M-REGRESYON VE İMKB100 ENDEKSİ ÜZERİNDE BİR UYGULAMA DENEMESİ

Alper BEKKİ

Yüksek Lisans Tezi

Fen Bilimleri Enstitüsü
İstatistik Anabilim Dalı

Eylül - 2001

ÖZET
Yüksek Lisans Tezi

M-REGRESYON VE İMKB100 ENDEKSİ ÜZERİNDE BİR UYGULAMA DENEMESİ

Alper BEKKİ

Anadolu Üniversitesi
Fen Bilimleri Enstitüsü
İstatistik Anabilim Dalı

Danışman: Yrd. Doç. Dr. Atilla ASLANARGUN
2001, 58 Sayfa

Regresyon analizinde yaygın kullanıma sahip En Küçük Kareler tekniğinin parametre tahmini için kullanımında kabul edilmesi gereken bazı varsayımlar vardır. Fakat günümüz koşullarında elde edilen veri setleri için istatistiksel modelin bu varsayımları sağlanmayabilir. Rassal hatalar anakütlesinin normallik varsayımı geçersiz olduğunda En Küçük Kareler tekniği ile elde edilecek tahminler yanlı sonuçlar verebilir. Özellikle aykırı değerlerin bulunduğu veri setlerinde hatalara ilişkin dağılım normal dağılıma uymamaktadır. Bu durum için Huber, alternatif olarak M regresyon fikrini 1964 yılında ortaya atmış ve o zamandan beri bilgisayar teknolojisindeki gelişmelere paralel olarak geniş bir uygulama alanına sahip olmuştur. Huber'ın M tahmincileri, istatistikte Robust Teknikler adı verilen bir sınıflama içerisinde yer almaktadır. Robust tekniklerde temel hedef, veri setinin genel yapısındaki bozuklukların veri setini temsil edecek istatistikler üzerindeki olumsuz etkilerini en aza indirgemektir.

Bu çalışmada, rassal hatalar anakütlesinin dağılımının normal olmadığı durumlarda alternatif bir teknik olan Huber'ın M-Regresyon tekniği için teorik detaylar ve algoritmalar ayrıntılı bir biçimde incelenmiştir. Tezin uygulama aşamasında İMKB100 endeks değerini etkileyen 6 değişken (Mevduat Faiz Oranı, Hazine Bonosu Faiz Oranı, Ortalama Dolar Fiyatı, Külçe Altın Satış Fiyatı, Tüketici Fiyat Endeksi ve M2Y) için Huber'ın M-Regresyon tekniği kullanılarak bir regresyon modeli elde edilmiştir.

Anahtar Kelimeler: Robust, M Regresyon, Aykırı Değer, İMKB100

ABSTRACT
Master of Science Thesis

M-REGRESSION AND A TRIAL APPLICATION ON THE ISE100 INDEX

Alper BEKKİ

Anadolu University
Graduate School of Natural and Applied Sciences
Statistics Program

Supervisor: Assist. Prof. Dr. Atilla ASLANARGUN
2001, 58 Pages

There are some assumptions to be accepted when the Least Squares Technique, widely used in regression analysis for parameter estimation. But these assumptions of the statistical model may not hold for today's data sets. If the assumption of normality for the population of random errors is invalid, the estimates that can be obtained by Least Squares Technique may be biased. The distribution of residuals may not follow a normal distribution in the presence of outliers in data sets. Huber has introduced the idea of M regression in 1964 as an alternative for this problem and since then it has gained wide spread usage parallel to the developments in computer technology. Huber's M estimators are part of what is called robust techniques in statistical terminology. The main aim of Robust Techniques is to reduce the negative effects on the parameter estimations due to defects in data set.

In this study, theoretical details and the algorithm for Huber's M Regression Technique, an alternative technique for situations when the distribution of the residuals is not normal, is investigated in detail. In the application part of the thesis, a regression model with 6 independent variables (Interest Rate on Deposits, Interest Rate on Treasury Bills, Average US Dollar Price, Bullion Gold Selling Price, Consumer Price Index and M2Y) order the dependent variable the ISE100 index value has been obtained by using Huber's M Regression Technique.

Keywords: Robust, M Regression, Outlier, ISE100

TEŞEKKÜR

Bu çalışmada, yaptığı katkı ve eleştirilerinden dolayı danışmanım Sayın Yrd. Doç. Dr. Atilla ASLANARGUN'a (Anadolu Üniversitesi), çalışmanın ortaya çıkmasında göstermiş oldukları özveri ve değerli katkılarından dolayı Sayın Prof. Dr. İlyas ŞIKLAR'a (Anadolu Üniversitesi) ve Sayın Yrd. Doç. Dr. Fikret ER'e (Anadolu Üniversitesi), çalışma boyunca manevi desteklerini esirgemeyen bölümdeki değerli hocalarıma ve arkadaşlarıma, göstermiş olduğu destek ve anlayıştan dolayı sevgili arkadaşım Hülya TAN'a ve sevgili aileme teşekkür ederim.

İÇİNDEKİLER

ŞEKİLLER DİZİNİ

ÇİZELGELER DİZİNİ

1. GİRİŞ

Yaklaşık 200 yıl önce başlayan en küçük kareler kullanımı birçok bilimsel araştırmada halen tüm sürati ile devam etmektedir. Geçen bu 200 yıllık süre, bu teknikte herhangi bir değişime yol açmamasına rağmen tekniğin uygulandığı veri setlerinde büyük bir değişim meydana gelmiştir. Günümüz dünyasında herhangi bir eğilimin uzun süre devam etmesi maalesef beklenmemektedir. Çoğu zaman ilgilenilen veri seti çeşitli sebeplerden ortaya çıkan ve serinin genel gidişatına uymayan gözlem değerlerini de içermektedir. Ayrıca en küçük kareler analizi temelde birçok varsayımı da beraberinde getirmektedir. Bu varsayımların gerçekleşmemesi durumunda ise en küçük kareler bizi yanıltıcı sonuçlara götürebilecektir.

En küçük kareler analizi hesaplama kolaylığı bakımından en çok tercih edilen regresyon tekniklerinden bir tanesidir. Bu hesaplama kolaylığı, en küçük kareler tekniğinin ortaya çıktığı dönemden itibaren geniş bir kullanım alanına sahip olmasını sağlamıştır. En küçük mutlak sapmalar tekniği, en küçük kareler tekniğinden yaklaşık olarak 50 yıl önce ortaya atılan bir tekniktir. Fakat hesaplama zorluklarından dolayı bu teknik gözardı edilmiş ve en küçük mutlak sapmalar tekniğine ilişkin gelişmeler ne yazık ki son 40 yıl içerisinde meydana gelmiştir. En küçük karelerin varsayımlarının sağlanmaması halinde parametre tahminlerinin yanlılığı 20. yüzyıl istatistikçilerini en çok uğraştıran konu haline gelmiştir.

Bir istatistiksel model için oluşturulan varsayımların doğru olmadığı durumlarda dahi uygun sonuçlar verebilen istatistiksel teknikler robust teknikler adı verilen bir sınıflama içerisinde yer alırlar. Eğer verimizin normal doğrusal regresyon modeline uyduğu varsayılırsa en küçük kareler tahminleri gayet iyi sonuçlar verecektir; fakat anakütle rassal hataları için normallik varsayımı geçersiz ise bu tahminler yanlı olabilir. M regresyon özellikle bu varsayımın gerçekleşmediği olaylarda robust parametre tahminlerinin yapılabilmesi için geliştirilmiştir. Peter Huber M tahmini fikrini 1964 yılında ilk olarak ortaya atmıştır. Huber'ın kullandığı bu M tahminlerinde özel bir fonksiyon yer almaktadır. Bu fonksiyonun özelliği artık kareler toplamları ve mutlak artık

sapmalar en küçüklemeleri arasında bir denge oluşturmasıdır. En küçük mutlak sapmalar tahminleri aykırı değerlere karşı en küçük kareler kadar duyarlı değildir. Yine de unutulmamalıdır ki gözlem değerleri arasında aykırı değer yoksa ve artıklar normal dağılıma sahipse en küçük kareler tahminleri daha iyi sonuçlar verebilir.

Bu tez çalışmasında, Huber tarafından geliştirilen M regresyon tekniği ele alınmıştır. İkinci bölümde, çeşitli regresyon teknikleri teorik olarak incelenmiştir. Üçüncü bölüm, tamamıyla Huber'ın M tahmincilerine ayrılmıştır. Ayrıca bu bölümde parametre tahminlerinin yapılabilmesi için gerekli algoritmalar da detaylı bir şekilde incelenmiştir. Son bölümde ise, teorik olarak incelenen bu tekniklerin gerçek hayattaki geçerliliklerini görebilmek amacıyla ekonomik bir veri seti üzerinde çalışılmıştır. Bu bölümün incelenmesiyle de kolaylıkla görülebileceği gibi artıklar için normallik varsayımı aykırı değerlerin varlığından dolayı büyük bir ölçüde bozulmaktadır. Dolayısıyla rassal hatalar anakütlesinin normallik varsayımı geçersiz olduğunda en küçük kareler tekniği kullanılarak elde edilecek tahminler yanlı sonuçlar verebileceği için bu türden veri setlerinin incelenmesinde Huber tarafından geliştirilen M regresyon tekniğinin kullanılması daha doğru olabilmektedir.

2. DOĞRUSAL REGRESYON KAVRAMI

2.1. Tarihsel Gelişmeler

Regresyonun genel tarihine baktığımız zaman ilk karşımıza çıkan isim genetik bilimci Francis Galton'dur. Galton 1877 yılında İngiltere'deki "kalıtımın tipik kanunları" adlı sunusunda ilk defa regresyon kavramına değinmiş ve bu dönemde uzun boylu babaların oğullarının da uzun boylu, kısa boylu babaların oğullarının da kısa boylu olma eğiliminde olduğunu gözlemlemiştir. Bir çok araştırmacı regresyon kelimesi ile karşılaştığında, bu analizin en popüler tekniği olan ve 200 yıl önce keşfedilmiş olan en küçük kareler tekniğini düşünmektedir. Oysa ki regresyon analizi için bugün "parametrik olmayan", "robust" regresyon analizi gibi sınıflamalar yapılmaktadır. Günümüzde teknolojinin bize sunduğu imkanlar sayesinde ve bilgisayarın tüm bilim dallarında etkin bir şekilde kullanımı sonucunda, her alanda elde edilen veri türü ve yapısındaki büyük değişikliklere paralel olarak, regresyonda uygulanacak analiz ve tekniklerde de değişimler ve gelişimler sağlanmıştır. Bugün etkin bir şekilde kullanılan regresyon tekniklerini; En Küçük Kareler, En Küçük Mutlak Sapmalar, Huber'in M Tahmincileri, Parametrik Olmayan, Bayesgil, Ridge, En Küçük Medyan Kareler, ... olarak sıralayabiliriz [1,2].

Genel anlamı ile regresyon bir şeyi başka bir şeye bağlamaktır. Daha açık bir ifadeyle, bir değişkenin bir veya birden fazla değişkenle olan ilişkisinin varlığını ve varolan ilişkinin matematiksel modelini kurma olarak verilebilir. İstatistiksel anlamda regresyon değişkenler arasındaki ilişkinin veya ele alınan değişkenlere ilişkin karşılıklı değişimlerin matematiksel bir fonksiyonla ifadesidir. Bir regresyon modeli değişkenler arasındaki nedensellik ilişkilerini göstermede kullanılacağı gibi değişkenlerin karşılıklı değişimlerini de tanımlamada kullanılabilir [3,4].

Oluşturulacak doğrusal bir regresyon denklemi bağımlı değişken (Y) ve bağımlı değişkeni %100 açıkladığı düşünülen d tane bağımsız değişkene $(X_j,\ j=1,2,...,d)$ dayalı olarak kurulmalıdır. Buna göre regresyon modelinin matematiksel ifadesi $Y=f(X_1,X_2,...,X_d)$ şeklinde olacaktır. Ancak gerçekte böyle bir model oluşturmak veya bağımlı değişkeni bütünüyle açıklayacak

bağımsız değişkenlerin hepsine modelde yer vermek mümkün değildir. Dolayısıyla oluşturulacak modelde bağımlı değişkeni %100 açıklayacak bağımsız değişkenlerin hepsine yer vermek mümkün olamayacağı için bu eşitlik $Y = f(X_1, X_2, \ldots, X_j) + \varepsilon \quad (j = 1, 2, \ldots, p)$ ve $p < d$ şeklinde oluşturulduğu düşünülür. Burada f regresyon fonksiyonu ve ε rassal hata olarak nitelendirilir. Böylece oluşacak doğrusal regresyon modeli,

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots\ldots + \beta_j X_j + \varepsilon \quad j = 1, 2, \ldots, p$$

şeklinde değerleri belirli olmayan, bilinmeyen parametreler olan ve regresyon katsayıları veya regresyon parametreleri olarak isimlendirilen β_j'lerden ve sıfır ortalamayla dağıldığı varsayılan rassal değişken ε hata teriminden meydana gelir.

Çizelge 2.1. Regresyon Analizi İçin Örnek Veri Setinin Genel Gösterimi

Gözlem No	Y	X_1	X_2	...	X_p
1	y_1	x_{11}	x_{12}		x_{1p}
2	y_2	x_{21}	x_{22}		x_{2p}
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
n	y_n	x_{n1}	x_{n2}	...	x_{np}

Doğrusal regresyon modeli, Çizelge 2.1.'de genel olarak gösterilen Y bağımlı değişkeni ve X_1, X_2, ..., X_p, p tane açıklayıcı değişken için belirlenen n gözleme göre;

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \ldots + \hat{\beta}_j x_{ij} + e_i \quad i = 1, 2, \ldots, n \quad j = 1, 2, \ldots, p$$

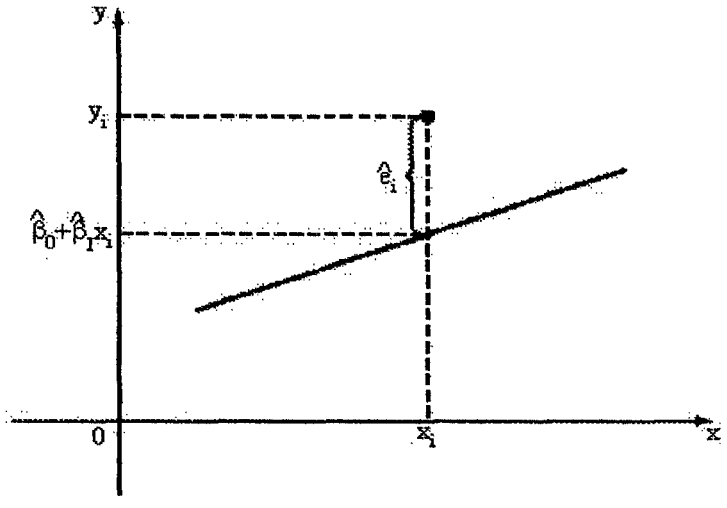
şeklinde ifade edebiliriz. Ayrıca modelde e_1, e_2, ..., e_n rassal hatalarının sıfır ortalamaya sahip bir anakütleden çekilmiş birbirinden bağımsız rassal örnekler olduğu varsayımı da bulunmaktadır.

Modelde β_j'ler bilinmeyen parametreler oldukları için bu değerler yerine bağımlı ve bağımsız değişkenlere ilişkin gözlem değerleri kullanılarak regresyon parametrelerine ilişkin tahmin değerleri olan $\hat{\beta}_j$'lar modelde kullanılarak,

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + ... + \hat{\beta}_j X_j \quad j = 1, 2, ..., p$$

regresyon denklemi oluşturulur.

Gözlem değerlerinin tahmin edilen doğruya olan uzaklıkları sapma veya hata olarak nitelendirilir. $Y = \beta_0 + \beta_1 X + \varepsilon$ şeklindeki bir basit doğrusal regresyon denklemi için tahmin edilecek regresyon denklemi $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$ şeklinde ifade edilebilir. Tahmin edilen bu regresyon denklemine verinin ne kadar iyi uyduğuna karar verebilmek için $\hat{e}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$ $i = 1,2,...,n$ hatalarının büyüklüklerine bakmak gerekmektedir. Veri setindeki bir (x_i, y_i) noktası için $\hat{e}_i$ artık değerinin grafik üzerinde gösterimi Şekil 2.1.'de verilmiştir.



Şekil 2.1. Veri Setindeki Bir (x_i, y_i) Noktası İçin $\hat{e}_i$ Artık Değeri

2.2. Parametre Tahmin Teknikleri

Bir regresyon denklemi oluşturmanın ve bu denklemin parametre tahminlerinin belirlenmesinin amacını üç başlık altında toplamak mümkündür. İlki, değişkenler için elde edilmiş verilerdeki ana kalıbın ortaya çıkarılması ve bu değişkenler arasındaki ilişkinin yaklaşık olarak tanımlanmasını sağlamaktır. İkincisi, bağımlı değişkenle bağımsız değişkenler arasında kurulan ilişkiye dayanarak bağımsız değişkenler için daha sonra elde edilecek gözlem verilerinden hareketle bağımlı değişkenin alacağı değeri tahmin edebilmektir. Üçüncüsü ise,

bağımlı değişkenle bağımsız değişkenler arasındaki bu ilişkinin önemliliğini tahminlenen regresyon denklemi sayesinde ortaya koymaktır. Bir regresyon denklemi tahminlenirken, oluşturulacak modelin dolayısıyla modelin ortaya koyduğu ilişkinin hayattaki gerçeklere ters düşmemesi gerekmektedir. Aynı zamanda kurulacak model alternatifleri arasından bu ilişkiyi en iyi şekilde ifade edecek en sade modelin seçilmesi gerekmektedir [5].

2.2.1. En Küçük Kareler Tekniği

En küçük kareler tekniği günümüzde en yaygın kullanılan regresyon tekniğidir. Bu teknik 1795 yılı civarlarında Alman Carl Frederich Gauss ve 1805 yılı civarlarında Fransız Adrien Marie Legendre tarafından birbirlerinden habersiz olarak geliştirdikleri bir tekniktir [6]. Tekniğe ilişkin ilk uygulamalar astronomik ve yerbilim verilerinin analizinde kullanılmıştır. En küçük kareler regresyon ilk olarak 1805 yılında Legendre tarafından yayınlanan bir kitapta kuyruklu yıldızların yörüngelerinin hesaplanmasında kullanılmıştır.

En küçük kareler, temel olarak bir en küçükleme problemidir. Bu teknikte amaç, örnek veri setindeki gözlem değerlerinin çizilecek olan bir regresyon doğrusuna olan uzaklıklarının kareleri toplamını minimum yapan bir doğru denklemini bulmaktır.

En küçük kareler tekniğinin uygulanabilirliği bazı varsayımların sağlanması durumunda gerçekleşebilmektedir. Daha açık bir ifade ile ilgilenilen regresyon modeline dahil edilen değişkenler arasındaki ilişkiyi ifade edecek olan parametre kestirimlerini yaparken şu varsayımların sağlanmış olması gerekmektedir [3,7-10] :

1. Hata terimi rassal bir değişkendir. Hata terimi değerleri tamamıyla şansa bağlıdır. Yani hata terimi stokastik bir değişkendir ve hatalı matematiksel model seçimi, modele dahil edilmeyen bağımsız değişken ve değişkenlerdeki ölçüm hataları gibi hataları ifade etmektedir. Stokastik bir değişken olması veya belirli olasılıklarla pozitif, negatif ve sıfır değerlerini alabilmesinden ötürü hata teriminin beklenen değerinin yani ortalamasının sıfır olduğu varsayımı yapılır.

2. Hata teriminin dağılımı, ortalaması sıfır olan ve sabit varyansa sahip normal dağılımdır. Parametre tahminlerinin ve bunlara ilişkin testlerin

yapılabilmesi için bağımsız değişkenin her bir değerine ilişkin hata değerleri kendi ortalamaları etrafında çan eğrisi şeklinde simetrik bir dağılım gösterir.

3. Hata terimi değerleri arasında bir ilişki yoktur. Hata değerlerinin ardışık değerleri birbirinden bağımsızdır yani otokorelasyon yoktur. Buna göre birbirini takip eden iki hata teriminin kovaryansı sıfıra eşit olur. Dolayısıyla bağımlı değişkenler de birbirinden istatistiksel olarak bağımsızdırlar ve her biri diğerinden bağımsız olarak elde edilmelidir.

4. Hata terimi varyansı bağımsız değişkenin her bir değeri için aynıdır. Hata terimi varyansı bağımsız değişkenin her bir değerine göre değişmemekte ve sabit kalmaktadır.

5. Bağımsız değişken ile hata terimi arasında ilişki yoktur. Yani kovaryansları sıfıra eşittir. Dolayısıyla bağımsız değişken stokastik değildir ve değerleri sabittir. Bağımsız değişkene ilişkin değerler daha önceden bilinen değerlerdir ve anakütleden çekilecek herbir örneklem için bu değerler değişmez.

6. Tahmin edilecek regresyon modeli, model belirleme hatası taşımamaktadır. Modele bazı değişkenlerin dahil edilmemesi, model fonksiyonunun yanlış seçimi, modeldeki değişkenler konusunda hatalı varsayımlar durumunda tahmin edilecek fonksiyon güvenilir olmayacak ve model belirlemesi hatalı olacaktır.

7. Çoklu doğrusal regresyon modelleri için bağımsız değişkenler arasında ilişki yoktur. Birden fazla bağımsız değişkenin yer aldığı modellerde bu değişkenler arasındaki ilişki çoklu doğrusal bağıntı olarak nitelendirilir. Değişkenler arasındaki ilişkinin derecesi korelasyon katsayısı ile ifade edilir. Korelasyon katsayısı 0 ile 1 arasında değer alır. Korelasyon katsayısının 0 değerini alması değişkenler arasında ilişkinin olmadığını, 1 değerini alması ise değişkenler arasında tam bir ilişkinin olduğunu gösterir.

En küçük kareler analizinin uygulanmasında yukarıda belirtilen 7 varsayımın sağlandığı kabul edilerek analizin yapılması gerekir. Bu varsayımların sağlanmadığı durumlarda dahi matematiksel olarak teknik uygulanabilir ancak bu durumda elde edilen sonuçlar yanlı olacaktır.

Regresyon analizinde kullanılan en küçük kareler tekniğinde hataların toplam büyüklüğü $\sum \hat{e}_i^2$ olarak ifade edilir. β_0 ve β_1 parametrelerinin en küçük

kareler tahminleri de artıkların kareleri toplamı olan $\sum \hat{e}_i^2$ değerini en küçük yapacak $\hat{\beta}_0$ ve $\hat{\beta}_1$ değeri olarak tanımlanır. Bu durumda en küçüklenen fonksiyon;

$$\sum_{i=1}^{n}\left(y_i-\left(\hat{\beta}_0+\hat{\beta}_1 x_i\right)\right)^2=\sum_{i=1}^{n}\hat{e}_i^2=q$$

şeklinde olacaktır [6]. Buradan regresyon katsayılarının en küçük kareler tahminlerini elde edebilmek için q fonksiyonunun $\hat{\beta}_0$ ve $\hat{\beta}_1$'e göre kısmi türevlerinin alınarak sıfıra eşitlenmesi gerekmektedir. Buna göre;

$$\frac{\partial q}{\partial \hat{\beta}_0}=\sum_{i=1}^{n}(-2)\left[y_i-\left(\hat{\beta}_0+\hat{\beta}_1 x_i\right)\right]=0$$

ve

$$\frac{\partial q}{\partial \hat{\beta}_1}=\sum_{i=1}^{n}(-2)x_i\left[y_i-\left(\hat{\beta}_0+\hat{\beta}_1 x_i\right)\right]=0$$

eşitlikleri ve buradan da;

$$\sum_{i=1}^{n} y_i=\hat{\beta}_0 n+\hat{\beta}_1\sum_{i=1}^{n} x_i$$

ve

$$\sum_{i=1}^{n} x_i y_i=\hat{\beta}_0\sum_{i=1}^{n} x_i+\hat{\beta}_1\sum_{i=1}^{n} x_i^2$$

normal denklemler olarak adlandırılan eşitlikler elde edilir. Elde edilen bu eşitlik sisteminin determinantlar yardımıyla veya değişken eleme yöntemi kullanılarak çözülmesi sonucunda $\hat{\beta}_1$ ve $\hat{\beta}_0$ formülleri;

$$\hat{\beta}_1=\frac{n\left(\sum_{i=1}^{n} x_i y_i\right)-\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n\left(\sum_{i=1}^{n} x_i^2\right)-\left(\sum_{i=1}^{n} x_i\right)^2}=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} \qquad (2.1.1)$$

$$\hat{\beta}_0=\frac{\sum_{i=1}^{n} y_i-\hat{\beta}\sum_{i=1}^{n} x_i}{n}=\bar{y}-\hat{\beta}_1\bar{x} \qquad (2.1.2)$$

şeklinde elde edilir. Burada $\bar{x}$ ve $\bar{y}$ sırasıyla x_i ve y_i değerlerinin ortalamalarını ifade etmektedir. $\hat{\beta}_1$ için elde edilen (2.1.1) eşitliğini daha basit bir şekilde ifade edebilmek için, varyans-kovaryans matrisinin elemanlarından yararlanabiliriz. Varyans-kovaryans matrisi elemanlarını;

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2$$

$$S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right)^2$$

$$S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)$$

şeklinde gösterirsek $\hat{\beta}_1$'ya ilişkin formülü;

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$$

şeklinde de ifade edebiliriz.

$\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}$ formülü, tahmin edilen regresyon doğrusunun, veri noktaları bulutunun merkezi olarak kabul edilen $(\bar{x}, \bar{y})$ noktasından geçeceğini ifade etmektedir. Bu durum, veri setinin genel yapısına uygun olacağı beklenen bir doğru için iyi bir özelliktir.

Ayrıca $\hat{\beta}_1$;

$$\hat{\beta}_1 = \sum w_i \left(\frac{y_i - \bar{y}}{x_i - \bar{x}}\right), \qquad w_i = \frac{(x_i - \bar{x})^2}{\sum (x_i - \bar{x})^2}$$

şeklinde de yazılabilir [6]. Burada $(y_i - \bar{y})/(x_i - \bar{x})$, merkezi nokta $(\bar{x}, \bar{y})$ ile veri noktası (x_i, y_i) arasındaki doğrunun eğimini ifade etmektedir. Böylece tahmin edilen regresyon doğrusunun $\hat{\beta}_1$ eğimi, bu eğimlerin bir tür ortalaması olmaktadır. Daha kesin bir ifadeyle bu bir ağırlıklı ortalamadır. w_i ağırlıkları negatif değildir ve toplamları 1'e eşittir. (Alışılagelen bir ortalama, bütün ağırlıkların 1/n 'e eşit olduğu durumdur.) Her bir veri noktası, kendisinin merkezden x-uzaklığının karesiyle orantılı olarak ağırlıklandırılır. Bu da veri merkezinden uzaktaki veri noktalarının, regresyon doğrusunun eğiminin tahmininde büyük bir etkisi

olduğunu göstermektedir. Dolayısıyla veri setinde bulunabilecek aykırı değerlerin tahmin edilecek doğru üzerindeki etkileri çok daha fazla olacaktır.

$$Y = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots\ldots + \beta_p X_{ij} + \varepsilon_i \quad (i = 1,2,\ldots,n \quad j = 1, 2, \ldots, p)$$

şeklinde oluşturulacak N gözlem birimi için, bir bağımlı ve p bağımsız değişkenden oluşan çoklu doğrusal regresyon denkleminin parametre tahminlerini yapabilmek için modelin matrisler biçiminde ifade edilerek çözümü daha basit bir yoldur. Modelin matris formunda ifadesi;

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & X_{21} & \cdots & X_{1p} \\ 1 & X_{12} & X_{22} & \cdots & X_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{1n} & X_{2n} & \cdots & X_{np} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \varepsilon_0 \\ \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

$$Y_{(n\times 1)} = X_{(n\times p)} \beta_{(p\times 1)} + \varepsilon_{(n\times 1)}$$

şeklinde olur [10]. Burada;

Y : Bağımlı değişkene ait nx1 boyutlu sütun vektörünü,

X : p bağımsız değişken için n adet gözlem değerini ifade eden nxp boyutlu matrisini (veri matrisi),

β : p tane bilinmeyen parametreleri ifade eden px1 boyutlu sütun vektörünü,

ε : n tane hata terimini ifade eden nx1 boyutlu sütun vektörünü ifade eder.

$Y = X\hat{\beta} + e$ şeklinde matris formunda gösterilen çoklu doğrusal regresyon denkleminde $\hat{\beta}$ parametre tahmin vektörü;

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

eşitliğinin çözülmesi sonucunda elde edilir.

2.2.2. Tartılı En Küçük Kareler Tekniği

Tartılı en küçük kareler, regresyon analizinde kullanılan bazı gözlemlerin diğerlerine göre daha az güvenilir olmasından yola çıkılarak ortaya atılan ve her bir gözleme farklı tartılar verilmek suretiyle bu tür gözlemlerin model üzerindeki ters etkilerini yok etmeyi amaçlayan bir tekniktir. Bunun anlamı, bu tür gözlem değerlerine ilişkin varyans değerlerinin farklı olmasıdır. Yani hata terimlerinin tekil olmayan varyans-kovaryans matrisindeki köşegen elemanlarının birbirine

eşit olmaması dolayısıyla varyans-kovaryans matrisinin $I\sigma^2$ şeklinde I birim matrisiyle çarpımı şeklinde ifade edilememesi anlamına gelmektedir. Bu durumda en küçük kareler minimizasyonu varsayımlarından bir tanesi hata terimlerinin varyanslarının sabit olması, değişmemesi varsayımının sağlanmaması durumuyla karşı karşıya kalınır. Varyansların birbirlerinden farklı olmaları en küçük kareler kestiricilerinin etkinliklerini yitirmelerine ve güvenilir olmaktan çıkmalarına neden olmaktadır. Böyle bir durumda varyansların eşitlenebilmesi için bu durumun ortaya çıkmasına neden olan değişken veya değişkenlere çeşitli dönüşümler uygulanması gerekmektedir. Bu dönüşümleri değişken gözlem değerlerine verilen tartılar olarak nitelendirebiliriz. Aslında bu dönüşümün amacı varyansların farklılığına neden olan değişkeni ilgili koşulları sağlayan ve en küçük kareler minimizasyonunun uygulanabileceği başka bir değişken şeklinde ifade edebilmektir. En küçük kareler tekniğinin genel işleyişinde de hata kareler toplamını minimize ederken tüm hatalara tartı değeri 1 olan eşit tartılar verilmektedir [7,8,11,12].

Tartılı en küçük kareler ile sıradan en küçük kareler regresyon işleyişindeki değişimleri şu şekilde ifade edebiliriz:

$Y = X\beta + \varepsilon$ (β, $p \times 1$ boyutlu katsayılar vektörünü ifade etmek üzere),

şeklindeki matris formu şeklinde ifade edilmiş doğrusal bir model için;

$$E(\varepsilon) = 0, \quad V(\varepsilon) = V\sigma^2 \quad \text{ve} \quad \varepsilon \sim N(0, V\sigma^2)$$

şeklindedir. Burada;

$$P'P = PP = P^2 = V$$

olan tekil olmayan ve simetrik bir P matrisi bulunabilir. Bundan hareketle,

$f = P^{-1}\varepsilon$, dolayısıyla $E(f) = 0$

yazabiliriz ki burada f, $E(f) = 0$ ve dolayısıyla $E(ff') = V(f)$ olan rassal değişken vektörüdür. $n \times n$ boyutlu ff' kare matrisindeki her bir değerin beklenen değerleri ayrı ayrı hesaplanmıştır. Buna bağlı olarak;

$$\begin{aligned} V(f) = E(ff') &= E(P^{-1}\varepsilon\varepsilon'P^{-1}), \qquad (P^{-1})' = P^{-1} \text{ 'den dolayı} \\ &= P^{-1}E(\varepsilon\varepsilon')P^{-1} \\ &= P^{-1}PPP^{-1}\sigma^2 \\ &= I\sigma^2 \end{aligned}$$

ε'un dağılımı normal olduğundan ve f de ε'un doğrusal kombinasyonlarından meydana geldiği için $f \sim N(0, I\sigma^2)$ şeklinde normal dağılıma sahiptir.

Bunlara dayanarak elde edilecek yeni model başlangıç modelinin her iki tarafının da P^{-1} matrisi ile ön çarpımı şeklinde elde edilecektir. Yeni oluşan modeli;

$$P^{-1}Y = P^{-1}X\beta + P^{-1}\varepsilon \quad \text{veya} \quad Z = Q\beta + f$$

şeklinde ifade edebiliriz. Oluşan bu yeni modelde artıkların dağılımları normal olduğu için ve oluşturulacak varyans-kovaryans matrisi $I\sigma^2$ şeklinde I birim matrisiyle çarpımı şeklinde ifade edilebildiği elde edilen yeni modele en küçük kareler minimizasyonunu rahatlıkla uygulanabilir [7,13,14].

2.2.3. En Küçük Mutlak Sapma (L_1 Regresyon) Tekniği

En küçük mutlak sapma tekniği, en küçük kareler tekniğinden yaklaşık 50 yıl önce 1757'de Roger Joseph Boscovich tarafından ortaya atılmıştır. Boscovich, bu tekniği dünyanın şeklini tahmin edebilmek için kullanmıştır. Teknik 30 yıl sonra Pierre Simon Laplace tarafından güncellenerek kullanılmaya başlanmış ancak en küçük kareler tekniğinin gölgesinde kalmıştır. En küçük kareler tekniğinin popülaritesi, hesaplamada sağladığı kolaylık ve teorik yapısının Gauss ve Laplace tarafından desteklenmesi sonucunda oldukça artmıştır [6].

$Y = \beta_0 + \beta_1 X_1 + \varepsilon$ şeklinde oluşturulacak bir basit doğrusal regresyon denkleminde $\hat{\beta}_0$ ve $\hat{\beta}_1$ katsayı tahmin değerleri bulunurken en küçük kareler tekniğinde hata kareler toplamı olan $\sum e_i^2$ $(i = 1,2,3,...,n)$ değerini minimum yapan katsayı değerleri seçilirken, en küçük mutlak sapma tekniğinde ise hata terimlerinin mutlak değerlerinin toplamı olan $\sum |e_i|$ $(i = 1,2,3,...,n)$ değerini minimum yapan katsayı değerleri seçilir.

Tekniğin amacı, mutlak sapmalar toplamını minimum yapan regresyon doğrusunun seçimine dayanır. Bu doğru ise herhangi bir (x_0, y_0) veri noktasından geçen doğrular arasından en iyi sonucu veren doğru olacaktır. En küçük mutlak sapma tekniği için regresyon doğrusu belirlenirken seçilen (x_0, y_0) veri noktasına karşı (x_i, y_i) $i > 0$ şeklindeki diğer veri noktalarından geçen doğrular belirlenir

ve oluşan doğrular arasından tahminlenen mutlak sapmalar toplamı $\sum|\hat{e}_i|$ $(i=1,2,3,...,n)$ değerini minimum yapan doğru ilk aşama için en iyi doğru olarak belirlenir. Seçilen doğrunun (x_0,y_0) noktası haricindeki diğer noktası bir sonraki adım için asıl nokta olan (x_0,y_0) noktası olur ve bu noktaya karşılık diğer veri noktalarından geçen doğrular belirlenerek yine bu doğrular arasından tahminlenen mutlak sapmalar toplamı $\sum|\hat{e}_i|$ $(i=1,2,3,...,n)$ değerini minimum yapan doğru en iyi doğru olarak seçilir. Bu işlemler elde edilen son iki doğruya ilişkin $\hat{\beta}_0$ ve $\hat{\beta}_1$ tahmin değerleri birbirine çok yakın değerler alıncaya kadar devam ettirilir. Son adımda elde edilen doğru en küçük mutlak sapma regresyon tahmin denklemini verir [6,13].

Yukarıda ifade edilen algoritmayı daha detaylı bir şekilde inceleyecek olursak; herhangi bir (x_0,y_0) veri noktasından her bir (x_i,y_i); $i>0$ veri noktası için çizilecek doğruların eğimleri $(y_i-y_0)/(x_i-x_0)$ $i>0$ şeklinde hesaplanır. Elde edilen eğim değerleri,

$$(y_1-y_0)/(x_1-x_0)\le(y_2-y_0)/(x_2-x_0)\le\ldots\le(y_n-y_0)/(x_n-x_0)$$

şeklinde küçükten büyüğe doğru sıralanır ve

$$T=\sum|x_i-x_0|$$

olmak üzere;

$$|x_1-x_0|+|x_2-x_0|+\ldots+|x_{k-1}-x_0|<\frac{1}{2}T$$

$$|x_1-x_0|+|x_2-x_0|+\ldots+|x_{k-1}-x_0|+|x_k-x_0|>\frac{1}{2}T$$

koşullarını sağlayan k noktası bulunur. Buna göre (x_0,y_0) noktasından geçen en iyi doğru;

$$\beta_1^*=\frac{y_k-y_0}{x_k-x_0} \text{ ve } \qquad \beta_0^*=y_0-\beta_1^*x_0$$

olmak üzere $\hat{Y}=\beta_0^*+\beta_1^*X$ olarak belirlenir.

En küçük mutlak sapma regresyon doğrusunun belirlenmesinde iki farklı durumla karşılaşmak mümkündür ki bunlardan birincisi belirlenen (x_0,y_0) veri noktasından geçen doğrulardan birden fazlasının bizim için en iyi doğru olması

veya birden fazla doğru için hesaplanılan mutlak sapmalar toplamı değerlerinin bu doğrular için birbirlerine eşit ve minimum olmasıdır. İkinci durum ise, belirlenen (x_0, y_0) veri noktasından geçen en iyi doğrunun iki veya daha fazla veri noktasından da geçmesidir. Birinci durumun geçerli olması yani belirlenecek en iyi doğrunun tek olmaması durumunda, mutlak sapmalar toplamı değerini minimum yapan tüm doğrular için tüm (x_i, y_i) veri çiftleri belirlenir ve bir sonraki adımda uygulanacak işlemler her bir nokta için ayrı ayrı uygulanır. Bu adım sonucunda elde edilen sonuçlar arasından en iyi (minimum) olanı seçilir. İkinci durumun geçerli olması yani en iyi doğru olarak belirlenen doğru üzerinde iki veya daha fazla veri noktasının çakışması durumunda ise bu veri noktalarından herhangi birisi seçilebileceği gibi çakışan tüm noktaların ortalaması alınarak ortalama nokta üzerinden de işlemler devam ettirilebilir.

$Y = X\beta + \varepsilon$ şeklinde matris formunda ifade edilen çoklu doğrusal regresyon modeli için en küçük mutlak sapma tekniğinde minimize edilmeye çalışılan fonksiyon;

$$\sum |Y - X\beta| \qquad (2.5.1)$$

şeklinde olacaktır. Basit doğrusal en küçük mutlak sapma regresyonda en iyi doğrunun seçimindeki aşamalar çoklu regresyon için de geçerlidir. Ancak çoklu regresyonda ilk aşamada tahminlenen bir $\hat{\beta}$ başlangıç vektörü, daha sonra (2.5.1) fonksiyon değerini daha küçük yapacak bir $\hat{\beta}^*$ vektörü ve sonuçta fonksiyon değerini minimum yapacak $\hat{\beta}^*$ vektörünün elde edilmesi şeklinde adımlar halinde en iyi regresyon doğrusu elde edilir. Çoklu doğrusal en küçük mutlak sapma regresyonda en iyi $\hat{\beta}^*$ tahmin vektörü bulunurken;

$$\hat{\beta}^* = \hat{\beta} + td$$

eşitliğinden yararlanılır. Burada d, en uygun yön vektörünü ifade ederken t değeri, d yönündeki tahminlenecek en iyi vektörün belirlenmesi için gereken katsayı değeridir. d yönünde tahminlenecek vektörler arasından en iyi vektörün seçimi için gereken t katsayı değeri;

$$\sum \left| Y - X\left(\hat{\beta} + td\right) \right|$$

fonksiyon değerini minimize eden t değeri olacaktır. Dolayısıyla bu ifade açılacak olursa;

$$\sum \left| Y - X\hat{\beta} - tXd \right|$$

ifadesi elde edilir. Burada $z_i = Y - X\hat{\beta}$ ve $w_i = Xd$ yazılırsa fonksiyon;

$$\sum \left| z_i - tw_i \right| \qquad (2.5.2)$$

şeklinde daha basit bir hale getirilir. Bu durumda en iyi vektörün elde edilebilmesi için gereken t değeri (2.5.2) değerini minimize eden değer olacaktır. (2.5.2)'de basit haliyle ifade edilen minimizasyon fonksiyonundan t değerini elde edebilmek için z_i / w_i oran değerleri bulunarak küçükten büyüğe doğru sıralanır. Elde edilen sıralamaya uygun olarak z_i ve w_i değerleri de sıralandıktan sonra $T = |w_i|$ olmak üzere;

$$|w_1| + |w_2| + \ldots + |w_{k-1}| < \frac{1}{2}T$$

$$|w_1| + |w_2| + \ldots + |w_{k-1}| + |w_k| > \frac{1}{2}T$$

koşullarını sağlayan k değeri bulunur. Buradan elde edilecek t değeri z_k / w_k olarak elde edilir.

2.3. Parametre Tahminlerini Etkileyen Faktörler

En küçük kareler minimizasyonu ile elde edilecek regresyon doğrusu ve buna ilişkin parametre tahminlerinin gerçeğe uygun bir biçimde elde edilebilmesi için Bölüm 2.2.1'de bahsedilen 7 adet varsayımın sağlanması gerekmektedir. Alternatif bir teknik olarak geliştirilen Huber-M tekniği;

- Veri seti içerisinde aykırı değerlerin olması durumunda
- Hataların normal dağılıma sahip olmaması durumunda

en küçük kareler tekniğine göre daha duyarsız kalmakta ve gerçeğe yakın sonuçlar vermektedir. Bu yüzden tezin genel amacı doğrultusunda yukarıda bahsedilen iki özellik detaylı bir şekilde incelenmeye çalışılacaktır [5,12].

2.3.3. Normallik Varsayımının Sağlanmaması

En küçük kareler tahmincilerinin olasılık dağılımları, hata teriminin olasılık dağılımı hakkında yapılan hata terimlerinin sıfır ortalamalı normal dağılıma sahip olmaları gerektiği varsayımına bağlıdır. Bundan dolayı yapılacak katsayı tahminleri ve bunlara ilişkin testlerin uygulanabilmesi için ilgili varsayımın sağlanması gerekmektedir. Şayet hata terimleri dağılımı normal ise parametre tahmincileri $\hat{\beta}_i$'ler de normal dağılıma sahip olacaklardır. Hata terimlerinin değerleri büyük olduğunda çizilecek dağılım fonksiyonu grafiği kalın kuyruklu olacaktır. Kuyrukların kalın olması veya beklenenden daha uzun bir kuyruğun olması hata değerlerinin büyük olmasını dolayısıyla da dağılımın normalliğinin yitirilmesi anlamına gelir. Bu durumda en küçük kareler tahminleri etkinliğini kaybeder. Normallik varsayımının gerçekleşmemesi halinde sıklıkla kullanılan yöntem bağımlı değişkene dönüşüm uygulamaktır. Ancak bağımlı değişkene yapılacak herhangi bir dönüşüm sabit varyanslılık ve doğrusallık varsayımlarından sapmaları da beraberinde getirecektir. Bu yüzden kullanılacak tekniğin direkt olarak herhangi bir dönüşüme ihtiyaç duymaksızın bu varsayımın ortaya çıkaracağı sonuçlara karşı daha duyarsız olması ve daha iyi sonuçlar vermesi istenir [15].

Anakütle hata terimi ε'un dağılımının tahmininde e hata terimlerinin dağılımından yararlanılır. Hata terimlerinin herhangi bir olasılık dağılımına uygunluğu Q-Q (Quantile-Quantile) çizimleri ile belirlenebilir. Q-Q çiziminde y ekseninde hata terimlerinin değerleri yer alırken x ekseninde ise incelenecek birikimli olasılık dağılım fonksiyonuna ilişkin kantil değerleri yer alır.

Hata terimlerinin normal dağılıma uyup uymadığının saptanabilmesi için çizilecek Q-Q çiziminde, veri noktaları x-y düzleminde 45 derecelik bir açı oluşturacak şekilde bir şekle sahiplerse hata terimlerinin normal dağılıma sahip oldukları söylenebilir.

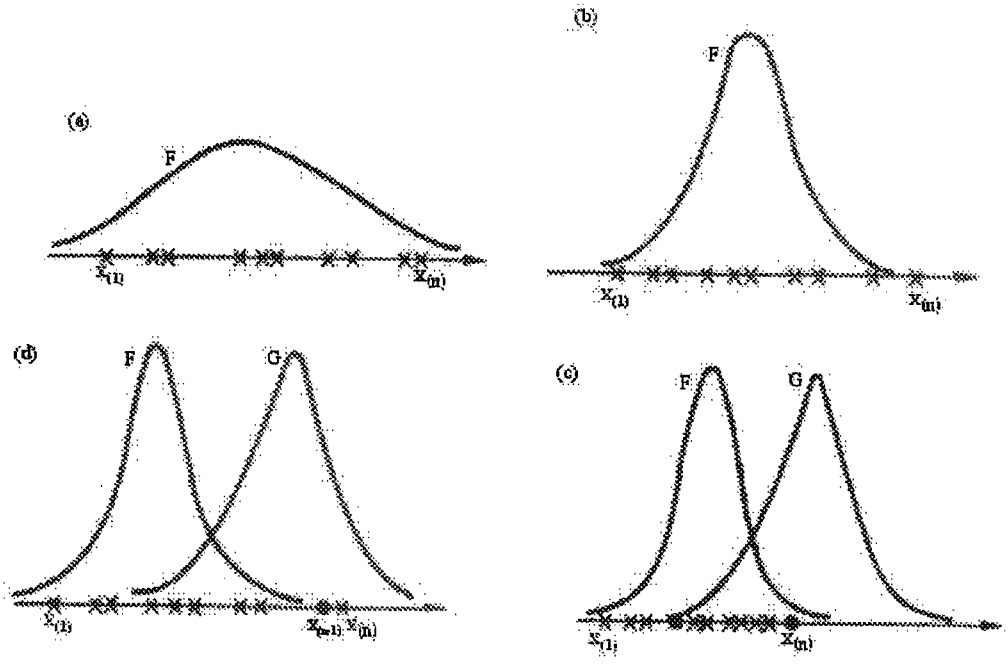
2.3.4. Aykırı Değer Problemi

Regresyonun kullanılmaya başlanmasından günümüze kadar veri setini temsil etmeyen, uyumsuz veya aykırı değerlerin varlığı araştırmacıyı her zaman kaygıya düşürmüştür. Veri yapısını bozan bu türden verilerin varlığı, veri

kaynağına ilişkin bilginin azalmasına yada çarpıtılmasına, kestirim mekanizmasının biçiminin bozulmasına neden olan ve sıkça karşılaşılan durumlardır. Uç değer (extreme), aykırı değer (outlier) ve kirletici değer (contaminant) ayrımları üzerinde durmak gerekir. İlk bakışta hepsi sanki aynı olguyu işaret ediyormuş gibi görünse de uç değer örneklem içinde yer alan ve örnek dağılımına ait olabilecek değerlerdir. Oysa ki aykırı değerler his ve mantık açısından o örneklemde yer alması mümkün olmayan değerler olarak tanımlanabilir. Kirletici değerler ise eldeki örneklemin dağılımına değil tamamen farklı bir dağılıma ait gözlem değeri olarak karşımıza çıkar [6,16]. Bu noktada aykırı değer kavramı için birbiriyle tamamen örtüşen iki tanım yapmak mümkündür. Bunlar;

1. Veri setindeki diğer gözlemlerle uyuşmayan aykırı bir gözlem veya gözlem grubu.
2. Örnekteki diğer birimlerden önemli derecede farklılık gösteren değer [16].

Uç değer, aykırı değer ve kirletici değer kavramları arasındaki ayrımın ne şekilde belirleneceğinin veya veri seti içerisinde bulunan ve durumundan şüphe edilen bir verinin bu üç kavramdan hangisi ile ve nasıl nitelendirileceğinin belirlenmesi gerekir. Bunun için $x_1, x_2, \ldots, x_n$ değerlerinin, F şeklinde ifade edebileceğimiz herhangi bir dağılımdan rassal olarak seçilmiş n boyutlu bir örneğin elemanları olduğu varsayımı altında, bu değerlerin küçükten büyüğe doğru dizilmiş halinin $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ şeklinde olduğunu kabul edelim. $x_{(1)}$ ve $x_{(n)}$ gözlem değerlerini örneklemdeki uç değerlerdir. Benimsenen dağılım F'e göre bu değerlerin aslında birer aykırı değer veya kirletici değer olup olmadıklarını Şekil 2.2. yardımıyla açıklayalım [16].



Şekil 2.2. Uç Değerler, Aykırı Değerler ve Kirletici Değerler

Şekil 2.2. (a)'da $x_{(1)}$ ve $x_{(n)}$ değerlerinin her ikisi de aykırı değer olarak görünmemektedir. Buna karşılık Şekil 2.2. (b)'de $x_{(n)}$ gözlem değeri üst aykırı değer olarak göze çarparken $x_{(1)}$ gözlem değerinin bir aykırı değer olarak kabul edilip edilmeyeceği konusunda kesin bir yargıya varmak mümkün değildir. $x_{(1)}$ gözlem değeri de alt aykırı değer olarak kabul edilebilir. Bu durumda $x_{(1)}$ ve $x_{(n)}$ gözlem değerlerini bu örnek için benimsenen F dağılımına göre aykırı değer çifti olarak değerlendirebiliriz. Aynı şekilde Şekil 2.2. (d) için $x_{(n-1)}$ ve $x_{(n)}$ değerlerini örneklemdeki üst aykırı değer çifti olarak nitelendirmek mümkündür. Sonuç olarak; örneklemdeki her bir aykırı değer aynı zamanda bir uç değerdir. Ancak uç değerler aykırı değer olabilir de olmayabilir de.

Kirletici değer ayrımını yapabilmek için, örneklemdeki tüm değerlerin F dağılımının elemanları değil birkaç değerin F dağılımının x ekseninde yukarıya kaydırılmış hali olan (daha büyük ortalamalı) G dağılımının elemanları olduğunu düşünelim. Bu durumda G dağılımına ait gözlem değerleri kirletici değer olarak adlandırılır. Şekil 2.2. (c)'de nokta (•) olarak ifade edilen iki değer F dağılımına

sahip olduğu düşünülen örneklem için kirletici değerlerdir. Bunlardan birisi üst uç değer, diğeri ise örneklemin ortasında bir değerdir. $x_{(n)}$ değeri, bir uç değer ve bir kirletici değer olmasına karşılık kesinlikle bir aykırı değer değildir. Buna karşılık Şekil 2.2. (d)'de $x_{(n-1)}$ ve $x_{(n)}$ değerleri birer uç değer olmamalarına rağmen kirletici değerlerdir ve aynı zamanda birer aykırı değerlerdir. Sonuç olarak; aykırı değerler aynı zamanda birer kirletici değer, kirletici değerler de aynı zamanda birer aykırı değer olabilirler de olmayabilirler de. Bu durumda bir aykırı değerin aynı zamanda bir kirletici değer olup olmadığını tespit etmek elbette ki mümkün değildir. Ancak aykırı değerlerin incelenmesinde kullanılan istatistiksel tekniklerde bu olasılık her zaman göz önünde bulundurulur.

İstatistiki analizi yapılacak veri seti için aykırı değerlerin sınıflandırılması veya yorumlanması ile ilgili kullanılacak metotların araştırılması gayet doğaldır. Bazen verinin genel gidişatını etkilememesi açısından aykırı değerlerin örneklemden çıkarılması söz konusu olacakken bazen de kullanılacak istatistiksel analizlerdeki etkilerini minimum yapacak metotların benimsenmesi ve uygulanması gerekir. Aykırı değer, temel olarak regresyon modellerinde, zaman serilerinde, deney planlamada, yapısal olmayan çok değişkenli veri analizinde, yönsel veri analizinde ve anket araştırmalarında sonuçlar açısından beklenen gidişatı bozan değer olarak ortaya çıkmaktadır.

Aykırı değer olarak gözlemlenen verilerin ortaya çıkış nedenlerini ise şu üç ana başlık halinde sıralamamız mümkündür [16]:

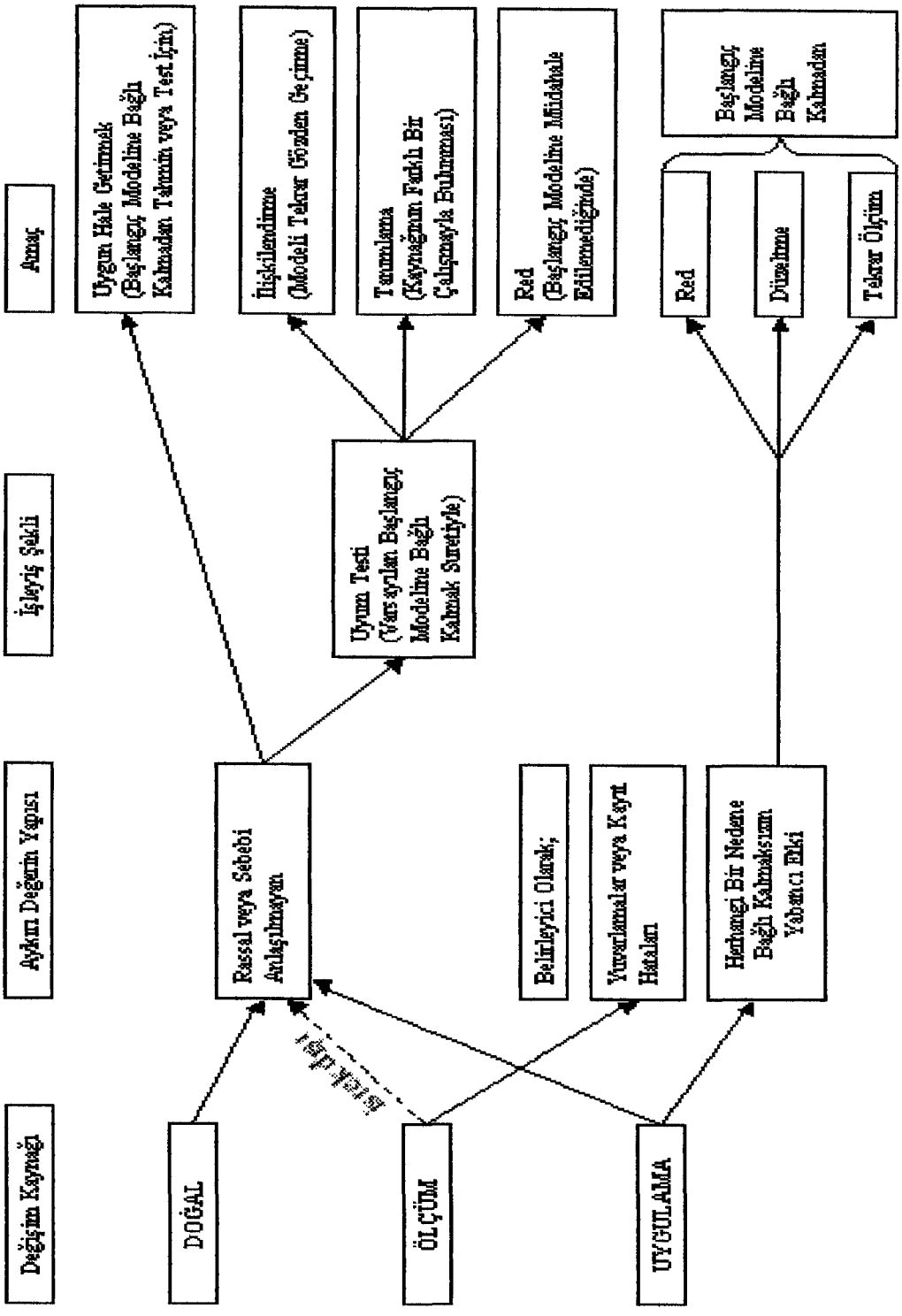
a. Doğal Değişkenlik (Inherent Variability): Anakütledeki birimlerin doğal nedenlerden kaynaklanan farklılıkları olarak tanımlanabilir. Örneğin erkeklerin boylarına ilişkin yapılacak bir analiz için farklı soylara mensup olmanın getirdiği doğal farklılık...

b. Ölçüm Hatası (Measurement Error): Elde edilen ölçüm değerlerinde yapılan yuvarlamalar, verideki süreklilik veya tekrar durumlarının dikkate alınmaması ve kayıt hatalarından kaynaklanan durumlardır.

c. Uygulama Hatası (Execution Error): Hatalı veya eksik veri toplama, reklam veya belirli bir amaç için özellikle yanlı bir kitlenin seçimi, anakütleyi gerçekte yansıtmayan kişisel yargıları içeren verilerden doğan hata olarak nitelendirilebilir.

Aykırı değerlerin varlığının tesbit edilmesi ile beraber, bu değerlerin etkilerini yok edebilmek için uygulanılan yaklaşımlara bakıldığında, iki ayrı teknik göze çarpmaktadır. Bunlardan ilki aykırı değerin ilgilenilen veri seti dahilinde kalıp kalmayacağının araştırılması diğeri ise veri setinden çıkarılmasıdır. Her iki durum için de aykırı değerlerin yapılacak parametre tahminlerindeki etkilerinin belirlenmesi gerekmektedir. Gözlem değerleri içerisinde aykırı değerlerin bulunması halinde, rassal olarak seçilen bir örnekten elde edilecek çıkarsama ve tahminlerin örneğin alındığı anakütle parametre değerlerini tam olarak yansıtması için kullanılacak istatistiksel tekniklerin uygulanış süreci, uyum süreci olarak adlandırılır. Bu teknikler ciddi güçlükler olmaksızın aykırı değerleri veri yapısına uygun hale getirirler. Bir başka deyişle uygulanılan süreç sonunda daha gerçekçi sonuçlar elde edilmesini sağlarlar. Aykırı değerlerin giderilmesinde izlenecek süreç Şekil 2.3.'de gösterilmektedir [16].

Veri setinde aykırı değerlerin olması durumunda hatalara ilişkin varyans değerleri arasındaki değişim miktarı artacaktır. Bu özellik temel alınarak veri seti içerisinde aykırı değerlerin olup olmadığı çeşitli şekillerde araştırılabilir. En sık rastlanan grafik türü, artık değerlerine karşılık tahmin değerlerinin yer aldığı nokta grafiğidir. Daha ileri düzeyde bir araştırmada ise student türü artıkların incelenmesidir. Student türü artıkların incelenmesinin en büyük faydası, artıklar üzerinde ölçek standartlaştırmasının sağlanmasıdır. Student artıklarının incelenmesi sonucu belirli bir gözlem biriminin aykırı değer olup olmadığına karar vermek kolaylaşmaktadır. Ayrıca student türü artıklar kullanılarak çizilecek Q-Q grafiği yardımıyla normallik varsayımının sağlanıp sağlanmadığını da incelemek mümkündür. Bilindiği gibi artıkların Q-Q grafiği sayesinde normallik varsayımının sağlanıp sağlanmadığını görmek mümkündü. Student türü artıklarda da benzer bir yapı sıra istatistiklerinden gelmektedir. Buna göre hesaplanan student türü artıklar küçükten büyüğe dizilerek Q-Q grafiği çizilmektedir. Şayet noktalar 45 derecelik bir açı ile doğru oluşturacak şekilde dağılmışlarsa dağılımın normal olduğu ve aykırı değerlerin olmadığı söylenebilir. Fakat daha çok S şeklinde bir görünüm elde ediliyorsa gözlem değerleri içerisinde aykırı değerlerin varlığından ve dolayısıyla normallik varsayımının sağlanmadığından sözedilebilir.



Şekil 2.3. Aykırı Değerlerin Giderilmesinde İzlenecek Süreç

3. HUBER'İN M TAHMİNCİLERİ

3.1. Regresyona Huber'ın Getirdiği M Yaklaşımı

M-regresyon robust istatistiğin bir parçasıdır. İstatistiksel bir modelin oluşumunda en küçük karelerin varsayımları sağlanmadığında, istatistiksel prosedüre göre robust tekniklerin kullanımı daha doğru olacaktır. Verimiz normal bir doğrusal regresyona uygun olduğunda en küçük kareler tahminleri ve test oldukça iyidir ama rassal hataların anakütle için normallik varsayımı geçersiz olduğunda en küçük kareler yapılması doğru olmaz. M-regresyon bu varsayıma uygun olarak geliştirilmiştir. M-tahmini fikrini ilk defa 1964'te Peter Huber ortaya atmıştır [6].

M-regresyon ismini en çok benzerlik tahminleri ile arasındaki ilişkiden almaktadır. Şayet hataların dağılımları belirli bir dağılıma uyuyorsa M tahminleri en çok benzerlik tahminleri olabilir. M tahminlerinin asıl amacı hataların sahip olabileceği dağılımlar için daha iyi tahminler elde edebilmektir. Genellikle M-regresyon hataların dağılımının normal dağılıma göre daha uzun kuyruklu simetrik bir dağılım göstermesi halinde kullanılır.

3.1.1. Basit Doğrusal Regresyonda Huber'ın M Yaklaşımı

$Y = \beta_0 + \beta_1 X + \varepsilon$ şeklindeki basit doğrusal regresyon modelini ele aldığımızda, tahmin edilecek $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X + e$ modeli için uygulanacak en küçük kareler tahmininde $\hat{\beta}_0$ ve $\hat{\beta}_1$ tahmin değerleri, hata kareler toplamı olan $\sum \hat{e}_i^2$ değeri olabildiğince küçük olacak şekilde seçilir. Yine robust regresyonun bir diğer parçası olan en küçük mutlak sapmalar tahmininde ise katsayı tahmin değerleri $\sum |\hat{e}_i|$ olabildiğince küçük olacak şekilde seçilir. Huber-M tahmininde ise bu iki düşünce genelleştirilmiştir. Buna göre $\hat{\beta}_0$ ve $\hat{\beta}_1$ katsayı tahminleri, $\sum \rho(\hat{e}_i)$ değerini olabildiğince küçük olacak şekilde seçilir ki burada $\rho(e)$ e hata teriminin bir fonksiyonudur. En küçük kareler ve en küçük mutlak sapmalar tahmini $\rho(e) = e^2$ ve $\rho(e) = |e|$'ye göre M-tahmininin özel bir durumları gibi dikkate alınmalıdır [6,11,12].

Huber M-tahminleri, e^2 ve $|e|$ arasında bir orta nokta olan $\rho(e)$ fonksiyonunu kullanan M-tahminleridir. En küçük mutlak sapmalar tahminlerinin en küçük kareler tahminlerinden asıl üstünlüğü uç değerlere ve aykırı değerlere duyarlı olmamasıdır. Uç değerler olmadığı zaman en küçük kareler tahminleri daha doğru olabilir. Bundan hareketle Huber her iki tekniğin üstünlüklerini biraraya getirmeyi amaçlayarak oluşturduğu modelde, e hata terimi, sıfıra uzak olduğunda $\rho(e)$ fonksiyonunun değerini $|e|$'ye; e hata terimi, sıfıra yakın olduğunda ise e^2 'ye eşitleyerek daha uzlaştırıcı bir teknik öne sürmüştür [6,17].

Buna göre Huber'in önerdiği teknikte $\rho(e)$ fonksiyonunu;

$$\rho(e)=\begin{cases} e^2 & -k \le e \le k \quad \text{ise} \\ 2k|e|-k^2 & e<-k \ \text{veya} \ k<e \quad \text{ise} \end{cases} \tag{3.1}$$

şeklinde tanımlamıştır. Huber'in diğer bir önerisi $k = 1.5\hat{\sigma}$ alınmasıdır ki buradaki $\hat{\sigma}$, rassal hatalar anakütlesinin standart sapması σ'nın bir tahminidir. $|e|$ yerine $2k|e|-k^2$ kullanılmasının nedeni $\rho(e)$ fonksiyonunu birinci mertebeden türevi sürekli bir fonksiyon (smooth function) haline getirmek içindir. $|\hat{e}_i|$ mutlak sapmalarının medyanını MSM olarak adlandırırsak; σ'nın tahmini için $\hat{\sigma}$=1.483MSM kullanırız. 1.483 çarpanı rassal hatalar dağılımının normal olduğu durumda $\hat{\sigma}$ 'nın, σ'nın iyi bir tahmini olduğunu garantilemek amacıyla seçilmiş bir sabittir. $\hat{\beta}_0$ ve $\hat{\beta}_1$ Huber M-tahminleri, (3.2) fonksiyonunda en küçüklenen a ve b değerleridir [6,15,16,17].

$$\sum \rho(y_i-(a+bx_i)) \qquad i=1,2,\ldots,n \tag{3.2}$$

a ve b, ρ'nun tanımı (3.1)'de kapalı olarak ifade edilmesine karşın (3.2)'de parantez içerisinde açık bir şekilde görülmektedir. ρ fonksiyonu, $k = 1.5\hat{\sigma}$ 'yı içermektedir ve $\hat{\sigma}$, $y_i-(a+bx_i)$ sapmasından hesaplanır. (3.2)'nin en küçüklemesi için uygulanacak algoritma şu şekildedir.

Algoritmanın ilk adımında β_0 ve β_1'in başlangıç tahminleri için en küçük kareler tahminlerini kullanılır. Başlangıç için en küçük kareler tahminlerinin kullanılmasının nedeni mümkün olabilecek en az adımda sonuca ulaşabilmek içindir. Elde edilen bu tahminler σ'nın bir tahmini ve sapmaların hesabında

kullanılır. Elde edilen σ'nın bir tahmini olan $\hat{\sigma}$ ve sapmalar ise ikinci adım için kullanılacak ve başlangıçta hesaplanmış olan β_0 ve β_1'in güncellenmiş tahminlerini elde etmede kullanılır. Bu güncellenen tahminler σ'nın güncellenmiş yeni bir tahmini ve yeni sapmaların hesabında kullanılır. Daha sonra yeni sapmalar ve yeni $\hat{\sigma}$, β_0 ve β_1'in bir sonraki adım olan üçüncü adımda kullanılacak başlangıç tahminlerinin bulunmasında kullanılır. Bu algoritma elde edilen son tahminlerle bir önceki adımda elde edilmiş olan tahminler birbirine eşit veya çok yaklaşık değerler alıncaya kadar yada artık kareler toplamı olan $\sum e_i^2$ değeri en küçük değerine ulaşıncaya kadar devam ettirilir [6,18,19].

Tahminlerin yapılabilmesi için kullanılacak algoritma adımları şu şekilde ifade edilebilir.

Algoritma 1

ADIM 1. β_0 ve β_1 tahmin değerlerini en küçük kareler ile hesapla.

ADIM 2. Bu tahminleri kullanarak $\hat{\sigma}$ ve e_i değerlerini bul.

ADIM 3. Bulunan e_i değerleri üzerinden ρ(e) fonksiyonunu kullanarak e_i^* düzeltilmiş sapma değerlerini bul.

ADIM 4. $\hat{\sigma}$ ve e_i^* değerlerini kullanarak en küçük kareler ile β_{00} ve β_{10} tahminlerini hesapla.

ADIM 5. β_0 ve β_1 için güncellenen tahminler ile bir önceki değerlerini karşılaştır.

ADIM 6. Tahminler arasındaki fark 0,001'den küçükse bitir.

ADIM 7. Değilse β_0 ve β_1 değerlerini β_{00} ve β_{10} olarak ata ve Adım 2'e dön.

Yukarıda verilen algoritmayı daha detaylı bir şekilde incelersek; algoritmanın ilk adımında en küçük kareler ile elde edilecek a^0 ve b^0 değerleri, β_0 ve β_1'in geçerli tahminleri olur. Bu tahminler yardımıyla $y_i - (a_0 + b_0 x_i)$ formülünden sapmalar hesaplanır ve buradan da $\hat{\sigma}^0 = 1.483MSM$ formülü yardımıyla ilk adım için geçerli olacak ve σ'nın bir tahmini olan $\hat{\sigma}^0$ değerine ulaşılır. Bu aşamada sapmaların büyük olmasını önlemek için şu düzeltmelerin yapılması gereklidir. Gerçek tahmin edilen regresyon doğrusundan y_i'nin sapması

$e_i^0 = y_i - (a^0 + b^0 x_i)$ 'dır. Buradan $y_i = a^0 + b^0 x_i + e_i^0$ 'dır. $y_i^* = a^0 + b^0 x_i + e_i^*$ tanımlanır ki burada e_i^* düzeltilmiş sapma, e_i^0 değerinden elde edilir ve sapmaların hiçbirisi mutlak değer içindeki $1.5\hat{\sigma}^0$ 'dan büyük değildir. Şayet e_i^0 sapma değeri, $-1.5\hat{\sigma}^0$ ve $1.5\hat{\sigma}^0$ arasında ise $e_i^* = e_i^0$ (ve dolayısıyla da $y_i^* = y_i$)'dır. Eğer e_i^0 sapma değeri, $-1.5\hat{\sigma}^0$ 'dan küçükse $e_i^* = -1.5\hat{\sigma}^0$ değerine eşit olarak alınır ve eğer e_i^0 sapma değeri, $1.5\hat{\sigma}^0$ 'dan büyükse $e_i^* = 1.5\hat{\sigma}^0$ olarak alınır. Bu düzeltmeler yardımıyla β_0 ve β_1'in en küçük kareler kullanılarak elde edilecek güncellenmiş tahminleri $y_1^*,.....,y_n^*$ düzeltilmiş verilerinden elde edilir [6,13,16].

Algoritmanın doğruluğu mantıklı gibi görünse de (3.2)'deki fonksiyon için yapılacak en küçüklemenin daha rahat anlaşılabilmesi için; $\hat{\sigma}$'yı sabit tutarak, (3.2)'nin a ve b'ye göre türevlerini alıp sıfıra eşitlemeliyiz. Buna göre; a ve b iki bilinmeyenine sahip şu iki eşitlik elde edilir:

$$\sum \rho'(y_i - (a + bx_i)) = 0$$

$$\sum x_i \rho'(y_i - (a + bx_i)) = 0 \qquad (3.3)$$

$\rho'(e)$ türevi, $-1.5\hat{\sigma}$'ya eşit veya daha küçük tüm e'ler için $-3\hat{\sigma}$'ya, $1.5\hat{\sigma}$'ya eşit veya daha büyük tüm e'ler için $3\hat{\sigma}$'ya eşittir. Bu nedenle $e_i = y_i - (a + bx_i)$ sapmalarının yerine e_i^* kırpılmış sapmaları konulur. Burada e_i sapma değeri, $-1.5\hat{\sigma}$ ile $1.5\hat{\sigma}$ arasında ise $e_i^* = e_i$; e_i sapma değeri, $-1.5\hat{\sigma}$'dan küçükse $e_i^* = -1.5\hat{\sigma}$ ve e_i sapma değeri, $1.5\hat{\sigma}$'dan büyükse $e_i^* = 1.5\hat{\sigma}$ olacaktır. Bu durumda (3.3)'e ilişkin çözümler değişmeden y_i yerine $y_i^* = a_i + bx_i + e_i^*$ düzeltilmiş değerlerini yazabiliriz. Sapmalar ilişkin yapılan bu düzeltmeler sonucunda (3.2)'yi minimize eden a ve b değerleri de değişmez. Düzeltmenin sonucunda $\rho(y_i^* - (a_i + bx_i)) = [y_i^* - (a + bx_i)]^2$ elde edilir. Tanıma göre $\sum [y_i^* - (a_i + bx_i)]^2$ en küçüklemesinin sonucu, en küçük kareler tahminlerinin düzeltilmesi ile elde edilen düzeltilmiş veriden elde edilir. Algoritmadaki y_i^* değerleri, en son elde edilen M-tahminine göre değil regresyon doğrusunun o anda elde edilen tahminine bağlı kalınarak düzeltilmiştir. Ancak bir çelişki gibi görünen bu durum

algoritma ilerledikçe elde edilen tahmin değerleri birbirine yaklaştığında ortadan kalkar. Güncellenen son tahmin değerleri ile bir önceki tahmin değerleri arasındaki fark yeterince az olduğu noktada adımlar durdurulur [6,19]. Artık son elde edilen tahmin değerleri için katsayı anlamlılık sınamaları yapılabilir.

$Y = \beta_0 + \beta_1 X + \varepsilon$ basit doğrusal regresyon modeli için $\beta_1 = 0$ katsayı anlamlılık testinin nasıl yapıldığına bir göz atacak olursak, en küçük kareler için test istatistiği;

$$F_{EKK} = \frac{AKT_i - AKT_F}{\hat{\sigma}^2_{EKK}}$$

şeklinde hesaplanmaktadır. Burada AKT_i değeri $Y = \beta_0 + \varepsilon$ indirgenmiş modeli için hesaplanan artık kareler toplamını, AKT_F değeri ise tüm model için hesaplanan artık kareler toplamını ifade eder. Burada gözden kaçırılmaması gereken önemli durum indirgenmiş model için hesaplanacak $\hat{\sigma}$ tahmin değerinin tekrarlamalar sonucu elde edilememesidir. Bunun için iterasyonlar sırasında kullanılacak $\hat{\sigma}$ tahmin değeri daha önceden tüm model için tahmin edilen $\hat{\sigma}$ değerine eşit olarak alınır. Yaklaşık bir p değeri,

$$P\,[F \geq F_{EKK}]$$

şeklinde hesaplanarak sonuca varılır. Burada F rassal değişkene ilişkin 1 ve n-2 serbestlik dereceli F dağılımını ifade eder. Huber'in M tahminine göre ise test istatistiği;

$$F_M = \frac{DHT_I - DHT_F}{\hat{\lambda}}$$

şeklinde hesaplanır. Burada DHT_I ve DHT_F değerleri sırasıyla indirgenmiş ve tüm modelin dönüştürülmüş hatalar toplamını ifade eder ve genel olarak;

$$DHT = \sum \rho(\hat{e}_i)$$

şeklinde hesaplanılır.

$$\hat{\lambda} = (n/m)\sum(\hat{e}_i^*)^2 \Big/ (n-2)$$

şeklinde hesaplanır ki n, toplam artık $(\hat{e}_i)$ sayısını ifade ederken m, toplam artık sayısından kırpılmış artık $(\hat{e}_i^*)$ sayısının çıkarılması sonucu elde edilir. Basit

doğrusal regresyon modeli için t ve F testleri birbirleriyle aynı sonuçlar verdikleri için birbirlerinin alternatifi iki testtir. Buna dayanarak yaklaşık bir p değeri,

$$P\,[F \geq F_M] \text{ veya } P[|t| \geq |t_M|]$$

şeklinde hesaplanabilir ki burada,

$|t_M| = \sqrt{F_M}$ 'dir ve t rassal bir değişkene ilişkin n-2 serbestlik dereceli t dağılımını ifade eder [6,11].

3.1.2. Çoklu Doğrusal Regresyonda Huber'ın M Yaklaşımı

Basit doğrusal regresyon modeli için M tahmin değerlerini elde etmek için uygulanacak algoritma adımlarını inceledikten sonra çoklu regresyon M-tahminleri için uygulanacak algoritmaya geçebiliriz. Aslında çoklu regresyon için uygulanacak algoritma basit regresyon için tanımlanan algoritmanın daha fazla bağımsız değişken için yapılacak bir genellemesidir.

$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_j X_j + \varepsilon \quad j = 1, 2, \ldots, p$ şeklinde oluşturulan çoklu doğrusal regresyon denkleminde, n gözlenen birim sayısını ifade etmek üzere, $\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_p$ Huber M-tahminleri,

$$\sum \rho\left(y_i - \left(\beta_0 + \beta_1 x_{i1} + ... + \beta_j x_{ij}\right)\right) \quad i = 1,2,\ldots,n \quad j = 1, 2, \ldots, p \qquad (3.4)$$

ifadesini minimize eden β_0, β_1, ..., β_p değerleridir. Burada $\rho(e)$ fonksiyonu, (3.1)'de tanımlanan e hata teriminin bir fonksiyonudur. Çoklu doğrusal regresyon denklemini;

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & X_{21} & \cdots & X_{1p} \\ 1 & X_{12} & X_{22} & \cdots & X_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{1N} & X_{2N} & \cdots & X_{np} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \varepsilon_0 \\ \varepsilon_p \\ \vdots \\ \varepsilon_p \end{bmatrix}$$

$$Y_{(n\times 1)} = \qquad X_{(n\times p)} \qquad \beta_{(p\times 1)} + \quad \varepsilon_{(n\times 1)}$$

$Y = X\beta + \varepsilon$ şeklinde matris formunda ifade edersek Huber M-tahminlerinin $\hat{\beta}$ parametre tahmin vektörü;

$$\sum \rho(Y - X\beta) \qquad (3.5)$$

ifadesini minimize eden β vektörü olarak tanımlanır [15,18,19].

(3.4) fonksiyon değerinin minimumunu ve buna bağlı olarak sabit bir standart sapma tahmin değeri $\hat{\sigma}$ değerini elde edebilmek için bu fonksiyonun sırasıyla β_0, β_1, ..., β_p'ye göre türevleri alınarak sıfıra eşitlenir. Sonuçta p+1 bilinmeyenli p+1 tane eşitlik elde edilir. Genel gösterimi ile elde edilen eşitlikler;

$$\sum x_{ij}\rho'\left(y_i - \left(\beta_0 + \beta_1 x_{i1} + ... + \beta_j x_{ij}\right)\right) = 0 \qquad (3.6)$$

$$i = 1,2,...,n \qquad j = 1,2,...,p$$

şeklinde ifade edilir. Burada i'nin tüm değerleri için $x_{i0} = 1$'dir. Ancak elde edilen bu eşitlikler $\beta_0, \beta_1, ..., \beta_p$ bilinmeyenleri için doğrusal olmayan eşitliklerdir. Ancak şu şekilde doğrusal hale yaklaştırılabilirler.

$\beta_0^0 + \beta_1^0 + ... + \beta_p^0$ başlangıç parametre tahmin değerlerini ve $\beta_0 + \beta_1 + ... + \beta_p$ değerleri de bu değerlerden elde edilecek güncellenmiş tahmin değerlerini ifade etmek üzere başlangıç aşamasında ve buna bağlı olarak ilk adım sonucunda elde edilecek artık değerleri;

$$e_i^0 = y_i - \left(\beta_0^0 + \beta_1^0 x_{i1} + ... + \beta_p^0 x_{ij}\right)$$

ve

$$e_i = y_i - \left(\beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ij}\right)$$

olarak elde edilir. Güncellenen tahmin değerlerinin çözümü için;

$$\rho'(e_i) = (\rho'(e_i)/e_i)e_i \approx \left(\rho'\left(e_i^0\right)/e_i^0\right)e_i \qquad (3.7)$$

olarak yazılabilir.

$$w_i = \rho'\left(e_i^0\right) \Big/ e_i^0$$

olarak yazarsak veya daha açık bir gösterimle;

$$w_i = \begin{cases} 2 & \left|e_i^0\right| \le 1{,}5\hat{\sigma} \quad \text{ise} \\ 3\hat{\sigma} \Big/ \left|e_i^0\right| & \left|e_i^0\right| > 1{,}5\hat{\sigma} \quad \text{ise} \end{cases}$$

olarak yazıldığında (3.7)'deki ifade;

$$\rho'(e_i) \approx w_i e_i$$

şeklinde ifade edilebilir. Bu ifade kullanılarak (3.6)'daki doğrusal olmayan eşitlikler şu şekilde yaklaşık doğrusal eşitlikler şeklinde ifade edilebilir;

$$\sum x_{ij} w_i \left[\left(y_i - \left(\beta_0 + \beta_1 x_{i1} + ... + \beta_j x_{ij}\right)\right)\right] = 0 \quad i = 1,2,...,N \quad j = 1,2,...,p$$

Köşegen elemanları w_i değerleri olan matrisi W olarak ifade edersek, β katsayı değerlerini;

$$\beta = (X'WX)^{-1} X'WY \qquad (3.8)$$

eşitliği ile elde edebiliriz. Burada elde edilen β vektör değerleri ağırlıklandırılmış en küçük kareler olarak adlandırılırlar [11,13,14,20].

Çoklu regresyon için (3.5)'in en küçüklemesi için uygulanacak algoritma basit regresyon için uygulanan Algoritma 1 ile özde aynıdır. Ancak çoklu regresyonda uygulanacak algoritma, basit regresyon için yazılan algoritmanın genellenmiş halidir. Algoritmanın başlangıcında, katsayı başlangıç tahmin vektörü β^0 en küçük kareler tekniği kullanılarak elde edilir. Elde edilen katsayı değerleri yardımıyla e^0 artıkları $e^0 = Y - X\beta^0$ eşitliği yardımıyla hesaplanır. Bulunan artık değerleri yardımıyla da standart sapma tahmin değeri olan $\hat{\sigma}$ ve ağırlık değerleri olan w_i değerleri elde edilir. Standart sapmanın tahmini olan $\hat{\sigma}$ değeri hesaplanırken bulunan artık değerleri için mutlak değerlerinin medyan değeri Huber tarafından önerilmiş 1,483 katsayısı ile çarpılır. Dolayısıyla standart sapma tahmin değeri $\hat{\sigma}$=1,483MSM olarak elde edilir. Son olarak da (3.8) eşitliği yardımıyla güncellenmiş katsayı tahmin değerleri vektörü elde edilir. Bulunan bu güncellenmiş katsayı tahmin vektörü bir sonraki adım için başlangıç vektörü olur ve aynı işlemler bir sonraki adımda da uygulanılır. Sonuçta elde edilen son iki katsayı tahmin vektörlerindeki değerler birbirlerine çok yakın değerler alıncaya kadar adımlar sürdürülür [6,21,22].

Bu durumda çoklu doğrusal regresyon için Huber-M parametre tahmin değerlerini bulabilmek için gereken algoritma adımları şu şekilde yazılabilir.

Algoritma 2

ADIM 1. En küçük kareler tekniği kullanılarak β^0 katsayı başlangıç tahmin vektörü ve bu tahminlere bağlı olarak da e_i^0 artık değerlerini hesapla.

ADIM 2. Elde edilen artık değerleri yardımıyla $\hat{\sigma}_0$ ve $w_i^0 = \rho'(e_i^0)/e_i^0$ başlangıç ağırlık değerlerini hesapla.

ADIM 3. β, Huber-M katsayı tahmin vektörünü ağırlıklandırılmış en küçük kareler tekniğini kullanarak $\beta = (X'WX)^{-1}X'WY$ eşitliği yardımıyla hesapla.

ADIM 4. Elde edilen güncellenmiş katsayı tahmin vektörü β değerleri ile β^0 katsayı başlangıç tahmin vektörü değerlerini karşılaştır. Tahminler arasındaki fark, ayrı ayrı tüm j değerleri için $|\beta_j^0 - \beta_j| / |\beta_j^0| < 0{,}0001 \quad j = 1,2,\ldots,n$ ise bitir.

ADIM 5. Değilse β vektörünü β^0 vektörü olarak ata ve bu değerler üzerinden yeni e_i^0 artık değerlerini hesapla.

ADIM 6. Adım 2'ye dön.

Çoklu doğrusal regresyon modeli için yapılacak katsayıların anlamlılık testi de basit doğrusal regresyon modeli için uygulanan testlerin genelleştirilmiş hali olarak gözümüze çarpar. $Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + \varepsilon$ çoklu genel doğrusal regresyon modeli için, $\beta_{q+1} = \ldots = \beta_p = 0$ testi için en küçük kareler test istatistiği;

$$F_{EKK} = \frac{AKT_I - AKT_F}{(p-q)\hat{\sigma}^2_{EKK}}$$

şeklinde hesaplanır. AKT_I için artık değerlerini $Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_q X_q + e$ indirgenen modele ve AKT_F için artık değerlerini ise $Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + e$ tüm modele en küçük kareler tekniği uygulanarak hesaplanır. Formülasyondaki AKT_I ve AKT_F değerleri sırasıyla indirgenmiş ve tüm model için hesaplanılan artık kareler toplamlarını ifade etmektedir. p değeri tüm modeldeki bağımsız değişken sayısını ifade ederken, q indirgenmiş modeldeki bağımsız değişken sayısını ifade etmektedir. Genel olarak;

$$AKT = \sum \hat{e}_i^2$$

ve

$$\hat{\sigma}^2_{EKK} = \sum \hat{e}_i^2 / (n-p-1)$$

olarak hesaplanılır. $\hat{\sigma}^2_{EKK}$'nin hesaplanmasında tüm modele ilişkin hatalar kullanılır [3,7,10].

Benzer şekilde M-regresyon için hesaplanacak test istatistiği ise;

$$F_M = \frac{DHT_İ - DHT_F}{(p-q)\hat{\lambda}} \tag{3.6}$$

şeklinde hesaplanmaktadır. Burada $DHT_İ$ ve DHT_F değerleri sırasıyla indirgenmiş ve tüm modelin dönüştürülmüş hatalar toplamını ifade eder ve genel olarak;

$$DHT = \sum \rho(\hat{e}_i)$$

şeklinde hesaplanılır.

$$\hat{\lambda} = (n/m)\sum(\hat{e}_i^*)^2 \Big/ (n-p-1)$$

şeklinde hesaplanır. Burada $\hat{e}_i^*$ değeri, mutlak değer olarak $1.5\hat{\sigma}$ değerinden büyük artıkları ifade etmektedir. Standart sapma tahmin değeri olan $\hat{\sigma}$, basit doğrusal regresyon Huber-M tahminlerinde olduğu gibi 1.483MSM şeklindedir. $DHT_İ$ ve DHT_F 'deki artıklar sırasıyla indirgenen ve tüm modele M-regresyon prosedürü uygulanarak hesaplanır. İndirgenen model için tahmin prosedürü basit doğrusal regresyon için uygulanılan prosedürden biraz farklılaşmaktadır. σ'nın tahmini tekrarsız bulunur. Buna karşılık indirgenen modeldeki regresyon katsayılarının M-tahmin vektörünün elde edilmesi için iterasyonlar gerekir ve $\hat{\sigma}$ tahmini tüm modelden hesaplanır. Yine $\hat{\lambda}$'nın hesabında da tüm modeldeki artıklar kullanılır.

Teste ilişkin yaklaşık bir p değeri $P[F \geq F_M]$, aynen en küçük karelerdeki gibi hesaplanır. Buradaki F, rassal bir değişkene ilişkin p-q ve n-p-1 serbestlik dereceli F dağılımını ifade eder. Özellikle belirtmek gerekir ki çoklu doğrusal regresyon modeli için t testi, F testi için bir alternatif oluşturmamaktadır. Dolayısıyla çoklu doğrusal regresyon modelinde katsayıların anlamlılık testleri için t testi uygulanmaz [11,14,17].

M tahmininde kullanılan F_M formülü ile en küçük kareler tahminleri için kullanılan F_{EKK} formülü birbirlerine çok benzemektedir. $\rho(e)$ fonksiyonunda şayet 1.5 değeri yerine ∞ konulursa F_M formülü ile F_{EKK} aynı olur. F_M formülünde 1.5 değeri yerine ∞ yazıldığında m = n ve tüm i'ler için $\hat{e}_i^* = \hat{e}_i$ olur. Böylece $\hat{\lambda}$ ve

$\hat{\sigma}^2_{EKK}$ değerleri de bir noktada birleşir. Bu durumda $\rho(e)=e^2$ olacağından dönüştürülmüş hatalar toplamı (DHT) ile artık kareler toplamı (AKT) değerleri de birbirlerine eşit olacaktır. Bundan dolayı da F_M ile F_{EKK} formül değerleri bir noktada eşitlenir [6,17,19].

4. UYGULAMA

4.1. Veri Seti

Yapılan çalışmada veriler, 7 adet ekonomik değişken için Ocak 1987 – Haziran 2001 tarihleri arasında aylık dönemler halinde elde edilen gözlem değerlerinden oluşmaktadır. Bu çalışmada İstanbul Menkul Kıymetler Borsası 100 endeksi (IMKB100) üzerinde etkili olduğu düşünülen 6 adet bağımsız değişken bulunmaktadır. Sırasıyla bu değişkenler; 3 ay vadeli mevduat faiz oranı (MEVDUAT), 3 aylık hazine bonosu faiz oranı (H.BONOSU), dolar alış ve satış değerlerinin ortalaması (DOLAR), külçe altın satış fiyatı (K.ALTIN), tüketici fiyat endeksi (TUFE) ve para arzı M2Y (nakit + vadesiz mevduat + vadeli mevduat + döviz tevdiat hesabı) (M2Y)'dir. Verinin derlenmesi aşamasında internet kullanılmıştır. İlgilenilen değişkenler için gerekli verilere ulaşmada oldukça detaylı bir veri tabanına sahip olan Türkiye Cumhuriyeti Merkez Bankası Veri Dağıtım Merkezi (www.tcmb.gov.tr) kullanılmıştır.

Regresyon denkleminin oluşturulmasında, bağımsız değişkenlerin seçimi yapılırken, bağımlı değişken olan IMKB100 için en çok açıklayıcılığı elde etmek amaçlanmış ve bu doğrultuda ekonomik hayatın gerçeklerine ters düşmeyecek şekilde gereken bağımsız değişkenlerin seçimi iktisatçı bir danışman gözetiminde belirlenmiştir. Modelde kullanılan 6 bağımsız değişkenin haricindeki tüm etkileri konjektür etkisi olarak nitelendirmek mümkündür.

Uygulamada Huber-M tahminlerinin yapılmasında S-Plus istatistik paket programının bir alt fonksiyonu olan rreg fonksiyonu kullanılmıştır. rreg fonksiyonunun açık hali Ek-1'de verilmiştir.

4.2. Hesaplamaların Yapılması

Çoklu regresyonda Huber-M tahminlerini elde etmede kullanılacak algoritma Bölüm 3'de verilen Algoritma 2'dir. Buna göre;

$$\begin{aligned} \text{İmkb100} = \hat{\beta}_0 + \hat{\beta}_1 \text{Mevduat} + \hat{\beta}_2 \text{H.Bonosu} + \hat{\beta}_3 \text{Dolar} + \\ \hat{\beta}_4 \text{K.Altın} + \hat{\beta}_5 \text{Tüfe} + \hat{\beta}_6 \text{M2Y} + e_i \end{aligned} \qquad (4.2.1)$$

şeklinde oluşturulacak regresyon tahmin denklemine ilişkin katsayı değerleri (3.4) veya (3.5) denklemini minimize eden değerler vektörü olacaktır. S-Plus paket

programının rreg fonksiyonu kullanılarak yapılan minimizasyon işlemine göre algoritmanın her bir adımına ilişkin hesaplanılan yakınsama değerleri Çizelge 4.1.'de verilmiştir.

Çizelge 4.1. Minimizasyon Algoritmasının Her Bir Adımı İçin Bulunan Yakınsama Değerleri

Adım No	Yakınsama Değeri
1	0,192754576
2	0,129088472
3	0,084648845
4	0,057955208
5	0,036670618
6	0,024342146
7	0,019167758
8	0,012800150
9	0,009761339
10	0,007261262
11	0,006574738
12	0,006059545
13	0,005431062
14	0,004911862
15	0,004455617
16	0,004041710
17	0,003286636
18	0,162759020
19	0,070137949

Minimizasyon algoritmasının 17. adımı sonunda elde edilen değer yeterince küçük olduğu için algoritma için gerekli yakınsama elde edilmiş olur. 17. adım sonunda elde edilen katsayı tahminleri Çizelge 4.2.'de verilmiştir.

Çizelge 4.2. Huber-M Tahmin Değerlerine İlişkin Katsayı Değerleri

SABİT	104.19910
H.BONOSU	-2.97563
K.ALTIN	0.00075
M2Y	0.00010
DOLAR	-0.02584
MEVDUAT	1.02110
TUFE	0.13538

Bu sonuçlara göre elde edilen Huber-M tahmin denklemi;

$$\begin{aligned} \text{İmkb100} = {} & 104{,}19910 - 2{,}97563\text{H.Bonosu} + 0{,}00075\text{K.Altın} \\ & + 0{,}00010\text{M2Y} - 0{,}02584\text{Dolar} + 1{,}02110\text{Mevduat} \\ & + 0{,}13538\text{Tüfe} \end{aligned} \qquad (4.2.2)$$

şeklinde olur. Tahminlenen regresyon modeli incelendiğinde İMKB100 bağımlı değişkeninin hazine bonosu faiz oranı ve ortalama dolar fiyatı değişkenleri ile ters yönlü bir ilişkiye sahip olduğu, diğer bağımsız değişkenler ile doğru yönlü bir ilişkiye sahip olduğu görülür. Model parametrelerinin tahmini sırasında ilgili algoritmanın başlangıç ve son adımları için hesaplanılan w_i ağırlık değerleri Ek-2'de verilmiştir.

Modeldeki tüm değişkenler için hesaplanılmış korelasyon matrisi Çizelge 4.3.'de verilmiştir. Ayrıca tüm değişkenlerin kendi aralarındaki ikili grafik çizimleri matris formatında Şekil 4.1.'de verilmiştir.

Çizelge 4.3. Tüm Değişkenlere İlişkin Korelasyon Matrisi

	IMKB100	H.BONOSU	K.ALTIN	M2Y	DOLAR	MEVDUAT	TUFE
IMKB100	1	-0,11552	0,89257	0,90221	0,88576	-0,06166	0,91395
H.BONOSU	-0,11552	1	0,02520	-0,07065	-0,00339	0,81819	-0,03318
K.ALTIN	0,89257	0,02520	1	0,98731	0,99687	0,18414	0,99088
M2Y	0,90221	-0,07065	0,98731	1	0,99348	0,11474	0,99397
DOLAR	0,88576	-0,00339	0,99687	0,99348	1	0,17083	0,99252
MEVDUAT	-0,06166	0,81819	0,18414	0,11474	0,17083	1	0,14535
TUFE	0,91395	-0,03318	0,99088	0,99397	0,99252	0,14535	1

Huber-M tahmin denklemi sonucunda elde edilen artıkların incelenmesi, bir diğer teknik olan ve çok daha yaygın bir kullanım alanına sahip En Küçük

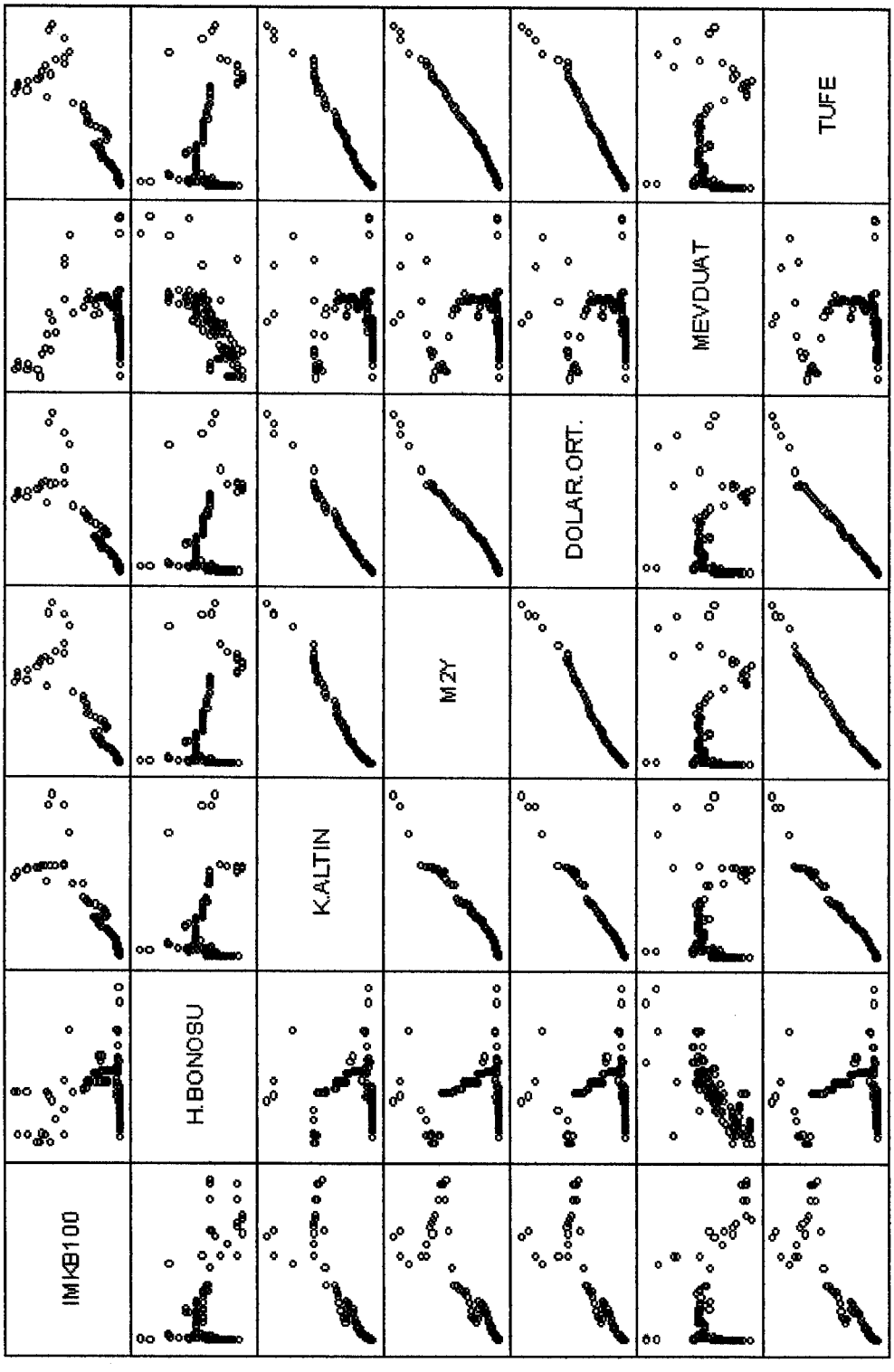
Kareler Tekniği sonucunda elde edilen tahmin denklemi artık değerleri ile karşılaştırmalı olarak şekiller yardımıyla incelenmiştir.

Buna göre her iki tekniğin kullanımı ile elde edilen artık değerlerine ilişkin çizgi grafiği, Şekil 4.2.'de verilmiştir. Bağımlı değişken İMKB100 değerlerine karşı artık değerlerinin grafiği Şekil 4.3.'de, modelden elde edilen $\hat{y}_i$ tahmin değerlerine karşı artıkların grafiği ise Şekil 4.4.'de verilmiştir.

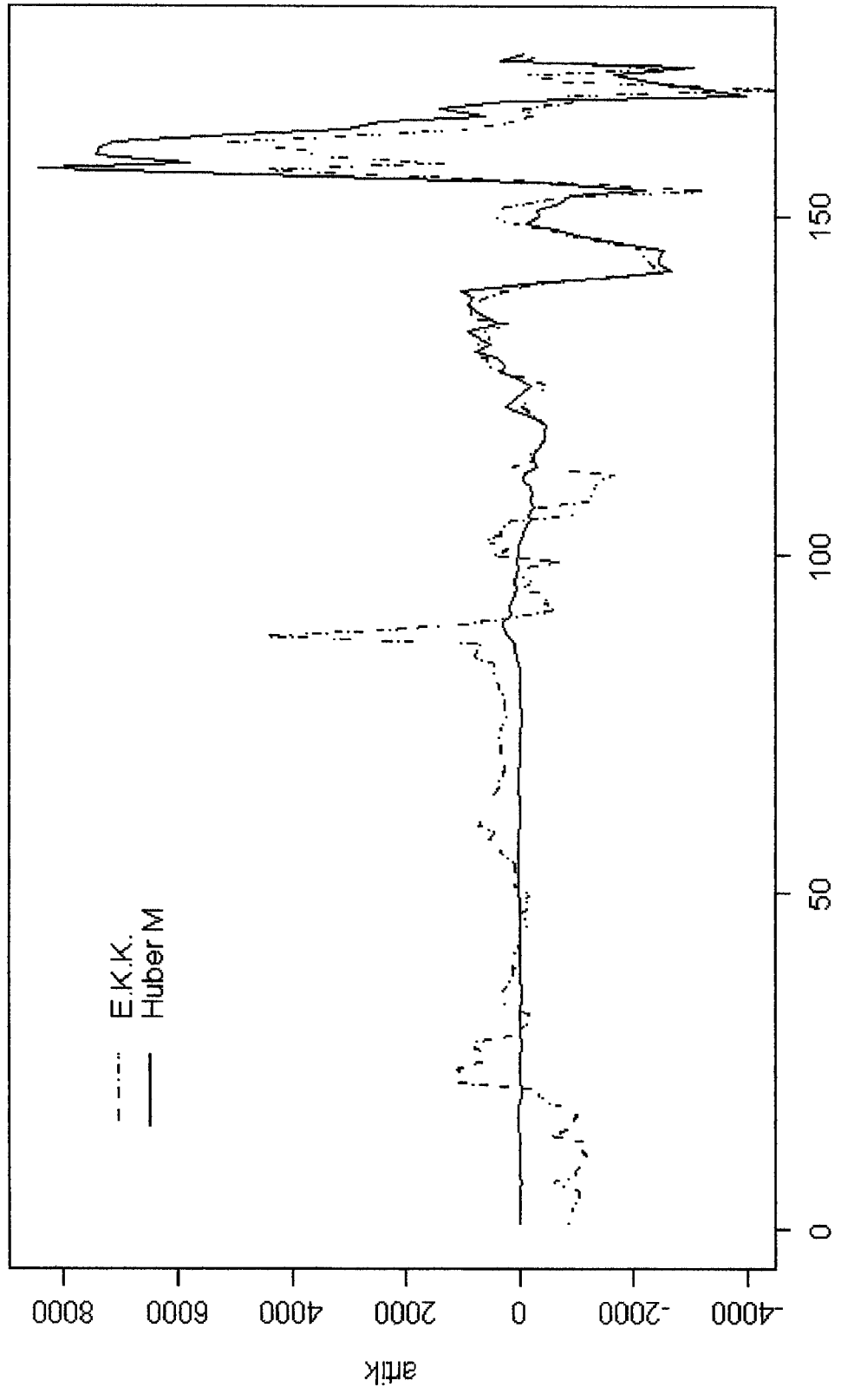
Elde edilen artık değerleri için tahmin edilen olasılık yoğunluk fonksiyonu grafiği Şekil 4.5.'de verilmiştir. En küçük kareler tekniğinin uygulanması sonucunda elde edilen Student türü artık değerleri ile normal dağılım kümülatif olasılık dağılım fonksiyonu kantil değerlerinin grafiksel karşılaştırması ise Şekil 4.6'da Q-Q grafiği şeklinde verilmiştir.

Artıklar için çizilen Şekil 4.5 'teki olasılık yoğunluk fonksiyonlarının incelenmesi sonucunda görülebileceği gibi Huber tahmincileri çok daha dar bir dağılıma sahiptir. Bu da aykırı değerlerinin varlığının bilindiği veri setleri için istenen bir özelliktir. Ayrıca dağılımın sağ taraf kuyruğunun bilinen ve beklenen dağılım formundan daha uzun olması sapma değerlerinin olması gerekenden daha büyük değerlere sahip olduğunu dolayısıyla normalliğin bozulduğunu da ifade etmektedir. Aynı şekilde Şekil 4.6'daki student türü artık değerleri üzerinden çizilen Q-Q grafiği incelendiğinde de artık değerlerinin normal dağılıma sahip olmadıkları anlaşılmaktadır.

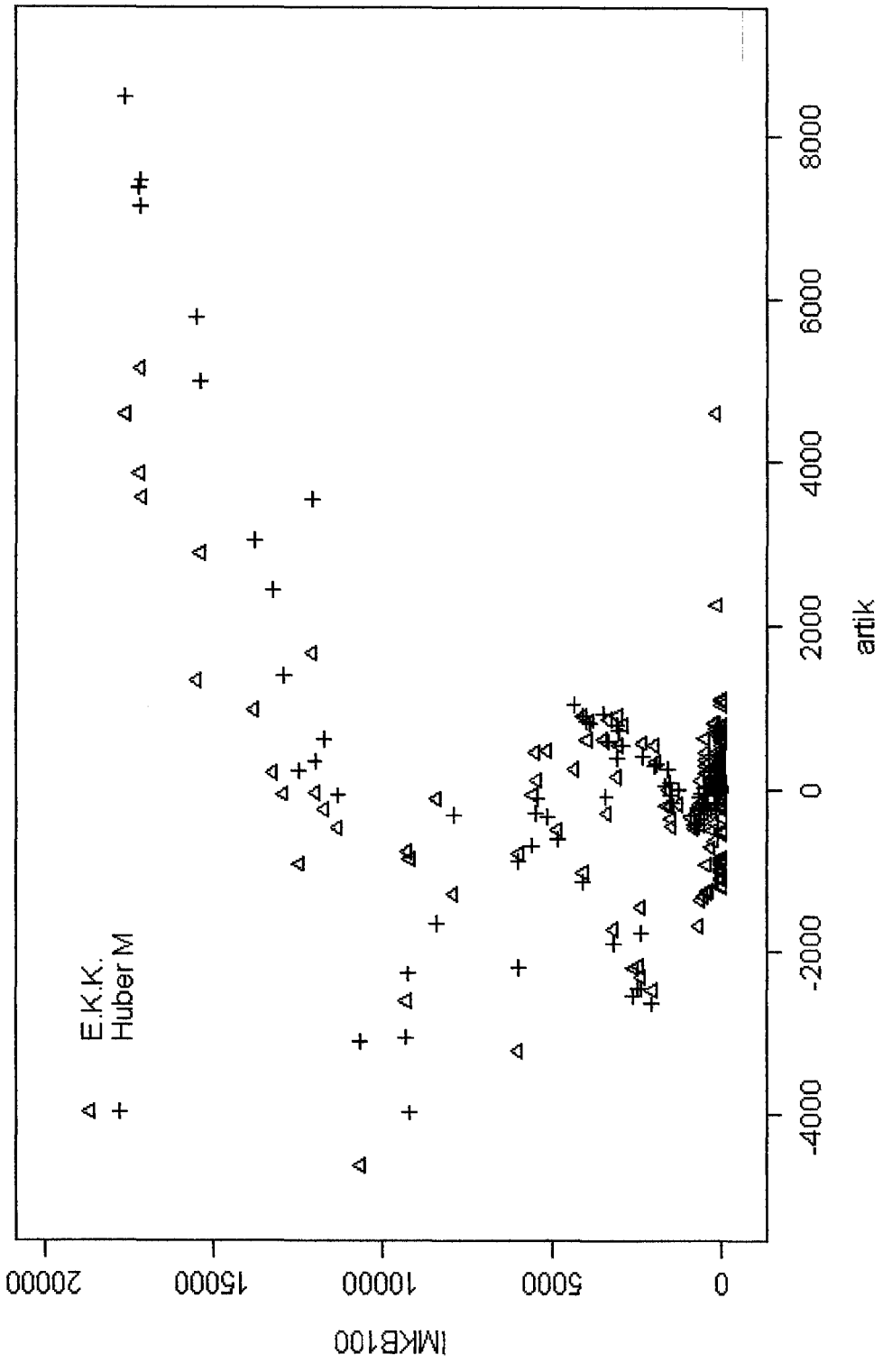
Grafiklerin yanısıra en küçük kareler artık değerlerinin dağılımının normal dağılıma uyup uymadığı tek örneklem Kolmogorov-Simirnov iyi uyum testi ile incelenebilir. Test sonucuna bakıldığında elde edilen test istatistiği değeri $ks = 0{,}1489$ ve buna ilişkin olasılık ise $p = 0{,}000$ olarak elde edilir. Buna göre en küçük kareler artık değerlerinin dağılımının standart normal dağılıma uymadığı söylenir. Dolayısıyla en küçük kareler tekniğinin uygulanabilmesi için gereken varsayımın sağlanmadığı görülür.



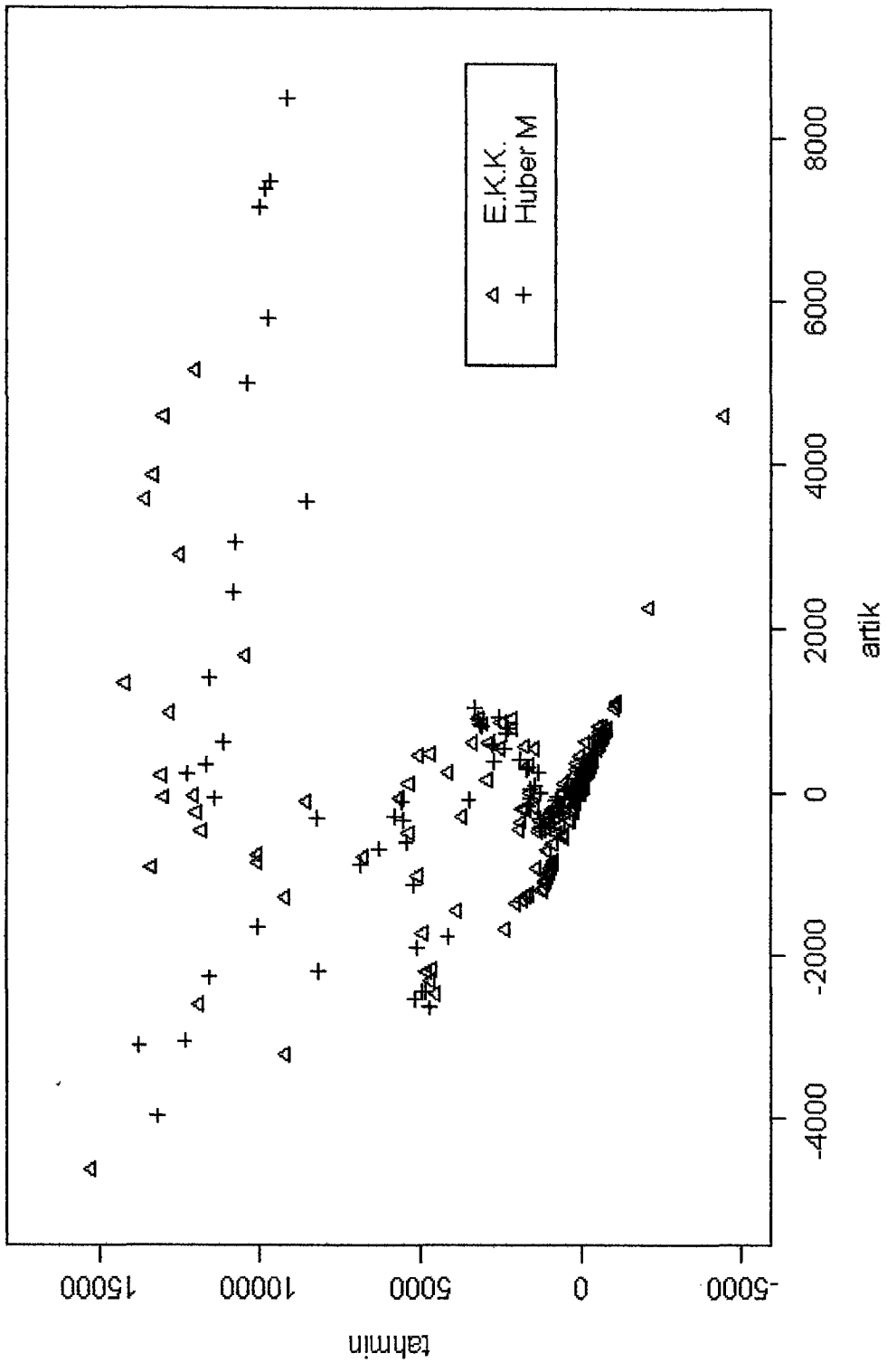
Şekil 4.1. Tüm Değişkenler İçin Matris Grafiği



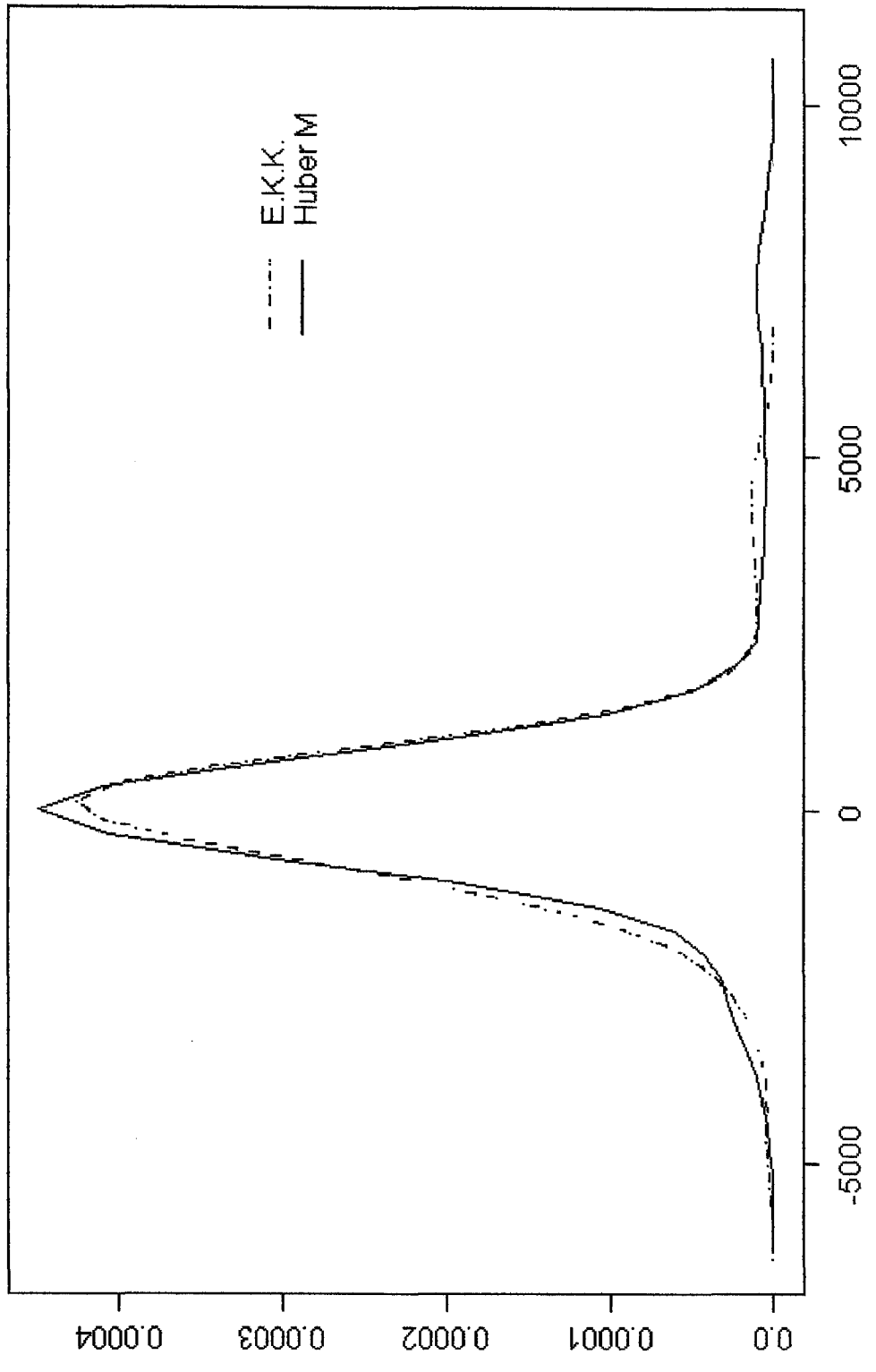
Şekil 4.2. Artık Değerleri



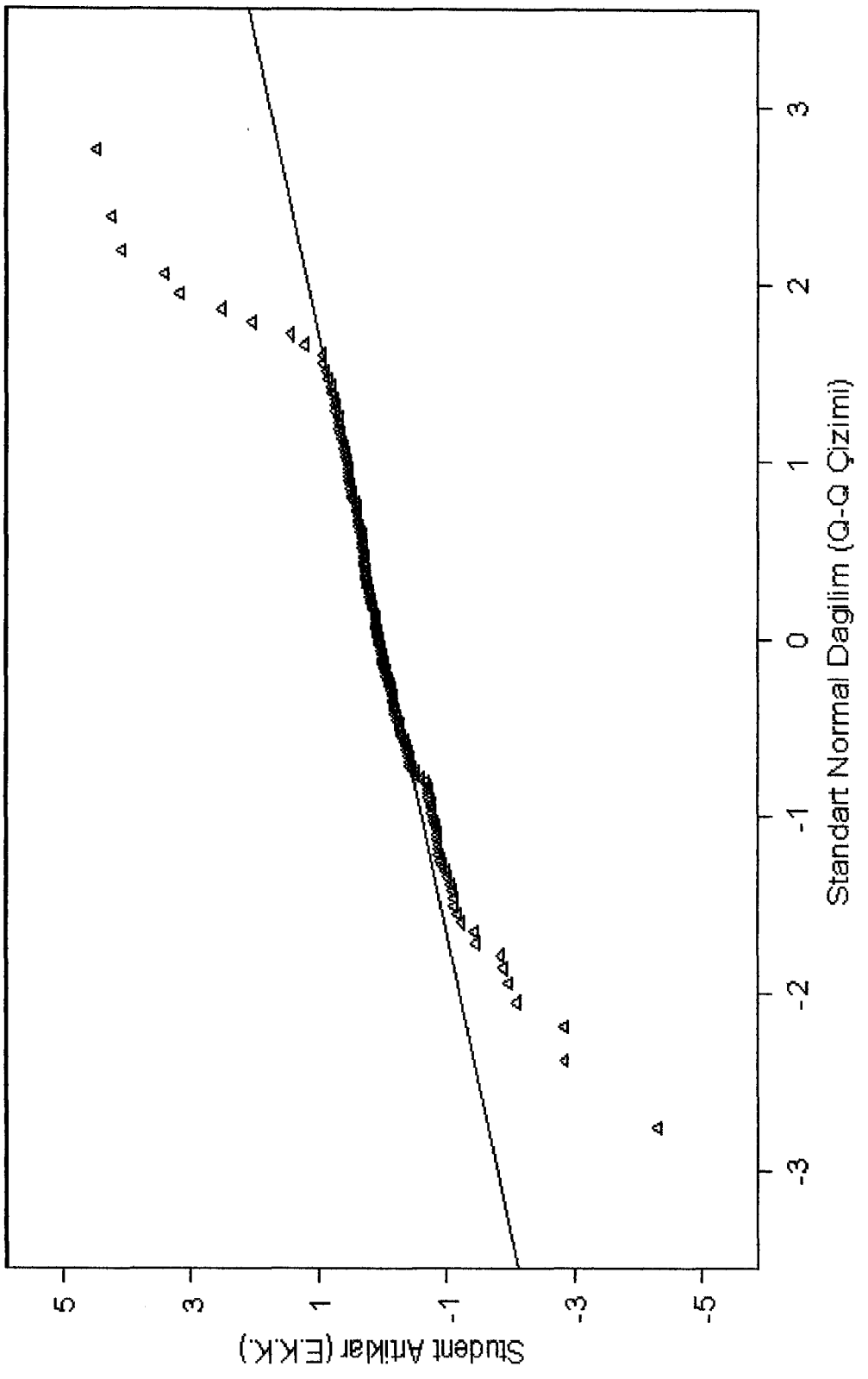
Şekil 4.3. Bağımlı Değişken İMKB100 Değerlerine Karşı Artıklar



Şekil 4.4. Tahmin Değerlerine Karşı Artıklar



Şekil 4.5. Artıklar İçin Olasılık Yoğunluk Fonksiyonu



Şekil 4.6. E.K.K. Student Artık Değerleri İçin Q-Q Grafiği

Katsayıların anlamlılık sınamaları için, tahmin edilen regresyon denklemi artık değerleri için Mutlak Sapmalar Medyanı (MSM) 454,2198 olarak hesaplanmıştır. Dolayısıyla standart sapma tahmin değeri MSM değerinin Huber tarafından önerilmiş olan 1,483 değeri ile çarpımına eşittir ki bu da $\hat{\sigma} = 673,6080$ olacaktır. Buna göre $\hat{e}_i^*$ artık değerleri, mutlak değer olarak , $1,5\hat{\sigma} = 1010,4120$ değerinden büyük olan artık değerleri olacaktır. Ele alınan model için 35 adet artık değeri $1,5\hat{\sigma}$ değerinden mutlak olarak daha büyük değere sahiptirler. İşlemler sonucunda elde edilen $\hat{\lambda}$ değeri 1521096,849 ve DHT_F değeri 144875644,8 olarak karşımıza çıkmaktadır. Tahmin edilen tüm modelin özet değerleri Çizelge 4.4.'de verilmiştir.

Çizelge 4.4. Tüm Model İçin Özet İstatistikler

MSM	$\hat{\sigma}$	DHT_F	$\hat{\lambda}$
454,22	673,608	144875644,8	1521096,849

Tahmin edilen asıl modelden bir bağımsız değişkenin çıkarılması sonucu oluşan indirgenmiş modellere ilişkin dönüştürülmüş hatalar toplamları, hesaplanan F istatistiği değerleri ve bunlara ilişkin olasılık değerleri Çizelge 4.5.'de verilmiştir.

Çizelge 4.5. Tüm Modelden Bir Bağımsız Değişkenin Çıkarılması İle Oluşan İndirgenmiş Modellere İlişkin Özet İstatistikler

Modelden Çıkarılan Değişken	DHTi	Serbestlik Dereceleri	F	P
MEVDUAT	149980245,4	1 ; 167	3,356	0,069
H.BONOSU	152005171,5	1 ; 167	4,687	0,032
DOLAR	173607165,7	1 ; 167	18,889	2,401 E-5
ALTIN	154019172,2	1 ; 167	6,011	0,015
TÜFE	174658562,6	1 ; 167	19,579	1,735 E-5
M2Y	148650229,6	1 ; 167	2,482	0,117

Aynı şekilde tahmin edilen tüm modelden iki bağımsız değişkenin çıkarılması sonucu oluşan indirgenmiş modellerin dönüştürülmüş hatalar toplamları, hesaplanmış F istatistiği değerleri ve bunlara ilişkin olasılık değerleri ise Çizelge 4.6.'da verilmiştir.

Çizelge 4.6. Tüm Modelden İki Bağımsız Değişkenin Çıkarılması İle Oluşan İndirgenmiş Modellere İlişkin Özet İstatistikler

Modelden Çıkarılan Değişkenler	DHTi	Serbestlik Dereceleri	F	P
MEVDUAT ve H.BONOSU	152777932,6	2 ; 167	5,195	0,006
MEVDUAT ve DOLAR	184915831,6	2 ; 167	26,323	1,156 E-10
MEVDUAT ve ALTIN	162523701,7	2 ; 167	11,602	1,914 E-5
MEVDUAT ve TÜFE	170263447,5	2 ; 167	16,691	2,465 E-7
MEVDUAT ve M2Y	159590847,9	2 ; 167	9,674	1,059 E-4
H.BONOSU ve DOLAR	183234004,1	2 ; 167	25,218	2,690 E-10
H.BONOSU ve ALTIN	169662596,6	2 ; 167	16,295	3,428 E-7
H.BONOSU ve TÜFE	175877120,5	2 ; 167	20,381	1,202 E-8
H.BONOSU ve M2Y	154073976,2	2 ; 167	6,047	2,914 E-3
DOLAR ve ALTIN	179080720,7	2 ; 167	22,487	2,250 E-9
DOLAR ve TÜFE	196104922,7	2 ; 167	33,679	5,150 E-13
DOLAR ve M2Y	176510161,8	2 ; 167	20,797	8,611 E-9

Çizelge 4.6. **(Devam)** Tüm Modelden İki Bağımsız Değişkenin Çıkarılması İle Oluşan İndirgenmiş Modellere İlişkin Özet İstatistikler

Modelden Çıkarılan Değişkenler	DHTi	Serbestlik Derecesi	F	P
ALTIN ve TÜFE	205080723,5	2 ; 167	39,580	8,514 E-15
ALTIN ve M2Y	154170499,4	2 ; 167	6,111	2,747 E-3
TÜFE ve M2Y	204864801,3	2 ; 167	39,438	9,376 E-15

Çizelge 4.5 incelendiğinde $\alpha = 0,05$ anlam düzeyinde Mevduat ve M2Y değişkenlerinin modelde yer almalarının istatistiksel olarak anlamlı olmadığı görülmektedir. Bağımsız değişkenlerin ikili kombinasyonlarının yer aldığı Çizelge 4.6.'nın incelenmesi sonucunda ise tahmin edilen modeldeki bütün bağımsız değişkenlerin model için istatistiksel olarak anlamlı ve önemli oldukları ve bu değişkenlerin bağımlı değişken olan İMKB100 değişkeni ile önemli derecede bir ilişkilerinin olduğunu söylemek mümkündür.

5. SONUÇ VE ÖNERİLER

Günümüz Türkiye ekonomisinde ekonomik göstergeler çok hızlı bir değişim göstermektedir. Ekonomik göstergelerde meydana gelen bu hızlı değişimler alışılagelen yapıdan çok daha farklı veri yapılarını da karşımıza çıkarmaktadır. Dolayısıyla bu tür veri yapıları kullanılarak yapılacak çalışmalarda artık kolaylıkla ulaşılabilen Robust regresyon tekniklerinin kullanımı da giderek yaygın hale gelmektedir. Bu çalışmada, aykırı değerlerin bulunduğu veri setlerinde Robust bir teknik olan Huber'ın M regresyon tekniği kullanılarak İMKB100 endeksini etkileyen 6 değişken için bir regresyon modeli oluşturulmuştur. Belirlenen model iktisadi açıdan ele alındığında şu şekilde yorumlanabilir.

Türkiye gibi kamu açıklarının finansmanında ağırlıklı olarak devlet iç borçlanma senetlerinin kullanıldığı ülkelerde hazine bonosuna yatırımla hisse senedi yatırımları hem bireysel hem de kurumsal yatırımcılar açısından yakın ikame oluşturmaktadır. Öyle ki, kurumsal yatırımcıların portföyleri içerisinde bu iki yatırım aracı neredeyse eşit ve büyük oranlı ağırlık oluşturmaktadır. Bu nedenle, hazine bonosu faiz oranları ile hisse senedi getirileri arasında ters yönlü bir ilişki söz konusudur. Altına yatırım yapanlar ile hisse senetlerine yatırım yapanların farklı bölgelerden farklı sosyal gruplar oluşturmalarından dolayı külçe altın fiyatının imkb100 endeksi üzerindeki etkisinin çok düşük olmasını da beraberinde getirmektedir. M2Y para arzının ise ilk bakışta İMKB endeksi üzerinde pozitif etkiye sahip olması gerekir. Çünkü, M2Y M2'den farklı olarak döviz tevdiat hesabını da içerir. Yoğun para ikamesinin (dolarlaşmanın) söz konusu olduğu ülkemizde, dolar alternatif bir yatırım aracı olarak genel kabul görmektedir. Ancak buradaki pozitif ilişki yabancı yatırımcıların TL pozisyonu alıp hisse senedi piyasasına yatırım yapmalarından kaynaklanmaktadır. Öte yandan, yukarıda belirtildiği gibi, yoğun para ikamesi (dolarlaşma) nedeniyle TL-Dolar kurundaki değişmeler ise negatif bir etki yaratmaktadır. Özellikle hisse senedi piyasalarında oluşan bir iktidarsızlık sonucu yabancı yatırımcıların hisse senedi satıp, dolara yönelmeleri bu durumu desteklemektedir. Mevduat faiz oranları bulunanın tam aksine negatif yönlü bir ilişki içerisinde olmalıdır. Tüketici

fiyat endeksindeki artışlar, İMKB100'e dahil şirketlerin gelirlerindeki bir artışı yansıtacağı için ayrıca enflasyon ile birlikte bazı şirketlerin kar marjları arttığından TÜFE ile İMKB100 arasındaki ilişkinin pozitif olması da ülkemiz koşullarında mantıklı görünmektedir. Ancak, iktisat teorisine göre değerlendirildiğinde bu ilişkinin negatif yönlü olması beklenir.

KAYNAKLAR

1. BEKKİ, A. ve ER, F., *S-Plus ve R Paket Programlarında Modern Regresyon Teknikleri*, V. Ulusal Biyoistatistik Kongresi, 13-15 Eylül 2000, Eskişehir.

2. BEKKİ, A. ve ER F., *Ekonomik Veriler Üzerinde Bazı Regresyon Tekniklerinin Kullanımı*, II. İstatistik Kongresi, 2-6 Mayıs 2001, Belek-Antalya.

3. ŞIKLAR, E., *Regresyon Analizine Giriş*, Anadolu Üniversitesi, Fen Fakültesi Yayınları No:16, Eskişehir, 2000.

4. ERAR, A., *Regresyon (Bağlanım) Çözümlemesi*, Ders Notları, Ankara, 1985.

5. MYERS, R.H., *Classical and Modern Regression with Applications*, Duxbury Press, 1986.

6. BIRKES, D. ve DODGE, Y., *Alternative Methods of Regression*, John Wiley & Sons Inc., New York, 1993.

7. DRAPER, N.R. ve SMITH, H., *Applied Regression Analysis*, John Wiley & Sons Inc, New York, 1981.

8. GUJARATI, D.N., *Basic Econometrics*, Literatür Yayıncılık, İstanbul, 1995.

9. ŞENESEN, Ü. ve ŞENESEN, G.G., *Ekonometri Kuramı*, İstanbul Teknik Üniversite Matbaası, Gümüşsuyu, 1992.

10. AKKAYA, Ş. ve PAZARCIOĞLU, V., *Ekonometri I*, Anadolu Matbaacılık, İzmir, 1995.

11. GREEN, P.J., *Iteratively Reweighted Least Squares for Maximum Likelihood Estimation and Some Robust and Resistant Alternatives*, J.R. Statist. Soc., **46**, 149-192, 1984.

12. KMENTA, J., *Elements of Econometrics*, Macmillan Publishing Company, New York, 1986.

13. KABE, D.G. ve GUPTA, R.P., *Alternatives to Least Squares in Multiple Regression*, Multivariate Statistical Inference Proceeding of the Research Seminar at Dalhousie University, **25-40**, Halifax, 23-25 March 1972.

14. WEISBERG, S., *Applied Linear Regression*, John Wiley, Canada, 1980.

15. BEHNKEN, D.W. ve DRAPER, N.R., *Residuals and Their Variance Patterns*, Technometrics, **14**, 101-111, 1972.

16. BARNETT, V. ve LEWIS T., *Outliers in Statistical Data*, John Wiley, Chichester, 1998.

17. CROW, E.L. ve SIDDIQUI, M.M., *Robust Estimation of Location*, American Statistical Association Journal, 353-389, 1967.

18. SIEGEL, A.F., *Robust Regression Using Repeated Medians*, Biometrica, **69**, 242-244, 1982.

19. HILL, M.A. ve DIXON, W.J., *Robustness In Real Life: A Study Of Clinical Laboratory Data*, Biometrics, **38**, 377-396, June-1982.

20. WANG, J.L., *Asymptotic Properties of M-Estimators Based on Estimating Equations and Cencored Data*, Scandinavian Journal of Statistics, **26**, 297-318, Oxford, 1999.

21. VENABLES, W.N. ve RIPLEY, B.D., *Modern Applied Statistics with S-Plus*, Springer-Verlag, New York, 1996.

22. MATHSOFT, *S-Plus 4 Guide To Statistics*, Mathsoft Inc., Washington, 1997.

Ek-1 : S-Plus İstatistik Paket Programında Huber-M Tahmincileri İçin Düzenlenmiş *rreg* Fonksiyonu

```
function(x, y, w = rep(1, n), int = TRUE, init = lsfit.simp(x, y, w, n,
        p)$coef, method = wt.default, wx, iter = 20, acc = 10 *
        .Machine$single.eps^0.5, test.vec = "resid")
{
        irls.delta <- function(old, new)
        {
            a <- sum((old - new)^2)
            b <- sum(old^2)
            if(b >= 1 || a < b * .Machine$double.xmax)
                    sqrt(a/b)
            else .Machine$double.xmax
        }
        irls.rrxwr <- function(x, w, r)
        {
            w <- sqrt(w)
            max(abs((as.vector(r * w) %*% x)/sqrt(as.vector(w) %*% (
                    x^2))))/sqrt(sum(w * r^2))
        }
        lsfit.simp <- function(x, y, wt, n, p)
        {
            wt.factor <- as.vector(wt^0.5)
            wt.zero <- wt.factor == 0
            x0 <- x[wt.zero, , drop = F]
            y0 <- y[wt.zero]
            x <- x * wt.factor
            y <- y * wt.factor
            inv.wt.factor <- 1/ifelse(wt.zero, 1, wt.factor)
            z <- .Fortran("dqrls",
                    qr = as.double(x),
                    as.integer(c(n, p)),
```

Ek-1 (Devam)

```
                pivot = as.integer(1:p),
                qraux = double(p),
                y,
                as.integer(c(n, 1)),
                coef = double(p),
                residuals = y,
                qt = y,
                tol = as.double(1e-007),
                double(2 * p),
                rank = as.integer(p))[c("coef", "residuals",
                "pivot", "rank")]
        if(z$rank < p) {
                xn <- names(z$coef)
                z$coef <- z$coef[z$pivot]
                names(z$coef) <- xn
        }
        z$residuals <- z$residuals * inv.wt.factor
        if(any(wt.zero)) {
                z$residuals[wt.zero] <- y0 - x0 %*% z$coef
        }
        z
}
if(!(any(test.vec == c("resid", "coef", "w", "NULL")) ||
        is.null(test.vec)))
        stop("invalid testvec")
if(int)
        x <- cbind("(Intercept)" = 1, x)
else x <- as.matrix(x)
cnames <- dimnames(x)[[2]]
n <- dim(x)[1]
p <- dim(x)[2]
```

Ek-1 (Devam)

```
if(length(y) != n)
        stop("length of y is not equal to number of rows in x")
specials.x <- !is.finite(x %*% rep(1, p))
specials.y <- !is.finite(y)
if(missing(wx)) {
        specials.wt <- F
}
else {
        if(length(wx) != n)
                stop("Length of wx must equal number of observat
ions"
                        )
        specials.wt <- !is.finite(wx)
        if(any(wx[!specials.wt] < 0))
                stop("Negative wx value")
        w <- w * wx
}
ok <- !(specials.x | specials.y | specials.wt)
if((bad.obs <- sum(!ok)) > 0)
        warning(paste(bad.obs,
                "observations with NA/NaN/Inf in x, y, or wx rem
oved."
                ))
fitted.out <- y
fitted.out[!ok] <- NA
resid.out <- wt.out <- y * NA
y <- y[ok]
x <- x[ok,  , drop = F]
w <- w[ok]
if(!missing(wx))
        wx <- wx[ok]
```

Ek-1 (Devam)

```
        n <- dim(x)[1]
        if(n < p)
            stop("not enough usable observations")
        coef <- init
        if(p != length(coef))
            stop("Must have same number of initial values as coeffic
ients"
                 )
        resid <- y - x %*% coef
        converged <- FALSE
        status <- "converged"
        conv <- NULL
        method.in.control <- method.exit <- FALSE
        if(iter > 0) {
            for(iiter in 1:iter) {
                if(!is.null(test.vec))
                    previous <- get(test.vec)
                scale <- median(abs(resid))/0.6745
                if(scale == 0) {
                    convi <- 0
                    method.exit <- TRUE
                    status <-
                      "could not compute scale of residuals"
                }
                else {
                    w <- method(resid/scale)
                    if(!missing(wx))
                      w <- w * wx
                    temp <- lsfit.simp(x, y, w, n, p)
                    coef <- temp$coef
                    resid <- temp$residuals
```

Ek-1 (Devam)

```
                    if(!is.null(test.vec))
                      convi <- irls.delta(previous, get(
                        test.vec))
                    else convi <- irls.rrxwr(x, w, resid)
                }
                conv <- c(conv, convi)
                converged <- convi <= acc
                done <- method.exit || (converged && !
                        method.in.control)
                if(done)
                        break
        }
        if(!done)
                warning(status <- paste("failed to converge in",
                        iter, "steps"))
    }
    if(!missing(wx)) {
        tmp <- (wx != 0)
        w[tmp] <- w[tmp]/wx[tmp]
    }
    resid.out[ok] <- resid
    fitted.out[ok] <- fitted.out[ok] - resid
    wt.out[ok] <- w
    names(coef) <- cnames
    list(coefficients = coef, residuals = resid.out, fitted.values
         = fitted.out, w = wt.out, int = int, conv = conv,
        status = status)
}
```

Ek-2 : Çoklu Huber-M Tahmin Denklemi Bulunurken Başlangıç ve Bitiş Adımları İçin Hesaplanan Ağırlık Değerleri

Gözlem No	Başlangıç Ağırlık Değerleri	Bitiş Ağırlık Değerleri
1	0.7882	0.9590
2	0.7876	0.9601
3	0.7344	0.9641
4	0.6845	0.9677
5	0.6480	0.9700
6	0.6491	0.9713
7	1.0000	0.9423
8	0.7129	0.9729
9	0.6716	0.9740
10	0.6028	0.9772
11	0.5692	0.9782
12	0.6167	0.9717
13	0.5573	0.9772
14	1.0000	0.9981
15	0.7908	1.0000
16	0.7820	1.0000
17	0.6617	0.9992
18	0.7484	0.9999
19	1.0000	0.9976
20	1.0000	0.9803
21	1.0000	0.9798
22	0.6197	0.9744
23	0.6591	0.9752
24	0.6243	0.9820
25	0.9878	0.9869
26	1.0000	0.9896
27	0.8826	0.9999
28	0.8573	0.9950
29	1.0000	0.9998
30	1.0000	0.9998
31	1.0000	0.9986
32	1.0000	0.9986
33	1.0000	1.0000
34	1.0000	0.9973
35	1.0000	0.9932
36	1.0000	0.9907
37	1.0000	0.9918
38	1.0000	0.9942
39	1.0000	0.9926
40	1.0000	0.9910
41	1.0000	0.9935
42	1.0000	0.9962
43	1.0000	0.9989
44	1.0000	0.9994

Ek-2 (Devam)

Gözlem No	Başlangıç Ağırlık Değerleri	Bitiş Ağırlık Değerleri
45	1.0000	0.9994
46	1.0000	0.9997
47	1.0000	0.9999
48	1.0000	0.9999
49	1.0000	1.0000
50	1.0000	0.9924
51	1.0000	0.9701
52	1.0000	0.9648
53	1.0000	0.9621
54	1.0000	0.9662
55	1.0000	0.9705
56	1.0000	0.9686
57	1.0000	0.9682
58	1.0000	0.9516
59	1.0000	0.9422
60	0.9465	0.9589
61	1.0000	0.9666
62	1.0000	0.9752
63	1.0000	0.9773
64	1.0000	0.9701
65	1.0000	0.9581
66	1.0000	0.9483
67	1.0000	0.9471
68	1.0000	0.9592
69	1.0000	0.9741
70	1.0000	0.9829
71	1.0000	0.9827
72	1.0000	0.9821
73	1.0000	0.9845
74	1.0000	0.9927
75	1.0000	0.9993
76	1.0000	0.9988
77	1.0000	0.9996
78	1.0000	0.9949
79	1.0000	0.9976
80	1.0000	0.9971
81	1.0000	0.9899
82	1.0000	0.9938
83	1.0000	0.9967
84	1.0000	0.9883
85	0.8494	0.9398
86	0.9089	0.7536
87	1.0000	0.8144
88	0.1464	0.0515
89	0.2980	0.0031

Ek-2 (Devam)

Gözlem No	Başlangıç Ağırlık Değerleri	Bitiş Ağırlık Değerleri
90	0.8410	0.0284
91	1.0000	0.7705
92	1.0000	0.8220
93	1.0000	0.9040
94	1.0000	0.9884
95	1.0000	0.9613
96	1.0000	0.9343
97	1.0000	0.9107
98	1.0000	0.9446
99	0.9326	0.9513
100	1.0000	0.9965
101	1.0000	0.9996
102	1.0000	1.0000
103	1.0000	0.9657
104	1.0000	0.9321
105	1.0000	0.7178
106	0.7254	0.7720
107	1.0000	0.6278
108	0.5406	0.8109
109	0.5274	0.6996
110	0.5119	0.7965
111	0.4900	0.9479
112	0.4003	0.8688
113	1.0000	0.3599
114	1.0000	0.5099
115	1.0000	0.5349
116	1.0000	0.1943
117	1.0000	0.0646
118	1.0000	0.0660
119	1.0000	0.0316
120	1.0000	0.2691
121	1.0000	0.9601
122	1.0000	0.8761
123	1.0000	0.9991
124	1.0000	0.8314
125	1.0000	0.5808
126	1.0000	0.9684
127	1.0000	0.7034
128	1.0000	0.7829
129	1.0000	0.5585
130	0.7485	0.0000
131	0.8725	0.1935
132	1.0000	0.0000
133	1.0000	0.0000
134	1.0000	0.6239

Ek-2 (Devam)

Gözlem No	Başlangıç Ağırlık Değerleri	Bitiş Ağırlık Değerleri
135	0.7888	0.1315
136	0.8047	0.0000
137	0.7544	0.0000
138	1.0000	0.0000
139	1.0000	0.0000
140	1.0000	0.7785
141	0.4602	0.0000
142	0.2726	0.0000
143	0.2912	0.0000
144	0.3102	0.0000
145	0.3042	0.0000
146	0.3906	0.0000
147	0.6548	0.0000
148	1.0000	0.0000
149	1.0000	0.8210
150	1.0000	0.4837
151	1.0000	0.6587
152	1.0000	0.0000
153	0.8439	0.0000
154	0.2087	0.0000
155	0.5139	0.1372
156	0.4066	0.0000
157	0.1472	0.0000
158	0.5131	0.0000
159	0.1888	0.0000
160	0.1748	0.0000
161	0.1310	0.0000
162	0.2344	0.0000
163	0.7081	0.0000
164	1.0000	0.0000
165	1.0000	0.0175
166	1.0000	0.0000
167	0.7253	0.7707
168	0.7772	0.0000
169	0.1460	0.0000
170	0.8575	0.0000
171	1.0000	0.0000
172	0.2584	0.0000
173	1.0000	0.2721
174	1.0000	0.9917