i')} (j = 0, 1, \dots, p)$	0,9940	0,9993	0,9998
	$DFFITs_{(i')}$	0,7844	0,8378	0,8869
	$COVRATIO_{(i')}$	0,9857	0,9407	0,8924
	$AP_{(i')}$	0,1008	0,1943	0,2776
	$LD_{(i')}$	0,0919	0,0801	0,0689
	$K_{(i')}$	0,1475	0,1745	0,1901
	$S_{(i')}$	0,7012	0,8644	0,9158
$p = 5$	SICDGY	0,2948	0,3399	0,3675
	BYDGY	0,1640	0,1911	0,2066
	$D_{(i')}$	0,1955	0,1646	0,1433
	$DFBETAS_{j(i')} (j = 0, 1, \dots, p)$	0,9994	0,9891	0,9349
	$DFFITs_{(i')}$	0,8716	0,8865	0,9198
	$COVRATIO_{(i')}$	0,9991	0,9867	0,9562
	$AP_{(i')}$	0,0851	0,1358	0,1698
	$LD_{(i')}$	0,1277	0,1110	0,0901
	$K_{(i')}$	0,1955	0,2487	0,2744
	$S_{(i')}$	0,7298	0,9093	0,9586

Çizelge 7.4'teki sonuçlar incelendiğinde örnek hacminin ve parametre sayısının farklı durumlarında DFBETAS, DFFITS, COVRATIO istatistiklerinin regresyon tahminleri üzerinde etkisi olmayan gözlemlerden en az birini etkili gözlem olarak saptaması oranlarının yüksek olduğu açıkça görülmektedir. Karar kuralında yer alan kritik değeri örnek hacmine göre azalan bir fonksiyon olan bu yöntemler için verinin geneline uyan gözlemlerden en az birinin etkili olarak saptanması beklenen bir sonuçtur.

Cook Uzaklığı İstatistiği için elde edilen sonuçlar incelendiğinde, parametre sayısı sabitken örnek hacmi arttıkça, etkili olmayan gözlemlerden en az birini etkili olarak saptama oranında azalma gözlenmektedir. Bununla birlikte n sabitken p arttıkça bu istatistik için elde edilen oranın değerce büyüdüğü görülmektedir.

Çizelge 7.4'te AP istatistiği için elde edilen sonuçlar incelendiğinde p sabitken n arttıkça, etkili olmayan gözlemlerden en az birini etkili olarak saptama oranında artış gözlenmekte, n sabitken p arttıkça bu saptama oranında bir azalış gözlenmektedir.

Hadi'nin K İstatistiği için sonuçlara göre örnek hacmi küçükken etkili olmayan gözlemlerden en az birini etkili olarak saptama oranlarının çok büyük olmadığı görülmektedir. Çizelge 7.4'te Cook ve Weisberg'in Olabilirlik İstatistiği LD için sonuçlar incelendiğinde, verinin geneline uyan gözlemlerden en az birini etkili olarak saptama oranlarının küçük olduğu gözlenmektedir. Bununla birlikte LD için Çizelge 7.2'de verilen sonuçlara göre örnek hacmi ve parametre sayısının farklı durumları için etkili gözlemi saptama oranları, sonuçların geneline göre küçüktür. Pena'nın S İstatistiği için elde edilen oranlar incelendiğinde örnek hacmi ve parametre sayısının tüm durumlarında bu oranların diğer istatistiklere ilişkin oranların bazılarına göre nispeten büyük olduğu görülmektedir.

Çizelge 7.4'te SICDGY için sonuçlar incelendiğinde p sabitken n arttıkça, etkili olmayan gözlemlerden en az birinin etkili olarak saptama oranında artış gözlenmektedir. Benzer biçimde bu artış n sabitken p arttıkça yine görülmektedir.

BYDGY için sonuçlar incelendiğinde ise p sabitken n arttıkça etkili olmayan gözlemlerden en az birini etkili olarak saptama oranında sayısal olarak küçük bir artış söz konusudur. Benzer biçimde bu küçük artış n sabitken p arttıkça tekrar gözlenmektedir.

Çizelge 7.4'te verilen sonuçlar genel olarak incelendiğinde, örnek hacmi ve parametre sayısının tüm durumlarında SICDGY ve BYDGY için simülasyonla elde edilen, regresyon tahminleri üzerinde etkisi olmayan gözlemlerin en az birinin etkili gözlem olarak saptama oranlarının DFBETAS, DFFITS COVRATIO ve Pena'nın S istatistiği için elde edilen oranlardan küçük olduğu açıkça görülmektedir. Biplot Yaklaşımına Dayalı Grafik Yöntemi parametre sayısının 3 ve 4 olduğu durumlarda Andrews ve Pregibon'un AP istatistiğine göre; parametre sayısının 3, 4 ve 5 olduğu durumlarda Hadi'nin K istatistiğine göre etkili olmayan gözlemlerden en az birini etkili olarak saptama bakımından daha küçük oranlara sahiptir. Örnek hacminin küçük olduğu durumlarda da BYDGY için Çizelge 7.4'te verilen oranlar Cook Uzaklığı İstatistiği için elde edilen oranlardan daha küçüktür.

Çizelge 7.4'te verilen simülasyon sonuçları yorumlama kolaylığı sağlamak amacıyla n ve p ana faktörler olarak ele alınıp verinin geneline uyan gözlemlerden en az birini etkili gözlem olarak saptama oranı bakımından her yöntem için ayrı ayrı iki yönlü varyans analizi (ANOVA) uygulanmıştır. Ayrıca ele alınan iki ana faktörün etkileşim etkisini incelemek amacıyla aynı simülasyon çalışması tekrar yapıp 9 farklı durum için ek olarak birer adet oran daha elde edilmiştir. Böylece ANOVA için bağımlı değişken olarak ele alınan saptama oranına ilişkin varyans analizi tasarımında iki tekrar söz konusu olup, iki ana faktörün etkileşim etkisinin istatistiksel olarak anlamlı olup olmadığının incelenmesi sağlanmıştır.

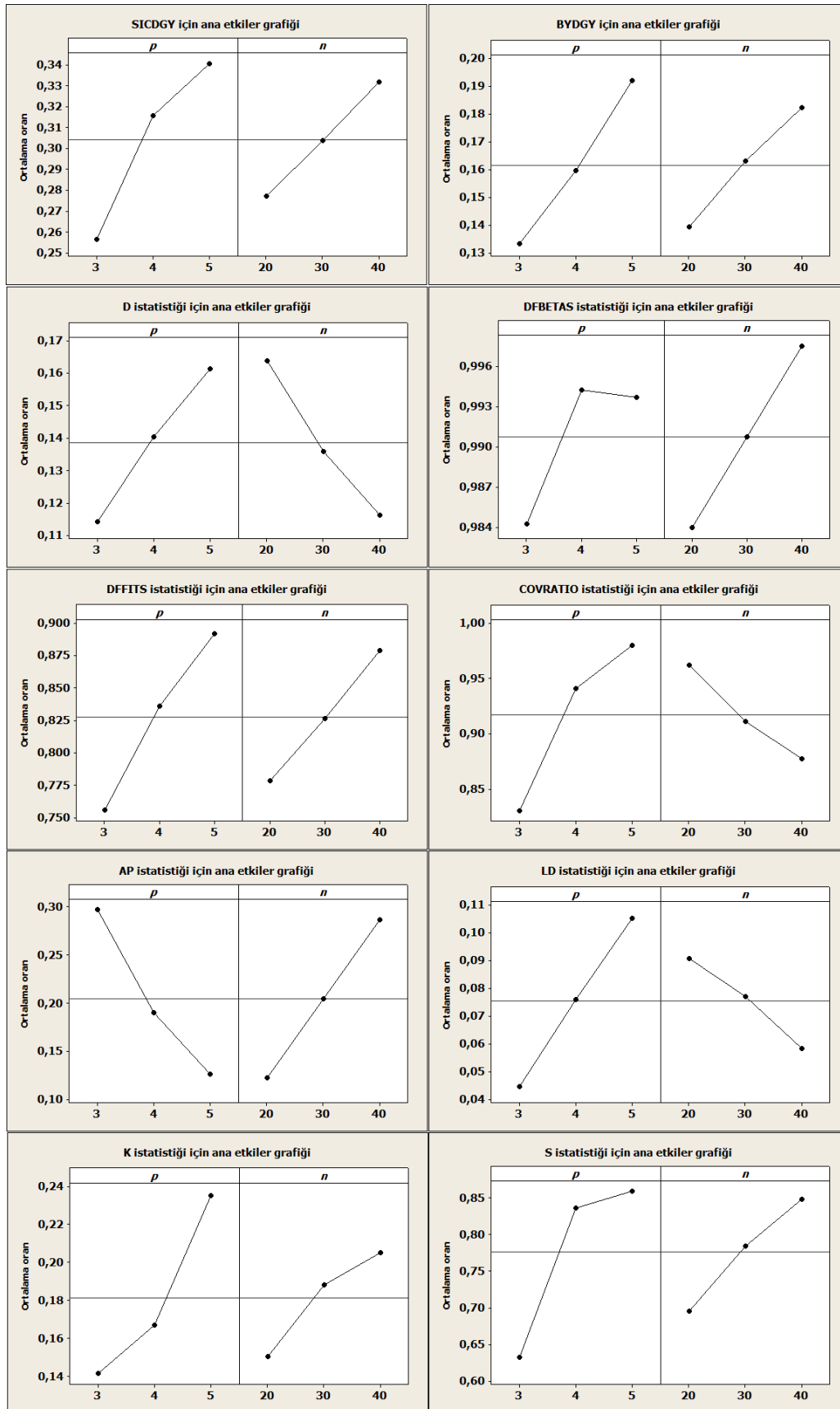
Varyans analizi için deney tasarımı oluşturulduktan sonra n ve p faktörleri ile bu iki faktörün etkileşim etkisinin istatistiksel olarak anlamlılığını test etmek amacıyla İki Yönlü ANOVA yöntemiyle p -değerleri elde edilmiştir. Sonuçlar Çizelge 7.5'teki gibi sunulmuştur.

Çizelge 7.5. Bir etkili gözlemin bulunduğu veri kümesinde etkili olmayan gözlemlerden en az birini saptama oranları bakımından iki yönlü varyans analiziyle elde edilen p-değerleri

Tanı yöntemi	Faktörler		Etkileşim
	p	n	
SICDGY	0,000*	0,000*	0,044*
BYDGY	0,000*	0,000*	0,469
$D_{(i)}$	0,000*	0,000*	0,475
$DFBETAS_{j(i)} (j = 0,1,...,p)$	0,001*	0,013*	0,000*
$DFFITS_{(i)}$	0,000*	0,000*	0,000*
$COVRATIO_{(i)}$	0,000*	0,000*	0,000*
$AP_{(i)}$	0,000*	0,000*	0,000*
$LD_{(i)}$	0,000*	0,000*	0,072
$K_{(i)}$	0,000*	0,000*	0,017*
$S_{(i)}$	0,000*	0,000*	0,001*

* p-değeri < 0,05 olduğu sonuçlar

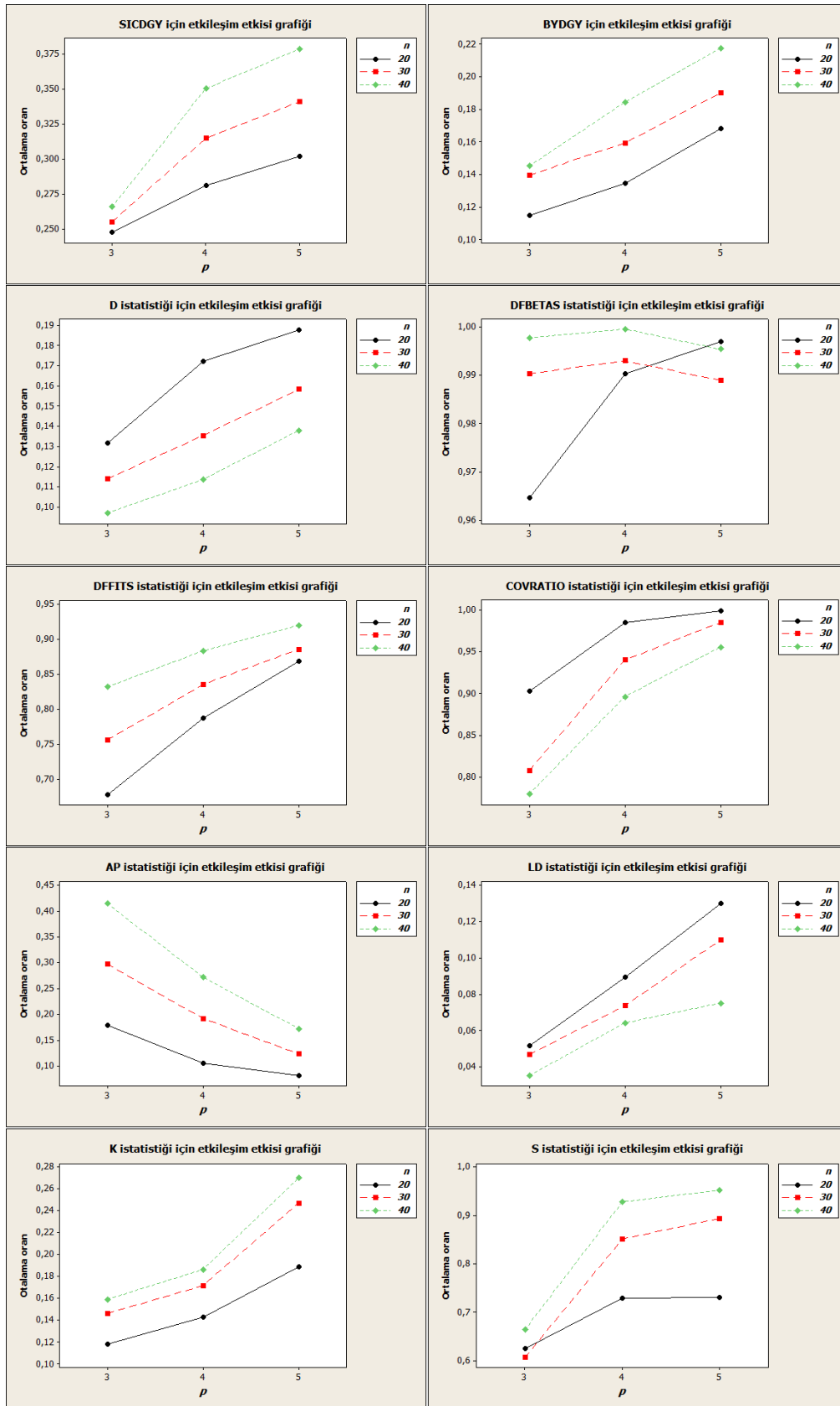
Çizelge 7.5'te İki Yönlü ANOVA yöntemiyle elde edilen p-değeri sonuçları incelendiğinde tüm yöntemler için, etkili olmayan gözlemlerden en az birini etkili olarak saptama oranı üzerinde parametre sayısının etkisinin istatistiksel olarak anlamlı olduğu açıkça görülmektedir. Buna ek olarak ele alınan tüm tanı yöntemleri için elde edilen oranlar üzerinde örnek hacminin etkisinin anlamlı olduğu da söylenebilir. Son olarak Çizelge 7.5'te etkileşim etkisi için p-değeri sonuçları incelendiğinde SIC'ye Dayalı Grafik Yöntemi, DFBETAS, DFFITS, COVRATIO, AP, Hadi'nin K ve Pena'nın S İstatistikleri için elde edilen oranlar üzerinde, n ve p faktörlerinin etkileşiminin etkisinin istatistiksel olarak anlamlı olduğu sonucuna varılmıştır. Ana etkilere ilişkin anlamlı sonuçlar dikkate alınırsa; örnek hacmi ve parametre sayısındaki artışa dayalı olarak tanı yöntemleriyle verinin geneline uyan gözlemlerden en az birini etkili gözlem olarak saptama oranlarındaki değişimi görsel olarak değerlendirmek amacıyla ana etki grafikleri oluşturulmuş ve Şekil 7.3'te verilmiştir.



Şekil 7.3. Bir etkili gözlemin bulunduğu veri kümesinde verinin geneline uyan gözlemlerden en az birini etkili olarak saptama oranları bakımından iki yönlü varyans analizinden sonra oluşturulan ana etki grafikleri

Şekil 7.3'ten de görüldüğü üzere n sabitken, SICDGY, BYDGY ile Cook Uzaklığı, DFBETAS, DFFITS, COVRATIO, LD, K ve S İstatistikleri için parametre sayısı arttıkça verinin geneline uyan gözlemlerden en az birini etkili olarak saptama oranlarının arttığı sonucuna varılır. n sabitken parametre sayısı arttıkça AP istatistiği için saptama oranının da azaldığı gözlenmektedir. p sabitken örnek hacmi arttıkça SICDGY, BYDGY ile DFBETAS, DFFITS, AP, K ve S İstatistikleri için verinin geneline uyan gözlemlerden en az birini etkili olarak saptama oranlarında artış gözleniyorken; D, COVRATIO ve LD İstatistikleri için bu saptama oranlarında azalış gözlenmektedir.

n ve p faktörlerinin etkileşim etkisini görsel olarak incelemek amacıyla ele alınan tanı yöntemlerine ilişkin etkileşim etkisi grafikleri oluşturulup; Şekil 7.4'te sunulmuştur.



Şekil 7.4. Bir etkili gözlemin bulunduğu veri kümesinde verinin geneline uyan gözlemlerden en az birini etkili olarak saptama oranları bakımından iki yönlü varyans analizinden sonra oluşturulan etkileşim etkisi grafikleri

Şekil 7.4 incelendiğinde SIC'ye Dayalı Grafik Yöntemi, DFBETAS, DFFITS, COVRATIO, AP, K ve S İstatistikleri için grafiklerde her üç çizginin veya uzantılarının ortak olarak kesiştiği görülmektedir. Dolayısıyla bu istatistikler için elde edilen etkili olmayan gözlemlerden en az birini saptama oranları bakımından n ve p faktörlerinin etkileşiminin mevcut olduğu sonucuna varılır.

Bölüm 5'te de bahsedildiği gibi veri kümesine standartlaştırma yapılmışken, Δ matrisinin i . satır j . sütun elemanı olan $\hat{\beta}_{j(i)} - \hat{\beta}_j$ değeri, i . gözlemin j . parametre tahmininde meydana getirdiği değişimin sayısal değeridir. Dolayısıyla Δ matrisinin satırlarına karşılık gelen elemanlardan genele uymayan bir eleman $\hat{\beta}_{j(i)} - \hat{\beta}_j$ bakımından aykırı bir noktadır. Bu matrisin sütun sayısı ikiden fazla iken, biplot yaklaşımı bu aykırı noktaların saptanmasında kullanılabilir. $\hat{\beta}_{j(i)} - \hat{\beta}_j$ bakımından saptanan bu aykırı nokta da doğrusal regresyon analizinde etkili gözleme karşılık gelir. Nitekim bu noktaya karşılık gelen gözlem etkili gözlem olarak saptanması bir koşul olarak biplotta elips biçiminde sınır oluşturulmuştur. Bu sınır da görsel olarak değerlendirme bakımından kolaylık sağlar.

Bu kısımda yapılan simülasyonla elde edilen sonuçlar genel olarak incelendiğinde, örnek hacminin büyük olduğu durumlarda Cook Uzaklığı İstatistiği, DFBETAS İstatistiği, DFFITS İstatistiği, AP İstatistiği, K İstatistiği, S istatistiği SIC'ye Dayalı Grafik Yöntemi ve Biplot Yaklaşımına Dayalı Grafik Yönteminin bir etkili gözlem içeren veride bulunan etkili gözlemi saptama oranı bakımından diğerlerinden yani, COVRATIO İstatistiği ve LD İstatistiğinden, daha iyi olduğu görülmektedir. Parametre sayısının büyük olduğu durumlarda ise DFFITS İstatistiği, AP İstatistiği, S İstatistiği ve Biplot Yaklaşımına Dayalı Grafik Yönteminin bir etkili gözlem içeren veride bulunan etkili gözlemi saptama oranı bakımından iyi olduğu gözlenmiştir.

7.2. Birden Fazla Gözlem Silme Yaklaşımına Dayalı Tanı İstatistikleri ile Önerilen İki Grafik Yönteminin Karşılaştırılması

Bu kısımda gerçekte veri kümesinde iki etkili gözlem varken, örnek hacmi ve parametre sayısının farklı durumlarında Cook Uzaklığı, COVRATIO ve K istatistiklerinin birden fazla gözlem silme yaklaşımına dayalı olan genelleştirilmiş biçimleriyle bu çalışmada önerilen SIC'ye Dayalı Grafik Yöntemi (SICDGY) ve Biplot Yaklaşımına Dayalı Grafik

Yöntemi (BYDGY) için etkili gözlem grubunu saptama oranları simülasyonla elde edilmiştir. Ayrıca birden fazla gözlem silme yaklaşımına dayalı klasik tanı yöntemleriyle önerilen iki yöntem, etkili gözlem grubu haricinde, regresyon tahminleri üzerinde etkisi olmamak üzere üretilen diğer gözlemlerin ikili alt kümelerinden en az birini etkili gözlem grubu olarak saptama oranları da simülasyonla elde edilmiştir. Daha sonra her bir yöntem için elde edilen simülasyon sonuçları ayrı ayrı ele alınarak iki yönlü ANOVA yöntemi uygulanmış ve sonuçlar yorumlanmıştır. Veri kümesinde iki etkili gözlemin olduğu durum için simülasyon aşağıdaki adımlar doğrultusunda yapılmıştır.

Adım 1: Verinin geneline uyan $n-2$ adet gözlemin $p-1$ adet bağımsız değişkenin değerleri, bu gözlemler yüksek dereceli kaldıraç noktası olmayacak biçimde, $[0,15]$ aralığında tekdüze dağılımdan üretilir.

Adım 2: Verinin geneline uyan $n-2$ adet gözlemin hata değişken değerleri, bu gözlemler artıklar bakımından aykırı bir değer olmayacak biçimde, $\varepsilon_i \sim N(0,1)$ dağılımından üretilir.

Adım 3: Etkili gözlemler sırasıyla i' ve i'' ile ifade edilmek üzere bu gözlem için $p-1$ adet bağımsız değişkenlerinin değerleri $x_{j i'} = 20$ ve $x_{j i''} = 21$ olarak, hata terimi değerleri de $\varepsilon_{i'} = -5$ ve $\varepsilon_{i''} = -5$ olarak belirlenir.

Adım 4: $j = 0, 1, \dots, p-1$ iken her $\beta_j = 1$ olmak üzere i' ve i'' ile ifade edilen etkili gözlemlerin de bulunduğu n adet gözlem için bağımlı değişken değerleri

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

modeline göre üretilir.

Adım 5: Üretilen veri kullanılarak $D_{(1)}$, $COVRATIO_{(1)}$ ve $K_{(1)}$ değerleri ve bunlara karşılık gelen kritik değerler hesaplanır. i' ve i'' ile ifade edilen gözlem çiftinin ve diğer gözlem çiftlerinin etkili gözlem grubu olup olmadığına karar verilir.

Adım 6: SICDGY ve BYDGY ile grafikler oluşturulur. i' ve i'' ile ifade edilen gözlem çiftinin ve diğer gözlem çiftlerinin etkili gözlem grubu olup olmadığına karar verilir.

Adım 7: Eğer deneme numarası 1000'in katı ise *Adım 1*'e gidilir, değilse *Adım 2*'den devam edilir.

Deneme sayısı 100 000 olmak üzere, n örnek hacminin ve p parametre sayısının farklı durumları için etkili gözlem grubunu saptamaya dayalı tanı istatistiklerinin ve önerilen iki grafik yönteminin etkili gözlem grubu olan $I' = \{i', i''\}$ kümesini saptama oranları MATLAB2011 programı kullanılarak hesaplanmıştır. Elde edilen sonuçlar Çizelge 7.6'da verilmiştir.

Çizelge 7.6. Birden fazla gözlem silme yaklaşımına dayalı tanı istatistikleri ile BYDGY ve SICDGY için etkili gözlem grubunu saptama oranları

Parametre sayısı	Tanı yöntemi	Örnek Hacmi		
		$n = 20$	$n = 30$	$n = 40$
$p = 3$	SICDGY	0,6010	0,7910	0,8910
	BYDGY	0,9143	0,9765	0,9956
	$D_{(I)}$	0,9465	0,9666	0,9899
	$COVRATIO_{(I)}$	0,8735	0,9359	0,9636
	$K_{(I)}$	0,7608	0,9733	0,9960
$p = 4$	SICDGY	0,4940	0,7810	0,8840
	BYDGY	0,6861	0,9851	0,9966
	$D_{(I)}$	0,9325	0,9605	0,9971
	$COVRATIO_{(I)}$	0,8896	0,9406	0,9620
	$K_{(I)}$	0,5654	0,9437	0,9965
$p = 5$	SICDGY	0,4755	0,7740	0,8650
	BYDGY	0,5001	0,9354	0,9963
	$D_{(I)}$	0,9212	0,9499	0,9942
	$COVRATIO_{(I)}$	0,8911	0,9438	0,9642
	$K_{(I)}$	0,3571	0,8828	0,9868

Çizelge 7.6'da verilen sonuçlardan ilk olarak Cook Uzaklığı İstatistiği için elde edilen oranlar incelendiğinde parametre sayısı p sabitken örnek hacmi arttıkça etkili gözlem grubunu saptama oranının da arttığı görülmektedir. Bu artışın nedeni, p sabitken örnek hacmi arttığında Cook Uzaklığı için kritik değer olan $F_{0.50,p,n-p}$ değerinin ikinci serbestlik derecesindeki artışın, bu değeri küçültmesi biçiminde yorumlanabilir. $n=20$ iken p arttıkça, $D_{(I)}$ istatistiği için elde edilen oranlarda bir düşme söz konusudur. Bunun nedeni ise p artarken $F_{0.50,p,n-p}$ kritik değerinin büyümesi ve $D_{(I)}$ istatistiğinin iki etkili gözlemden oluşan etkili gözlem grubunu saptama eğiliminin azalması biçiminde yorumlanabilir.

$COVRATIO_{(I)}$ İstatistiği için elde edilen oranlar incelendiğinde p sabitken örnek hacmi arttıkça bu istatistiğin etkili gözlem grubunu saptama oranının arttığı görülmektedir. Kritik

değeri örnek hacminin azalan bir fonksiyonu olan $COVRATIO_{(n)}$ İstatistiği için bu sonuç doğaldır. Bununla birlikte kritik değerin azalması, etkili olmayan bir gözlemi içeren herhangi bir I kümesinin de etkili gözlem grubu olarak saptanmasına neden olabilir. Bu durumda araştırmacının yanlış karar almasına neden olur. Ayrıca n sabitken p arttıkça, $COVRATIO_{(n)}$ İstatistiği için elde edilen oranlarda küçük bir artışın olduğu görülmektedir.

Bir gözlem silme yaklaşımına dayalı $K_{(i)}$ istatistiğine ilişkin oranlarda gözlemlendiği gibi bu istatistiğin etkili gözlem grupları için genelleştirilmiş hali olan $K_{(n)}$ istatistiğine karşılık gelen oranlar da p sabitken örnek hacmi arttığında artmaktadır. Fakat n sabitken p arttığında bu istatistik için elde edilen oranlarda bir düşme gözlenmiştir. Özellikle bu düşme $n = 20$ iken açıkça görülmektedir.

Çizelge 7.6'dan SICDGY için elde edilen sonuçlar incelendiğinde p sabitken n arttıkça bu yöntem için elde edilen oranların değer olarak büyüdüğü görülmektedir. Fakat n sabitken p arttıkça SICDGY için elde edilen oranlarda bir azalış da gözlenmektedir. Bu azalış $n = 40$ olduğunda küçüktür.

Son olarak BYDGY için oranlar incelendiğinde p sabitken n arttıkça bu oranların büyüdüğü açıkça görülmektedir. Bu büyüme $n=30$ durumundan sonra daha küçük artışlarla gözlenmektedir. $n=20$ olduğunda parametre sayısı arttıkça BYDGY için oranların küçüldüğü gözlenmesine rağmen $n=30$ ve $n=40$ iken bu oranlarda dikkate değer bir değişim yoktur.

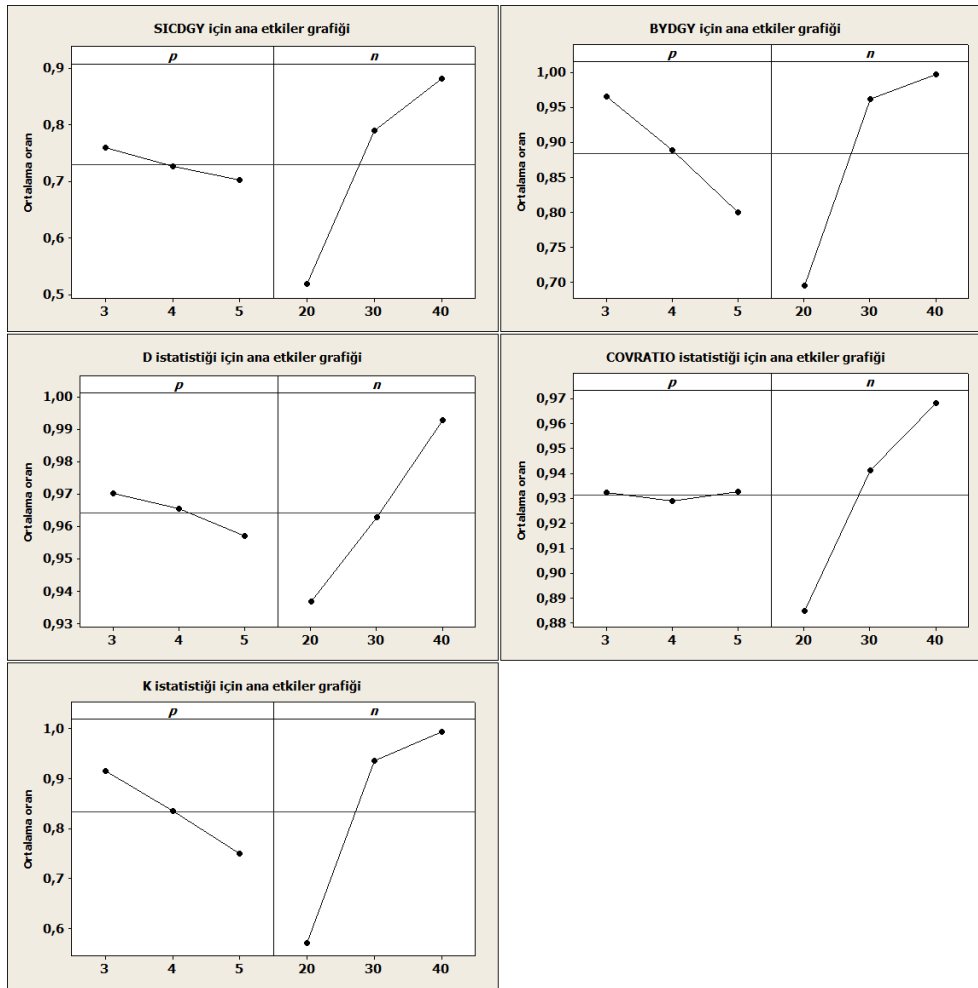
Çizelge 7.6'da verilen simülasyon sonuçları yorumlama kolaylığı sağlamak amacıyla n ve p ana faktörler olarak ele alınıp etkili gözlem grubunu saptama oranı bakımından her yöntem için ayrı ayrı iki yönlü varyans analizi (ANOVA) uygulanmıştır. Ayrıca ele alınan iki ana faktörün etkileşim etkisini incelemek amacıyla aynı simülasyon çalışması tekrar yapıp 9 farklı durum için ek olarak birer adet oran daha elde edilmiştir. Böylece ANOVA için bağımlı değişken olarak ele alınan saptama oranına ilişkin varyans analizi tasarımında iki tekrar söz konusu olup, iki ana faktörün etkileşim etkisinin istatistiksel olarak anlamlı olup olmadığının incelenmesi sağlanmıştır. Varyans analizi için deney tasarımı oluşturulduktan sonra n ve p faktörleri ile bu iki faktörün etkileşim etkisinin istatistiksel olarak anlamlılığını test etmek amacıyla İki Yönlü ANOVA yöntemiyle p -değerleri elde edilmiştir. Sonuçlar Çizelge 7.7'deki gibi sunulmuştur.

Çizelge 7.7. İki etkili gözlemin bulunduğu veri kümesinde etkili gözlem grubunu saptama oranları bakımından iki yönlü varyans analiziyle elde edilen p-değerleri

Tanı yöntemi	Faktörler		Etkileşim
	p	n	
SICDGY	0,000*	0,000*	0,000*
BYDGY	0,000*	0,000*	0,000*
$D_{(I)}$	0,005*	0,000*	0,037*
$COVRATIO_{(I)}$	0,733	0,000*	0,123
$K_{(I)}$	0,000*	0,000*	0,000*

* p-değeri < 0,05 olduğu sonuçlar

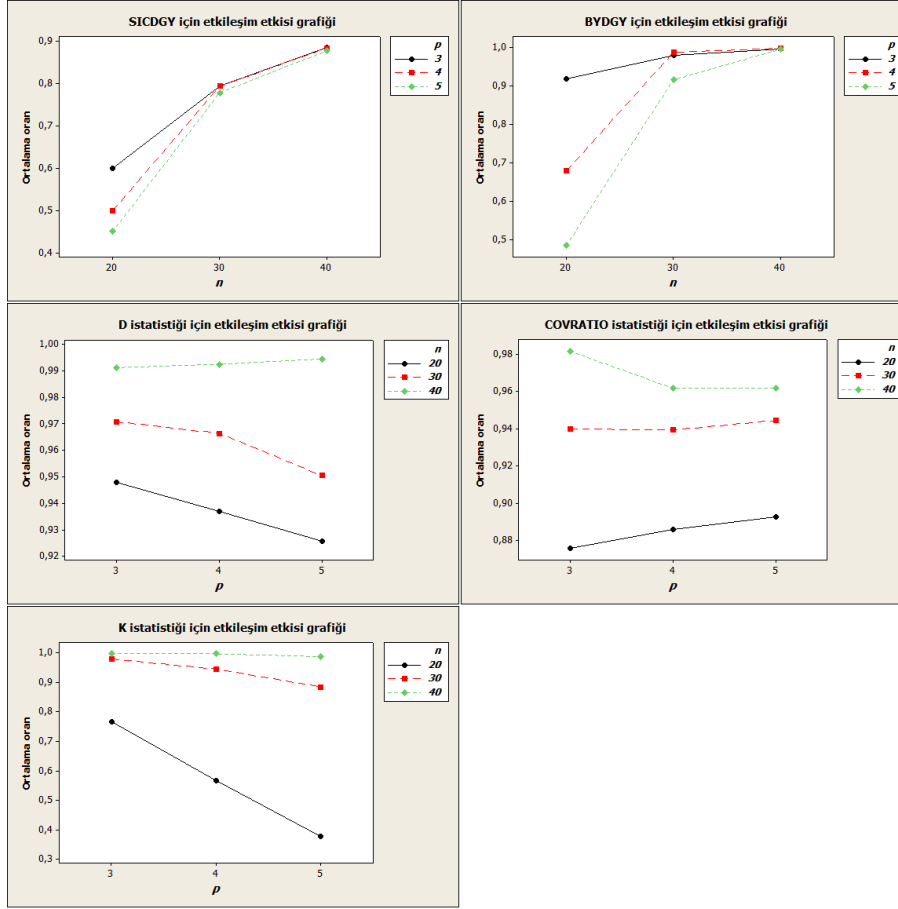
Çizelge 7.7’de İki Yönlü ANOVA yöntemiyle elde edilen p-değeri sonuçları incelendiğinde SIC’ye Dayalı Grafik Yöntemi, Biplot Yaklaşımına Dayalı Grafik Yöntemi, Cook Uzaklığı ve Hadi’nin K İstatistikleri için elde edilen oranlar üzerinde parametre sayısının etkisinin istatistiksel olarak anlamlı olduğu açıkça görülmektedir. Buna ek olarak ele alınan tüm tanı yöntemleri için elde edilen oranlar üzerinde örnek hacminin etkisinin anlamlı olduğu söylenebilir. Son olarak Çizelge 7.7’de etkileşim etkisi için p-değeri sonuçları incelendiğinde SIC’ye Dayalı Grafik Yöntemi, Biplot Yaklaşımına Dayalı Grafik Yöntemi, Cook Uzaklığı İstatistiği ve K İstatistiği için n ve p faktörlerinin etkileşiminin etkisinin istatistiksel olarak anlamlı olduğu sonucuna varılmıştır. Ana etkilere ilişkin anlamlı sonuçlar dikkate alınır; örnek hacmi ve parametre sayısındaki artışa dayalı olarak tanı yöntemleriyle etkili gözlem grubunu saptama oranlarındaki değişimi görsel olarak değerlendirmek amacıyla ana etki grafikleri oluşturulmuş ve Şekil 7.5’te verilmiştir.



Şekil 7.5. İki etkili gözlemin bulunduğu veri kümesinde etkili gözlem grubunu saptama oranları bakımından iki yönlü varyans analizinden sonra oluşturulan ana etki grafikleri

Şekil 7.5'ten de görüldüğü üzere n sabitken SICDGY, BYDGY, Cook Uzaklığı İstatistiği, ve Hadi'nin K İstatistiği yöntemleri için parametre sayısı arttıkça etkili gözlem grubunu saptama oranlarının azaldığı sonucuna varılır. p sabitken örnek hacmi arttığında ele alınan beş yöntem için etkili gözlem grubunu saptama oranlarında artış gözlenmektedir.

n ve p faktörlerinin etkileşim etkisini görsel olarak incelemek amacıyla ele alınan tanı yöntemlerine ilişkin etkileşim etkisi grafikleri oluşturulup, Şekil 7.6'da sunulmuştur.



Şekil 7.6. İki etkili gözlemin bulunduğu veri kümesinde etkili gözlem grubunu saptama oranları bakımından iki yönlü varyans analizinden sonra oluşturulan etkileşim etkisi grafikleri

Şekil 7.6 incelendiğinde SIC'ye Dayalı Grafik Yöntemi, Biplot Yaklaşımına Dayalı Grafik Yöntemi, Cook Uzaklığı İstatistiği ve Hadi'nin K İstatistiği için grafiklerde her üç çizginin veya uzantılarının ortak olarak kesiştiği görülmektedir. Dolayısıyla bu istatistikler için elde edilen etkili gözlem grubunu saptama oranları bakımından n ve p faktörlerinin etkileşiminin mevcut olduğu sonucuna varılır.

Bu çalışmada önerilen iki yöntemle birden fazla gözlem silme yaklaşımına dayalı klasik tanı yöntemlerinin etkili olmamak üzere üretilen gözlemlerin ikili alt kümelerinin en az birini etkili olarak saptama bakımından incelenmesi bu yöntemlerin karşılaştırılması için bir başka bakış açısı sağlayabilir. Bu amaçla Çizelge 7.8'de, önerilen yöntemlerle birden fazla gözlem silme yaklaşımına dayalı klasik yöntemler etkili olmayan gözlem çiftlerinden en az birini etkili gözlem grubu olarak saptama oranları sunulmuştur.

Çizelge 7.8. Birden fazla gözlem silme yaklaşımına dayalı tanı istatistikleri ile BYDGY ve SICDGY için etkili olmayan gözlem çiftlerinin en az birini etkili olarak saptama oranları

Parametre sayısı	Tanı yöntemi	Örnek Hacmi		
		$n = 20$	$n = 30$	$n = 40$
$p = 3$	SICDGY	0,2500	0,2960	0,3340
	BYDGY	0,1195	0,1588	0,1953
	$D_{(r)}$	0,2302	0,1960	0,1618
	$COVRATIO_{(r)}$	0,9891	0,9895	0,9900
	$K_{(r)}$	0,8335	0,8798	0,8894
$p = 4$	SICDGY	0,2910	0,3321	0,3781
	BYDGY	0,1588	0,1791	0,2135
	$D_{(r)}$	0,2632	0,2265	0,1848
	$COVRATIO_{(r)}$	0,9817	0,9799	0,9912
	$K_{(r)}$	0,9672	0,9828	0,9818
$p = 4$	SICDGY	0,3101	0,3358	0,3998
	BYDGY	0,1825	0,2031	0,2455
	$D_{(r)}$	0,3040	0,2772	0,2155
	$COVRATIO_{(r)}$	0,9856	0,9845	0,9955
	$K_{(r)}$	0,9947	0,9985	0,9978

Çizelge 7.8'deki sonuçlar incelendiğinde, örnek hacmi ve parametre sayısının tüm durumlarında COVRATIO ve Hadi'nin K istatistikleri için etkili gözlemler haricinde etkili olmamak üzere üretilen diğer gözlemlerin çiftlerinin en az birini etkili gözlem grubu olarak saptanma oranlarının büyük değerler olduğu açıkça görülmektedir. Cook Uzaklığı İstatistiği için oranlar incelendiğinde p sabitken n arttıkça bu oranlarda bir azalma söz konusudur. Bununla birlikte n sabitken p arttıkça bu oranlarda bir artış gözlenmektedir.

Bu çalışmada önerilen SICDGY için sonuçlar incelendiğinde p sabitken n arttıkça verinin geneline uymak üzere üretilen diğer gözlemlerin ikili alt kümelerine karşılık gelen gözlem gruplarından en az birinin etkili gözlem grubu olarak saptanma oranlarının arttığı, n sabitken p arttıkça yine bu oranların arttığı gözlenmektedir. Bu artış büyük bir artış değildir. Ayrıca SICDGY için elde edilen etkili olmayan gözlem gruplarından en az birinin etkili olarak saptanması oranları COVRATIO ve Hadi'nin K istatistikleri için elde edilen oranlara göre küçüktür.

Çizelge 7.8'deki sonuçlardan bu çalışmada önerilen ikinci yöntem olan BYDGY için simülasyonla elde edilen etkili olmayan gözlem gruplarından en az birinin etkili gözlem grubu olarak saptanması oranları incelendiğinde, bu oranların diğer yöntemler için elde edilen oranlara göre genel olarak daha küçük oranlar olduğu gözlenmiştir. Fakat örnek hacmi 40 iken parametre sayısının tüm durumlarında Cook Uzaklığı İstatistiği için elde edilen oranlar BYDGY için elde edilen oranlara göre küçüktür.

Çizelge 7.8’de verilen simülasyon sonuçları yorumlama kolaylığı sağlamak amacıyla n ve p ana faktörler olarak ele alınıp verinin geneline uyan gözlemlerin ikili gruplarından en az birini etkili gözlem grubu olarak saptama oranı bakımından her yöntem için ayrı ayrı iki yönlü varyans analizi (ANOVA) uygulanmıştır. Ayrıca ele alınan iki ana faktörün etkileşim etkisini incelemek amacıyla aynı simülasyon çalışması tekrar yapıp 9 farklı durum için ek olarak birer adet oran daha elde edilmiştir. Böylece ANOVA için bağımlı değişken olarak ele alınan saptama oranına ilişkin varyans analizi tasarımında iki tekrar söz konusu olup, iki ana faktörün etkileşim etkisinin istatistiksel olarak anlamlı olup olmadığının incelenmesi sağlanmıştır.

Varyans analizi için deney tasarımı oluşturulduktan sonra n ve p faktörleri ile bu iki faktörün etkileşim etkisinin istatistiksel olarak anlamlılığını test etmek amacıyla İki Yönlü ANOVA yöntemiyle p -değerleri elde edilmiştir. Sonuçlar Çizelge 7.9’daki gibi sunulmuştur.

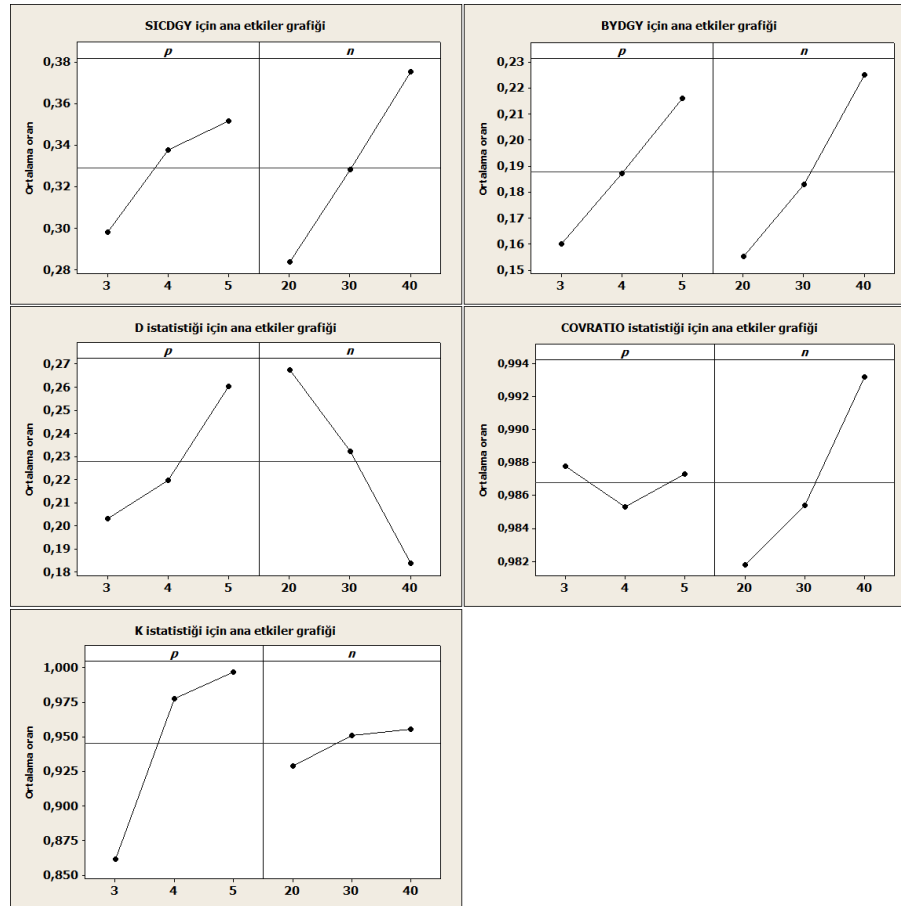
Çizelge 7.9. İki etkili gözlemin bulunduğu veri kümesinde verinin geneline uyan gözlemlerin ikili gruplarından en az birini etkili gözlem grubu olarak saptama oranları bakımından iki yönlü varyans analiziyle elde edilen p -değerleri

Tanı yöntemi	Faktörler		Etkileşim
	p	n	
SICDGY	0,000*	0,000*	0,068
BYDGY	0,000*	0,000*	0,218
$D_{(I)}$	0,000*	0,000*	0,625
$COVRATIO_{(I)}$	0,593	0,004*	0,324
$K_{(I)}$	0,000*	0,000*	0,000*

* p -değeri $< 0,05$ olduğu sonuçlar

Çizelge 7.9’da İki Yönlü ANOVA yöntemiyle elde edilen p -değeri sonuçları incelendiğinde SIC’ye Dayalı Grafik Yöntemi, Biplot Yaklaşımına Dayalı Grafik Yöntemi, Cook Uzaklığı ve Hadi’nin K İstatistikleri için elde edilen, verinin geneline uyan gözlem çiftlerinden en az birinin etkili gözlem grubu olarak saptanması oranları üzerinde parametre sayısının etkisinin istatistiksel olarak anlamlı olduğu açıkça görülmektedir. Buna ek olarak ele alınan tüm tanı yöntemleri için elde edilen oranlar üzerinde örnek hacminin etkisinin anlamlı olduğu söylenebilir. Son olarak Çizelge 7.9’da etkileşim etkisi için p -değeri sonuçları incelendiğinde, bu yöntemlerden sadece K İstatistiği için n ve p faktörlerinin etkileşim etkisinin istatistiksel olarak anlamlı olduğu sonucuna varılmıştır. Ana etkilere ilişkin anlamlı sonuçlar dikkate alınırsa; örnek hacmi ve parametre

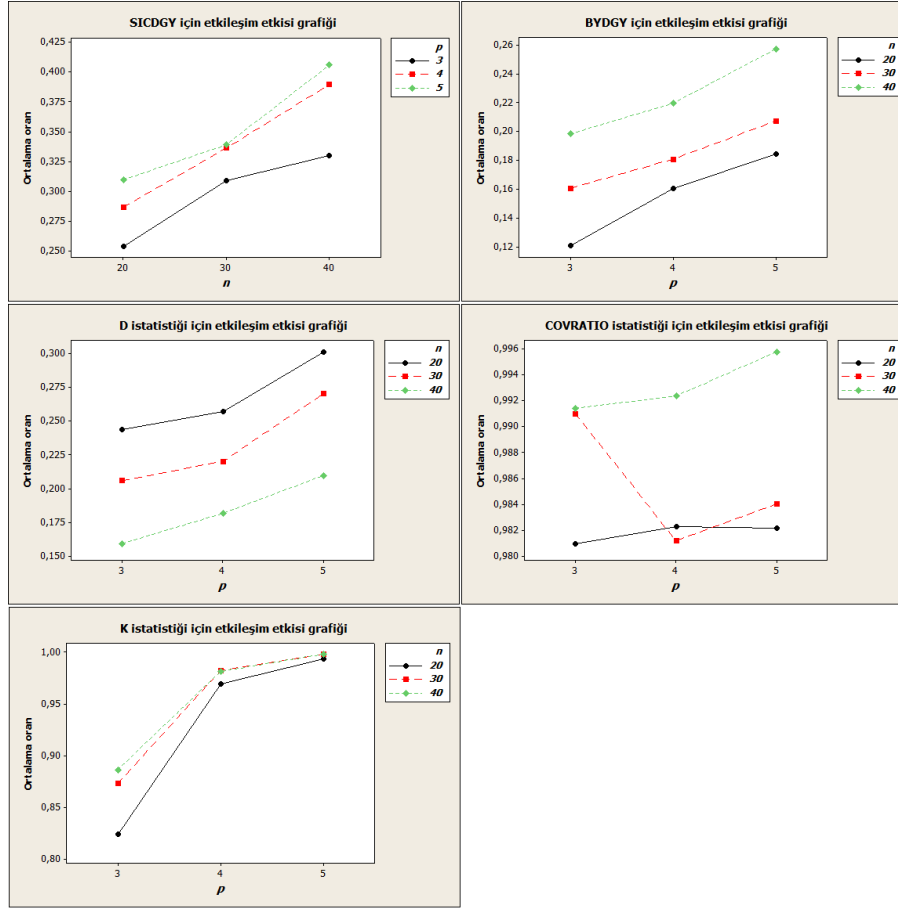
sayısındaki artışa dayalı olarak tanı yöntemleriyle verinin geneline uyan gözlemlerin ikili gruplarından en az birini etkili gözlem grubu olarak saptama oranlarındaki değişimi görsel olarak değerlendirmek amacıyla ana etki grafikleri oluşturulmuş ve Şekil 7.7’de verilmiştir.



Şekil 7.7. İki etkili gözlemin bulunduğu veri kümesinde verinin geneline uyan gözlemlerin ikili gruplarından en az birini etkili gözlem grubu olarak saptama oranları bakımından iki yönlü varyans analizinden sonra oluşturulan ana etki grafikleri

Şekil 7.7’den de görüldüğü üzere n sabitken SICDGY, BYDGY ile Cook Uzaklığı ve K İstatistikleri için parametre sayısı arttıkça verinin geneline uyan gözlemlerin ikili gruplarından en az birini etkili gözlem grubu olarak saptama oranlarının arttığı sonucuna varılır. p sabitken örnek hacmi arttıkça SICDGY, BYDGY ile COVRATIO ve K İstatistikleri için verinin geneline uyan gözlemlerin ikili gruplarından en az birini etkili gözlem grubu olarak saptama oranlarında artış gözleniyorken; D İstatistiği için bu saptama oranlarında azalış gözlenmektedir.

n ve p faktörlerinin etkileşim etkisini görsel olarak incelemek amacıyla ele alınan tanı yöntemlerine ilişkin etkileşim etkisi grafikleri oluşturulup; Şekil 7.8’de sunulmuştur.



Şekil 7.8. İki etkili gözlemin bulunduğu veri kümesinde verinin geneline uyan gözlemlerin ikili gruplarından en az birini etkili gözlem grubu olarak saptama oranları bakımından iki yönlü varyans analizinden sonra oluşturulan etkileşim etkisi grafikleri

Şekil 7.8 incelendiğinde tanı yöntemlerinden sadece K İstatistiği için grafikte her üç çizginin veya uzantılarının ortak olarak kesiştiği görülmektedir. Dolayısıyla bu istatistik için elde edilen etkili gözlemi saptama oranı bakımından n ve p faktörlerinin etkileşiminin etkisinin anlamlı olduğu sonucuna varılır.

Etkili gözlem gruplarının saptanması için birden fazla gözlem silme yaklaşımına dayalı tanı yöntemlerinde temel sorun, $1 < k < n/2$ olmak üzere k büyüklüğünde gözlem grupları

tanı istatistikleriyle incelenirken k ' nın değeri $n/2$ ' ye yaklaştıkça, $\binom{n}{k}$ kadar olan mümkün gözlem grubu sayısının artmasıdır. Böylece çok sayıda gözlem grubunun

incelenmesi durumu ortaya çıkar ve yöntem pratik olmaz. Bunu bir örnek durumla açıklanması yol göstericidir. $k = 2$ ve $n = 40$ iken incelenmesi gereken ikili kombinasyon sayısı 780 olarak hesaplanır. Bu ikili kombinasyon sayısı oldukça büyüktür ve 780 farklı ikili gözlem grubu için gözlem silme yaklaşımıyla regresyon parametrelerinin EKK tahminlerinin hesaplanmasını gerektirir. Özellikle k değerinin 2'den büyük olduğu durumlar araştırılmak istenirse, incelenecek gözlem kümelerine ilişkin kombinasyon sayısı daha da artmakta ve ikili kombinasyonlardan sonra bunların da incelenmesi zorunluluğu ortaya çıkmaktadır. Bu durumda $n = 40$ iken örnek hacminin birli, ikili ve üçlü kombinasyonlarının toplam sayısı 10700 olarak belirlenir. Buna dayalı olarak bu belirlenen sayı kadar gözlem silme yaklaşımıyla regresyon parametrelerinin EKK tahminlerinin hesaplanmasını gerekir. Ayrıca $k > 1$ iken herhangi bir gözlem grubunda en az bir adet etkili gözlem ve bir adet etkili olmayan gözlem bulunursa, bu durumda etkili gözlemden dolayı ilgili tanı istatistiğinin değeri büyük olacağından bu gözlem kümesinin etkili bir gözlem kümesi olduğu sonucuna varılabilir. Böylece yanılgıya düşürme sorunuyla karşılaşmış olunur. Dolayısıyla bu çalışmada önerilen SIC'ye Dayalı Grafik Yöntemi ve Biplot Yaklaşımına Dayalı Grafik Yöntemi etkili gözlem gruplarının saptanması bakımından Cook Uzaklığı, COVRATIO ve Hadi'nin K istatistiğine göre daha pratiktir. Özellikle önerilen bu yöntemler etkili gözlem veya gözlemlerin grafik üzerinden incelenmesi olanağını da araştırmacıya sunmaktadır.

Biplot yöntemine dayalı grafik yöntemi hem etkili gözlemlerin saptanması bakımından pratik bir yöntemdir, hem de bağımsız değişken sayısı 2 veya daha fazla iken etkili gözlemlerin grafik yardımıyla görsel olarak saptanması imkanını vermektedir. Birinci simülasyon çalışmasının sonuçları incelendiğinde gerçekte veri kümesinde bir etkili gözlem varken BYDGY ile bu etkili gözlem iyi bir şekilde saptanmaktadır. İkinci simülasyon çalışmasının sonuçları incelendiğinde özellikle $n > 20$ iken Biplot Yaklaşımına Dayalı Grafik Yöntemi etkili gözlem grubunu saptama bakımından başarılıdır. BYDGY'nin etkili gözlem veya gözlem grubunu saptamada başarısının nedeni, her bir gözlemin $\hat{\beta}_j$ tahmininde meydana getirdiği değişimin klasik Biplot yaklaşımıyla daha iyi bir şekilde gözlenebilmesidir. Ayrıca bu durum bu yöntemin diğer klasik yöntemlere göre daha pratik olduğunu da göstermektedir.

Bu kısımda yapılan simülasyonla elde edilen sonuçlar genel olarak incelendiğinde örnek hacminin büyük olduğu durumlarda Cook Uzaklığı İstatistiği, COVRATIO İstatistiği, K

İstatistiği, SIC'ye Dayalı Grafik Yöntemi ve Biplot Yaklaşımına Dayalı Grafik Yönteminin iki etkili gözlem içeren veride bulunan etkili gözlem grubunu saptama oranı bakımından iyi sonuçlar verdiği görülmektedir. Parametre sayısının büyük olduğu durumlarda ise Cook Uzaklığı İstatistiği ve COVRATIO İstatistiğinin iki etkili gözlem içeren veride bulunan etkili gözlem grubunu saptama oranı bakımından iyi olduğu gözlenmiştir. Özellikle $p = 5$ durumunda $n = 30$ ve $n = 40$ iken Cook Uzaklığı İstatistiği, COVRATIO İstatistiği, K İstatistiği ve Biplot Yaklaşımına Dayalı Grafik Yöntemi iki etkili gözlemden oluşan etkili gözlem grubunu saptama oranı bakımından diğer yonteme göre daha iyi sonuçlar vermiştir.

7.3. Parametre Sayısı ve Örnek Hacmi Sabitken Önerilen İki Yöntemin Klasik Tanı Yöntemleriyle Kaldıraç Değerindeki Değişim Bakımından Karşılaştırılması

Simülasyon çalışmasının ilk kısmında örnek hacmi ve parametre sayısının farklı durumları için önerilen iki yöntemle bir gözlem silme yaklaşımına dayalı klasik tanı yöntemleri karşılaştırılırken simülasyon tasarımında, i' ile ifade edilen gözlemin bağımsız değişken değerleri ve hata değeri sabit olmak üzere etkili gözlem üretilmeye çalışılmıştır. İkinci kısımda ise önerilen iki yöntemle birden fazla gözlem silme yaklaşımına dayalı klasik tanı yöntemlerinin karşılaştırması için simülasyon tasarımında, bu etkili gözlem üretme yaklaşımının iki gözlem için genişletilmiş hali kullanılmıştır. Bu kısımda ilk olarak örnek hacmi ve parametre sayısı sabit iken önerilen iki yöntemle bir gözlem silme yaklaşımına dayalı klasik tanı yöntemleri ele alınıp, i' ile ifade edilen gözlemin bağımsız değişken değerleri verinin geneline göre daha uç olmak üzere değerce büyütüldüğünde, bu yöntemlere ilişkin etkili gözlemi saptama oranları elde edilmiştir. Daha sonra örnek hacmi ve parametre sayısı sabit iken önerilen iki yöntemle birden fazla gözlem silme yaklaşımına dayalı klasik tanı yöntemleri ele alınmış ve bu yöntemler için, $I' = \{i', i''\}$ ile ifade edilen etkili gözlem grubunu oluşturan gözlemlerin bağımsız değişken değerleri verinin geneline göre daha uç olmak üzere, değerce büyütüldüğünde etkili gözlemi saptama oranları elde edilmiştir. Buradaki amaç etkili gözlem veya gözlem grubunu oluşturan gözlemler kaldıraç değerleri bakımından verinin genelinden farklılaştırıldığında tanı yöntemlerini karşılaştırmaktır.

İlk olarak $n = 20$ ve $p = 3$ sabit olarak alındığında, önerilen iki yöntemle bir gözlem silme yaklaşımına dayalı klasik tanı yöntemleri ele alınsın. Simülasyon sonuçlarını elde etmek

amacıyla bölümün ilk kısmında verilen simülasyon tasarımından yararlanılmış ve bu tasarımda adım 3'te belirtilen i' ile ifade edilen gözlemin bağımsız değişkenlerinin değerleri sırasıyla $x_{ij} = 20$, $x_{ij} = 30$ ve $x_{ij} = 40$ ($j = 1, 2, \dots, p-1$) olarak alınmıştır. Sonra bölümün ilk kısmında verilen simülasyon tasarımında yer alan adımlar doğrultusunda her 1000 tekrarda bir, verinin geneline uyan gözlemlere ilişkin bağımsız değişkenlerin değerleri yeniden üretilmek üzere 100 000 deneme yapıp, etkili gözleme ilişkin bağımsız değişkenlerin farklı değerleri için tanı yöntemlerinin etkili gözlemi saptama oranları elde edilmiştir. Simülasyon sonuçları Çizelge 7.10'da verilmiştir.

Çizelge 7.10. $n = 20$ ve $p = 3$ iken önerilen iki yöntemle bir gözlem silme yaklaşımına dayalı tanı yöntemlerinin, farklı bağımsız değişken değerlerine göre üretilen etkili gözlemi saptama oranları

Tanı Yöntemi	Etkili olarak üretilen gözlemin bağımsız değişkenlerinin değerleri		
	$x_{ij} = 20$ ($j = 1, \dots, p-1$)	$x_{ij} = 30$ ($j = 1, \dots, p-1$)	$x_{ij} = 40$ ($j = 1, \dots, p-1$)
SICDGY	0,7830	0,8400	0,8600
BYDGY	0,9794	0,9816	0,9906
$D_{(i')}$	0,9424	0,9396	0,9162
$DFBETAS_{j(i')}$ ($j = 0, 1, \dots, p$)	0,9586	0,9316	0,8602
$DFFITS_{(i')}$	0,9696	0,9510	0,9470
$COVRATIO_{(i')}$	0,6594	0,7086	0,9405
$AP_{(i')}$	0,9554	0,9855	0,9965
$LD_{(i')}$	0,9209	0,8741	0,8114
$K_{(i')}$	0,9742	0,9856	0,9998
$S_{(i')}$	0,8872	0,9067	0,9318

Çizelge 7.10'daki sonuçlar $n = 20$ ve $p = 3$ iken i' ile ifade edilen etkili gözlemin bağımsız değişken değerlerini büyütmek suretiyle bu gözlemin, kaldıraç değeri bakımından daha uç bir gözlem olması amaçlanarak elde edilmiştir. Doğal olarak i' ile ifade edilen gözlemin bağımsız değişken değerlerini büyütmek suretiyle bu gözlemi daha uç bir gözlem yapmak, ilgili gözleme ilişkin kaldıraç değerini de büyütme karşılık gelir.

Kaldıraç değerinin büyümesiyle klasik tanı yöntemlerinden Cook Uzaklığı, DFBETAS, DFFITS ve LD istatistikleri için elde edilen etkili gözlemi saptama oranlarda bir azalma söz konusudur. Bu azalış Cook Uzaklığı ve DFFITS için dikkate değer bir azalış değildir. Klasik tanı yöntemlerinden COVRATIO, AP, Hadi'nin K ve Pena'nın S istatistikleri için oranlar incelendiğinde kaldıraç değerinin büyümesiyle bu oranlarda bir artış söz

konusudur. Benzer biçimde SICDGY ve BYDGY için elde edilen oranlarda da bir artış gözlenmektedir.

İkinci olarak önerilen iki yöntemle birden fazla gözlem silme yaklaşımına dayalı tanı istatistikleri ele alınmıştır. Simülasyon sonuçlarını elde etmek amacıyla bölümün ikinci kısmında belirtilen algoritma ele alınıp, $n=20$ ve $p=3$ olmak üzere, adım 3'te belirtilen i' ile ifade edilen gözlemin bağımsız değişkenlerinin değerleri sırasıyla $x_{ij} = 20$, $x_{ij} = 30$ ve $x_{ij} = 40$; i'' ile ifade edilen gözlemin bağımsız değişkenlerinin değerleri ise sırasıyla $x_{ij} = 21$, $x_{ij} = 31$ ve $x_{ij} = 41$ ($j = 1, 2, \dots, p-1$) olarak belirlenmiştir. Sonra bölümün ikinci kısmında verilen simülasyon tasarımında yer alan adımlar doğrultusunda, her 1000 tekrarda bir, verinin geneline uyan gözlemlere ilişkin bağımsız değişkenlerin değerleri yeniden üretilmek üzere 100 000 deneme yapıp etkili gözlem grubunu saptama oranları elde edilmiştir. Böylece $I' = \{i', i''\}$ ile ifade edilen etkili gözlem grubunun elemanları kaldıraç değeri bakımından verinin genelinden farklılaştırıldığında, birden fazla gözlem silme yaklaşımına dayalı tanı yöntemlerinin bu etkili gözlem grubunu saptama oranları elde edilip sonuçlar Çizelge 7.11'de verilmiştir.

Çizelge 7.11. $n=20$ ve $p=3$ iken önerilen iki yöntemle birden fazla gözlem silme yaklaşımına dayalı tanı yöntemlerinin, farklı bağımsız değişken değerlerine göre üretilen iki etkili gözlemin oluşturduğu etkili gözlem grubunu saptama oranları

Tanı Yöntemi	Etkili olarak üretilen gözlem grubunu oluşturan gözlemlerin bağımsız değişkenlerinin değerleri		
	$x_{ij} = 20, x_{i'j} = 21$ ($j = 1, \dots, p-1$)	$x_{ij} = 30, x_{i'j} = 31$ ($j = 1, \dots, p-1$)	$x_{ij} = 40, x_{i'j} = 41$ ($j = 1, \dots, p-1$)
SICDGY	0,6010	0,7430	0,7570
BYDGY	0,9143	0,9356	0,9468
$D_{(I')}$	0,9465	0,9778	0,9982
$COVRATIO_{(I')}$	0,8735	0,8744	0,8806
$K_{(I')}$	0,7608	0,9658	0,9987

Çizelge 7.11'deki sonuçlar incelendiğinde, etkili gözlem grubunu oluşturan iki gözlem kaldıraç değeri bakımından verinin genelinden farklılaştırıldığında SICDGY, BYDGY, Cook Uzaklığı İstatistiği ve Hadi'nin K İstatistiği gibi tanı yöntemleri için elde edilen etkili gözlem grubunu saptama oranlarında bir artış gözlenmektedir. COVRATIO istatistiği için elde edilen sonuçlar incelendiğinde ise etkili gözlem grubu kaldıraç değerleri bakımından

verinin genelinden farklılaştırıldığında saptama oranına küçük bir artışın olduğu görülmektedir.

Veri kümesinde bir etkili gözlem varken, etkili gözlem kaldırma değeri bakımından verinin genelinden daha da farklılaştırıldığında, COVRATIO İstatistiği, AP İstatistiği, K İstatistiği, S İstatistiği, SIC'ye Dayalı Grafik Yöntemi ve Biplot Yaklaşımına Dayalı Grafik Yöntemi etkili gözlemi saptama bakımından diğer tanı yöntemlerine göre daha iyi sonuçlar vermiştir. Veri kümesinde iki gözlemden oluşan bir etkili gözlem grubunun olduğu durumda ise bu gözlemler kaldırma değeri bakımından verinin genelinden daha da farklılaştırıldığında, Cook Uzaklığı İstatistiği, COVRATIO İstatistiği, K İstatistiği, SIC'ye Dayalı Grafik Yöntemi ve Biplot Yaklaşımına Dayalı Grafik Yöntemi etkili gözlem grubunu saptama bakımından iyi sonuçlar vermiştir.

7.4. Parametre Sayısı ve Örnek Hacmi Sabitken Önerilen İki Yöntemin Klasik Tanı Yöntemleriyle Hata Değişkeni Değerindeki Değişim Bakımından Karşılaştırılması

Bu kısımda önce önerilen iki tanı yöntemiyle bir gözlem silme yaklaşımına dayalı klasik tanı yöntemleri ele alınıp, etkili gözlemin üretiminde kullanılan hata değişkeni değerinin mutlak değerce büyütüldüğü durumda bu yöntemlerin etkili gözlemi saptama oranları üzerinden tanı yöntemleri karşılaştırılmıştır. Sonra önerilen iki tanı yöntemiyle birden fazla gözlem silme yaklaşımına dayalı klasik tanı yöntemleri ele alınıp, iki elemanlı bir etkili gözlem grubunu oluşturan gözlemlerin üretiminde kullanılan hata terimlerinin değerlerinin mutlak değerce büyütüldüğü durumda, bu gözlem grubunu saptama oranları üzerinden tanı yöntemleri karşılaştırılmıştır. Buradaki amaç, etkili gözlem veya etkili gözlem grubunu oluşturan gözlemler, hata değişkeni bakımından verinin genelinden farklılaştırıldığında tanı yöntemlerini saptama oranı bakımından incelemektir.

İlk olarak $n=20$ ve $p=3$ sabit alındığında, önerilen iki yöntemle bir gözlem silme yaklaşımına dayalı klasik tanı yöntemleri ele alınsın. Simülasyon sonuçlarını elde etmek amacıyla bölümün ilk kısmında verilen simülasyon tasarımından yararlanılmış ve bu tasarımda adım 3'te belirtilen i' ile ifade edilen gözlemin hata terimi değeri sırasıyla $\varepsilon_{i'} = -5$, $\varepsilon_{i'} = -7.5$ ve $\varepsilon_{i'} = -10$ olarak alınmıştır. Sonra bölümün ilk kısmında verilen simülasyon tasarımında yer alan adımlar doğrultusunda her 1000 tekrarda bir, verinin

geneline uyan gözlemlere ilişkin bağımsız değişkenlerin değerleri yeniden üretilmek üzere 100 000 deneme yapıp, hata değişken değerinin farklı olduğu durumlarda, etkili gözlemi saptama oranları elde edilmiştir. Simülasyon sonuçları Çizelge 7.12’de verilmiştir.

Çizelge 7.12. $n = 20$ ve $p = 3$ iken önerilen iki yöntemle bir gözlem silme yaklaşımına dayalı tanı yöntemlerinin, farklı hata değişken değerlerine göre üretilen etkili gözlemi saptama oranları

Tanı Yöntemi	Etkili olarak üretilen gözlemin hata değişkeni değerleri		
	$\varepsilon_{i'} = -5$	$\varepsilon_{i'} = -7.5$	$\varepsilon_{i'} = -10$
SICDGY	0,7830	0,7900	0,8180
BYDGY	0,9794	0,9980	0,9998
$D_{(i')}$	0,9424	0,9895	0,9967
$DFBETAS_{j(i')} (j = 0, 1, \dots, p)$	0,9586	0,9768	0,9701
$DEFFITS_{(i')}$	0,9696	0,9917	0,9925
$COVRATIO_{(i')}$	0,6594	0,9088	0,9879
$AP_{(i')}$	0,9554	0,9895	0,9923
$LD_{(i')}$	0,9209	0,9982	0,9986
$K_{(i')}$	0,9742	0,9891	0,9989
$S_{(i')}$	0,8872	0,9140	0,9331

Çizelge 7.12’deki sonuçlar incelendiğinde $n = 20$ ve $p = 3$ iken etkili gözlem hata teriminin değeri bakımından verinin genelinden farklılaştırıldığında, COVRATIO istatistiği başta olmak üzere SICDGY, Pena’nın S İstatistiği ve LD İstatistiği için elde edilen etkili gözlemi saptama oranlarında hızlı bir artış gözlenmektedir. Ayrıca Çizelge 7.12’den gözlemlendiği üzere BYDGY, Cook Uzaklığı, DFBETAS, DEFFITS, AP ve Hadi’nin K İstatistiği için elde edilen oranlarda artış söz konusudur ve bu oranlar dikkate değer biçimde yüksektir. Bu oranlar etkili gözleme ilişkin hata teriminin değeri mutlak değerce büyüdüğünde 1 değerine yaklaşmaktadır.

İkinci olarak $n = 20$ ve $p = 3$ olmak üzere sabit alınmışken önerilen iki yöntemle birden fazla gözlem silme yaklaşımına dayalı klasik tanı yöntemleri ele alınmıştır. Simülasyon sonuçlarını elde etmek amacıyla bölümün ikinci kısmında verilen simülasyon tasarımından yararlanılmış ve bu tasarımda adım 3’te belirtilen i' ve i'' ile ifade edilen gözlemlerin hata terimlerinin değeri sırasıyla $\varepsilon_{i'} = \varepsilon_{i''} = -5$, $\varepsilon_{i'} = \varepsilon_{i''} = -7,5$ ve $\varepsilon_{i'} = \varepsilon_{i''} = -10$ olarak alınmıştır. Daha sonra bölümün ikinci kısmında verilen simülasyon tasarımında yer alan adımlar doğrultusunda her 1000 tekrarda bir, verinin geneline uyan gözlemlere ilişkin bağımsız değişkenlerin değerleri yeniden üretilmek üzere 100 000 deneme yapıp, farklı

hata değişken değerlerine göre üretilen iki etkili gözlemin oluşturduğu etkili gözlem grubunu saptama oranları elde edilmiştir. Simülasyon sonuçları Çizelge 7.13'te verilmiştir.

Çizelge 7.13. $n = 20$ ve $p = 3$ iken önerilen iki yöntemle birden fazla gözlem silme yaklaşımına dayalı tanı yöntemlerinin, farklı hata değişken değerlerine göre üretilen iki etkili gözlemin oluşturduğu etkili gözlem grubunu saptama oranları

Tanı Yöntemi	Etkili olarak üretilen gözlem grubunu oluşturan gözlemlerin hata değişkenlerinin değerleri		
	$\varepsilon_{i'} = \varepsilon_{i''} = -5$	$\varepsilon_{i'} = \varepsilon_{i''} = -7.5$	$\varepsilon_{i'} = \varepsilon_{i''} = -10$
SICDGY	0,6010	0,6105	0,6091
BYDGY	0,9143	0,9752	0,9997
$D_{(1)}$	0,9465	0,9789	0,9956
COVRATIO ₍₁₎	0,8735	0,9401	0,9987
$K_{(1)}$	0,7608	0,9366	0,9866

Çizelge 7.13'teki sonuçlar incelendiğinde $n = 20$ ve $p = 3$ iken etkili gözlem grubunu oluşturan iki gözlemin hata değişken değerleri mutlak değerce büyüdüğünde, SICDGY için elde edilen oranlarda dikkate değer bir değişim gözlenmemektedir. Bununla birlikte etkili gözlem grubunu oluşturan gözlemlere ilişkin hata değerleri mutlak değerce büyüdüğünde BYDGY, Cook Uzaklığı, COVRATIO ve Hadi'nin K İstatistiği için elde edilen oranlarda bir artış gözlenmektedir ve bu oranlar 1 değerine yaklaşmaktadır.

Veri kümesinde bir etkili gözlem varken etkili gözlem, hata değişken değeri bakımından verinin genelinden daha da farklılaştırıldığında, ele alınan tüm klasik tanı istatistikleri ile önerilen iki grafiksel yöntem etkili gözlemi saptama oranı bakımından iyi sonuçlar vermiştir. Veri kümesinde iki gözlemde oluşan bir etkili gözlem grubu varken etkili gözlemler, hata değişken değerleri bakımından verinin genelinden daha da uzaklaştırıldığı durumda ise Cook Uzaklığı İstatistiği, COVRATIO İstatistiği, K İstatistiği ve BYDGY etkili gözlem grubunu saptama oranı bakımından diğer tanı yöntemine göre daha iyi sonuçlar vermiştir.

8. SONUÇ VE ÖNERİLER

Bu çalışmada birinci bölümde aykırı ve etkili gözlemlerin tanımı verilmiş ve etkili gözlemlerin saptanmasının öneminden bahsedilmiştir. Ayrıca literatürde yer alan, etkili gözlem veya gözlem gruplarının saptanması için bilinen, tanı istatistiklerinin dezavantajları da belirtilmiştir.

İkinci bölümde çoklu doğrusal regresyon analizine ilişkin temel kavramlar verilmiştir. Ayrıca tanı yöntemlerinin temelini oluşturan bazı istatistikler ve kavramlar tanıtılmıştır. Bu kavramlardan student tipi artıkların ve kaldıraç değerlerinin önemi vurgulanmıştır, çünkü etkili bir gözlem genellikle artığı mutlak değerce büyük olduğundan ve genellikle yüksek dereceli kaldıraç noktası olduğundan tanı istatistiklerinin temeli kaldıraç değerlerine, artıklara ve artıkların standart biçimlerine dayalıdır.

Üçüncü bölümde literatürde bilinen ve sıklıkla kullanılan tanı istatistikleri verilmiştir. Ayrıca bunların karar kurallarında yer alan kritik değerlerden bahsedilmiştir.

Dördüncü bölümde etkili gözlemlerin saptanması için örnek etki eğrisi (SIC) kavramına dayalı bir grafik yöntem (SICDGY) önerisi yer almaktadır. Bu yöntemle SIC fonksiyonundan oluşturulan bir amaç fonksiyonunun optimum çözümü elde edilmeye çalışılır. Bu çözüm $n \times 1$ boyutlu bir vektör ile gösterilmiştir ve bu vektörün i . elemanı, ilgili gözlemin regresyon tahminlerini değiştirme eğilimine ilişkin katsayıdır. i . gözleme karşılık gelen katsayı önerilen kritik değerlerin oluşturduğu aralığın dışında ise ilgili gözlem etkili gözlem olarak belirlenir. Yöntemle elde edilen sonuçlar aynı zamanda grafik üzerinden de incelenebilir.

Beşinci bölümde etkili gözlemlerin saptanması için biplot yaklaşımına dayalı bir başka grafik yöntemi önerisi (BYDGY) yer almaktadır. Önerilen grafik yöntemi sırasıyla $\hat{\beta}_{j(i)}$ ve $\hat{\beta}_{j(i)} - \hat{\beta}_j$ değişkenleri için oluşturulan Ψ ve Δ matrisleri kullanılarak çizilen biplota dayalıdır. Ayrıca bu bölümde etkili gözlem veya gözlemlerin görsel olarak saptanmasına yardımcı olmak amacıyla elips biçiminde bir sınır da önerilmiştir.

Altıncı bölümde bu çalışmada tanıtılmış klasik tanı istatistikleriyle önerilen iki grafik yönteminin iki veri kümesi üzerinden uygulamasına yer verilmiştir. Literatürde tanı

yöntemlerinin incelenmesi için sıklıkla kullanılan bu veri kümeleri etkili gözlem veya gözlemler içerdiği bilinen veri kümeleridir.

Yedinci bölümde ise bu çalışmada tanıtılmış klasik tanı istatistikleriyle önerilen iki grafik yönteminin karşılaştırılmasına ilişkin simülasyon çalışması bulunmaktadır. Simülasyon çalışması dört kısımdan oluşmaktadır. İlk kısımda örnek hacmi ve parametre sayısının farklı durumlarında veri kümesinde gerçekte bir etkili gözlem varken, bir gözlem silme yaklaşımına dayalı klasik tanı istatistikleri ile önerilen iki grafik yöntemi için simülasyonla elde edilen, etkili gözlemi saptama oranları ve verinin geneline uyan diğer gözlemlerin en az birini etkili olarak saptama oranları verilip, bu oranlar yorumlanmıştır. Ayrıca bu sonuçlara dayalı olarak İki Yönlü ANOVA yöntemi uygulanıp elde edilen p-değerleri üzerinden tanı yöntemleri karşılaştırılmıştır.

Simülasyon çalışmasının ikinci kısımda örnek hacmi ve parametre sayısının farklı durumlarında veri kümesinde gerçekte etkili gözlem grubunu oluşturan iki etkili gözlem varken, etkili gözlem gruplarını saptamaya dayalı tanı istatistikleriyle önerilen iki grafik yöntemi için simülasyonla elde edilen, etkili gözlem grubunu saptama oranları ve verinin geneline uyan diğer gözlemlerin ikili kümelerinin en az birini etkili gözlem grubu olarak saptama oranları verilip, bu oranlar yorumlanmıştır. Ayrıca simülasyon çalışmasının birinci kısmında yapıldığı gibi, bu sonuçlara dayalı olarak İki Yönlü ANOVA yöntemi uygulanıp elde edilen p-değerleri üzerinden tanı yöntemleri karşılaştırılmıştır.

Simülasyon çalışmasının üçüncü kısmında örnek hacmi ve parametre sayısı sabit iken etkili gözlem veya etkili gözlem grubunu oluşturan gözlemlerin bağımsız değişkenlerinin değerleri daha da büyütülmek suretiyle, bu etkili gözlem veya gözlem grubunu saptama oranları elde edilmiş ve bu oranlar üzerinden önerilen yöntemlerle klasik tanı yöntemleri karşılaştırılmıştır. Buradaki amaç etkili gözlem veya etkili gözlem grubunu oluşturan gözlemler kaldıraç değeri bakımından verinin genelinden farklılaştırıldığında tanı yöntemlerini karşılaştırmaktır.

Simülasyon çalışmasının son bölümünde örnek hacmi ve parametre sayısı sabitken etkili gözlem veya etkili gözlem grubunu oluşturan gözlemlerin hata değişken değeri mutlak değerce daha da büyütülmek suretiyle, bu etkili gözlem veya gözlem grubunu saptama

oranları elde edilmiş ve bu oranlar üzerinden önerilen yöntemlerle klasik tanı yöntemleri karşılaştırılmıştır.

Bir gözlem silme yaklaşımına dayalı simülasyon sonuçları genel olarak değerlendirildiğinde, örnek hacminin büyük olduğu durumda birçok tanı istatistiğinin etkili gözlemi saptama oranının büyük olduğu gözlenmiştir. Örnek hacmi büyükken DFBETAS, DFFITS ve COVRATIO istatistiklerine ilişkin oranların büyük olmasının nedeni bu istatistiklerin kritik değerlerinin örnek hacminin bir fonksiyonu olup, n arttıkça bu değerlerin sayısal olarak küçülmesidir. Bununla birlikte büyük hacimli bir örnekte bu istatistikler için kritik değerin küçülmesi, verinin geneline uyan ve etkili olmayan bir gözlemin de etkili gözlem olarak saptanmasına neden olabilir. Dolayısıyla bu durum araştırmacıyı yanlış karar almaya yönlendirir.

Veri kümesinde bir etkili gözlem olduğu durumda, Cook Uzaklığı İstatistiği için elde edilen oranlar da sayısal olarak büyük değerlerdir. Bununla birlikte örnek hacminin küçük değerleri ve parametre sayısının büyük değerleri için bu istatistiğe ilişkin etkili gözlemi saptama oranlarında bir düşme gözlenmiştir. Buna ek olarak Cook Uzaklığı için elde edilen etkili olmayan gözlemlerden en az birini etkili olarak saptama oranları sonuçların geneline göre düşük seviyededir. Cook ve Weisberg Olabilirlik İstatistiği için elde edilen etkili gözlemi saptama oranları incelendiğinde ise örnek hacmi ve parametre sayısının artmasıyla bu oranların düştüğü sonucuna varılmıştır.

AP istatistiği için veri kümesinde bir etkili gözlem varken, bu etkili gözlemi saptama oranları büyük değerleridir. Fakat parametre sayısının küçük olduğu durumlarda bu istatistik için elde edilen etkili olmayan gözlemlerden en az birini etkili olarak saptama oranları nispeten büyük değerlerdir.

Hadi'nin K İstatistiği için bir etkili gözlemi saptama oranlarının p sabitken örnek hacmi arttığında sayısal olarak büyüdüğü, n sabitken parametre sayısı arttığında sayısal olarak küçüldüğü gözlenmiştir. K İstatistiğinin kritik değeri bu istatistik için örnek ortalaması ve örnek varyansına dayalı olduğundan bu kritik değer, etkili gözlemin büyük değere sahip K_1 değerinden etkilenecek sayısal olarak büyümektedir. Bu durum yöntemin bir dezavantajıdır. Hadi'nin K İstatistiği için etkili olmayan gözlemlerden en az birini etkili

olarak saptama oranları incelendiğinde, bu oranların genele göre düşük olduğu görülmektedir.

n sabit olduğunda, parametre sayısının küçük olduğu durumlarda Pena'nın S İstatistiği için elde edilen, etkili gözlemi saptama oranlarının sonuçların geneline göre sayısal olarak küçük olduğu gözlenmiştir. Ayrıca p sabitken, örnek hacmi arttığında bu istatistik için oranlarda da bir düşme söz konusudur. Pena'nın S İstatistiği için etkili olmayan gözlemlerden en az birini saptama oranları incelendiğinde, bu oranların genele göre daha büyük olduğu gözlenmiştir.

Bu çalışmada önerilen Biplot Yaklaşımına Dayalı Grafik Yöntemi için bir etkili gözlemi saptama oranları incelendiğinde, tüm durumlarda bu oranların yüksek olduğu açıkça görülmektedir. Ayrıca veri kümesinde bir etkili gözlem varken, BYDGY için elde edilen, etkili gözlem haricinde diğer gözlemlerden en az birini etkili gözlem olarak saptama oranlarının da düşük değerler olduğu gözlenmiştir.

SIC'ye Dayalı Grafik Yöntemi için elde edilen bir etkili gözlemi saptama oranları incelendiğinde, örnek hacminin büyük olduğu durumlarda etkili gözlemi saptama oranının da büyük olduğu gözlenmiştir. Ayrıca bu yöntem için etkili gözlem haricinde diğer gözlemlerden en az birini etkili gözlem olarak saptama oranları incelendiğinde, bu oranların nispeten küçük oranlar olduğu görülmektedir.

Veri kümesinde iki etkili gözlem varken, Cook Uzaklığı İstatistiği'nin etkili gözlem grubunu saptamak için genelleştirilmiş biçimiyle elde edilen oranların, örnek hacmi ve parametre sayısının tüm durumlarda genel olarak büyük olduğu açıktır. Diğer bir istatistik olan COVRATIO istatistiğinin genelleştirilmiş biçimi için elde edilen etkili gözlemi saptama oranları incelendiğinde, örnek hacminin küçük olduğu durumlarda, parametre sayısı arttıkça bu oranların büyüdüğü görülmektedir. Ayrıca p sabitken örnek hacmi arttıkça bu oranların arttığı gözlenmiştir. Kritik değeri örnek hacminin azalan bir fonksiyonu olan bu istatistik için örnek hacminin büyük olduğu durumlarda etkili gözlem grubunu saptama oranının da büyük olması doğaldır. Örnek hacmi büyükken COVRATIO istatistiği için kritik değerin küçülmesi, gerçekte etkili olan gözlem grubunun saptanması için bir avantaj olmasına rağmen, bu kritik değerin küçülmesi etkili olmayan gözlemlerin oluşturduğu grupların da etkili gözlem grubu olarak saptanmasına neden olabilir. Nitekim

Çizelge 7.8'deki sonuçlardan bu açıkça görülmektedir. Böylece bu durum araştırmacıyı yanlış karar almaya yönlendirir.

Hadi'nin K istatistiğinin gözlem grupları için genişletilmiş biçimi için elde edilen sonuçlara göre büyük hacimli örneklerde bu istatistiğin etkili gözlem grubunu saptama oranının büyük olduğu gözlenmiştir. Bununla birlikte örnek hacminin küçük, parametre sayısının büyük olduğu durumda K istatistiğinin etkili gözlem grubunu saptama oranının düşük olduğu görülmüştür. Kritik değeri K istatistiğinin ortalamasına ve varyansına dayalı olan bu yöntem için bu sonuç beklenen bir sonuçtur. Buna ek olarak Hadi'nin K istatistiği için etkili gözlem grubunu oluşturan gözlemler haricinde verinin geneline uyan diğer gözlem çiftlerinin en az birini etkili gözlem grubu olarak saptama oranları dikkate değer biçimde büyüktür.

Bu çalışmada önerilen Biplot Yaklaşımına Dayalı Grafik Yöntemi için etkili gözlem grubunu saptamaya dayalı simülasyon sonuçları incelendiğinde, örnek hacminin 20 değerinden büyük olduğu durumlarda, bu yöntemin etkili gözlem grubunu saptamada iyi olduğu sonucuna varılır. Ayrıca önerilen bu yöntem için etkili olmayan gözlemlerin en az birini etkili gözlem olarak saptama oranları da düşük oranlardır.

SIC'ye Dayalı Grafik Yöntemi büyük hacimli örnekler için etkili gözlem grubunu saptamada başarılıdır. Ayrıca önerilen bu yöntem için etkili olmayan gözlemlerin en az birini etkili gözlem olarak saptama oranları da genele göre nispeten düşük oranlardır. Bununla birlikte bu çalışmada etkili gözlem veya gözlem grubunu saptamak amacıyla SIC'ye Dayalı Grafik Yöntemi'nde yer alan amaç fonksiyonunun çözümü için MATLAB2011 programında *patternsearch* komut kullanılmıştır. Bu yöntem için farklı optimizasyon yaklaşımları daha iyi sonuçlar elde etmek amacıyla kullanılabilir.

Bu çalışmada yapılan diğer iki simülasyon, örnek hacmi ve parametre sayısı sabitken, etkili gözlem veya etkili gözlem grubunu oluşturan gözlemler, kaldırma değeri ve hata değişken değeri bakımından verinin geneline göre farklılaştırıldığında, önerilen iki yöntemle klasik tanı yöntemlerinin karşılaştırmasına ilişkindir. Veri kümesinde bir etkili gözlem varken, bu gözlem kaldırma değeri bakımından verinin genelinden uzaklaştırıldığında, klasik tanı istatistiklerinden Cook Uzaklığı, DFBETAS, DFFITS ve LD istatistikleri için elde edilen etkili gözlemi saptama oranlarında bir düşüş

gözleniyorken; COVRATIO, AP, K ve S istatistikleri için elde edilen oranlarda bir artış gözlenmiştir. Ayrıca SICDGY ve BYDGY için de bir artış söz konusudur. Veri kümesinde etkili gözlem grubunu oluşturan iki etkili gözlem varken, bu gözlemler kaldıraç değeri bakımından verinin genelinden uzaklaştırıldığında klasik tanı yöntemlerinden Cook Uzaklığı, COVRATIO ve K istatistikleri için elde edilen etkili gözlem grubunu saptama oranlarında bir artış gözlenmiştir. Benzer artış bu çalışmada önerilen iki yöntem olan BYDGY ve SICDGY için de söz konusudur.

Son olarak örnek hacmi ve parametre sayısı sabitken, veri kümesinde yer alan etkili gözlem ve etkili gözlem grubunu oluşturan gözlemlere ilişkin hata değişkeninin değerleri mutlak değerce arttırıldığında, bu gözlem veya gözlemler verinin genelinden hata değişken değeri bakımından uzaklaştırıldığında, saptama oranları elde edilmiştir. Hata değişken değeri mutlak değerce arttırıldığında klasik tanı yöntemleri ve önerilen iki yöntem için etkili gözlemi saptama oranlarının da büyüdüğü gözlenmiştir.

Gerçekte birden fazla etkili gözlem içeren veri kümelerinde etkili gözlem grubunun, gözlem silme yaklaşımına dayalı tanı istatistikleriyle saptanması için temel sorun incelenecek gözlem gruplarının k hacimli kombinasyonlarının çok olmasıdır. Böylece toplam kombinasyon sayısı kadar EKK yönteminin uygulanması gerekir. Bundan dolayı bu tanı istatistikleri pratik değildir. Bu çalışmada önerilen iki yöntemle bu kombinasyonların her biri için ayrı bir işlem yapılması gereksinimi duyulmaz. Dolayısıyla önerilen bu iki yöntem klasik tanı yöntemlerine göre daha pratiktir.

KAYNAKLAR

- Akdeniz, F., Öztürk, F. (1996). *Lineer Modeller*. Ankara: A.Ü.F.F. Döner Sermaye İşletmesi Yayınları, 34.
- Andrews, D.F., Pregibon, D. (1978). Finding Outliers That Matter. *Journal of Royal Statistical Society, Series B*. 40(1), 85-93.
- Andrews, D.F., Bickel, P., Hampel, F., Huber, F. Rogers, W.H., Tukey, J.W. (1972). *Robust Estimates of Location*. New Jersey: Princeton University Press, 1-369.
- Atkinson, A.C. (1981). Two Graphical Displays for Outlying and Influential Observations in Regression. *Biometrika*, 68(1), 13-20.
- Atkinson, A.C. (1986). Masking Unmasked. *Biometrika*, 73 (3), 533-541.
- Atkinson, A.C. (2004). Plots, Transformations and Regression. Oxford: Oxford Science Publications, 62-78.
- Atkinson, A.C., Riani M. (2000). *Robust Diagnostic Regression Analysis*. New York: Springer, 20-22.
- Belsley, D.A., Kuh, E., Welsch, R.E. (1980). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. New York: Willey Series in Probability and Mathematical Statistics, 6-84.
- Caussinus, H., Fekri, M., Hakam, S., Gazen, A.R. (2003). A Monitoring Display of Multivariate Outliers. *Computational Statistics and Data Analysis*, 44 (2), 237-252.
- Chatterjee, S., Hadi, A.S. (1986). Influential Observations, High Leverage Points and Outliers in Linear Regression. *Statistical Science*, 1(3): 379-393.
- Chatterjee S., Hadi A.S. (2006). *Regression Analysis by Example*. New Jersey: Willey Series in Probability and Mathematical Statistics, 85-115.
- Cook, R.D. (1977). Detection of Influential Observations in Linear Regression. *Technometrics*, 19 (1): 15-18.
- Cook, R.D. (1998). *Regression Graphics*. New York: Willey Series in Probability and Mathematical Statistics, 49-51.
- Cook, R.D., Weisberg, S. (1982). *Residuals and Influence in Regression*. New York: Chapman and Hall, 10-190.
- Ezekiel, M., Fox, K.A. (1959). *Methods of Correlation and Regression Analysis, Linear and Curvilinear*. (3rd Edition), New York: John Willey & Sons, 151-346.

- Gabriel, K.R. (1971). The Biplot Graphics Display of Matrices with Application to Principal Component Analysis. *Biometrika*, 58 (3), 453-467,
- Gabriel, K.R. (1995). *MANOVA Biplots for Two-way Contingency Tables, Recent Advantages in Descriptive Multivariate Analysis*. Oxford: Clarendon Press, 227-268.
- Galpin, J.S., Hawkins, D.M. (1984). The Use of Recursive Residuals in Checking Model Fit in Linear Regression. *Journal of The American Statistician*, 38 (2), 94-105,
- Gardner, S. (2001). *Extensions of Biplot Methodology to Discriminant Analysis with Applications of Non-parametric Principal Components*. Stellenbosch: PhD thesis, University of Stellenbosch, 1-473.
- Gloub, G.H., Van Loan, C.F. (1996). *Matrix Computations*. London: The Johns Hopkins University Press, 48-85.
- Gower, J.C. Hand, D.J. (1996). *Biplots*. London: Chapman & Hall, 1-264.
- Gower, J., Lubbe, S., Roux, N. (2011). *Understanding Biplot*. West Sussex: John Wiley & Sons, 1-67.
- Gray, J.B., Ling, R.F. (1984). K-Clustering as a Detection Tool for Influential Subsets in Regression. *Technometrics*, 26(4), 305-318.
- Greenacre, M. (1984). *Theory and Applications of Correspondence Analysis*. London: Academic Press, 1-364.
- Greenacre, M. (2010). *Biplots in Practice*. Bilbao: Fundacion BBVA, 15-198.
- Hadi A.S. (1992). A New Measure of Overall Potential Influence in Linear Regression. *Computational Statistics and Data Analysis*, 14(1), 1-27.
- Hadi, A.S., Simonoff, J.S. (1993). Procedures for the Identification of Multiple Outliers in Linear Models. *Journal of the American Statistical Association*, 88(424), 1264-1272.
- Hampel F.R. (1974). The Influence Curve and Its Role in Robust Estimation. *Journal of The American Statistical Association*, 69(346), 383-393.
- Hoaglin, D.C., Welsch, R.E. (1978). The Hat Matrix in Regression and ANOVA. *The American Statistician*, 32(1), 17-22.
- Huber, P.J., Ronchetti, E.M. (2009). *Robust Statistics*. (2nd Edition), New Jersey: Willey Series in Probability and Mathematical Statistics, 105-115.

- Kim, C., Storer, B.E. (1996). Reference Values for Cook's Distance. ***Communications in Statistics – Simulation and Computation***, 25(3), 691-708.
- Lawrance, A.J. (1995). Deletion Influence and Masking in Regression. ***Journal of the Royal Statistical Society. Series B (Methodological)***, 57(1), 181-189.
- Li, B., Martin, E.B., Morris, A.J. (2001). A Graphical Technique for Detecting Influential Cases in Regression Analysis. ***Communications in Statistics – Theory and Methods***, 30(3), 463-483.
- Li, B., Morris, A.J., Martin, E.B. (2003). Detection of Multiple Influential Cases in Principal Component Analysis: A Graphical Technique. ***Communications in Statistics – Simulation and Computation***, 32(2), 489-503.
- Margolin, B.H. (1977). The Distribution of Internally Studentized Statistics via Laplace Transform Inversion. ***Biometrika***, 64(3), 573-582.
- Mallows, C.L. (1975). *On Some Topics in Robustness*. New Jersey: Unpublished Bell Telephone Laboratories Report.
- Meloun, M., Militky, J. (2001). Detection of Single Influential Points in OLS Regression Model Buildings. ***Analytica Chimica Acta***, 439, 169-191.
- Montgomery, D.C., Peck, E.A., Vining, G.G. (2001). *Introduction to Linear Regression Analysis*. New York: Wiley Series in Probability and Mathematical Statistics, 150-219.
- Mosteller, F., Tukey, J.W. (1977). *Data Analysis and Regression*. Reading: Addison-Wesley Publishing Company, 1-570.
- Pena, D. (2005). A New Statistics for Influence in Linear Regression. ***American Statistical Association and the American Society for Quality***, 47(1), 1-12.
- Rousseeuw, P.J., Leroy, A.M. (2003). *Robust Regression and Outlier Detection*. New York: Wiley, 216-244.
- Ververidis D., Kotropoulos, C. (2008). Gaussian Mixture Modelling by Exploiting the Mahalanobis Distance. ***IEEE Transactions on Signal Processing***, 56(7), 2797-2811.
- Von Mises, R. (1947). On The Asymptotic Distribution of Differentiable Statistical Functions. ***Ann. Math. Statist.***, 18(3), 309-348.
- Welsch, R.E. (1982). *Influence Functions and Regression Diagnostics*. In *Modern Data Analysis*. New York: Academic Press, 149-169.
- Welsch, R.E., Kuh, E. (1977). Linear Regression Diagnostics. ***Massachusetts Institute of Technology and NBER Computer Research Center Working Paper***, No. 173.

EKLER

EK 1: Fare Verisi

Gözlem no	X_1 Vücut ağırlığı (gr)	X_2 Karaciğer ağırlığı (gr)	X_3 Doz oranı	Y Karaciğerdeki dozun oranı
1	176	6,5	0,88	0,42
2	176	9,5	0,88	0,25
3	190	9	1	0,56
4	176	8,9	0,88	0,23
5	200	7,2	1	0,23
6	167	8,9	0,83	0,32
7	188	8	0,94	0,37
8	195	10	0,98	0,41
9	176	8	0,88	0,33
10	165	7,9	0,84	0,38
11	158	6,9	0,80	0,27
12	148	7,3	0,74	0,36
13	149	5,2	0,75	0,21
14	163	8,4	0,81	0,28
15	170	7,2	0,85	0,34
16	186	6,8	0,94	0,28
17	146	7,3	0,73	0,30
18	181	9	0,90	0,37
19	149	6,4	0,75	0,46

EK 2: New York'taki Nehirler Verisi

Gözlem no	X_1 Tarım alanı yüzdesi	X_2 Ormanlık alan yüzdesi	X_3 Yerleşim alanı yüzdesi	X_4 Endüstriyel alan yüzdesi	Y nitrojen konsantrasyonu oranı (mg/litre)
1	26,00	63,00	1,20	0,29	1,10
2	29,00	57,00	0,70	0,09	1,01
3	54,00	26,00	1,80	0,58	1,90
4	2,00	84,00	1,90	1,98	1,00
5	3,00	27,00	29,40	3,11	1,99
6	19,00	61,00	3,40	0,56	1,42
7	16,00	60,00	5,60	1,11	2,04
8	40,00	43,00	1,30	0,24	1,65
9	28,00	62,00	1,10	0,15	1,01
10	26,00	60,00	0,90	0,23	1,21
11	26,00	53,00	0,90	0,18	1,33
12	15,00	75,00	0,70	0,16	0,75
13	6,00	84,00	0,50	0,12	0,73
14	3,00	81,00	0,80	0,35	0,80
15	2,00	89,00	0,70	0,35	0,76
16	6,00	82,00	0,50	0,15	0,87
17	22,00	70,00	0,90	0,22	0,80
18	4,00	75,00	0,40	0,18	0,87
19	21,00	56,00	0,50	0,13	0,66
20	40,00	49,00	1,10	0,13	1,25

ÖZGEÇMİŞ



Kişisel Bilgiler

Soyadı, adı : ACARLAR, Irmak
 Uyruğu : T.C.
 Doğum tarihi ve yeri : 18.04.1984, Ankara
 Medeni hali : Bekâr
 Telefon : 0 (555) 992 50 31
 e-mail : irmakacarlar@gazi.edu.tr

Eğitim

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

İş Deneyimi

Yıl	Yer	Görev
2007-2014	Gazi Üniversitesi	Araştırma Görevlisi
2014-Halen	Antalya Muratpaşa Belediyesi	İstatistikçi

Yabancı Dil

İngilizce

Yayınlar

Uluslar Arası Dergilerde Yayınlanan Makaleler

ACARLAR, I., KINACI, H., NAJJARI, V. (2014). A New Measure for Detecting Influential DMUs in DEA. *Journal of Optimization*, Volume 2014, Article ID 567692, 7 pages (Basım aşamasında).

Ulusal Hakemli Dergilerde Yayınlanan Makaleler

ACARLAR, I., ALTUNKAYNAK, B. (2012). İki Dağılış Parametresinin Eşitliği için Doğrusal Sıra İstatistiğine Dayalı Bir Test. *TÜİK İstatistik Araştırma Dergisi*, 1(9), 29-36.

ACARLAR, I., KINACI, H. (2012). Farlie-Gumbel-Morgenstern Kapula Ailesi, Bazı Genişletmeleri ve Bir Uygulama. *TÜİK İstatistik Araştırma Dergisi*, 1(9), 37-44.

ACARLAR, I. (2011). Regresyonda Bir Etkili Gözlemin Saptanması için Kullanılan Tanı Yöntemlerinin Karşılaştırılması. *Anadolu Üniversitesi Bilim ve Teknoloji Dergisi*, 1(2), 105-116.

ACARLAR, I. (2011). Varyansın Sağlam Tahmin Edicilerine Dayalı Cook Uzaklığı İstatistiğinin İncelenmesi. *Süleyman Demirel Üniversitesi Fen Bilimleri Enstitüsü Dergisi*, 1(15), 8-14.

ACARLAR, I., GAMGAM, H. (2010). Çoklu Doğrusal Regresyonda Etkili Gözlem Gruplarının Saptanması için Kullanılan Tanı Yöntemlerinin Karşılaştırılması. *TÜİK İstatistik Araştırma Dergisi*, 1(7), 83-99.

Ulusal ve Uluslar Arası Kongrelerde Sunulmuş Bildiriler

ACARLAR, I., GAMGAM, H., (2008). Çoklu Regresyonda Etkili Gözlemlerin Saptanması için Kullanılan Tanı Yöntemlerinin Karşılaştırılması, *VI. İstatistik Günleri Sempozyumu*, SAMSUN.

Hobiler

Yüzme, Gitar, Bilgisayar Teknolojileri