
GAZİ ÜNİVERSİTESİ
Gazi
1926

DOĞRUSAL REGRESYON MODELLERİNDE AYKIRI GÖZLEMLERİN TESPİTİ İÇİN SAĞLAM TAHMİN EDİCİLERE DAYALI ETKİLİ UZAKLIĞIN PERFORMANSININ İNCELENMESİ

Fulya KARAKOCA

YÜKSEK LİSANS TEZİ
İSTATİSTİK ANA BİLİM DALI

GAZİ ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

OCAK 2020

Fulya KARAKOCA tarafından hazırlanan "DOĞRUSAL REGRESYON MODELLERİNDE AYKIRI GÖZLEMLERİN TESPİTİ İÇİN SAĞLAM TAHMİN EDİCİLERE DAYALI ETKİLİ UZAKLIĞIN PERFORMANSININ İNCELENMESİ " adlı tez çalışması aşağıdaki jüri tarafından OY BİRLİĞİ ile Gazi Üniversitesi İSTATİSTİK Ana Bilim Dalında YÜKSEK LİSANS TEZİ olarak kabul edilmiştir.

Danışman: Doç. Dr. Meltem EKİZ

İstatistik Ana Bilim Dalı, Gazi Üniversitesi

Bu tezin, kapsam ve kalite olarak Yüksek Lisans Tezi olduğunu onaylıyorum. ...……

Başkan: Prof. Dr. Sibel ATAN

Ekonometri Ana Bilim Dalı, Ankara Hacı Bayram Veli Üniversitesi

Bu tezin, kapsam ve kalite olarak Yüksek Lisans Tezi olduğunu onaylıyorum. ……...

Üye: Doç. Dr. Hülya OLMUŞ

İstatistik Ana Bilim Dalı, Gazi Üniversitesi

Bu tezin, kapsam ve kalite olarak Yüksek Lisans Tezi olduğunu onaylıyorum. ……...

Tez Savunma Tarihi: 15/01/2020

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

……..…….

Prof. Dr. Sena YAŞYERLİ

Fen Bilimleri Enstitüsü Müdürü

ETİK BEYAN

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

- 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 çalışmasında 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.

Fulya KARAKOCA

15/01/2020

DOĞRUSAL REGRESYON MODELLERİNDE AYKIRI GÖZLEMLERİN TESPİTİ İÇİN SAĞLAM TAHMİN EDİCİLERE DAYALI ETKİLİ UZAKLIĞIN PERFORMANSININ İNCELENMESİ

(Yüksek Lisans Tezi)

Fulya KARAKOCA

GAZİ ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

Ocak 2020

ÖZET

Çoklu doğrusal regresyon analizinde verinin bütününden farklılık gösteren ve literatürde düzgün olmayan gözlemler olarak adlandırılan gözlemlerle karşılaşılabilir. Aykırı gözlem, etkili gözlem ve kaldıraç noktaları olarak sınıflandırılan bu gözlemlerin tespit edilmesi doğru istatistiksel çıkarsamaların yapılabilmesi bakımından önemlidir. Nurunnabi ve arkadaşları (2016) düzgün olmayan gözlemlerin belirlenmesinde etkili uzaklık (*EU*) ölçüsünü önermiştir. Bu çalışmada, en çok olabilirlik, en küçük medyan kareler, yeniden ağırlıklandırılmış en küçük kareler, *M* ve *S* tahmin edicilerine dayalı *EU*'ın kullanılması ile düzgün olmayan gözlemlerin tespit edilmesi amaçlanmıştır. Gerçekleştirilen simülasyon çalışması ile bu yöntemlerin performansları karşılaştırılmıştır.

Bilim Kodu : 20513
Anahtar Kelimeler : Aykırı gözlem, etkili gözlem, kaldıraç nokta, etkili uzaklık, en çok olabilirlik, en küçük medyan kareler, yeniden ağırlıklandırılmış en küçük kareler, M tahmin edicisi, S tahmin edicisi
Sayfa Adedi : 59
Danışman : Doç. Dr. Meltem EKİZ

PERFORMANCE ANALYSIS OF THE INFLUENCE DISTANCE BASED ON ROBUST ESTIMATORS FOR THE IDENTIFICATION OF OUTLIERS IN LINEAR REGRESSION MODELS

(M. Sc. Thesis)

Fulya KARAKOCA

GAZİ UNIVERSITY

GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

January 2020

ABSTRACT

In multiple linear regression analysis, it could be possible to encounter with observations that differ from the bulk of the data and are named as unusual observations in the literature. It is important to identify these observations, which are classified as outliers, influential and leverage points, in order to make accurate statistical inferences. Nurunnabi et al. (2016) suggested influence distance (*ID*) used to create suspicious observations. In this study, it is intended to determine unusual observations by using *ID* based on maximum likelihood, least median of squares, re-weighted least squares, *M* and *S* estimators. A comparison of performance was performed with a simulation study.

Science Code : 20513
Key Words : Outlier, influential observation, leverage point, influence distance, maximum likelihood estimation, least median of squares, re-weighted Least Square, M estimator, S estimator.
Page Number : 59
Supervisor : Assoc. Prof. Dr. Meltem EKİZ

TEŞEKKÜR

Çalışmalarım boyunca bilgi birikimi ve desteğini esirgemeyen hocam Doç. Dr. Meltem EKİZ ve Dr. Öğr. Üyesi Ufuk EKİZ'e, destekleri ile her zaman yanımda olan aileme teşekkürü bir borç bilirim.

İÇİNDEKİLER

Sayfa

Sayfa

ÇİZELGELERİN LİSTESİ

Çizelge **Sayfa**

SİMGELER VE KISALTMALAR

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

Simgeler	**Açıklamalar**
n	Örnek hacmi
p	Bağımsız değişken sayısı
H	Hat matris
$\boldsymbol{h_{ii}}$	Hat matrisin köşegen elemanları
$\boldsymbol{x_i}$	Gözlem değeri
$\boldsymbol{\overline{x}}$	Örnek ortalaması
$\boldsymbol{\sigma^2}$	Yığın varyansı
$\boldsymbol{\sigma}$	Yığının standart sapması
$\boldsymbol{T}$	Tahmin edici
$\boldsymbol{\hat{\sigma}^2_{(i)}}$	*i*. gözlem değerinin veri kümesinden silinmesi ile hesaplanan varyansın tahmin edicisi

Kısaltmalar	**Açıklamalar**
EÇOB	En çok olabilirlik
EKK	En küçük kareler
EKKK	En küçük kırpılmış kareler
EKMD	En küçük mutlak değerler
EKMK	En küçük medyan kareler
EU	Etkili uzaklık
GM	Genelleştirilmiş *M* tahmin edicisi
HKO	Hata kare ortalaması

1. GİRİŞ

Regresyon analizi değişkenler arasındaki ilişkinin araştırılması ve modellenmesi için kullanılan istatistiksel bir yöntemdir (Montgomery, Peck ve Vining, 2012). Ancak genellikle bu yönteme ilişkin teorik gereklilikler ve uygulamalar arasında önemli düzeyde farklılıklar bulunmaktadır. Regresyon analizinde model parametrelerinin tahmini için en sık kullanılan yöntem en küçük kareler (*EKK*) yöntemidir. Fakat *EKK* yönteminin veri kümesi içerisinde yer alan bazı gözlemlerden oldukça etkilendiği ve regresyon modelinin değişkenler arasındaki ilişkiden ziyade bu gözlemlerin etkisini yansıttığı bilinmektedir (Cook ve Weisberg, 1982: 1).

Bu tür gözlemler veri kümesinde yer alan diğer gözlemler içerisinde farklılık gösteren ve veri kümesinin genel davranışına uyum sağlamayan birimler olarak tanımlanan aykırı gözlemlerdir. Bazı durumlarda araştırmacılar gözlemlerin grev, savaş yılları vb. çeşitli nedenlerle yanlış olduğunu bilirler. Ancak bazen tespit edilemeyen ve öngörülemeyen bazı veriler artık kareler toplamının büyümesine neden olur (Belsley, Kuh ve Welsch, 1980: 12). Artık kareler toplamı her bir gözlemin ortalamadan farkının karelerinin toplamıdır ve *EKK* yönteminin temel amacı artık kareler toplamını minimize etmektir (Wallach, Jones, Makowski ve Brun, 2019: 164). Sağlanmayan varsayımlar ve veri kümesinde yer alan aykırı değerler model parametrelerine ilişkin tahmin değerlerini ve bu tahmin değerleri kullanılarak gerçekleştirilen analiz sonuçlarını dramatik bir şekilde değiştirebilmektedir (Chatterjee ve Hadi, 1986). Bu nedenle regresyon analizinde veri kümesi ile uyumlu olan düzgün gözlemlerin yanında yer alan; aykırı gözlemler, kaldıraç noktaları ve etkili gözlemlerin ve bu gözlemlerin regresyon modeline olan etkisinin tespit edilmesi oldukça önemlidir.

Etkili gözlemleri Belsley ve arkadaşları (1980), tahmin değerleri üzerinde diğer gözlemlere göre çok daha fazla etkiye sahip olan gözlemler olarak tanımlamışlar ve etkili gözlemlerin tespit edilmesinin sağlıklı bir analiz süreci için gerekliliğini vurgulamışlardır (Imon, 2005). Etkili gözlemlerin tespit edilmesi çalışması ilk olarak 1977 senesinde R. Dennis Cook tarafından gerçekleştirilmiştir. 1966 yılında ilk kez Draper ve Smith tarafından tanıtılan Student türü artıklar ve Student türü artıklardan daha kullanışlı olan standartlaştırılmış artıklar aykırı değerlerin ve etkili gözlemlerin tespit edilmesinde kullanılmaktadır. Hoaglin ve Welsh (1978) aykırı değer ve etkili gözlemlerin belirlenmesinde bir gösterge olan

kaldıraç noktalarının tespit edilmesi için bu tez çalışmasının dördüncü bölümünde detaylı olarak anlatılacak olan hat matrisin köşegen elemanlarının kullanılabileceğini ifade etmişlerdir. Veri kümesinin merkezinden *x* ya da *y* yönünde ya da her iki yönünde uzak olan gözlemler ise kaldıraç noktası olarak tanımlanırlar. Ancak, Welsch (1980) etkili gözlemlerin belirlenmesinde ne kaldıraç noktalarının ne de Student türü artıkların tek başına yeterli olmadığını savunmuş ve bu yöntemlerin yerine hem kaldıraç noktalarını hem de artık bileşenlerini kullanan *DFBETAS* ve *DFFITS* yöntemini önermiştir. Bahsedilen tüm bu yöntemler veri kümesi içerisinde tek bir etkili gözlem yer aldığında başarılıdır ve genellikle bu gözlemin yerinin ve anlamlılığının tespit edilmesini sağlamaktadır (Hawkins, 1980: 51). Ancak, veride birden fazla sayıda etkili gözlemin yer alması durumunda bazı istenmeyen durumlar ortaya çıkmaktadır. Bunlardan bir tanesi analizin başında etkili olmadığı halde, etkili bir gözlemin veriden atılmasından sonra etkili bir gözleme dönüşen veri bulunma durumu olarak tanımlanan maskeleme etkisidir (Rousseeuw ve Leroy, 1987: 81). Maskeleme etkisinin tam tersi olarak tanımlanan bir diğeri ise etkili olmayan bir gözlemin etkili gözlem olarak tanımlandığı süpürme etkisidir (Lawrance, 1995).

Birçok tanısal teknik; aykırı değer, kaldıraç noktası ve etkili gözlemlerin ayrı ayrı tespit edilmeleri üzerine kurulmuştur. Bu prensiple tek bir gözlemin silinmesi tekniğine dayalı yöntemler maskeleme ve süpürme etkisinden olumsuz olarak etkilenirler (Atkinson, 1986). Çoklu düzgün olmayan gözlemlerin aynı anda tespit edilmeleri gerekli olsa da, veri kümesinde yer alan böyle bir gözlem bir diğerinin tespitini zorlaştıracağından mümkün olamamaktadır (Pena ve Yohai, 1995). Nurunnabi, Nasser ve Imon (2016), etkili uzaklık (*EU*) olarak isimlendirilen ve grup silme tekniğine dayanan yeni bir yöntem önermişlerdir.

Veri kümesi içerisinde yer alan aykırı değer ve etkili gözlemlerin neden olduğu diğer bir handikap sağlanmayan varsayımlardır. Klasik istatistiksel teknikler uyulması gereken varsayımların sağlanması koşuluyla en iyi sonuçları verirler. Uygulama tecrübeleri ve bazı ileri çalışmalar bu varsayımların sağlanmadığı durumlarda ise bu tekniklerin doğru sonuçlar vermediğini ileri sürmüşlerdir (Hoaglin, Mosteller ve Tukey, 1983). En yaygın olarak kullanılan varsayım ise gözlenen verilerin normal dağılımdan geldiği varsayımıdır. Ancak aykırı ve etkili gözlemlerin varlığı veri kümesinin normal dağılımdan uzaklaşmasına neden olur. Bu durumda dayanıklı tahmin yöntemleri önerilmektedir (Maronna, Martin ve Yohai, 2006).

Dayanıklı yöntemler ilk olarak 19.yy sonlarında Simon Newcomb tarafından araştırılmıştır. İlk önemli adımlar 1960'lı yıllarda atılmış ve 1970'li yılların başında; John Tukey (1960, 1962), Peter Huber (1964, 1967) ve Frank Hampel (1971, 1974) temel çalışmaları gerçekleştirmiştir. Daha sonra Peter Huber (1981), Frank Hampel, Elvezio Ronchetti, Peter J. Rousseeuw ve Annick M. Leroy (1986), Robert G. Staudte ve Simon J. Sheather (1990) tarafından önemli kitaplar yazılmıştır.

Yohai ve Zamar (1988) dayanıklı regresyonun amaçlarını; eş zamanlı olarak yaklaşık % 50 değerini alan yüksek bir kırılma noktası elde etmek, sınırlı etki fonksiyonuna ve yüksek etkinliğe sahip olmak olarak ifade etmişlerdir. Etki fonksiyonu büyük bir örnek içerisine eklenen tek bir aykırı gözlemin etkisini ölçmek için kullanılmaktadır (Welsch, 1982). *EKK* yöntemi tahmin edicileri diğer tahmin edicilerle kıyaslandığında varyans değeri daha küçüktür yani yüksek etkinliğe sahiptir. Ancak kırılma noktaları sıfırdır. Kırıma noktalarının sıfır olması sadece bir aykırı değerin bile tüm analizi etkileyebileceği anlamına gelmektedir. Bu aşırı hassasiyet *EKK* yönteminin hatanın normal dağıldığı yönündeki varsayımından kaynaklanmaktadır.

Dayanıklı regresyon yöntemleri, farklı özellikteki aykırı değerlere karşı koruma sağlamakta farklı yeteneklere sahiptir. En eski dayanıklı yöntemlerden biri en küçük mutlak değerler (*EKMD*) dayanıklı yöntemidir. Bu yöntem mutlak artık değerlerin toplamını minimize etmektedir. *EKMD* yöntemi, *EKK* yöntemine göre daha yüksek bir etkinlik sağlamasına rağmen, düşük bir kırılma noktasına sahiptir (1/*n*). Bu durum *EKMD* yönteminin diğer dayanıklı yöntemlere göre daha az tercih edilmesine neden olmaktadır (Armstrong ve Kung, 1978). Bu sınırlılık *x* yönlü aykırı değerlerin dikkate alınmamasından kaynaklanmaktadır (Bagheri, Midi, Ganjali ve Eftekhari, 2010).

Huber (1973), *M* tipi tahmin edici adında farklı bir dayanıklı yöntem türü önermiştir. Bu yöntem yüksek etkinlik sağlamasına rağmen, kırılma noktası (1/*n*)'dir ve bu değer oldukça düşüktür. *M* tipi tahmin ediciler *y* yönlü aykırı değerlerin üstesinden gelme konusunda oldukça başarılı olmakla birlikte, *x* yönlü aykırı değerlerden fazlasıyla etkilenmektedir. *M* tipi tahmin edicilerin kaldıraç noktası varlığındaki bu zafiyeti nedeniyle Hill (1977) tarafından genelleştirilmiş *M* tahmin edicisi (*GM*) tanımlanmıştır. *GM* tahmin edicisi, bazı ağırlık fonksiyonlarını kullanarak *x* yönlü aykırı değerlerin etkisini sınırlamaktadır. Simpson, bu yöntemin yüksek etkinlik ve sınırlı etkiye sahip olduğunu ifade etmektedir.

Ancak kırılma noktası yine $1/p$'den yüksek olmamaktadır. Bunun sonucu olarak aykırı değer sayısı, veri kümesi içerisinde küçük bir oran ise bu yöntem aykırı değerlere karşı direnç göstermektedir. Ancak, değişken sayısının artması kırılma noktası değerini azaltmaktadır.

Rousseeuw (1984), daha önce önerilen dayanıklı yöntemlerin olumsuz özelliklerini aşmak amacıyla, yüksek kırılma noktasına sahip en küçük medyan kareler (*EKMK*) yöntemini önermiştir. Bu yöntem artık karelerinin medyanlarını minimize etmektedir. Ancak oldukça yavaş bir yakınsama hızına sahiptir. Ayrıca kümelenmiş aykırı değer içeren örnekler olduğunda kırılma noktası düşüktür.

Rousseeuw (1985), kırılma noktası mümkün olan en yüksek değer olan % 50 değerini elde etmek amacıyla en küçük kırpılmış kareler (*EKKK*) yöntemini önermiştir. Bununla birlikte yüksek bir kırılma noktası en uygun çözümün bulunduğu anlamına gelmemektedir. Etki fonksiyonu ve etkinlik robust yöntemler araştırılırken kullanılan diğer önemli ölçütlerdir. *EKKK* ve *EKMK* yöntemleri sınırsız etki fonksiyonu üretmekte ve sırasıyla % 8 ve % 37 gibi düşük bir göreli etkinliğe ulaşmaktadırlar.

Rousseeuw ve Yohai (1984) yüksek etkinlik sağlayan *S* tipi tahmin edicisini tanıtmışlardır. *S* tipi tahmin edicisi, *EKMK* ve *EKKK* tahmin edicilerinin genelleştirilmiş bir halidir ve değişken sayısına bağlı olmaksızın % 50' ye yakın bir kırılma noktası değerine sahiptir (Tyler, 1999: 659-662; Donoho, Johnstone, Rousseeuw ve Stahel, 1985: 496, 500).

Bu tez çalışmasında, düzgün olmayan gözlemlerin tespit edilmesinde sağlam ve sağlam olmayan tahmin edicilere dayalı *EU* kullanılmıştır. Daha sonra kullanılan bu tahmin edicilerin performansları karşılaştırılmıştır. Çalışmanın ikinci bölümünde temel kavramlardan regresyon analizine, bazı tahmin yöntemlerine ve tahmin edicilerin özelliklerine yer verilmiştir. Üçüncü bölümde veri kümesinde yer alabilecek gözlem türlerinden, dördüncü bölümde ise bahsedilen bu gözlemlerin belirlenmesine yönelik bazı yöntemlerden bahsedilmiştir. Beşinci bölümde *EU* yöntemi tanıtılmıştır. Altıncı bölümde en çok olabilirlik, en küçük medyan kareler, yeniden ağırlıklandırılmış en küçük kareler, *M* ve *S* tahmin edicilerine dayalı *EU*'lar kullanılarak bir simülasyon çalışması gerçekleştirilmiştir. Gerçekleştirilen simülasyon çalışması ile aykırı ve etkili gözlemlerin 10 000 tekrara dayalı olarak doğru belirleme oranları (*DBO*) tespit edilmiştir.

2. TEMEL KAVRAMLAR

2.1. Regresyon Analizi

Regresyon kelimesi ilk olarak Francis Galton tarafından, kalıtım üzerine yapmış olduğu çalışmalarla ilgili olarak kullanılmıştır. Galton, yapmış olduğu bir çalışmasında, uzun boylu ana babaların uzun, kısa boylu ana babaların kısa çocukları olur eğiliminin geçerliliğine karşın, belli bir boydaki ana babaların çocuklarının ortalama boyunun genel nüfustaki ortalama boya doğru yaklaşma eğiliminde olduğunu bulmuştur (Galton, 1886: 42, 72).

Regresyon analizi, bağımlı değişken ve bağımsız değişkenler arasındaki yaklaşık ilişkinin tahmin edilmesini amaçlamaktadır (Dhhan, Rana, Midi, 2016). Bu analizde ilk olarak bağımlı ve bağımsız değişkenlere ilişkin gözlem değerleri incelenmekte ve bu ilişkinin matematiksel biçimini ortaya koyan bir regresyon eğrisi elde edilmektedir. İkinci aşamada ise bazı varsayımlara dayanarak parametre tahmini ve model parametrelerine ilişkin hipotez testleri yapılmaktadır.

Y_i, *i.* bağımlı rassal değişken, x_i, gözlemlenebilen *i.* bağımsız değişken (kestirim değişkeni), ε_i rassal hata terimi, *n* örnekteki gözlem sayısı, *i=1,2,….,n* olmak üzere basit doğrusal regresyon modeli,

$$Y_i = \theta_1 + \theta_2 x_i + \varepsilon_i \qquad i = 1,2,\dots,n \tag{2.1}$$

ile ifade edilir. Modelde θ_1; regresyon doğrusunun *y* eksenini kestiği nokta, θ_2 regresyon doğrusunun eğimidir. Hata teriminin, ortalaması sıfır ve varyansı σ_ε^2 olan normal dağılıma uyduğu ve rassal değişkenler arasındaki ilişkinin gücünü gösteren bir ölçüt olan kovaryans değerinin sıfıra eşit olduğu varsayılmaktadır. Normal dağılım varsayımı yaygın olarak kullanılmaktadır. Ancak olmazsa olmaz değildir. Diğer dağılımlar da kullanılabilmektedir (Casella ve Berger, 2001: 578).

Eş. 2.1 ile verilen regresyon modelindeki θ_1 ve θ_2 parametrelerine ilişkin tahmin değerleri örnek verileri aracılığıyla hesaplanır. Hesaplanan bu tahmin değerleri ile elde edilen

doğrusal regresyon modeli; $\hat{\theta}_1$ tahmin edilen sabit değeri, $\hat{\theta}_2$ tahmin edilen regresyon doğrusunun eğim değerini ve e_i ise artık terimi göstermek üzere;

$$Y_i = \hat{\theta}_1 + \hat{\theta}_2 x_i + e_i \qquad i=1,2,\ldots,n \tag{2.2}$$

şeklinde gösterilmektedir (Casella ve Berger, 2001: 552).

Regresyon modeli konusunda sıkça kullanılan "tahmin" ile "tahmin edici" kavramları birbirinden farklı kavramlardır. Tahmin ile tahmin edici kavramı arasındaki farkı açıklamak için basit bir ayrım yapmak gerekmektedir. Tahmin gerçek bir örnek ele alınarak elde edilen bir değerken, tahmin edici ise rassal örneğin bir fonksiyonu olarak tanımlanmaktadır (Casella ve Berger, 2001: 312).

Aşağıdaki bölümlerde ilk olarak tahmin edicilerin bulunmasında kullanılan bazı yöntemlerden, daha sonra bu yöntemler kullanılarak elde edilen tahmin edicilerin değerlendirilmesini sağlayan bazı temel kriterlerden bahsedilecektir.

2.2. Doğrusal Regresyon Analizinde En Küçük Kareler Yöntemi

Örnek üzerinden regresyon doğrusunun tahminini elde etmek için çeşitli yöntemler vardır, ancak bunlardan en yaygın olarak kullanılanı *EKK* yöntemidir. *EKK* yönteminin Alman matematikçi Carl Friedrich Gauss tarafından ortaya atıldığı kabul edilir (Gujarati, 2010: 52).

Regresyon analizi ile bir araştırma yapılırken, daha önce ifade edilen bağımlı ve p tane bağımsız değişken arasındaki matematiksel ifadenin biçimini ortaya koyan regresyon eğrisinde yer alan parametrelerin tahmin edilmesi gerekmektedir. $\theta_1, \theta_2, \ldots, \theta_p$ parametreler, n rassal örnekteki gözlem sayısı, $i=1,2,\ldots,n$ olmak üzere $x_{i1}, x_{i2}, \ldots, x_{ip}$ doğrusal bağımsız değişkenler, x_{i1} birler vektörü, Y_i bağımlı değişken ve ε_i hata terimi olmak üzere çoklu doğrusal regresyon modeli;

$$Y_i = x_{i1}\theta_1 + x_{i2}\theta_2 + \cdots + x_{ip}\theta_p + \varepsilon_i \qquad i = 1,2,\ldots,n \tag{2.3}$$

$$j = 1,2,\ldots,p$$

şeklinde ifade edilir (Rousseuw ve Leroy, 1987: 1).

$\theta_1, \theta_2, \dots, \theta_p$ parametrelerinin tahmin değerleri örnek üzerinden, $\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_p$ olarak tanımlandığında $i = 1,2, \dots, n$ ve $j = 1,2, \dots, p$ iken $\hat{Y}_i$ tahmin değeri;

$$\hat{Y}_i = \hat{\theta}_1 x_{i1} + \hat{\theta}_2 x_{i2} + \cdots + \hat{\theta}_p x_{ip} \quad (2.4)$$

ile gösterilir. $i = 1,2, \dots, n$ iken Y_i,

$$Y_i = \hat{Y}_i + e_i \quad (2.5)$$

ile gösterilir. Eş. 2.5'de kullanılan e_i ;

$$e_i = Y_i - \hat{Y}_i \quad (2.6)$$

olarak ifade edildiğinde, *i.* gözlem için artık değer olarak yorumlanır (Rousseuw ve Leroy, 1987: 2).

EKK yönteminin kullanılabilmesi için iki önemli koşulun sağlanması gerekmektedir. Bu koşullar aşağıda gösterilmektedir.

$$\sum_{i=1}^{n} e_i = 0 \quad (2.7)$$

$$\sum_{i=1}^{n} e_i^2 = min \quad (2.8)$$

Bu yöntem ile regresyon doğrusundaki parametre tahmin edicilerinin elde edilmesi için Eş. 2.4 ile gösterilen modelde tek bir bağımsız değişken olması durumunda;

$$Y_i = \hat{\theta}_1 + \hat{\theta}_2 x_i + e_i \quad (2.9)$$

ile gösterilsin. Eş. 2.9' da verilen basit doğrusal regresyon modelindeki $\hat{\theta}_1$ ve $\hat{\theta}_2$ tahmin edicilerinin elde edilmesinde artıkların karelerinin toplamının minimum yapılması;

$$\min_{\hat{\theta}} \sum_{i=1}^{n} e_i^2 \tag{2.10}$$

ile ifade edilir. Bu minimizasyon sonucunda θ_1 ve θ_2 parametrelerinin tahmin edicileri, $\hat{\theta}_1$ ve $\hat{\theta}_2$ bulunur. Parametre tahminlerinin bulunmasında Eş. 2.8 ifadesinin θ_1 ve θ_2'ye göre kısmi türevleri alınır ve sıfıra eşitlenirse;

$$\frac{\partial \sum_{i=1}^{n} e_i^2}{\partial \theta_1} = -2 \sum_{i=1}^{n} (Y_i - \theta_1 - \theta_2 x_i) = 0 \tag{2.11}$$

$$\frac{\partial \sum_{i=1}^{n} e_i^2}{\partial \theta_2} = -2 \sum_{i=1}^{n} (Y_i - \theta_1 - \theta_2 x_i) x_i = 0 \tag{2.12}$$

Eş. 2.11 ve Eş. 2.12 normal denklemler olarak adlandırılır ve aşağıdaki şekilde de ifade edilebilirler.

$$\sum_{i=1}^{n} Y_i = n\theta_1 + \theta_2 \sum_{i=1}^{n} x_i \tag{2.13}$$

$$\sum_{i=1}^{n} Y_i x_i = \theta_1 \sum_{i=1}^{n} x_i + \theta_2 \sum_{i=1}^{n} x_i^2 \tag{2.14}$$

Normal denklemlerin eşanlı olarak çözülmesi ile $\hat{\theta}_1$ ve $\hat{\theta}_2$ ile ifade edilen *EKK* tahmin edicileri;

$$\hat{\theta}_1 = \frac{\sum_{i=1}^{n} Y_i - \hat{\theta}_2 \sum_{i=1}^{n} x_i}{n} \tag{2.15}$$

$$\hat{\theta}_2 = \frac{n \sum_{i=1}^{n} Y_i x_i - \sum_{i=1}^{n} Y_i \sum_{i=1}^{n} x_i}{n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \tag{2.16}$$

şeklinde elde edilir. Klasik doğrusal regresyon modeline ilişkin varsayımlar sağlanıyorken, *EKK* tahmin edicileri ideal ya da en uygun özellikleri taşımaktadır (Dobson, 2002: 31,32).

Klasik doğrusal regresyon modelinde *EKK* yönteminin uygulanabilirliği için gerekli varsayımlar;

- Doğrusal regresyon modeli θ_i katsayılarına göre doğrusaldır,

- x_i değerleri yinelenen örneklemelerde değişmez,
- $E\ (e_i/x_i) = 0$'dır,
- $V(e_i/x_i) = \sigma^2$'dir,
- Artık terimler arasında ardışık bağımlılık yoktur (otokorelasyon),
- Artık terimi ile *x* değişkeni ilişkisizdir,
- Gözlem sayısı *n* tahmin edilecek parametre tahminlerinin sayısından fazla olmalıdır ($n > p$),
- Belirli bir örneklemdeki x_i değerlerinin hepsi aynı olmamalıdır,
- Regresyon modeli doğru kurulmalıdır,
- Bağımsız değişkenler arasında tam doğrusal ilişki (collinearity) yoktur,

şeklindedir (Gujarati, 2010: 59).

Rousseuw ve Leroy (1987) çalışmalarında *EKK* tahmin edicilerinin tercih edilme sebeplerinden ve bazı olumsuzluklarından bahsetmiştir. "Yöntemin bu kadar yaygın olarak kullanılmasının nedeni kolay anlaşılabilir ve uygulanabilir olmasındandır. Yöntemin bulunduğu zamanlarda (1800'lü yıllarda) bilgisayarların olmaması ve *EKK* tahmin edicilerinin veriler üzerinde bazı matris cebri uygulamaları ile kolaylıkla hesaplanabilirliği, sadece *EKK* tahmin edicisini uygulanabilir bir yöntem yapmıştır. Günümüzde bile, birçok istatistik paket programı gelenekselleşmiş bir yöntem olduğundan ve hesaplama kolaylığından dolayı sadece *EKK* yöntemi kullanmaktadır. Gauss daha sonra hata terimlerinin dağılımları normal dağılıma uyduğunda *EKK*'nin en iyi olduğunu göstermiştir. Bu mükemmel matematik teorisinin akabinde, normal dağılım varsayımları ve *EKK* yöntemi istatistiksel uygulamalar için standart bir uygulama haline gelmiştir. Ancak, son zamanlarda bu istatistiksel varsayımların gerçek veriler üzerinde çoğu durumda sağlanmadığı anlaşılmıştır.

2.3. En Çok Olabilirlik Yöntemi

Parametre tahmin etmede kullanılan bir diğer önemli yöntem en çok olabilirlik (*EÇOB*) yöntemidir. Bu yöntem 1922 yılında istatistikçi ve genetikçi Ronald A. Fisher tarafından ortaya atılmıştır.

EÇOB yöntemi tahmin edicilerin türetilmesinde en sık kullanılan yöntemdir. Bir yığından seçilmiş n çaplı rassal örnek $X_1, X_2, \ldots\ldots, X_n$ ve değişken sayısı p olmak üzere, bu örneğe ait olasılık yoğunluk fonksiyonu;

$$f(x_1, x_2, \ldots, x_n \,/\theta_1, \theta_2, \ldots\ldots, \theta_p) \tag{2.17}$$

şeklinde ifade edilir. Olabilirlik fonksiyonu ise;

$$L(\theta \backslash X) = L\big(\theta_1, \theta_2, \ldots\ldots, \theta_p / x_1, x_2, \ldots\ldots, x_n\big) = \prod_{i=1}^{n} f(x_i/\theta_1, \theta_2, \ldots\ldots, \theta_p) \tag{2.18}$$

şeklinde ifade edilir.

Eğer olabilirlik fonksiyonu, θ_i' ye göre türevlenebilir bir fonksiyon ise EÇOB tahmin edicileri aşağıda verilen eşitliğin çözülmesi ile belirlenir (Casella ve Berger, 2001: 315, 316).

$$\frac{\partial}{\partial\, \theta_i}\, L(\theta/X) = 0 \qquad\qquad i = 1, 2, \ldots, p. \tag{2.19}$$

Bu bölümde tahmin edicilerin bulunmasına yönelik sıklıkla tercih edilen yöntemlerden bahsedilmiştir. Kullanılan bu yöntemlerin dışında Bayes yöntemi, momentler yöntemi beklenti-maksimizasyonu algoritması gibi farklı tahmin yöntemleri bulunmaktadır. Bir sonraki bölümde farklı tahmin yöntemleri ile de elde edilebilen tahmin edicilerin kıyaslanmasında kullanılabilen bazı kriterlerden bahsedilecektir.

2.4. Tahmin Edicilerin Özellikleri

Bir önceki bölümde parametrelere ilişkin tahmin edicilerin bulunması için kullanılan iki yönteme değinilmiştir. Ancak, çalışmalarda çok sayıda tahmin yöntemi ile farklı tahmin ediciler elde edilebilmektedir. Böyle durumlarda elde edilen tahmin ediciler içerisinden bir tanesinin seçilmesi gerekmektedir (Casella ve Berger, 2001: 330). Bu bölümde tahmin edicilerin değerlendirilmesinde kullanılabilen bazı temel kriterlerden bahsedilecektir.

2.4.1. Yansızlık

Yansızlık, istatistik ile yığın parametresi arasındaki ilişkiyi gösterir. İyi bir tahmin edicide aranan özelliklerden biri yansızlıktır. θ parametresinin bir tahmin edicisi $\hat{\theta}$ olsun, $\hat{\theta}$'nın beklenen değeri parametre değerine eşit ise bu tahmin ediciye yansız tahmin edici adı verilir.

$$E(\hat{\theta}) = \theta \tag{2.20}$$

Eş. 2.20'den hareketle bir tahmin edicinin yanlılık değeri;

$$Yan\,(\hat{\theta}) = E(\hat{\theta}) - \theta \tag{2.21}$$

ifadesi ile hesaplanabilir (Casella ve Berger, 2001: 214-334).

2.4.2. Tutarlılık

$\hat{\theta}$, θ parametresine ilişkin tutarlı bir tahmin edici ve Ω parametre uzayı olmak üzere, her bir $\varepsilon > 0$ ve $\theta \in \Omega$ için,

$$\lim_{n\to\infty} P_\theta(|\hat{\theta} - \theta| < \varepsilon) = 1 \tag{2.22}$$

eşitliği sağlanır (Casella ve Berger, 2001: 468).

2.4.3. Etkinlik

Yansız tahmin ediciler içerisinde en küçük varyansa sahip olan tahmin edici, etkin tahmin edici olarak tanımlanmaktadır. θ parametresinin yansız iki tahmin edicisi $\hat{\theta}$ ve $\tilde{\theta}$ olmak üzere $var\,(\hat{\theta}) < var\,(\tilde{\theta})$ ise $\hat{\theta}$, $\tilde{\theta}$'dan daha etkin bir tahmin edicidir denir.

$\hat{\theta}$ tahmin edicisinin, $\tilde{\theta}$ tahmin edicisine göre göreli etkinliği ise;

$$\frac{var(\hat{\theta})}{var(\tilde{\theta})} \tag{2.23}$$

şeklinde ifade edilir (Casella ve Berger, 2001: 472).

2.4.4. Yeterlilik

Yeterlilik kavramı R.A. Fisher tarafından geliştirilen bir ölçüttür. Bir tahmin edicinin yeterli olabilmesi için, tahmin edicinin tahmin edilecek parametre hakkında örnekte mevcut bulunan bütün bilgiyi kullanması gerekmektedir (Yamane, 1973: 245).

$X_1, X_2, \dots, X_n;\ \theta \in \Omega$ iken $f_1(x_n:\theta)$ olasılık yoğunluk fonksiyonuna sahip bir dağılımdan çekilen n çaplı rassal bir örneği göstersin. $Y_1 = u_1(X_1, X_2, \dots, X_n)$ olasılık yoğunluk fonksiyonu $f_2(y_1:\theta)$ şeklinde gösterilen bir istatistik iken Y_1'in yeterli bir istatistik olabilmesi için;

$$\frac{f_1(x_1:\theta) f_1(x_2:\theta) \dots f_1(x_n:\theta)}{f_2[u_1(X_1, X_2, \dots, X_n);\theta]} = H(x_1, x_2, \dots x_n) \qquad (2.24)$$

olmak üzere $H(x_1, x_2, \dots x_n)$'ın $\theta \in \Omega$'dan bağımsız olması gerekmektedir (Hogg ve Craig, 2004: 316, 317).

θ parametresini tahmin etmek için kullanılan yeterli istatistik, örnek uzayını alt kümelere ayırırken başka hiçbir yeterli istatistik bu sınıflamadan daha ayrıntılı bir sınıflama yapamaz ise istatistik minimal yeterli istatistik adını alır. Bu alt kümelerin her birindeki örneklerden herhangi biriyle karşılaşma olasılığı θ'dan bağımsızdır (Casella ve Berger, 2001: 279, 280).

2.4.5. Hata kare ortalaması

Hata kare ortalaması (*HKO*), $\hat{\theta}$ tahmin değerinin parametre değerinden ortalamada ne kadar uzakta olduğu ile ilgilenmektedir. $\hat{\theta}$'nın *HKO*'sı;

$$HKO = E_\theta \left(\hat{\theta} - \theta\right)^2 \qquad (2.25)$$

şeklinde ifade edilebilir. Bir tahmin edicinin değerlendirilmesinde genellikle $\left|\hat{\theta} - \theta\right|$ uzaklığının artan herhangi bir fonksiyonu iyilik ölçütü olarak yorumlanmaktadır. Ancak, diğer karşılaştırma ölçütleri ile kıyaslandığında *HKO* iki önemli avantaja sahiptir.

Bunlardan ilki analitik olarak çözülebilir olması, ikincisi ise yorumlanabilir olmasıdır (Casella ve Berger, 2001: 330).

Uygulamalarda iki adet yansız tahmin edici etkinlikleri bakımından karşılaştırılırken tahmin edicilerin varyanslarına bakmak yeterlidir. Ancak, biri yanlı biri yansız iki tahmin edici ya da yanlılık miktarları aynı olmayan iki yanlı tahmin edici karşılaştırılırken *HKO* kriteri kullanılabilir. Yanlı ve yansız iki tahmin ediciden düşük *HKO*'sına sahip yanlı tahmin edici θ parametresini tahmin için kullanılabilir. $\hat{\theta}$'nın yansız bir tahmin edici olduğu durumda *HKO* tahmin edicinin varyansına eşittir. *HKO*'nın varyansını tahmin edicinin yanlı ya da yansız olduğu durumlarda;

$$Var_{\theta}(\hat{\theta}) = \begin{cases} E_{\theta}\left(\hat{\theta} - \theta\right)^2, & \hat{\theta} \text{ yansız iken} \\ E_{\theta}\left(\hat{\theta} - \theta\right)^2 - \left(E_{\theta}\hat{\theta} - \theta\right)^2, & \hat{\theta} \text{ yanlı iken} \end{cases} \tag{2.26}$$

ile gösterilebilir (Casella ve Berger, 2001: 330).

Bu bölümde regresyon analizi, bazı tahmin yöntemleri ve tahmin edicilerin özelliklerinden bahsedilmiştir. Bir istatistiksel analizde gözlem değerlerinin etkisi de oldukça önemlidir. Chatterjiee ve Hadi (1986) çalışmalarında bu durumu "İstatistiksel bazı nicelikler bir ya da birkaç gözlem tarafından önemli derecede etkilenebilmektedir. Ancak, tüm gözlemler *EKK* doğrusu üzerinde eşit etkiye sahip değildir. Bazı gözlem tipleri analiz üzerinde daha fazla etkilidir. Sonuç olarak bir araştırmacı için böyle etkiye sahip gözlemlerin tanımlanması ve bu gözlemlerin analize olan etkilerinin değerlendirilmesi çok önemlidir." şeklinde ifade etmişlerdir.

Bir sonraki bölümde regresyon analizi için büyük önem taşıyan farklı gözlem türlerinden bahsedilecektir.

3. REGRESYON ANALİZİNDE AYKIRI (X VE Y YÖNÜNDE) VE ETKİLİ GÖZLEMLER

Dhhan, Rana ve Habshah (2016) çalışmalarında aykırı değerleri, diğer gözlemlerden oldukça uzak olan ve bu nedenle farklı bir mekanizma tarafından üretildiği şüphesini doğuran gözlemler olarak tanımlamışlardır. Veri kümesi içerisinde yer alan aykırı değerler regresyon eğrisini önemli düzeyde etkileyebilmekte ve model parametrelerinin tahminleri gerçek değerlerinden oldukça farklı olabilmektedir.

Aykırı değerler veri kümesinde yer alan gözlemlerin çoğunluğunun yer aldığı merkezden, *x* ya da *y* eksenine göre uzak olabilmektedir. Eğer aykırı gözlemler *y* ekseni yönünde bir sapma gösteriyorsa bu aykırı değerler *y* yönlü aykırı değerler olarak tanımlanmaktadır. *y* yönlü aykırı değerler genellikle büyük değer almaktadır. Ancak, *EKK* yöntemi ile elde edilen regresyon doğrusu üzerinde önemli düzeyde bir değişikliğe neden olmayabilirler. Aynı şekilde aykırı gözlemler *x* ekseni yönünde bir sapma gösteriyorsa bu aykırı değerler *x* yönlü aykırı değerler ya da iyi kaldıraç noktası olarak tanımlanmaktadır. Genellikle iyi kaldıraç noktalarının regresyon doğrusu üzerinde önemli düzeyde bir değişime neden olmadığı bilinmektedir. Hatta bazı durumlarda iyi kaldıraç noktaları analize katkı sağladığından yararlıdır (Rousseuw ve Leroy, 1987:6).

Aykırı değer veri kümesinde yer alan diğer gözlemlerden, hem *x* hem de *y* yönünde sapma gösteriyorsa bu aykırı değerlere kötü kaldıraç noktası adı verilmektedir. Kötü kaldıraç noktaları regresyon doğrusunun eğimini değiştirmekte ve doğruyu bir kaldıraç gibi kendine yaklaştırmaktadır. Eğimi değişen regresyon doğrusu bu noktaya çekildiğinden, bu noktaya ilişkin artık değeri de küçük olmaktadır.

Etkili bir gözlem tek olarak ya da diğer gözlemler ile birlikte değerlendirildiğinde, çeşitli tahmin değerleri üzerinde, diğer gözlemlerle kıyaslandığında, çok daha belirgin bir biçimde büyük değerde etkiye sahip olan gözlemdir (Chatterjee, Hadi, 1986).

Etkili gözlemleri belirlemeye yönelik yöntemlerde amaç parametre tahmini yapmak değildir. Amaç gözlemin tahminler üzerinde ne kadar büyük değişiklikler meydana getirdiğini tespit etmektir. Dolayısı ile aykırı gözlemin ya da gözlemlerin etkili olma ya da

olmama durumları söz konusu olabilir. Ayrıca gözlemlerin birbiri ile olan ilişkileri etkililiği etkiler.

Özetle veri setleri dört çeşit gözlem türü içerebilmektedir: veri kümesinde yer alan gözlemlerle ve regresyon eğrisi ile uyumlu olan düzgün gözlemler, *y* yönlü aykırı değerler, iyi kaldıraç noktaları, hem *x* hem de *y* yönlü sapma gösteren aykırı değerler (kötü kaldıraç noktaları).

Sonuç olarak aykırı ve etkili değerler parametre tahminlerini etkileyebildiğinden, *EKK* yöntemi ile elde edilen regresyon doğrusu tahmini üzerinde ciddi bir tehdit oluştururlar. Bu problemin çözülmesi için kullanılan iki temel yöntem bulunmaktadır. Bunlardan ilki ve muhtemelen en çok bilinenleri, tanısal olarak adlandırılan yaklaşımlardır. Tanılar, etkili noktaları tespit etmek amacıyla veriden hesaplanan niceliklerdir. Bu yaklaşım ile tespit edilen aykırı değerler veri kümesinden çıkarılır ya da hatalı kaydedildi ise düzeltilir, daha sonra kalan verilere *EKK* yöntemi uygulanır. Veri kümesinde sadece tek bir aykırı değer yer aldığında silmenin etkisiyle başarılı bir performans yakalanmaktadır. Ancak veri kümesinde birden fazla aykırı değer yer aldığında bunların belirlenmesi oldukça zor olacaktır ve bu durumda regresyon tanıları yöntemi kapsamlı hesaplamalar gerektirecektir. Ayrıca aykırı gözlemlerin düzgün gözlem olarak düşünülmesi durumu olan maskeleme etkisi ve düzgün gözlemlerin aykırı gözlem olarak düşünülmesi durumu olan süpürme etkisi gibi bazı problemlerin ortaya çıkması etkili gözlemlerin doğru olarak tespit edilmesini zorlaştıracaktır (Serfling ve Wang, 2014). Diğer bir yaklaşım ise bu çalışmanın ilerleyen bölümlerinde detaylı olarak anlatılacak olan sağlam regresyon yöntemleridir (Rousseeuw ve Leroy, 1987: 8).

4. AYKIRI (X VE Y YÖNÜNDE) VE ETKİLİ GÖZLEMLERİN BELİRLENMESİNE YÖNELİK YÖNTEMLER

Tek bağımsız değişkenli bir analizde yani iki boyutlu durumlarda çizilen serpilme diyagramları aracılığıyla aykırı ve etkili gözlemlerin tespit edilmesi kolaylıkla gerçekleşmektedir. Ancak, uygulamalarda çoğunlukla ikiden fazla değişken yer almakta ve yüksek boyutlu analizler söz konusu olmaktadır. Böyle yüksek boyutlu analizlerde görsel diyagramlar aracılığıyla gözlem türlerinin belirlenmesi pek mümkün değildir (Rousseuw ve Leroy, 1987: 3).

Bir önceki bölümde veri kümesi içerisinde yer alan aykırı ve etkili değerlerin regresyon modelindeki parametre tahminlerini önemli ölçüde etkilediği ifade edilmiştir. Bu nedenle doğru tahmin değerleri elde etmek için aykırı ve etkili gözlemlerin doğru olarak tespit edilmesi büyük önem taşımaktadır.

Regresyon analizinde veride yer alan aykırı değerler (*x* ya da *y* yönlü) *EKK* yöntemi ile model parametrelerinin tahmininde ciddi bir tehdit oluşturmaktadır. Bu problemin çözümü için temel olarak kullanılan iki yöntem bulunmaktadır. İlki ve muhtemelen en iyi bilineni tanısal yöntemlerdir. İkincisi ise sağlam istatistiksel yöntemlerdir (Rousseuw ve Leroy, 1987: 8). Bu yöntemlerden literatürde sık kullanılan bazı yaklaşımların özeti bu bölümde ele alınmıştır. Aşağıda sırasıyla bu iki yönteme ilişkin bazı ölçütlere yer verilmiştir.

4.1. Tanısal Yaklaşımlar

4.1.1. Hat matris

Eş 2.3 ile verilen çoklu doğrusal regresyon modeli,

$$Y = X\theta + \varepsilon \tag{4.1}$$

şeklinde tanımlansın. *X*; ilk sütunu birler vektöründen oluşan *nxp* boyutlu veri matrisini, *Y*; *nx1* boyutlu bağımlı değişken vektörünü, θ bilinmeyen parametreler vektörünü ve ε ise *nx1* boyutlu hata vektörünü ifade etsin. Hat matris ise;

$$H = X(X^T X)^{-1} X^T \tag{4.2}$$

olarak tanımlanmaktadır (Maronna, Martin ve Yohai, 2006: 95). Eş. 4.2 içerisinde yer alan "*T*" transpozu ifade etmektedir ve $X^T X$ tersi alınabilir bir matristir.

H matrisi idempotent *(HH=H)* ve simetrik $(H^T = H)$ bir matristir. *H* matrisinin rankı, köşegen elemanlarının toplamına karşılık gelir. Yani, *rank(H)=iz(H)*'dır. Ayrıca *rank(H)=p*'dir.

Hat matrisin köşegen elemanları h_{ii}'ler *i*. gözlem biriminin parametre tahmini üzerindeki etkisini ölçmektedir. Hat matrisin köşegen elemanları $0 \leq h_{ii} \leq 1$ aralığında değer alır. Genellikle köşegen elemanlarının değeri 1'e yaklaştıkça gözlemlerin veri merkezine olan uzaklığının arttığı düşünüldüğünden söz konusu gözlemler kaldıraç noktası olarak isimlendirilir (Rousseuw ve Leroy, 1987: 220).

4.1.2. Mahalanobis uzaklığı

Mahalanobis uzaklığı sabit terimli bir regresyon modelinde bir gözlemin kaldıraç noktası olup olmadığının belirlenmesinde kullanılmaktadır. X_i vektörleri gözlem değerlerinden oluşmak üzere, $\bar{X}$ ve *pxp* boyutlu $\hat{\Sigma}_n$ kovaryans matrisi;

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{4.3}$$

ve

$$\hat{\Sigma}_n = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^T(X_i - \bar{X}) \tag{4.4}$$

ile hesaplanır (Gath ve Hayes, 2011).

Bu durumda $\forall$ X_i değişkeninin, ortalamasından $(\bar{X})$ ne kadar uzak olduğu kovaryans matrisi $\hat{\Sigma}_n$'nin yardımı ile aşağıdaki formül kullanılarak hesaplanmaktadır.

$$MD_i^2 = (x_i - \bar{x})\hat{\Sigma}_n^{-1}(x_i - \bar{x})^T \qquad i=1,2,\ldots,n \tag{4.5}$$

MD_i^2'ler; p serbestlik dereceli ki kare dağılımına uyar. Uygulamalarda genellikle, α birinci tip hatayı göstermek üzere, $MD_i^2 > \chi^2_{p,1-\alpha}$ ise *i*. gözlemin kaldıraç noktası olduğu düşünülür. Hat matrisi ve Mahalanobis uzaklığı veri kümesinde sadece tek bir kaldıraç noktası yer aldığında kullanışlıdır. (Rousseeuuw ve Leroy, 1987: 224).

4.1.3. Standartlaştırılmış artıklar

Aykırı değerlerin tespit edilmesi için artık değerler yeterli değildir. Bu durumda, standarlaştırılmış artık değerlerin kullanılması ile aykırı değerler tespit edilebilmektedir. Standartlaştırılmış artıkların hesaplanması *HKO*'sının karekökünün kullanılmasına dayalı olarak,

$$e_i^* = \frac{e_i}{\sqrt{HKO}} \qquad i = 1,2,\dots,n \tag{4.6}$$

ile hesaplanır. Büyük değerli standart artık değerleri potansiyel aykırı değer olarak nitelendirilirler (Chatterjee ve Hadi, 1988).

4.1.4. Student türü artıklar

Artıkların varyansı σ^2 parametresinin ve h_{ii} istatistiğinin bir fonksiyonudur ve artıklar ölçek istatistiğine dayalı bir dağılıma sahiptir. Çoğu tanısal yöntem için bu iki niceliğe bağlı olmayan Student türü artıklar kullanılabilir. Student türü artıklar dahili Student türü artıklar ve harici Student türü artıklar olmak üzere ikiye ayrılırlar. *EKK* yönteminde dahili Student türü artıklar; $\hat{\sigma} = \sqrt{\sum_{i=1}^{n} \frac{e_i^2}{n-p}}$ olmak üzere;

$$r_i = \frac{e_i}{\hat{\sigma}\sqrt{(1-h_{ii})}} \qquad i = 1,2\dots,n \tag{4.7}$$

ile elde edilir (Cook ve Weisberg, 1982: 14-18).

Hataların normal dağıldığı varsayımı altında, *i*. gözlem değerinin veri kümesinden silinmesi ile hesaplanan varyansın tahmin edicisi olan $\hat{\sigma}^2_{(i)}$ istatistiği;

$$\hat{\sigma}^2_{(i)} = \frac{(n-p)\hat{\sigma}^2 - e_i^2/(1-h_{ii})}{n-p-1} \qquad i = 1,2,\dots,n \tag{4.8}$$

ya da,

$$\hat{\sigma}^2_{(i)} = \hat{\sigma}^2 \left(\frac{n-p-r_i^2}{n-p-1}\right) \qquad i = 1,2,\dots,n \tag{4.9}$$

iken harici Student türü artık değeri;

$$t_i = \frac{e_i}{\hat{\sigma}^2_{(i)}\sqrt{(1-h_{ii})}} \qquad i = 1,2,\dots,n \tag{4.10}$$

ile hesaplanır (Cook ve Weisberg, 1982: 20).

4.1.5. DFBETAS ölçütü

Besley, Kuh ve Welsch (1980) tarafından önerilen *DFBETAS* ölçütü, *i*. gözlem veriden silindiğinde, *j*. regresyon katsayısına ilişkin tahmin değerinin ne kadar değiştiğini göstermektedir. *i*. gözlem veriden silindikten sonra θ parametre vektörünün *EKK* tahmin edicisi $\hat{\theta}_{(i)}$ vektörü ile gösterilsin. Ayrıca *X*, *nxp* boyutlu bağımsız değişken matrisini x_i, *X* matrisinin *i*. sütununu e_i, artık vektörünü ve h_{ii} hat matrisinin köşegen elemanları olsun. *i*. gözlem silinmeden önceki veri ile elde edilen θ parametre vektörünün *EKK* tahmin edicisi $\hat{\theta}$ olmak üzere, $\hat{\theta}$ ile $\hat{\theta}_{(i)}$ arasındaki fark,

$$DFBETA_i = \hat{\theta} - \hat{\theta}_{(i)} = \frac{(X^TX)^{-1}x_ie_i}{1-h_{ii}} \tag{4.11}$$

şeklinde tanımlanmaktadır.

$DFBETA_i$ sadece vektörel bir sonuç verdiğinden etkili gözlemlerin saptanmasında skaler değerli bir sonuç veren $DFBETAS_{ij}$ istatistiği kullanılabilir. Bu ölçüt; $\hat{\theta}$ vektörünün *j*. elemanı $\hat{\theta}_j$, $\hat{\theta}_{(i)}$ vektörünün *j*. elemanı $\hat{\theta}_{j(i)}$ ve $s_{(i)} = \sqrt{\frac{1}{n-p-1}\sum_{k\neq i}(y_k - x_k\,\hat{\theta}_{(i)})^2}$ olmak üzere;

$$DFBETAS_{ij} = \frac{\hat{\theta}_j - \hat{\theta}_{j(i)}}{s(i)\sqrt{(X^TX)_{jj}^{-1}}} \tag{4.12}$$

ile hesaplanır. Bu ölçüt her bir gözlemin j. regresyon katsayısı üzerindeki etkisini gösterir. $\left|DFBETAS_{ij}\right| > 2/\sqrt{n}$ olan gözlem değerleri etkili gözlem (influental observation) olarak kabul edilir (Belsley, Kuh ve Welsch, 1980: 13,75).

4.1.6. DFFITS ölçütü

Etkili gözlemleri tespit etmek için kullanılan bir diğer istatistik Belsley ve diğerleri (1980) tarafından önerilmiş olan *DFFITS* ölçütüdür. Bu ölçüt gözlem silme tekniğine dayanmaktadır.

DFFITS ölçütü tüm birimler üzerinden elde edilen parametre tahmin değeri ve *i.* gözlemin veri kümesinden silinmesi ile elde edilen yeni tahmin değeri arasındaki farka dayanarak söz konusu gözlemin etkili bir gözlem olup olmadığını belirlemek için kullanılır (Belsley ve diğerleri 1980: 15).

$$DFFITS_i = \frac{e_i}{s(i)} \frac{(h_{ii})^{1/2}}{1-h_{ii}} \tag{4.13}$$

şeklinde tanımlanan bu ölçüt *i.* gözlem değerinin $\hat{y}_i$ tahmin değeri üzerindeki etkisini ölçmektedir. Bu ölçüt ile, $|DFFITS_i| > 2\sqrt{\frac{p}{n}}$ şartını sağlayan gözlemler etkili gözlem olarak değerlendirilmektedir.

Yukarıda bahsedilen yöntemler ile kaldıraç noktaları ya da etkili gözlemlerin tespit edilmesi mümkündür. Ancak, bir araştırma yapılırken çalışılan veri kümesinin dağılımı varsayılan dağılımdan sapmalar gösteriyorsa tahmin ediciler istenilen nitelikte olamamaktadır. Böyle durumlarda bu yöntemlere alternatif olarak kullanılan sağlam yöntemler mevcuttur. Aşağıdaki bölümlerde sırasıyla bu yöntemlerin kullanım tekniklerinden ve sağlam yöntemlerin karşılaştırılmasında kullanılan ölçütlerden bahsedilecektir.

4.2. Sağlam Yöntemler

İlk olarak 1963 yılında Box "sağlam" sözcüğünü istatistiksel anlam için kullanmıştır. Bilim adamları varsayımlara bağlı olmayan, özellikle normal dağılım varsayımına karşı hassasiyet göstermeyen yaklaşımları "sağlam" olarak tanımlamıştır (Stigler, 1973).

İstatistiksel çıkarımlar yalnızca gözlemler aracılığıyla yapılabilmektedir. En az gözlemler kadar önemli bir diğer husus önsel bazı varsayımların sağlanması gerekliliğidir. Örneğin dağılımlara ilişkin; rassallık, bağımsızlık ya da bilinmeyen parametrelere ilişkin bazı önsel varsayımların sağlanıyor olması gerekmektedir. Ancak, bu varsayımlar her zaman sağlanmaz. Zamanla, yaygın olarak kullanılan yöntemlerin varsayımlardan küçük sapmalara dahi aşırı hassasiyet gösterdiği gözlenmiş ve bu yöntemlerin önemli bir alternatifi olarak "sağlam" yöntemler önerilmiştir (Huber, 1981: 1).

Bir istatistiksel yöntem, istatistiksel modele ilişkin varsayımlar sağlanmadığında dahi iyi bir performans gösteriyorsa "sağlam" olarak kabul edilmektedir. Verinin doğrusal regresyon modeline uyduğu varsayıldığında, *EKK* tahminleri ve test performansı oldukça iyi olacaktır, fakat hata rasgele değişkeninin dağılımı varsayılandan biraz sapma gösterdiğinde "sağlam" olmayacaktır (Birkes ve Dodge, 1993: 85).

Sağlam istatistik yöntemler veri kümesi içerisinde yer alan aykırı değerlerin tespit edilmesi amacıyla da yaygın olarak kullanılmaktadır. Ryan (2009) yapmış olduğu çalışmasında neredeyse tüm veri setlerinin aykırı değer içerdiğini ve bu durumun belirli uygulama alanlarında problemlere neden olduğunu, bu nedenle sağlam tahmin yöntemlerine ihtiyaç duyulduğunu belirtmiştir.

Sağlam istatistiklerin kullanılmasındaki temel amaçlar aşağıdaki şekilde özetlenebilir;

i) Veri kümesine en iyi uyumu gösteren yapıyı tespit etmek,
ii) Gerekli görülürse daha ileri analizler gerçekleştirmek için aykırı değerleri ya da yapının temelinden sapmaları tespit etmek,
iii) Önemli etkili gözlemleri tespit etmek ve bu noktalara dikkat çekmek,
iv) Ardışık bağımlılık veya korelasyondan sapmaları tespit etmek (Hampel ve diğerleri, 1986: 84).

Yukarıda tanımı yapılan ve kullanılma nedenlerinden bahsedilen sağlam tahmin edicilerin bazı özellikleri sağlıyor olması gerekmektedir. Hampel (1974) çalışmasında bu özellikleri aşağıdaki şekilde özetlemiştir:

Sağlam tahmin ediciler düzgün gözlemlerin içerisine düzgün olmayan ancak, veri kümesinin genelinden de büyük ölçüde farklı olmayan gözlemlerin dahil olması sonucunda oluşan kirlenmelere karşı küçük tepkiler göstermelidir.

i) Büyük kirlenmeler söz konusu olsa da güvenilir olmalıdır. Yani kırılma noktaları yüksek olmalıdır.
ii) Sabit miktardaki kirlenmenin göreli etkinliği sınırlı olmalıdır. Yani kirletici bir gözlemin tahmin edici üzerinde asimptotik olarak oluşturacağı en büyük etkinin küçük olması istenir.
iii) Yuvarlama ve gruplama gibi işlemler karşısında düzgün tepki göstermelidir.
iv) Açıkça görünen aykırı değerleri veri kümesinden ayıklayabilmelidir.
v) Alabileceği en küçük varyans değerine sahip olmalıdır.

Aşağıdaki bölümlerde sağlam yöntemlere ilişkin sıklıkla kullanılan bazı kriterler ve bu kriterlerin gerekliliğinden bahsedilecektir.

4.2.1. Kırılma noktası ve etki fonksiyonu kavramı

Tahmin edicilerin sağlam olup olmadıklarının tespit edilmesinde aşağıdaki ölçütler kullanılabilmektedir.

Kırılma noktası

Kırılma noktası kavramı ilk olarak 1967 senesinde Hodges tarafından tek boyutlu konum tahmin edicileri ile sınırlı olarak kullanılmıştır. Daha sonra 1968 ve 1971 yıllarında, Hampel tarafından yapılmış olan çalışmalar ile genelleştirilmiştir (Zuo, 2001). En yaygın kullanıldığı şekli ise 1983 yılında Donoho ve Huber tarafından aşağıdaki şekilde tanımlanan versiyonudur.

Z, n tane gözlem biriminden oluşan bir örneklem olarak tanımlansın;

$Z=\{(X_{11}, \ldots X_{1p}, Y_1) \ldots (X_{n1}, \ldots, X_{np}, Y_n)\}.$ (4.14)

T bir regresyon tahmin edicisi ise, Z örneklemi için T tahmin edicisi kullanılarak regresyon katsayılarına ilişkin $T(Z) = \hat{\theta}$ vektörü hesaplanabilir.

Orijinal gözlem değerlerinden m tanesinin yerine keyfi değerler yazılarak bozulmuş bir Z' örneklemi tanımlansın. Böyle bir kirlenme ile ortaya çıkan maksimum sapma, $Yan\ (m:T,Z)$ ile gösterildiğinde bu sapma;

$$Yan\ (m:T,Z) = \sup_{Z'} \|T(Z') - T(Z)\| \quad (4.15)$$

eşitliği ile elde edilir. Burada *sup* üst sınırların en küçüğünü göstermektedir. Eğer $Yan(m:T,Z)$ sonsuz ise, m tane aykırı değerin T üzerinde büyük bir etkisinin olduğu, ve bu nedenle de tahmin edicinin iyi olmadığı söylenir. Bu durumda Z sonlu örneğine dayalı T tahmin edicisinin kırılma noktası,

$$\varepsilon_n^* = min\left\{\frac{m}{n}, Yan(m:T,Z) sonsuz\right\} \quad (4.16)$$

şeklinde tanımlanır (Rousseeuw ve Leroy, 1987: 9,10).

Bir başka ifadeyle kırılma noktası, tahmin edicilerin başa çıkabilecekleri aykırı değerlerin oranına ilişkin bir sınırdır (Huber, 1981:13).

Etki fonksiyonu

Büyük çaplı bir örneğe x değerine sahip bir gözlemin eklendiği varsayılsın. Bu gözlemin tahmin edici ve test istatistiği üzerindeki sınırlı etkisini değerlendirmede etki fonksiyonu kullanılabilir. Bu fonksiyon 1968 ve 1974 yıllarında Hampel tarafından etki eğrisi ya da etki fonksiyonu ismiyle tanıtılmıştır ve belki de bir istatistiğin sağlamlığını değerlendirmek için keşfedilmiş en kullanışlı araçlardandır (Huber, 1981: 13, 14).

Etki fonksiyonu, aykırı bir değerin neden olduğu sapmayı formülleştirmek amacıyla kullanılır. Büyük örneklemler içerisinde göz ardı edilemeyecek miktardaki kırılmaların tahmin edici üzerindeki etkisini ölçer (Hampel, Ronchetti, Rouseeuw ve Stahel, 1986: 84).

Asıl dağılımı F olan yığına ε oranında ve Δ_x dağılımından gelen düzgün olmayan verilerden oluşan bir yığın dahil edildiğinde, $F_{x,\varepsilon} = (1-\varepsilon)F + \varepsilon\Delta_x$ şeklinde karma bir dağılım elde edilir. Bu durumda dağılımı $F_{x,\varepsilon}$ olan bir yığından çekilen her bir birimin; Δ_x dağılımından çekilmesi olasılığı ε (küçük bir değer), F dağılımından çekilmesi olasılığı ise *1-*ε*'*dur. Δ_x dağılımından çekilen örnek "kötü", *F'*den çekilen örnek "iyi" örnek olarak adlandırılır. Etki fonksiyonu bir tahmin edicinin düzgün olmayan gözlemler karşısındaki yanıtı olduğundan;

$$EF(x) = \lim_{\varepsilon \to 0} \frac{T\left(F_{x,\varepsilon}\right) - T(F)}{\varepsilon}, \tag{4.17}$$

ya da daha açık bir ifadeyle;

$$EF(x,T,F) = \lim_{\varepsilon \to 0} \frac{T[(1-\varepsilon)F + \varepsilon\Delta_x] - T(F)}{\varepsilon} \tag{4.18}$$

şeklinde gösterilir (Hampel ve diğerleri1986: 84). Burada *T* tutarlı bir tahmin ediciyi ifade eder.

Etki fonksiyonunun iki temel kullanımı bulunmaktadır. Biricisi, tahmin edicinin ve test istatistiğinin değeri üzerinde tek bir gözlemin göreli etkisini elde etmektir. İkincisi ise, asimptotik varyansın tahmin edilmesini sağlayarak tahmin ediciye ilişkin asimptotik özellikleri sezgisel olarak hızla ve kolaylıkla değerlendirmek için imkan sunmaktır (Huber, 1981: 15). Asimptotik varyans eşitliği;

$$var\ (T,F) = \int EF\ (x,T,F)^2 dF(x) \tag{4.19}$$

ile ifade edilir (Hampel ve diğerleri, 1986: 85).

Aşağıdaki bölümlerde bazı sağlam tahmin ediciler ve özelliklerinden bahsedilecektir.

4.2.2. M tipi tahmin ediciler

Konum tahmini

Huber (1964), Eş 2.10 ile verilen *EKK* amaç fonksiyonundaki kareli ifadeyi değiştirme yolunu seçmiştir. *EKK* amaç fonksiyonunun düzgün olmayan gözlemlere karşı olan duyarlılığını azaltmak için uygun bir ρ fonksiyonu kullanmıştır. 1964'de Huber'in çalışması ile *M* tipi tahmin edicileri, sonuç odaklı yaklaşımı ve asimptotik özellikleri nedeniyle giderek daha da önemli hale gelmiştir (Boos ve Stefanski, 2001: 339). Huber'e göre;

$$min \sum_{i=1}^{n} \rho(x_i, T_n) \tag{4.20}$$

ya da $\psi(x_i, \theta) = \left(\frac{d}{d\theta}\right) \rho(x: \theta)$ olmak üzere,

$$\sum_{i=1}^{n} \psi(x_i; T_n) = 0 \tag{4.21}$$

eşitliğini sağlayan T_n, *M* tipi tahmin edici olarak adlandırılır (Huber, 1981: 43).

M tipi tahmin edicinin hesaplanmasında Huber tarafından kullanılması önerilen ρ fonksiyonu *c* sabiti 5,2 olmak üzere;

$$\rho(x) = \begin{cases} x^2/2\,, & |x| \leqq c \\ c|x| - c^2/2\,, & x > c \end{cases} \tag{4.22}$$

ile ifade edilmektedir. ρ fonksiyonunun türevi ise,

$$\psi = \begin{cases} [x]_{-c}^{c} = c, & x < -c \\ x, & -c \leqq x \leqq c \\ c, & x > c \end{cases} \tag{4.23}$$

şeklinde gösterilir. Konum parametresinin tahmini ise;

$$min \sum_{i=1}^{n} \rho(x_i - T_n) \tag{4.24}$$

ya da,

$$\sum_{i=1}^{n} \psi(x_i - T_n) = 0 \tag{4.25}$$

ın iteratif çözümüne dayanır (Huber, 1981: 43). Eş. 4.25 i. gözlemin ağırlığı $w_i = \frac{\psi(x_i - T_n)}{(x_i - T_n)}$ şeklinde tanımlandığında $\sum w_i(x_i - T_n) = 0$ olarak da ifade edilebilir. Bu ağırlıklar,

$$w_i = \begin{cases} 1, & |x_i - T_n| \leqq c \\ \frac{c}{|x_i - T_n|}, & |x_i - T_n| > c \end{cases} \tag{4.26}$$

ile tanımlanır (Huber, 1981: 44). *M* tipi tahmin edicilere ilişkin olarak önerilen birçok farklı fonksiyon olmakla birlikte uygulama adımları tüm yaklaşımlar için aynıdır. Bu çalışmada Tukey'in bi-weight ağırlığı kullanılacaktır. % 95 değerine sahip bir asimptotik etkinlik elde etmek için *k* değerinin 4,685 olarak kullanılması gerekmektedir (Chang, Roberts ve Welsh, 2017). Tukey'in bi-weigt ağırlığı için kullanılan ρ fonksiyonu;

$$\rho(x) = \begin{cases} 1 - \left[1 - \left(\frac{x}{k}\right)^2\right]^3, & |x| \leq k \\ 1, & |x| > k \end{cases} \tag{4.27}$$

ρ fonksiyonunun türevi,

$$\psi(x) = \begin{cases} x\left[1 - \left(\frac{x}{k}\right)^2\right]^2, & |x| \leq k \\ 0, & dd \end{cases} \tag{4.28}$$

Gözlemler için kullanılan ağırlıklar,

$$w(x) = \begin{cases} \left[1 - \left(\frac{x}{k}\right)^2\right]^2, & |x| \leq k \\ 0, & dd \end{cases} \tag{4.29}$$

ifadeleri ile gösterilmektedir (Maronna, 2006: 29,30).

Sağlam tahmin edicilerin bir çoğunda, tahminlere ilişkin nümerik değerler ancak iteratif süreçlerin uygulanması sonucu elde edilebilmektedir. Ayrıca, sağlam tahmin ediciler için,

varyans ya da ortalama değerleri hesaplanırken kullanılana benzer matematiksel bir model bulunmamaktadır. Sağlam tahmin ediciler üzerinden parametre tahminlerini elde etmek için bir fonksiyonun minimize edilmesine dayalı bir algoritma ile oluşturulan modellere gereksinim duyulmaktadır. Genel olarak, algoritmalarda öncelikle T_0 gibi bir başlangıç değeri seçilir daha sonra T^* gibi yeni bir tahmine ulaşmak amacıyla hesaplamalar yapılır. Sonraki adımda elde edilen son tahmin değeri başlangıç değeri olarak kullanılır. Bu süreç, iterasyonlar sonucu elde edilen tahminler arasındaki farkın yeterince küçük olması sağlanana kadar devam eder (Hoaglin, Mosteller ve Tukey, 1983: 291).

Eş 4.21'in türevi alınarak türetilen $T(F)$ kullanılarak M tipi tahmin edicilere ilişkin etki fonksiyonu;

$$EF(x, F, T) = \frac{\psi[x, T(F)]}{-\int\left(\frac{\delta}{\delta\theta}\right)\psi'[x, T(F)]F(\delta x)} \tag{4.30}$$

şeklinde tanımlanır (Huber, 1981: 45). M tipi tahmin ediciler için kırılma noktası $\eta = \min\{-\frac{\psi(-\infty)}{\psi(+\infty)}, -\frac{\psi(+\infty)}{\psi(-\infty)}\}$ olmak üzere,

$$\varepsilon^* = \frac{\eta}{1+\eta} \tag{4.31}$$

şeklindedir (Huber, 1981: 14-16).

Sağlam regresyon yöntemi için *M* tahmin edicileri

M tipi regresyon tahmin edicisi;

$$\min_{\hat{\theta}} \sum_{i=1}^{n} \rho(y_i - x_i\theta) = \min_{\hat{\theta}} \sum_{i=1}^{n} \rho(e_i) \tag{4.32}$$

ile elde edilir. ρ fonksiyonu; simetrik, sürekli, sıfır noktasında tek bir minimum noktasına sahip ve türevi alınabilir bir fonksiyondur (Huber, 1981:162). ρ fonksiyonu,

- $\rho(e) \geq 0$
- $\rho(0) = 0$
- $\rho(e) = \rho(-e)$

- $\rho(e_i) \geq \rho(e'_i), \qquad |e_i| > |e'_i|$

özelliklerini sağlamalıdır.

$w_i = w(e_i)$ olarak tanımlanmışken ve ağırlık fonksiyonu $w(e_i) = \frac{\psi(e)}{e}$ iken amaç fonksiyonu,

$$\sum w_i \left(y_i - x_i'\hat{\theta}\right)x_i' = 0 \tag{4.33}$$

şeklinde yazılır. Bu fonksiyona ilişkin çözüm ağırlıklandırılmış bir *EKK* tahminidir. Ağırlıklar artıklara, artıklar tahmin edilen katsayılara ve tahmin edilen katsayılar ağırlıklara bağlıdır. Bu nedenle iteratif çözüm gereklidir (Fox, 2002). *M* tipi tahmin edicilerin uygulanmasında kullanılan algoritma aşağıdaki şekildedir.

Adım 1- *EKK* yöntemi kullanılarak regresyon katsayılarının tahminleri elde edilir.

Adım 2- Regresyon modeline ilişkin varsayımlar test edilir.

Adım 3- Veri içerisinde yer alan aykırı değerler tespit edilir.

Adım 4- *EKK* yöntemi ile $\hat{\theta}^0$ hesaplanır.

Adım 5- $e_i = y_i - \hat{y}_i$ ile artık değerler hesaplanır.

Adım 6- $\hat{\sigma}_i$*=1.4826MAD* hesaplanır ($\hat{\sigma} = \frac{MAD}{0{,}6745} = \frac{Medyan|e_i - medyan(e_i)|}{0.6745}$). Burada *MAD* (mean absolute deviation) ortalama mutlak sapmayı ifade etmektedir.

Adım 7- $u_{i=}\, e_i/\hat{\sigma}_i$ değerleri hesaplanır.

Adım 8- w_i ağırlıkları,

$$w_i = \begin{cases} \left[1 - \left(\frac{u_i}{4{,}685}\right)^2\right], & |u_i| \leq 4{,}685; \\ 0 \qquad\qquad\quad , & |u_i| > 4{,}685. \end{cases}$$

ile hesaplanır.

Adım 9- w_i ağırlıkları ile ağırlıklı en küçük kareler yöntemi kullanılarak $\hat{\theta}^M$ hesaplanır.

Adım 10- $\hat{\theta}^M$'nın yakınsadığı değeri bulununcaya kadar 5-8. adımlar tekrarlanır.

Adım 11- Bağımsız değişkenlerin bağımlı değişken üzerinde anlamlı bir etkisinin olup olmadığı istatistiksel olarak test edilir (Susanti, Praitiwi, Sulistijowti ve Liana, 2014).

4.2.3. En küçük medyan kareler tahmin edicisi

En küçük medyan kareler tahmin edicisi (*EKMK*) 1984 yılında Rousseeuw tarafından ortaya atılmıştır. Bu tahmin edici,

$$\min_{\hat{\theta}} Medyan_i\, e_i^2 \tag{4.34}$$

şeklinde tanımlanır. Yani sabit terimli bir modelde basit regresyon için *EKMK* çözümü,

$$\min_{\hat{\theta}_1\hat{\theta}_2} Medyan_i\ (y_i - \hat{\theta}_1 x_i - \hat{\theta}_2\, x_i)^2 \tag{4.35}$$

şeklinde ifade edilir. Bu yöntem geometrik olarak gözlemlerin yarısını kapsayan en dar bandın bulunması anlamına gelir. Gözlemlerin yarısı ise *(n/*2*)*+1 sayıda gözlemi ifade etmektedir. Aslında bu tahmin edici en büyük artık değere sahip *n/*2 sayıda gözlem değerini hesaplamanın dışında bırakır. Bandın kalınlığı ise dikey yönde ölçülür. *EKMK* doğrusu ise bu bandın tam ortasından geçen doğrudur (Rousseeuw ve Leroy, 1987: 24).

EKMK tahmin edicisi hesaplanırken, *p* parametre sayısı kadar farklı gözlem içeren alt küme seçilerek işleme başlanır. Bu alt kümeler *i*=1,2,…,*n* olmak üzere *J* = $\{i_1, \ldots\ldots, i_p\}$ şeklinde gösterilir. Daha sonra *p* noktadan geçen düzlemin katsayıları olan θ_J elde edilir. Bu da *p* bilinmeyenli *p* tane denklemin çözülmesi anlamına gelir. Deneme tahminleri θ_J olarak ifade edilmişken, tüm veri kümesi ele alınarak, her bir θ_J'ye karşılık gelen;

$$Medyan_{i=1,\ldots,n}(Y_i - x_i\theta_J)^2 \tag{4.36}$$

amaç fonksiyonu elde edilir. Yukarıdaki ifadeyi minimum yapan regresyon doğrusundan elde edilen sonuç çözüm olarak alınır (Rousseeuw ve Leroy, 1987: 197, 198). Ancak bu süreçte asıl önemli olan altküme sayısının belirlenmesidir. En ideal olanı mümkün olan tüm altkümelerin kullanılmasıdır. Küçük veri setleri için tüm p gözlemli alt kümelerin seçilmesi mümkünken büyük veri setleri için bu işlem neredeyse imkansızdır. Böyle bir durumda ele alınacak m tane alt kümeden en az bir tanesinin "iyi" olma olasılığının bire yakın olmasını sağlayacak sayıda rassal seçim yapılmalıdır. Buradaki bir alt küme p tane düzgün gözlem içeriyorsa "iyi" olarak tanımlanmaktadır. Tüm veri kümesinde ε oranında aykırı değer olduğu varsayılırsa, m tane alt kümeden en az bir tanesinin "iyi" olma olasılığı;

$$P = 1 - [1 - (1 - \varepsilon)^p]^m \tag{4.37}$$

şeklinde tanımlanır. Buradaki P olasılığının 1'e yakın bir değer olması gerekliliği dikkate alınarak çekilecek alt örneklem sayısı,

$$m = \frac{log(1-P)}{log[1-(1-\varepsilon)^p]} \tag{4.38}$$

ile hesaplanır (Zhang, 1997).

EKMK yöntemi için ölçek tahmin edicisi de sağlam olarak tanımlanmıştır. Bu amaç doğrultusunda n ve p'ye dayalı ve bir düzeltme terimi ile çarpılmış başlangıç ölçek tahmini;

$$s^0 = 1.4826(1 + \frac{5}{n-p})\sqrt{\underset{i}{Medyan}\, e_i^2} \tag{4.39}$$

ile hesaplanır. Bu ölçek tahmini kullanılarak standartlaştırılmış ağırlıklar;

$$w_i = \begin{cases} 1, & \left|\frac{e_i}{s^0}\right| \leq 2{,}5 \\ 0, & \left|\frac{e_i}{s^0}\right| \geq 2{,}5 \end{cases} \tag{4.40}$$

ile elde edilir. Böylece standartlaştırılmış ağırlıklara ve artıklara dayalı *EKMK* ölçek tahmini;

$$\sigma^* = \sqrt{\frac{\sum_{i=1}^{n} w_i e_i^2}{\sum_{i=1}^{n} w_i - p}} \tag{4.41}$$

ifadesi ile hesaplanır (Rousseeuw ve Leroy, 1987: 44).

4.2.4. En küçük kırpılmış kareler tahmin edicisi

En küçük kırpılmış kareler tahmincisi (*EKKK*), asimptotik etkinlik bakımından *EKMK* tahmin edicisine göre daha yüksek etkinlik gösteren bir tahmin edici olarak 1984 yılında Rousseeuw tarafından önerilmiştir. Bu tahmin ediciye ilişkin kırılma noktası oldukça yüksektir. EKK yöntemine oldukça benzeyen bir minimizasyon yöntemi ile elde edilir. h değeri Giloni tarafından $\left[\frac{n}{2}\right] + 1$ olarak tanımlanmışken *EKKK* tahmin edicisi;

$$min \sum_{i=1}^{h} e_{(i)}^2 \tag{4.42}$$

olarak ifade edilir. Bu yöntem artık karelerinin h tanesinin toplamının küçükten büyüğe sıralanarak en küçüğünün belirlenmesi temeline dayanır. $h = \left[\frac{n}{2}\right] + [\frac{(p+1)}{2}]$ olduğunda *EKKK* tahmin edicisinin kırılma noktası alabileceği en yüksek değerini alır (Rousseeuw ve Leroy, 1987: 132).

EKKK yönteminin en büyük dezavantajı Eş. 4.42 ile verilen amaç fonksiyonunun matematiksel optimizasyona elverişli olmamasıdır. Ayrıca, uygulamada h değerinin belirlenmesi oldukça güçtür. Eğer varsayılandan fazla kirli veri ile karşı karşıya kalınırsa *EKKK* tahmin edicisi sağlamlık yönünden zafiyet gösterir. Bu durumun aksine, varsayılandan az kirlilik varsa, parametre tahminlerinin doğruluğu ve tüm gözlemlerin hesaplamaya dahil edilmemesinden kaynaklanan etkinlik zaafiyeti ortaya çıkar. h değerinin doğru olarak belirlenmesine ve veri yapısına bağlı olarak *EKKK* oldukça etkin olabilir. Aykırı değerlerin tümü kırpılırsa *EKKK* ile *EKK*'nın etkinliği eşit olur (Schumacker, Monathan ve Mount, 2002).

EKMK tahmin edicisi ile karşılaştırıldığında *EKKK* tahmin edicisinin amaç fonksiyonunun daha düzgün olduğu görülmektedir. *EKKK* tahmin edicisi, asimptotik normal dağılım durumunda *EKMK* tahmin edicisinden daha etkindir. *EKMK* tahmin edicisi ise veri kümesi normal dağılım varsayımını sağlanmadığından düşük bir yakınsama hızına sahiptir. Tüm

bu üstünlüklerine rağmen *EKKK* tahmin edicisinin hesaplanması *EKMK* tahmin edicisi ile kıyaslandığında oldukça zordur (Rousseeuw ve Driessen, 1999). *EKKK* tahmin edicisinin kırılma noktası, $h = \left[\frac{n}{2}\right] + 1$ iken;

$$\varepsilon_n^*(T, Z) = (\left[\frac{n-p}{2}\right] + 1)/n \tag{4.43}$$

şekilde tanımlanır (Rousseeuw ve Leroy, 1984: 132).

EKKK tahmin edicisinin hesaplanmasına ilişkin süreç bir dışlama yöntemi üzerine kurulmuştur. Yani veri kümesinde yer alan bir gözlem tahmin sürecinde ya tamamen hesaba dahil edilir, ya da tamamen sürecin dışında bırakılır. *EKKK* tahmin edicisinin hesaplanmasında, öncelikle veri kümesi içerisinden *h* gözlem değerine sahip alt örneklemlerin tümü çekilir. Bu işlem ile $\binom{n}{h}$ kombinasyonu kadar alt küme elde edilir. Daha sonra *h* gözlemli $\binom{n}{h}$ kombinasyonu kadar sayıda alt kümenin her biri için regresyon doğrusunun tahmini elde edilir. Elde edilen her tahmini regresyon doğrusu üzerinden hesaplanan *n* tane artıktan, kareleri toplamı en küçük olan *h* tanesinin toplamı bulunur. Toplamlar arasından minimumu seçilerek *EKKK* çözümü olarak kabul edilir. *EKKK* yönteminde karşılaşılacak en büyük problem, uygulamalarda çekilecek alt küme sayısının çok büyük boyutlara ulaşabilmesidir. Böyle durumlarda *EKKK* tahmin edicilerinin hesaplanmasında kesin algoritmalar yerine, *EKMK* tahmin edicilerinde olduğu şekilde yaklaşık çözümler veren bir algoritmanın kullanılması önerilmektedir (Rousseeuw ve Hubert; 1997).

4.2.4. S tipi tahmin ediciler

S tipi tahmin ediciler, *M* tipi tahmin edicilerinin kırılma noktasının düşük olması problemini çözmek üzere ilk olarak Hampel (1975) tarafından önerilmiş, daha sonra Rousseeuw ve Yohai (1984) tarafından geliştirilmiştir. *S* tipi tahmin edicileri *EKMK* ve *EKKK* tahmin edicilerinin genelleştirilmesi ile elde edilir (Donoho, Johnstone, Rousseeuw ve Stahel, 1985: 496, 500).

M tipi tahmin ediciler, veri kümesine ilişkin olarak yalnızca medyan değerini ağırlıklandırır ve bu nedenle veri kümesinin dağılımını ya da veri kümesi aracılığıyla elde

edilmiş olan bir fonksiyonu kullanmamaktadır. Bu durum yönteme ilişkin bir zaafiyet oluşturmaktadır. *S* tipi tahmin ediciler bu zaafiyeti çözmek amacıyla artıkların standart sapmalarını ve *M* tipi tahmin edicileri üzerinden hesaplanan artıkları kullanmaktadır. $n^{-\frac{1}{2}}$ yakınsama oranına ve yüksek kırılma noktasına sahip ve affine eş değişimli tahmin edicilerdir (Sibbertsen, 2001). Affine eş değişimli tahmin edici olmasından dolayı veri kümesine uygulanan herhangi bir dönüşüm tahmine de uygulanmış olur. Bu özellik *S* tipi tahmin edicilerin tercih edilmesine önemli bir katkı sağlamaktadır (Barrera ve Yohai, 2006).

S tipi tahmin edicileri artıkların dağılımlarının minimizasyonunu sağlamaya yönelik bir sürece dayanmaktadır. Bu tahmin edici $\hat{\theta}_n\,(x_1, x_2 \ldots., x_n)$; konum parametresinin *S* tipi tahmin edicisi olarak tanımlanmışken ve s; ölçek parametresinin *M* tipi tahmin edicisi olmak üzere,

$$\min_{\hat{\theta}_n} s(x_1 - \theta, \ldots, x_n - \theta\,), \tag{4.44}$$

amaç fonksiyonunun çözümüne dayalıdır. Bu fonksiyonun çözümü ile ölçek parametresinin tahmini,

$$\hat{\sigma}_n = s(x_1 - \hat{\theta}_n, \ldots, x_n - \hat{\theta}_n) \tag{4.45}$$

de eş zamanlı olarak elde edilir (Hampel ve diğerleri, 1986: 115). Daha sonra K,

$$\frac{1}{n}\sum_{i=1}^{n} \rho\left(\frac{e_i}{\hat{\sigma}_n}\right) = K \tag{4.46}$$

eşitliğinin çözümü elde edilir. *K* genellikle $E_\phi(\rho)$'ye eşit olarak belirlenir. Φ ise standart normal dağılımı ifade etmektedir. ρ fonksiyonun ise sınırlı olduğu varsayılmaktadır (Hampel ve diğerleri, 1986: 115). Rousseeuw ve Yohai, *c*=1,547 olmak üzere (1984) ρ fonksiyonu için;

$$\rho(e) = \begin{cases} \frac{e^2}{2} - \frac{e^4}{2c^2} + \frac{e^6}{6c^4}, & |e| \le c \\ \frac{c^2}{6}, & |e| \ge c \end{cases} \tag{4.47}$$

ifadesini önermişlerdir. Regresyon parametreleri tahmin edilirken, artık değerlerin değişkenliğinin minimize edilmesinden faydalanılır.

S tipi tahmin edicisinin kırılma noktası ise;

$$\varepsilon^* = \frac{K}{\rho(c)} \tag{4.48}$$

eşitliği ile elde edilir (Rousseeuw ve Leroy, 1987: 136).

Ne yazık ki, *S* tipi tahmin ediciler aynı anda hem yüksek etkinliğe hem de yüksek kırılma noktasına sahip olamazlar. Hössjer (1992), özellikle 0.5 kırılma noktasına sahip *S* tipi tahmin edicisinin hata terimleri normal dağıldığında sahip olduğu asimptotik etkinliğin 0.33'ten büyük olmadığını göstermiştir (Maronna ve diğerleri, 2006: 130,131). *S* tipi tahmin ediciler için kullanılan algoritma aşağıdaki şekildedir.

Adım 1- *EKK* yöntemi kullanılarak regresyon katsayıları elde edilir.

Adım 2- Regresyon modeline ilişkin varsayımların sağlanıp sağlanmadığı test edilir.

Adım3- Veri içerisinde aykırı değer varsa tespit edilir.

Adım 4- *EKK* yöntemi ile $\hat{\theta}^0$ hesaplanır.

Adım 5- e_i=y_i-$\hat{y}_i$ formülü ile artık değerleri hesaplanır.

Adım 6- i gözlem değerini göstermek üzere, $\hat{\sigma}_i$ değerleri heaplanır.

$$\hat{\sigma}_i = \begin{cases} \frac{\text{medyan}|e_i - \text{medyan}e_i|}{0{,}6745}, & \text{iterasyon} = 1 \\ \sqrt{\frac{1}{nK}\sum_{\text{İ}=1}^{n} w_i e_i^2}, & \text{iterasyon} > 1 \end{cases}$$

Adım 7- Standartlaştırılmış artık değerleri hesaplanmalıdır.

$$u_i = \frac{e_i}{\hat{\sigma}_i}$$

Adım 8- Ağırlıklandırılmış değerler Tukey'in bi-weight fonksiyonu ile hesaplanmalıdır. $\psi(e_i) = \rho'(e_i)$, ağırlıklar $w_i(e_i) = {\psi(e_i)}/{e_i}$ ve c sabit değeri 1,547 olmak üzere, bu fonksiyon;

$$\psi(e_i) = \begin{cases} e_i\left\{1 - \frac{e_i}{c}\right\}^2, & |e_i| \leq c \\ 0, & |e_i| > c \end{cases}$$

ifadesi ile ele edilir (M.Ekiz ve O. U.Ekiz, 2018; Rousseeuw ve Leroy, 1987:129).

Adım 9- w_i'ler kullanılarak ağırlıklı en küçük kareler (*AEKK*) yöntemi ile $\hat{\theta}_s$ hesaplanır.

Adım 10- $\hat{\theta}_s$ yakınsak bir değer alana kadar 5-8. adımlar tekrarlanmalıdır.

Adım 11- Bağımsız değişkenlerin, bağımlı değişken üzerinde anlamlı bir etkiye sahip olup olmadığı belirlenmelidir (Susanti ve diğerleri, 2014).

4.2.5. Yeniden ağırlıklandırılmış en küçük kareler yöntemi

Chatterjee ve Machler (1997) tarafından önerilen ağırlıklı en küçük kareler yöntemi maskelemenin varlığında etkili değildir. Billor ve Hadi (2006) maskeleme etkisinin varlığında da zafiyet göstermeyen Yeniden ağırlıklandırılmış en küçük kareler yöntemi (YAEKK) yöntemi önerilmiştir. *YAEKK* tahmin edicisinin hesaplanmasında kullanılan amaç fonksiyonu, l iterasyon sayısını ve w_i *i*'nci gözlem değerine ilişkin ağırlık değerini göstermek üzere;

$$min_{\hat{\theta}_l} \sum_{i=1}^{n} w_i e_{(i)}^2 \qquad i=1,2,\ldots,n \qquad (4.49)$$

ile ifade edilmektedir. *YAEKK* tahmin edicisine ilişkin varyans,

$$\hat{\sigma}^2_{YAEKK} = \frac{\sum_{i=1}^{n} w_i e_i^2}{\sum_{i=1}^{n} w_i - p} \qquad i=1,2,\ldots,n \qquad (4.50)$$

ve ağırlıklar,

$$w_i = \begin{cases} 1, & \left|\frac{r_i}{\hat{\sigma}_{YAEKK}}\right| \leq 2.5 \\ 0, & \left|\frac{r_i}{\hat{\sigma}_{YAEKK}}\right| \geq 2.5 \end{cases} \qquad i=1,2,\ldots,n \qquad (4.51)$$

ile gösterilmektedir (M.Ekiz ve O.U.Ekiz, 2018).

5. AYKIRI VE ETKİLİ GÖZLEMLERİN BELİRLENMESİNDE ETKİLİ UZAKLIK

Bu bölümde; çoklu etkili gözlemlerin tespitinde kullanılabilen sağlam etkili uzaklık (*EU*) ile doğrusal regresyon analizinde düzgün, aykırı, etkili gözlemler ve kötü kaldıraç noktalarının sınıflandırılmasını grup silme yöntemiyle sağlayan altı adımlı bir grafiksel yöntem tanıtılmıştır. Grup silme yöntemine ilişkin temel yaklaşım, düzgün olmayan tipte gözlem içermediğine inanılan veriler ile elde edilmiş bir alt küme tanımlamak ve bu kümedeki verilerin dışında kalan gözlemlerin veri merkezine olan uzaklığını araştırmaktır. Nurunnabi, Imon ve Nasser (2011) kaldıraç değerleri içeren veri setlerinde artık değerlerin çok büyük değerler aldığını, çoklu aykırı değer ve kötü kaldıraç noktalarının aynı anda veri kümesinde yer alması durumunda ise maskeleme ve süpürme etkisi dolayısıyla etkili gözlemlerin tespit edilememe riskinin bulunduğunu göstermişlerdir. Genellikle maskeleme ve süpürme etkisi nedeniyle yapılacak ilk işlemle birlikte etkili gözlem şüphesi taşıyan tüm gözlemlerin tespit edilmesi mümkün olamamaktadır. Eğer bu silme yöntemi ile, bir etkili gözlem veri kümesi dışında bırakılırsa tanısal yöntemler elverişsiz hale gelecektir. Çalışmalarda tanısal yöntemler uygulanmadan önce tüm potansiyel şüpheli gözlemlerin tespit edilmesi tercih edilmektedir. Düzgün bir gözlemin yanlışlıkla silinmesi, özellikle de bu gözlem iyi kaldıraç noktası ise, tüm analizi olumsuz olarak etkileyecektir (Habshah, Norazan ve Imon, 2009). Bu nedenle şüpheli gözlemlerin etkili gözlem olup olmadıklarının tespitinde iki aşamalı grup silme yönteminin kullanılması tercih edilebilir. Tanıtılacak *EU* yöntemine ilişkin adımlar aşağıda detaylı olarak açıklanmaktadır.

EU yöntemine ilişkin uygulama adımları

Adım 1- Etkin sağlam yöntemler ya da tanısal yaklaşımlar kullanılarak düzgün olmayan gözlemlerden oluşan şüpheli gözlem grubu bulunur. *n* sayıda gözlem içerisinden düzgün olmadığından şüphelenilen (aykırı değer, kötü kaldıraç noktası ya da etkili gözlem) *d* sayıda gözlemin belirlenebilmesi için bazı yöntemler uygulanır. Şüpheli aykırı değerler, kötü kaldıraç noktaları ve etkili gözlemler sırasıyla; standartlaştırılmış Student türü artıklar ve/veya standartlaştırılmış *EKMK* artıklar, kaldıraç değerleri ve *DFFITS* değerleri aracılığıyla tespit edilir. Bu şüpheli noktalar daha sonra veri kümesi dışarısında bırakılarak

yeni bir grup oluşturulur. Bu yeni grup *D* ile gösterilir. Geriye kalan *n-d* sayıda düzgün gözlemin sayısı *r* olmak üzere bu gözlemleri içeren küme *R* olsun. Böylece *X* ve *Y* veri matrisi, *R* düzgün gözlemlerin setini, *D* ise şüpheli gözlemlerin setini göstermek üzere,

$$X = \begin{bmatrix} X_R \\ X_D \end{bmatrix} \ , \ Y = \begin{bmatrix} Y_R \\ Y_D \end{bmatrix} \tag{5.1}$$

şeklinde ifade edilebilir.

Adım 2- Şüpheli gözlem grubu ihmal edilerek parametre tahminleri hesaplanır. $\hat{\theta}_R$ parametre tahmin vektörü, düzgün verilerden oluşan *R* veri kümesi kullanılarak,

$$\hat{\theta}_R = (X_R^T X_R)^{-1} X_R^T Y_R \tag{5.2}$$

ile elde edilir.

Adım 3- Çoklu aykırı değerler ve kötü kaldıraç noktalarının belirlenmesi için sırasıyla r_i^* genelleştirilmiş Student artıkları ve h_{ii}^* genelleştirilmiş kaldıraç değerleri hesaplanır (Imon, 2002, 2005). *R* veri kümesi kullanılarak r_i^* genelleştirilmiş Student artıkları;

$$r_i^* = \begin{cases} \frac{r_{i(R)}}{\hat{\sigma}_R \sqrt{1-h_{ii(R)}}} & i \epsilon R \\ \frac{r_{i(R)}}{\hat{\sigma}_R \sqrt{1+h_{ii(R)}}} & i \epsilon D \end{cases} \tag{5.3}$$

ile hesaplanır. Burada, *i.* artık $r_{i(R)} = Y_i - X_i \hat{\theta}_R$ *i.* kaldıraç değeri, $h_{ii(R)}$=$x_i^T (X_R^T X_R)^{-1} x_i \hat{\sigma}_R$ ve $\hat{\sigma}_R$ artık standart hatası *R* veri kümesindeki gözlemler üzerinden bulunur. $MAD(r_i^*) = medyan\{|r_i^* - medyan(r_i^*)|\}/0{,}6745$ olmak üzere, $medyan(r_i^*) \pm 3MAD(r_i^*)$ değerlerinin dışında kalan gözlemler aykırı değer olarak nitelendirilir. Genelleştirilmiş kaldıraç değerleri;

$$h_{ii}^* = \begin{cases} \frac{h_{ii(R)}}{1-h_{ii(R)}} & i \epsilon R \\ \frac{h_{ii(R)}}{1+h_{ii(R)}} & i \epsilon D \end{cases} \tag{5.4}$$

olmak üzere; h_{ii}^{*} değeri $medyan(h_{ii}^{*}) + 3MAD(h_{ii}^{*})$ değerlerinden daha büyük olan gözlem kötü kaldıraç noktası olarak nitelendirilir.

Adım 4- Adım 3'te elde edilen genelleştirilmiş student artıkları *y* eksenindeki noktaları ve genelleştirilmiş kaldıraç değerleri *x* eksenindeki noktaları temsil etmek üzere bir serpme grafiği elde edilir. Ayrıca grafikte çoklu aykırı değerlerin tespiti için *y* eksenini yatay olarak kesen noktası $(r_i^{*}) \pm 3MAD(r_i^{*})$ ve kötü kaldıraç noktalarını tespit için *x* eksenini dikey olarak kesen $(h_{ii}^{*}) + 3MAD(h_{ii}^{*})$ güven sınırları tespit edilir.

Adım 5- Etkili gözlemleri tespit etmede Mahalonobis uzaklığına benzer *EU*'dan yararlanılır. Adım 3'te elde edilen genelleştirilmiş student artıkları ilk sütun ve genelleştirilmiş kaldıraç değerleri ikinci sütun olmak üzere iki sütundan oluşan bir matris elde edilir. Bu matrise genelleştirilmiş artık ve kaldıraç (*GLAK*) matrisi adı verilir ve kısaca *G* ile gösterilir. *G* matrisinin *R* grubunda yer alan gözlemlerin ortalamalarına ilişkin matris $\bar{G}_R$ ve kovaryans matrisinin tersi ise $\hat{\Sigma}_R^{-1}$ olmak üzere *i.* gözlem için *EU*,

$$EU_i = \sqrt{(G_i - \bar{G}_R)\hat{\Sigma}_R^{-1}(G_i - \bar{G}_R)} \qquad i = 1,2,\dots,n \tag{5.5}$$

ile hesaplanır. G_R matrisi, Adım 3'te tespit edilen genelleştirilmiş student artıklar ve genelleştirilmiş kaldıraç değerlerine göre belirlenen çoklu aykırı değerler ve kötü kaldıraç noktaları haricindeki verilerden oluşmaktadır. Düzgün veriler ile elde edilmiş olan *GLAK* matrisine dayalı olarak ortalama ve kovaryans matrisinin kullanılmasıyla *EU* uzaklığı sağlamlık niteliği kazanmış olur.

Mahalanobis uzaklığının χ^2 dağılımına uyduğundan daha önceki bölümlerde bahsedilmişti. Hardin ve Rocke'ye (2005) göre, uç noktaların belirlenmesinde düzeltilmiş *F* dağılımı χ^2 dağılımından daha açıklayıcıdır. Ayrıca, sağlam uzaklıklar için *F* dağılımının diğer dağılımlara göre daha uygun olduğunu göstermişlerdir. Bu çalışmada tanıtılan *EU* için de Hardin ve Rocke'nin (2005) önerileri dikkate alınmış ve *i.* gözlem için,

$$EU_i > \sqrt{\frac{(n-1)p}{(n-p)} F_{\alpha,(p,n-p)}}, \qquad i=1,2,\dots,n \tag{5.6}$$

iken söz konusu gözlemin etkili olduğu yorumlanmıştır. Burada p değişken sayısı G matrisindeki değişken sayısı (3) ve α anlamlılık düzeyi % 5'dir.

Adım 6- Son olarak, Adım 4'te elde edilen serpme grafiği üzerinde hesaplanan *EU*'lar kullanılarak elipsoide benzer bir güven bölgesi çizilir. Bu grafik kaldıraç artık etkin (*KAE*) grafiği olarak isimlendirilir. Güven elipsoidinin dışında kalan gözlemler etkili gözlem olarak nitelendirilir. Bu grafikte veri kümesinin; düzgün gözlemler, aykırı değerler, kötü kaldıraç noktaları ve etkili gözlemler olarak dört grupta sınıflanması sağlanır. Bu çalışmada *EU*'ın değiştirilmiş bir versiyonunun düzgün olmayan gözlemleri doğru tespit etme performanslarına dayanan bir ölçüt kullanılacağından bu grafikler kullanılmamıştır.

Nurunnabi, Nasser ve Imon (2016) doğrusal regresyon analizinde düzgün olmayan gözlemlerin tespit edilmesi için, standartlaştırılmış Student türü artıklar, standartlaştırılmış *EKMK* artıklar, kaldıraç değerleri ve *DFFITS* değerlerine dayalı *EU* ölçütünün kullanımını önermişlerdir. Bu çalışmanın sonraki bölümünde düzgün olmayan gözlemlerin belirlenmesinde *EU* ölçütü; sağlam olmayan *EÇOB* yöntemi, sağlam *EKMK*, *YAEKK*, *M* ve *S* tahmin edicilere dayalı olarak hesaplanmış ve bu yöntemlere dayalı *EU* ölçütlerinin performanslarının incelenmesi amaçlanmıştır.

6. SİMÜLASYON ÇALIŞMASI

Veri kümesi içerisinde aykırı ve etkili gözlemlerin aynı anda yer alması durumunda, bu gözlemlerin belirlenmesinin zorluğu daha önceki bölümlerde detaylı olarak ele alınmıştır. Bu çalışmanın beşinci bölümünde tanıtılan *EU*, bu sorunun çözümü için Nurunnabi ve arkadaşları (2016) tarafından önerilmiştir. *EU*'a ilişkin hesaplamalar yapılırken veri kümesi, düzgün gözlemleri içeren (*R*) ve düzgün olmayan gözlemleri içeren (*D*) veri kümesi olmak üzere iki ayrı kümeye ayrılır ve daha sonraki hesaplamalar bu iki veri kümesi üzerinden ayrı ayrı gerçekleştirilir. *D* veri kümesi belirlenirken standartlaştırılmış Student türü artıklar, kaldıraç değerleri ve *DFFITS* değerleri gibi sağlam olmayan yöntemler kullanılmaktadır. Bu nedenle *EU* tam olarak sağlam bir yöntem olarak tanımlanamamaktadır. Bu çalışmada *D* veri kümesi belirlenirken sağlam olmayan *EÇOB* yöntemi ve düzgün olmayan gözlemlerden etkilenmeyen sağlam *EKMK*, *YAEKK*, *M* ve *S* tahmin edicilere dayalı bir *EU* tanıtılmıştır. Gerçekleştirilen simülasyon çalışmasında *EU* algoritmasında *D* veri kümesini belirlemek amacıyla *EÇOB*, *EKMK*, *YAEKK*, *M* ve *S* yöntemleri kullanılarak *EÇOB* tahmin edicisine dayalı *EU* (*EU-EÇOB*), *EKMK* yöntemine dayalı *EU* (*EU-EKMK*), *YAEKK* yöntemine dayalı *EU (EU-YAEKK)* , *M* tipi tahmin ediciye dayalı *EU* (*EU-M*) ve *S* tipi tahmin ediciye dayalı *EU* (*EU-S*) ölçütleri elde edilmiştir. Bu ölçütler veri kümesine farklı oranlarda ($0.02 \leq \gamma \leq 0.20$) ve türde (etkili gözlem (γ_{etk}) ve iyi kaldıraç noktası (γ_k)) düzgün olmayan gözlem eklenerek 10 000 tekrara dayalı olarak hesaplanmıştır. 10 000 tekrarda çoklu düzgün olmayan gözlemlerin kaç kez doğru olarak tespit edildiğine dayalı olarak tanımlanan doğru belirleme oranı (*DBO*) elde edilmiştir. *EU-EÇOB, EU-EKMK, EU-YAEKK*, *EU-M* ve *EU-S*'in *DBO*'ları üzerinden karşılaştırılması amacıyla *n*=20, 30, 50, 100 için sonuçlar tablolaştırılmıştır. Bağımsız değişken sayısı 3 olan regresyon modeli için veri;

$$\begin{bmatrix} Y \\ X_1 \\ X_2 \\ X_3 \end{bmatrix} \sim N_4 \left[\mu = \begin{bmatrix} 10 \\ 2 \\ 3 \\ 4 \end{bmatrix}, \Sigma_{XY} = \begin{bmatrix} 1 & 0.90 & 0.90 & 0.90 \\ 0.90 & 1 & 0.99 & 0.90 \\ 0.90 & 0.99 & 1 & 0.90 \\ 0.90 & 0.90 & 0.90 & 1 \end{bmatrix} \right]$$

olacak şekilde üretilmiştir.

Çizelge 6.1. n=20 için *EÇOB, EKMK, YAEKK, M* ve *S* tahmin edicilere dayalı *EU* üzerinden *DBO* değerleri

n	γ_k	γ_{etk}	γ	*EU-EÇOB*	*EU-EKMK*	*EU-YAEKK*	*EU-M*	*EU-S*
20	0	0,05	0,05	0,5216	0,2370	0,1465	0,5526	0,6658
	0	0,10	0,10	0,5656	0,1725	0,1475	0	0
	0	0,15	0,15	0,1351	0,0856	0,0435	0	0
	0	0,20	0,20	0	0,0297	0	0	0
	0,05	0	0,05	0,5310	0,2367	0,1523	0,5502	0,6652
	0,05	0,05	0,10	0,5275	0,3674	0,1487	0,5543	0,6670
	0,05	0,10	0,15	0,1316	0,4916	0,8380	0,6427	0,8386
	0,05	0,15	0,20	0,0041	0,0718	0,0954	0	0
	0,05	0,20	0,25	0	0,0251	0	0	0
	0,10	0	0,10	0,0336	0,0152	0,0066	0,1304	0,1440
	0,10	0,05	0,15	0,5206	0,4654	0,2733	0,0166	0,1698
	0,10	0,10	0,20	0,2405	0,0882	0,1505	0	0,0019
	0,10	0,15	0,25	0,0313	0,1219	0,8528	0	0
	0,10	0,20	0,30	0,0020	0,0185	0,0895	0	0
	0,15	0	0,15	0,1351	0,0856	0,0435	0	0
	0,15	0,05	0,20	0,3780	0,2254	0,1981	0	0
	0,15	0,10	0,25	0,0691	0,0444	0,0959	0	0
	0,15	0,15	0,30	0,0052	0,0234	0,0218	0	0
	0,15	0,20	0,35	0	0,0392	0,1913	0	0
	0,20	0	0,20	0	0,0342	0,1001	0	0
	0,20	0,05	0,25	0	0,0181	0,0935	0	0
	0,20	0,10	0,30	0,0487	0,0178	0,0212	0	0
	0,20	0,15	0,35	0	0,0152	0	0	0
	0,20	0,20	0,40	0	0,0082	0	0	0

Çizelge 6.1'de örnek çapı 20 iken, veri kümesi içerisinde farklı oranlarda iyi kaldıraç ve etkili gözlemlerin yer alması durumları için *EU-EÇOB, EU-EKMK*, *EU-YAEKK* ve *EU-S* ölçütlerine ilişkin 10 000 tekrarda hesaplanan *DBO*'ları yer almaktadır.

Veri kümesi iyi kaldıraç içermiyorken ve etkili gözlem oranı % 5 olduğunda, *DBO*'ları bakımından en iyi sonucu *EU-S* ölçütü vermektedir. *EU-EÇOB* ve *EU-M* ölçütlerine ilişkin *DBO*'ları, *EU-EKMK* ve *EU-YAEKK* ölçütlerine göre oldukça yüksektir. Veri kümesi iyi

kaldıraç içermiyor ve etkili gözlem oranı; % 10 ve % 15 iken *EU-EÇOB* ölçütü *DBO'*ları açısından en iyi sonucu vermektedir. Veri kümesi iyi kaldıraç içermiyor ve %20 oranında etkili gözlem içeriyorken ise tüm yöntemler oldukça düşük *DBO*'na sahiptir.

Veri kümesi içerisinde yer alan iyi kaldıraç oranı % 5 ve etkili gözlem oranı % 0 ve % 5 iken *EU-S* ölçütü *DBO* bakımından en iyi sonucu vermektedir. *EU-M* ve *EU-EÇOB* ölçütleri ise *DBO* bakımından *EU-S* ölçütünden sonra en iyi performansı göstermişlerdir. İyi kaldıraç oranı % 5 ve etkili gözlem oranı % 0 ve % 5 iken *EU-S* ölçütü *DBO* bakımından en iyi sonucu vermektedir. *EU-EÇOB* ve *EU-M* ölçütleri de oldukça yüksek *DBO*'larına sahiptir. İyi kaldıraç oranı % 5 ve etkili gözlem oranı % 10 iken *DBO* bakımından *EU-S* ölçütü en iyi performansı göstermekte, *EU-YAEKK* ve *EU-M* ölçütleri de oldukça yüksek *DBO*'larına sahip olmaktadır. *EU-EÇOB* ölçütüne ilişkin *DBO* ise oldukça düşüktür. İyi kaldıraç oranı % 5 ve etkili gözlem oranı % 15 iken tüm ölçütler düşük *DBO*'na sahiptir. Ancak, *EU-YAEKK* diğer ölçütlerle kıyaslandığında *DBO* bakımından daha iyi performans göstermektedir. İyi kaldıraç oranı % 5 ve etkili gözlem oranı % 20 iken tüm ölçütler sıfıra yakın *DBO*'na sahiptir.

Veri kümesi % 10 oranında iyi kaldıraç içeriyor ve etkili gözlem bulunmuyorken tüm ölçütler oldukça düşük *DBO*'larına sahiptir. Bununla birlikte *EU-M* ve *EU-S* ölçütleri, *EU-EÇOB, EU-EKMK* ve *EU-YAEKK* ölçütlerine kıyasla daha iyi performans göstermektedir. İyi kaldıraç oranı % 10 ve etkili gözlem oranı % 5 iken *DBO* bakımından en iyi sonucu *EU-EÇOB* ve *EU-EKMK* ölçütleri vermektedir. İyi kaldıraç ve etkili gözlem oranı % 10 iken en yüksek *DBO EU-EÇOB* ölçütünündür. İyi kaldıraç oranı % 10 ve etkili gözlem oranı % 15 iken en iyi performansı oldukça yüksek bir *DBO*'na sahip *EU- YAEKK* gösterirken *EU-M* ve *EU-S* ölçütlerine ilişkin *DBO*'ları sıfırlanmıştır. İyi kaldıraç oranı % 10 iken, etkili gözlem oranının % 20 olduğu durumda tüm ölçütlerin *DBO*'ları ya sıfıra yaklaşmış ya da sıfırlanmıştır. İyi kaldıraç oranı % 15'in üzerinde iken tüm etkili gözlem oranları için *DBO*'ları ya sıfıra yaklaşmış ya da sıfırlanmıştır.

Çizelge 6.2. n=30 için *EÇOB, EKMK, YAEKK*, *M* ve *S* tahmin edicilere dayalı *EU* üzerinden *DBO* değerleri

n	γ_k	γ_{etk}	γ	*EU-EÇOB*	*EU-EKMK*	*EU-YAEKK*	*EU-M*	*EU-S*
30	0	0,035	0,035	0,2179	0,0698	0,0319	0,3146	0,3990
	0	0,070	0,070	0,2137	0,0711	0,0362	0,3264	0,3985
	0	0,100	0,100	0,2781	0,0941	0,0309	0,0217	0,0333
	0,035	0	0,035	0,1815	0,1442	0,1782	0,6517	0,7220
	0,035	0,035	0,070	0,1941	0,1479	0,0301	0,3145	0,3947
	0,035	0,070	0,105	0,0304	0,3334	0,1382	0,3445	0,4450
	0,035	0,100	0,135	0,0638	0,3403	0,1597	0,1990	0,2350
	0,070	0	0,070	0,1546	0,0264	0,1096	0,0339	0,0491
	0,070	0,035	0,105	0,1826	0,3737	0,8625	0,7115	0,8953
	0,070	0,070	0,140	0,2165	0,0926	0,0402	0,3636	0,4464
	0,070	0,100	0,170	0,1693	0,0744	0,2683	0,2677	0,5110
	0,100	0	0,100	0,0284	0,0027	0,0332	0	0
	0,100	0,035	0,135	0,1092	0,0864	0,3776	0,0258	0,0800
	0,100	0,070	0,170	0,1961	0,2120	0,8703	0,6884	0,8760
	0,100	0,100	0,200	0,0742	0,0277	0,0491	0,0288	0,1186
	0,150	0,150	0,300	0,0270	0,0304	0,0113	0	0
	0,170	0,170	0,340	0	0,0022	0,0020	0	0

Çizelge 6.2'de, örnek çapının 30 olduğu durumda veri kümesi içerisinde yer alan iyi kaldıraç ve etkili gözlem oranları değiştikçe farklı ölçütlerin 10 000 tekrarda hesaplanan *DBO*'ları bakımından iyi performans gösterdikleri görülmektedir.

Veri kümesi iyi kaldıraç içermiyor ve % 3,5 ve % 7 oranlarında etkili gözlem içeriyorken *EU-M* ve *EU-S* ölçütlerine ilişkin *DBO*'ları en yüksektir. Veri kümesi iyi kaldıraç içermiyor ve % 10 oranında etkili gözlem içeriyorken *EU-EÇOB* ölçütü *DBO* bakımından en iyi performansı göstermektedir.

Veri kümesinde % 3,5 oranında iyi kaldıraç yer alıyor ve etkili gözlem bulunmuyorken *EU-M* ve *EU-S* ölçütleri, *EU-EÇOB, EU-EKMK* ve *EU-YAEKK* ölçütlerine kıyasla oldukça yüksek *DBO*'larına sahiptir. Veri kümesi % 3,5 oranında iyi kaldıraç içeriyor ve etkili gözlem oranı % 3,5 ve % 7 iken *EU-M* ve *EU-S* ölçütleri *EU-EÇOB, EU-EKMK* ve *EU-YAEKK* ölçütlerinden daha iyi performans göstermektedir. Veri kümesinde % 3,5

oranında iyi kaldıraç ve % 10 oranında etkili gözlem yer alıyor iken en yüksek *DBO*'na *EU-EKMK* ölçütü sahiptir.

Veri kümesinin içerdiği iyi kaldıraç oranı % 7 ve veri kümesi; etkili gözlem içermiyorken *EU-EÇOB* ölçütü, % 3,5 oranında etkili gözlem içerdiğinde *EU-YAEKK, EU-M* ve *EU-S* ölçütleri *DBO*'ları bakımından en iyi performansı göstermektedir. İyi kaldıraç oranı % 7 ve etkili gözlem oranı % 7 ve % 10 iken *EU-S* ölçütü en yüksek *DBO*'na sahiptir.

Veri kümesi % 10 oranında iyi kaldıraç içeriyorken; etkili gözlem bulunmayan ve % 3,5 oranında etkili gözlem bulunan durumda tüm ölçütlere ilişkin *DBO*'ları düşüktür. % 10 oranında iyi kaldıraç ve % 7 oranında etkili gözlem içeren veri kümesi için *EU-YAEKK, EU-M* ve *EU-S* ölçütleri oldukça yüksek *DBO*'na sahiptir.

Veri kümesinin % 10' un üzerinde iyi kaldıraç içerdiği ve % 7'nin üzerinde etkili gözlem içerdiği durumlarda tüm ölçütler sıfıra yakın ya da sıfır *DBO*'na sahiptir.

Çizelge 6.3. n=50 için *EÇOB, EKMK, YAEKK, M* ve *S* tahmin edicilere dayalı *EU* üzerinden *DBO* değerleri

n	γ_k	γ_{etk}	γ	*EU-EÇOB*	*EU-EKMK*	*EU-YAEKK*	*EU-M*	*EU-S*
50	0	0,02	0,02	0,0312	0,0079	0,0013	0,1309	0,1432
	0	0,04	0,04	0,0291	0,0151	0,0023	0,1306	0,1467
	0	0,06	0,06	0,0359	0,0241	0,0015	0,1336	0,1443
	0	0,08	0,08	0,0424	0,0268	0,0022	0,1200	0,1346
	0	0,10	0,10	0,0270	0	0	0,0724	0,0746
	0,02	0	0,02	0,0250	0,0260	0,0175	0,5202	0,5489
	0,02	0,02	0,04	0,0274	0,0294	0,0025	0,1291	0,1551
	0,02	0,04	0,06	0,0272	0,2268	0,1084	0,3171	0,4110
	0,02	0,06	0,08	0,0269	0,3374	0,1393	0,3491	0,4412
	0,02	0,08	0,10	0,0395	0,1333	0,2556	0,4526	0,5522
	0,02	0,10	0,12	0,0616	0,0332	0,1594	0,1975	0,2409
	0,04	0	0,04	0,0292	0,0110	0,0218	0,2157	0,2620
	0,04	0,02	0,06	0,0259	0,2501	0,7864	0,6947	0,8057
	0,04	0,04	0,08	0,0256	0,0532	0,0066	0,1710	0,2152
	0,04	0,06	0,10	0,0293	0,1164	0,1774	0,3882	0,4935

Çizelge 6.3. (devam) n=50 için *EÇOB, EKMK, YAEKK*, *M* ve *S* tahmin edicilere dayalı *EU* üzerinden *DBO* değerleri

	0,04	0,08	0,12	0,0344	0,3696	0,1580	0,3421	0,4358
	0,04	0,10	0,14	0,0517	0,1292	0,3028	0,4204	0,4901
	0,06	0	0,06	0,0821	0,0024	0,0106	0,0243	0,0365
	0,06	0,02	0,08	0,0687	0,1633	0,6018	0,2260	0,3975
	0,06	0,04	0,10	0,0312	0,3056	0,7899	0,7047	0,8065
	0,06	0,06	0,12	0,0323	0,0292	0,0101	0,2072	0,2767
	0,06	0,08	0,14	0,0688	0,0318	0,2181	0,4404	0,5494
	0,06	0,10	0,16	0,0970	0,0262	0,3368	0,5137	0,5900
	0,08	0	0,08	0,0974	0	0,0057	0,0011	0,0011
	0,08	0,02	0,10	0,1198	0,0653	0,2744	0,0123	0,0383
	0,08	0,04	0,12	0,0771	0,2382	0,6033	0,2312	0,3770
	0,08	0,06	0,14	0,0341	0,2065	0,8033	0,7334	0,8230
	0,08	0,08	0,16	0,0930	0,0126	0,0175	0,2463	0,3259
	0,08	0,10	0,18	0,1496	0,0086	0,2513	0,4221	0,5431
	0,10	0	0,10	0,0265	0	0,0026	0	0
	0,10	0,02	0,12	0,0690	0,0315	0,0600	0	0
	0,10	0,04	0,14	0,1241	0,1082	0,2790	0,0139	0,0369
	0,10	0,06	0,16	0,0630	0,1743	0,6014	0,2453	0,3702
	0,10	0,08	0,18	0,1314	0,1143	0,8049	0,6357	0,7625
	0,10	0,10	0,20	0,1023	0,0044	0,0235	0,1741	0,2410
	0,12	0,12	0,24	0,0102	0,0086	0,0053	0	0
	0,20	0,20	0,40	0	0	0	0	0

Çizelge 6.3'de görüldüğü üzere, örnek çapı 50 iken veri kümesinin içerdiği çeşitli miktarlardaki iyi kaldıraç ve etkili gözlem oranları için 10 000 tekrarda hesaplanan *DBO*'ları bakımından farklı ölçütler öne çıkmaktadır.

Veri kümesi iyi kaldıraç içermiyor ve etkili gözlem oranı % 2, 4, 6, 8 ve 10 iken tüm ölçütler oldukça düşük *DBO*'na sahiptir. Bununla birlikte *EU-M* ve *EU-S* ölçütü, *EU-EÇOB, EU-EKMK* ve *EU-YAEKK* ölçütlerine göre daha yüksek *DBO*' na sahiptir.

Veri kümesi % 2 oranında iyi kaldıraç içeriyor ve etkili gözlem oranı % 0, 2, 4, 6, 8 ve 10 iken *DBO*'ları bakımından *EU-M* ve *EU-S* ölçütleri *EU-EÇOB, EU-EKMK* ve *EU-YAEKK* ölçütlerine kıyasla daha iyi performans göstermektedir.

Veri kümesi; % 4 oranında iyi kaldıraç içeriyor ve etkili gözlem içermiyorken *EU-M* ve *EU-S* ölçütleri, % 2 oranında etkili gözlem içeriyorken *EU-YAEKK, EU-M* ve *EU-S* ölçütleri, % 4, 6, 8 ve 10 oranında etkili gözlem içeriyorken *EU-M* ve *EU-S* ölçütleri *DBO* bakımından iyi performans göstermektedir.

Veri kümesinin içerdiği iyi kaldıraç oranı % 6 ve veri kümesi etkili gözlem içermiyorken tüm ölçütler düşük *DBO*'na sahiptir. Veri kümesi % 6 oranında iyi kaldıraç içeriyor; etkili gözlem oranı % 2 ve 4 iken, *EU-YAEKK, EU-M* ve *EU-S* ölçütü, etkili gözlem oranı % 6, 8 ve 10 iken *EU-M* ve *EU-S* ölçütleri *DBO* bakımından iyi performans göstermektedir.

Veri kümesi; % 8 oranında iyi kaldıraç içeriyor ve etkili gözlem içermiyorken tüm yöntemler sıfıra yakın *DBO*'na sahiptir. Veri kümesinin içerdiği iyi kaldıraç oranı % 8 iken; % 2 ve 4 etkili gözlem oranı için *EU-YAEKK* ölçütü, % 6 etkili gözlem oranı için *EU-YAEKK, EU-M* ve *EU-S* ölçütü, % 8 ve %10 etkili gözlem oranı için *EU-M* ve *EU-S* ölçütü *DBO* bakımından iyi performans göstermektedir.

Veri kümesinin içerdiği iyi kaldıraç oranı % 10 iken; veri kümesi etkili gözlem içermiyor ve % 2 oranında etkili gözlem içeriyorken tüm ölçütler sıfıra yakın *DBO*'na sahiptir. İyi kaldıraç oranı % 10 ve etkili gözlem oranı % 4, 6 ve 8 iken *EU-YAEKK* ve *EU-S* ölçütleri iyi performans göstermektedir. Etkili gözlem oranı % 10 olduğunda *EU-M* ve *EU-S* ölçütleri *EU-EÇOB, EU-EKMK, EU-YAEKK* ölçütlerinden daha yüksek *DBO*'na sahiptir.

Veri kümesinin % 10' un üzerinde iyi kaldıraç içerdiği ve etkili gözlem içerdiği durumlarda tüm ölçütler sıfıra yakın ya da sıfır *DBO*'na sahiptir.

Çizelge 6.4. n=100 için *EÇOB, EKMK, YAEKK, M* ve *S* tahmin edicilere dayalı *EU* üzerinden *DBO* değerleri

n	γ_k	γ_{etk}	γ	*EU-EÇOB*	*EU-EKMK*	*EU-YAEKK*	*EU-M*	*EU-S*
100	0	0,01	0,01	0	0	0	0,0135	0,0136
	0	0,02	0,02	0	0	0	0,0126	0,1370
	0,01	0,01	0,02	0	0	0	0,0114	0,0115
	0,01	0,02	0,03	0	0	0	0,0123	0,0265
	0,02	0	0,02	0	0	0	0,2066	0,2086
	0,02	0,02	0,04	0	0	0	0,0294	0,0405
	0,10	0,10	0,20	0,0521	0	0,0211	0,2260	0,2296
	0,15	0,15	0,30	0	0	0,0100	0	0

Çizelge 6.4’ de görüldüğü üzere örnek çapı 100 iken, veri kümesi içerisindeki tüm kaldıraç noktası ve etkili gözlem oranları için *EU-M* ve *EU-S* ölçütleri, *DBO* bakımından çok yüksek değerler almasa da *EU-EÇOB, EU-EKMK, EU-YAEKK* ölçütlerine göre daha iyi performans göstermektedir.

7. SONUÇ VE ÖNERİLER

Bu çalışmanın birinci bölümünde, doğrusal regresyon analizi, düzgün olmayan gözlem türleri ve bunların tespit edilmesinde kullanılan sağlam ve sağlam olmayan bazı yöntemlere yer verilmiştir.

İkinci bölümde, değişkenler arasındaki ilişkilerin incelenmesinde kullanılan doğrusal regresyon analizine ilişkin bazı temel bilgilerden, tahmin yöntemlerinden ve tahmin edicilerin özelliklerinden bahsedilmiştir.

Üçüncü bölümde, veri kümesi içerisinde yer alan düzgün olmayan gözlem türleri ve bu gözlemlerin tespit edilmesinin doğrusal regresyon analizi için önemi açıklanmıştır.

Dördüncü bölümde, düzgün olmayan gözlemlerin tespit edilmesinde kullanılan tanısal yaklaşımlar ve sağlam yöntemlere detaylı olarak yer verilmiştir.

Beşinci bölümde, veri kümesi içerisinde çoklu düzgün olmayan gözlemlerin yer alması durumunda bu gözlemlerin tespiti için Nurunnabi ve arkadaşları (2016) tarafından önerilen *EU* ölçüt tanıtılmış ve bu yöntemin uygulanmasına ilişkin adımlardan bahsedilmiştir.

Altıncı bölümde ise *EU'*ın algoritması *EÇOB*, *EKMK*, *YAEKK*, *M* ve *S* tahmin edicilere dayalı olarak değiştirilerek simülasyon çalışması yapılmıştır. Bu çalışmada çeşitli örnek çapları ve düzgün olmayan gözlemlerin farklı oranları için bu yöntemler *DBO'*ları bakımından karşılaştırılmıştır.

Elde edilen tüm sonuçlar değerlendirildiğinde, n=20 örnek çapı için genellikle *EU-EÇOB, EU-M* ve *EU-S* ölçütleri *DBO*'ları bakımından daha iyi performans göstermektedir. Ancak, veri kümesinde yer alan düzgün olmayan gözlem oranı arttığında bu ölçütlerin *DBO*'ları sıfıra yaklaşmakta ya da sıfırlanmaktadır. Örnek çapı n=30 ve 50 iken *EU-M* ve *EU-S* ölçütleri tüm düzgün olmayan gözlem oranları için daha yüksek *DBO*'na sahip olmakla birlikte düzgün olmayan gözlemlerin oranı % 28 seviyesine ulaştığında sıfırlanmaktadır. Örnek çapı n=100 olduğunda *EU-M* ve *EU-S* ölçütleri *DBO* bakımından daha iyi performans göstermektedir.

Sonuç olarak bu çalışmada, veri kümesi 20, 30, 50 ve 100 gözlem içerdiğinde ve bu gözlemlerin içerisinde farklı tür ve oranlarda düzgün olmayan gözlem yer aldığında, şüpheli gözlemlerin tespit edilmesinde sağlam tahmin edicilere dayalı *EU* ölçütlerinin genellikle daha iyi performans gösterdiği gözlemlenmiştir. Özellikle veri kümesi 50 ve 100 gözlem değerine sahipken genel olarak sağlam *EU-S* ve *EU-M* ölçütleri düzgün olmayan gözlemlerin tespit edilmesinde diğer sağlam ve sağlam olmayan yöntemlerden daha iyi performans göstermektedir. Farklı tür ve oranlarda düzgün olmayan gözlem içerdiği düşünülen veri kümeleri için şüpheli gözlemlerin tespitinde sağlam *EU-S* ve *EU-M* ölçütlerinin kullanılması sağlam olmayan yöntemlere göre daha iyi bir alternatif olarak düşünülebilir.

KAYNAKLAR

Armstrong, R. D. and Kung, M. T. (1978). Least absolute value estimates for a simple linear regression problem. *Journal of the Royal Statistical Society*, 27(3), 363–366.

Atkinson, A.C. (1986). Masking unmasked. *Biometrica*, 73(3), 533-541.

Imon, A. H. M. R. (2005). Identifying multiple influential observations in linear regression, *Journal of Applied Statistics*, 32(9), 929–946.

Bagheri, A., Midi, H., Ganjali, M. and Eftekhari, S. (2010). A comparison of various influential points diagnostic methods and robust regression approaches: Reanalysis of interstitial lung disease data. *Applied Mathematical Sciences*, 4(28), 1367–1386.

Barrera, M. S. and Yohai, V. J. (2006). A fast algorithm for S regression estimates. *Journal of Computational and Graphical Statistics*, 15(2), 414-427.

Belsley, D. A., Kuh, E. and Welsch, R. E. (1980). *Regression diagnostics identifing influential data and sources of collinearity*. (First edition). New Jersey: John Wiley and Sons, 13-75.

Billor, N. and Hadi, A. (2006). A re-weighted least squares method for robust regression estimation. *American Journal of Mathematical and management Sciences*, 26(3-4), 229-252.

Birkes, D. and Dodge,Y. (1993). *Alternative methods of regression*. (First Published), New York: John Wiley and Sons, 85.

Boss, D. D. and Stefanski, L. A. (2001). The calculus of m-estimation. *The American Statistician*, 12(1), 29-42.

Box, G. E. P. (1953). Non-Normality and tests on variances. *Biometrica*, 40, 318-335.

Casella, G. and Berger, R. (2001). *Statistical inference*. (Second edition). United States of America: Duxbury, 214-578.

Chang, L., Roberts, S. and Welsh, A. (2017). Robust Lasso regression using Tukey's biweight criterion. *Technometrics*, 60, 36-47.

Chatterjee, S. and Hadi, A. S. (1986). Influential observations, high leverage points, and outliers in linear regression. *Statistical Science*, 1(3), 379-393.

Chatterjee, S. and Hadi, A.S. (1988). *Sensitivity analysis in linear regression*. (First edition). New York: John Wiley and Sons, 73.

Chatterjee, S. and Machler, M. (1997). Robust regression: A weighted least squares approach. *Communications in Statistics, Theory and Methods*, 26, 1381–1394.

Cook, R. D. (1977). Detection of influential observation in linear regression. *Technometrics*, 19(1), 15-18.

Cook, R. D. and Weisberg, S. (1982). *Residuals and influence in regression*. (First edition). New York: Chapman and Hall, 14-20.

Dhhan, W., Rana, S. and Midi, H. (2016). A high breakdown, high efficiency and bounded influence modified GM estimator based on support vector regression. *Journal of Applied Statistics,* 44(4), 700-714.

Dobson, A. J. (2002). *An Introduction to generalized linear models* (Second edition). Washington: Chapman and Hall, 31-32.

Donoho, D. and Huber, P. J. (1983). *The notion of breakdown point.* Wadsworth: A Festschirift for Erich Lehmann, 157-184.

Donoho, D., Johnstone, L., Rousseeuw, P. and Stahel, W. (1985). Discussion: Projection pursuit. *The Annals of Statistics*, 13(2), 496-500.

Draper, N. R. and Smith, H. (1966). *Applied regression analysis*. New York: John Wiley and Sons, 1.

Ekiz, M. and Ekiz, U. (2018). Impact of interactions between collinearity, leverage points and outliers on ridge, robust, and ridge-type robust estimators. *International Journal of Statistics and Applications,*8(2), 88-102.

Galton, F. (1886). Family likeness in stature. *Proceedings of Royal Society*, 7(40), 42-72.

Gath, E. G. and Hayes, K. (2011). Bounds for a multivariate extension of range over standard deviation besed on the Mahalanobis distance. *Linear Algebra and Its Applications*, 435(6), 1267-1276.

Gujarati, D. N. (2010). *Temel ekonometri*. (Yedinci baskı). İstanbul: Literatür Yayıncılık, 52-72.

Habshah, M. Norazan, R. and Imon, A. H. M. R. (2009). The performance of diagnostic–robust generalized potentials for the identification of multiple high leverage points in linear regression. *Journal of Applied Statistics,* 36(5), 507-520.

Hampel, F. R. (1968). *Contributions to theory of robust estimation*. Ph. D. dissertation, University of California, Berkeley.

Hampel, F. R. (1971). A general qualitative definition of robustness. *The Annal of Mathematical Statistics,* 42(6), 45-49.

Hampel, F. R. (1974). The influence curve and its role in robust estimation. *Journal of the American Association*, 69(346), 383-393.

Hampel, F. R. (1975). Beyond location parameters: Robust concepts and methods (with discussion). *Bulletin of the International Statistical Istitute*, 46(1), 375–391.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J. and Stahel, W. A. (1986). *Robust statistics*. (First edition). New York: John Wiley and Sons, 11.

Hardin, J. and Rocke, D.M. (2005). The distribution of robust distances. *Journal of Computational and Graphical Statistic*s, 14, 928–946.

Hawkins, D. M. (1980). *Identification of outliers*. (First edition). London: Chapman and Hall, 51.

Hill, R. W. (1977). *Robust regression when there are outliers in the carriers*. Ph.D. Thesis, Harvard University, Cambridge.

Hoaglin, D. C. and Welsch, R. E. (1978). The hat matrix in regression and anova. *The American Statisticion*, 32(1), 17-22.

Hoaglin, D. C., Mosteller, F. and Tukey, J. W. (1983). *Understanding robust and explaratory Data analysis*. (First edition). Canada: John Wiley and Sons, 291.

Hodges, J. L. (1967). Efficiency in normal samples and tolerance of extreme values for some estimates of location. *Mathematical Statistics and Probability*, 1, 163-168.

Hogg, R. V. and Craig, A. (2014). *Introduction to mathematical statistics.* (Fifth edition). Hong Kong and Macou: Pearson Education Asia Limited and Higher Education Press, 316-317.

Hössjer, O. (1992). On the optimality of S estimators. *Statistics & Probability Letters*, 14(5), 413-419.

Huber, P. J. (1964). Robust estimation of allocation parameter. *The Annal of Mathematical Statistics*, 35(1), 73-101.

Huber, P. J. (1967). In proceedings of the fifth berkeley symposium. *Mathematical Statistics and Probability*, 1(1), 221-233.

Huber, P. J. (1973). Robust estimation: Asymptoties, conjectures and Monte Carlo. *Annal Statistics*, 1(5), 799-821.

Huber, P. J. (1981). *Robust statistic*. (First edition). New York: John Wiley and Sons, 1-162.

Huber, P. J. (1987). Robust statistical procedures. (Second edition). Bayreuth: *Society for Industrial and Applied Mathematics*, 13-14.

İnternet: Fox, J. (January, 2002). Robust regression: Appendix to an R and S-Plus companion to applied regression. URL: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.182.5736, Son Erişim Tarihi: 15.08.2019.

Lawrance, A. J. (1995). Deletion influence and masking in regression. *Journal of the Royal Statistical Society,* 57(1), 181-189.

Maronna, R. A., Martin, R. D. and Yohai, V. J. (2006). *Robust statistics: Theory and methods* (First edition). Chiester: John Wiley and Sons, 29-131.

Montgomery, D. C., Peck, E. A. and Vining, G. G. (2012). *Introduction to linear regression analysis*. (Beşinci baskı). New York: John Wiley and Sons, 1.

Newcomb, S. (1878). *Researches on the motion of the moon, I. Washington observations for 1875–appendix II*. Washington: The U.S. Naval Observatory.

Nurunnabi A. M., Nasser, M. A. and Imon, A. H. M. R. (2016). Identification and classification of multiple outlers, high leverege points and influential observations in linear regresion. *Journal of Applied Statistics*, 43(3), 509-525.

Nurunnabi, A., Imon, A. H. M. R. and Nasser, M. A. (2011). Diagnostic measure for influential observation in linear regression. *Communications in Statistics – Theory and Methods*, 40(7), 1169-1183.

Pena, D. and Yohai, V. J. (1995). The detection of influential subsets in linear regression using an influence matrix. *Journal of the Royal Statistical Society*, 57(1), 145-156.

Rousseeuw, P. J. (1984). Least median of squares regression. *Journal of American Statistic Association*, 79(388), 871-880.

Rousseeuw, P. J. and Yohai, V. J. (1984). *Robust regression by means of S-estimators. Robust and Nonlinear Time Series Analysis* (J. Franke, W. Härdle and R. Martin, eds.). Springer, New York: Lecture Notes in Statist. 26, 256–272.

Rousseeuw, P. J. (1985). Multivariate estimation with high breakdown point. *Institute Mathematical Statistics and Applications*,1(1), 283-297.

Rousseeuw, P. L. and Leroy, A. M. (1987). *Robust regression and outlier detection*. (First edition). Belgium: John Wiley and Sons, 1-220.

Rousseeuw, P. J. and Hubert, M. (1997). Recent developments in progress, L_1 statistical procedures and related topics. *Institute of Mathematical Statistics Lecture Notes – Monograph Series*, 31(1), 201-214.

Rousseeuw, P. J. and Driessen, K. V. (1999). Computing LTS regression for large data sets, Technical Report University of Antwerp 1. *Belgium*, 212-223.

Ryan, T. P. (2009). *Modern regression methods*. (Second Edition). New York: John Wiley and Sons, 421.

Serfling, R. and Wang, S. (2014). General foundations for studying masking and swamping robustness of outlier identifiers. *Statistical Methodology*, 20(1), 79-90.

Schumacker, R. E., Mount, R.E., Monahan, M. P. (2002). Factors affecting multiple regression and discriminant analysis with a dichotomous dependent variable: Prediction, explanation and classification. *Multiple Linear Regression Viewpoints*, 28(2), 23-39.

Sibbertsen, P. (2001). S estimation in the linear regression model with long memory error terms under trend. *Journal of Time Series Analysis*, 22(3), 353-363.

Staudte, R. G. and Sheather, S. J. (1990). *Robust estimation & testing*. (First editions). Canada: John Wiley & Sons, 1.

Stigler, S. M. (1973). Simon Newcomb, Perscy Daniell, and the history of robust estimation 1885-1920. *Journal of the American Statistical Association*, 68(344), 872-879.

Susanti, Y., Pratiwi, H., Sulistijowati, S. H. and Liana, T. (2014). Mestimation, S estimation and MM estimation in robust regression. *International Journal of Pure and Applied Mathematics*, 91(3), 349-360.

Tukey, J. W. (1960). *A survey of sampling from contaminated distributions*. (First editions). New Jersey: Stanford University Press, 448-485.

Tukey, J. W. (1962). The future of data analysis. *Annals of Mathematical Statistics*, 33(1), 1–67.

Tyler, D. E. (1999). S Estimators, In S. Kotz, C. B. Read, D. L. Banks (Eds.), *Encyclopedia of Statistical Sciences***,** New York: John Wiley and Sons, Inc., 659-662.

Wallach, D., Jones, J. W., Makowski, D. and Brun, F. (2019). *Working with dynamic crop models*. (First editions). London: Elsevier, 161.

Welsch, R. E. (1980). Regression sensitivity analysis and bounded influence estimation. *Evaluation of Econometric Models,* 153-167.

Welsch, R. E. (1982). Influence functions and regression diagnostics. *Modern Data Analysis*, 149-165.

Yamane, T. (1973). *Statistics: An introductory analysis*. (Third edition). New York: Harper and Row: 245.

Yohai, V. J. and Zamar, R. H. (1988). High breakdown-point estimates of regression by means of the minimization of an efficient scale. *Journal of the American Statistical Association*, 83(402), 406-413.

Zhang, Z. (1997). Parameter estimation techniques: A tutorial with application to conic fitting. *Image and Vision Computing*, 15(1), 59-76.

Zuo, Y. (2001). Some quantitative relationships between two types of finite sample breakdown point. *Statitics & Probability Letters*, 51(4), 369-375.

ÖZGEÇMİŞ

Kişisel Bilgiler

Soyadı, adı : KARAK OCA, Fulya

Uyruğu : T.C.

Doğum tarihi ve yeri : 14.05.1984

Medeni hali : Evli

Telefon : 05079294761

e-mail : fulyakarakocay@gmail.com

Eğitim

Derece	Eğitim Birimi	Mezuniyet Tarihi
Yüksek lisans	Gazi Üniversitesi / İstatistik	Devam ediyor
Lisans	Gazi Üniversitesi / İstatistik	2007
Lise	Rauf Denktaş Lisesi	2002

İş Deneyimi

Yıl	Yer	Görev
2009-Halen	Kültür ve Turizm Bakanlığı	Programcı
2009	Ankara Doğalgaz Üretim A.Ş.	Programcı

Yabancı Dil

İngilizce

Yayınlar

Karakoca, F., Ekiz, M. ve Ekiz, O. U. (2018). *Sağlam M ve S tahmin edicilere dayalı etkili uzaklık.* 3. Uluslararası Bilimsel Araştırmalar Kongresi, Nevşehir.

Hobiler

Yüzme, Kitap okumak.

GAZİ GELECEKTİR...