T.C.
MİMAR SİNAN
GÜZEL SANATLAR ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ
İSTATİSTİK ANABİLİM DALI
YÜKSEK LİSANS TEZİ

ÇOKLU DOĞRUSAL REGRESYONDA AYKIRI, ETKİLİ DEĞERLERİN ARAŞTIRILMASI ve BİR UYGULAMA

Barış AŞIKGİL
DANIŞMAN: Prof. Dr. Nalan CİNEMRE

İSTANBUL, 2006

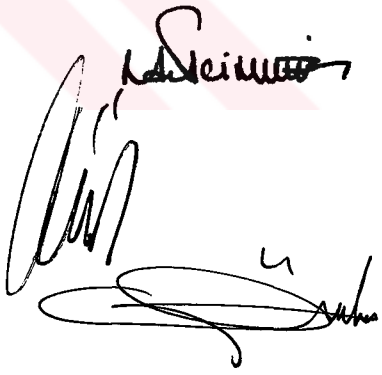
Barış Aşıkgil tarafından hazırlanan "Çoklu Doğrusal Regresyonda Aykırı, Etkili Değerlerin Araştırılması ve Bir Uygulama" adlı araştırmanın Yüksek Lisans Tezi olarak uygun olduğunu onaylarım.

Bu çalışma Mimar Sinan Güzel Sanatlar Üniversitesi Fen Bilimleri Enstitüsü İstatistik Anabilim Dalında Yüksek Lisans Tezi olarak kabul edilmiştir.

Danışman : Prof. Dr. Nalan CİNEMRE

Jüri Üyesi : Prof. Dr. Gülay KIROĞLU

Jüri Üyesi : Yrd. Doç. Dr. Özlem YILMAZ



ÖZET

Bu çalışmanın amacı, çoklu doğrusal regresyonda kuşkulu gözlemleri, bir başka deyişle aykırı, uç değerleri ve etkili gözlemleri incelemek; uygulama verileri üzerinde çeşitli yöntemlerle kuşkulu gözlemleri saptayıp hangi yöntemin daha iyi sonuç verdiğini araştırmaktır.

Beş bölümden oluşan çalışmanın birinci bölümünde, çoklu doğrusal regresyonla ilgili ön bilgilerin yanısıra "artık" kavramı üzerinde duruldu. Artıklar çeşitli sınıflara ayrılıp açıklandı. Daha sonra, sırasıyla aykırı değerler, uç değerler ve etkili gözlemler açıklanıp aralarındaki ilişkiler belirtildi.

İkinci bölümde, tek kuşkulu gözlemleri saptamada kullanılan çeşitli grafikler ve çeşitli istatistikler tanıtıldı.

Üçüncü bölümde, gizleme ve sürükleme etkileri tanımlandı ve bu etkilerin varlığında çoklu kuşkulu gözlemlerin, tek kuşkulu gözlemleri saptamada kullanılan yöntemler ile doğru biçimde belirlenemeyeceği vurgulandı. Bu nedenle, çoklu kuşkulu gözlemleri saptamada kullanılan sağlam yöntemler açıklanıp bu yöntemlerden elde edilen sonuçların çeşitli grafiklerle gösterimi sunuldu.

Dördüncü bölümde, iki ayrı gerçek veri kümesi için tek ve çoklu kuşkulu gözlemler, anlatılan yöntemler ile incelendi ve hangi yöntemin daha iyi sonuç verdiğini saptamak üzere geçerlilik çözümlemesi yapıldı.

Sonuç olarak, iki ayrı gerçek veri kümesi için farklı sonuçlar elde edildiğinden kuşkulu gözlemlerin saptanmasında hangi yöntemin daha iyi olduğunun çalışılan veri kümesine bağlı olarak değiştiği belirlendi.

Anahtar Kelimeler: Regresyon, aykırı değer, uç değer, etkili gözlem, çoklu aykırı değer.

SUMMARY

The aim of this study is to examine suspicious observations i.e. outlier, leverage, influential observations in multiple linear regression and to investigate which method gives better result on determining suspicious observations.

In the first chapter of the study consisting of five chapters, residual as a concept was explained in multiple linear regression. Then, outlier, leverage, influential observations were defined respectively and relationships among them were stated.

In the second chapter, various graphs and statistics used for determining single suspicious observations were introduced.

In the third chapter, masking and swamping effects were defined and it was stressed that multiple suspicious observations can't be determined correctly by methods used for determining single suspicious observations in the presence of these effects. Therefore, robust methods used for determining multiple suspicious observations were explained and results obtained from these methods have been displayed by using various graphs.

In the fourth chapter, single and multiple suspicious observations were examined on two real data sets and validation analysis has been applied to determine which method gives better result.

Finally, because of different results obtained from two different real data sets it has been determined that a good method used for determining suspicious observations changes according to data sets.

Keywords: Regression, outlier, leverage, influential observation, multiple outlier.

TEŞEKKÜR

Bu çalışmanın gerçekleşmesinde, her türlü öğüt ve yardımlarından dolayı danışmanım Sayın Prof. Dr. Nalan CİNEMRE'ye, değerli katkı ve yapıcı eleştirilerinden dolayı ikinci danışmanım Sayın Prof. Dr. Aydın ERAR'a, tez konusunun belirlenmesi sırasında yönlendirmelerinden dolayı hocam Sayın Prof. Dr. Gülay KIROĞLU'na, beni her zaman destekleyen sevgili AİLEME ve tüm bölüm arkadaşlarıma teşekkür ederim.

İÇİNDEKİLER DİZİNİ

ÇİZELGELER DİZİNİ

ŞEKİLLER DİZİNİ

BİRİNCİ BÖLÜM
ÇOKLU DOĞRUSAL REGRESYON ve ARTIK İNCELEMESİ

1.1. Çoklu Doğrusal Regresyon Modeli

Regresyon çözümlemesinin amacı, elde edilen veri yardımıyla bağımlı (Y) ve bağımsız değişkenler (X_j) arasındaki ilişkinin açıklanmasını sağlayacak uygun bir modelin oluşturulmasıdır. Ekonomi ve işletmecilik gibi çoğu alanda herhangi bir ekonomik değişkeni birden çok bağımsız değişkenle açıklamak için çoklu regresyon kullanılır (Orhunbilge, 2002). Bir rastlantı değişkeni olan bağımlı değişkenin, k sayıda bağımsız değişkenle açıklandığı doğrusal model,

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik} + \varepsilon_i, \qquad i = 1,2,...,n \tag{1.1}$$

biçiminde verilir. Burada doğrusal sözcüğü, Y'nin β_j parametrelerinin (j = 0, 1, ..., k) doğrusal bir fonksiyonu olduğu anlamında kullanılmıştır. (1.1) modelinde β_j parametresi (j = 0, 1, ..., k) kısmi regresyon katsayısıdır ve diğer bağımsız değişkenler sabit tutulduğunda X_j'deki bir birimlik değişimin Y bağımlı değişkenindeki beklenen değişimini gösterir. ε_i, bir hata terimidir ve $\varepsilon_i = y_i - E(y_i)$ olarak verilir.

Çoklu doğrusal regresyonda iyi parametre kestirimleri için aşağıdaki model varsayımlarının sağlanması gerekir (Draper and Smith, 1966):

- $E(\varepsilon_i) = 0, \quad i = 1,2,...,n$,
- $V(\varepsilon_1) = V(\varepsilon_2) = ... = V(\varepsilon_n) = \sigma^2$ (Eşit varyanslılık),
- $i \neq j$ için $Cov(\varepsilon_i, \varepsilon_j) = 0$ (Gözlemlerin bağımsızlığı),
- Eğer X_j bir rastlantı değişkeni ise $E(\varepsilon_i, x_{ij}) = 0$,
- Eğer çıkarsama yöntemi kullanılacak ise $\varepsilon_i \sim N(0, \sigma^2)$,
- X matrisinin tam ranklı olması, yani X'in sütunları arasında tam ya da yaklaşık doğrusal bağımlılık (çoklu bağlantı) olmaması.

Çoklu doğrusal regresyon modeli matris gösterimiyle,

$$\boldsymbol{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{1.2}$$

biçiminde verilir. Burada, $k' = k + 1$ iken $\boldsymbol{Y}$, gözlemleri içeren (n x 1) boyutlu bağımlı değişken vektörü; X, bağımsız değişkenleri içeren (n x k') boyutlu girdi matrisi; $\boldsymbol{\beta}$, regresyon katsayılarından oluşan (k' x 1) boyutlu parametreler vektörü ve $\boldsymbol{\varepsilon}$, (n x 1) boyutlu hata vektörüdür.

$\boldsymbol{\beta}$ 'nın en küçük kareler kestiricileri ise,

$$S(\boldsymbol{\beta}) = \left(\boldsymbol{Y} - X\boldsymbol{\beta}\right)' \left(\boldsymbol{Y} - X\boldsymbol{\beta}\right) \tag{1.3}$$

fonksiyonunun en küçüklenmesi ile

$$\hat{\boldsymbol{\beta}} = \left(X'X\right)^{-1} X'\boldsymbol{Y} \tag{1.4}$$

biçiminde elde edilir. Burada $\hat{\boldsymbol{\beta}}$, (k' x 1) boyutlu kestirim vektörüdür.

Ayrıca, y_i gözlenen değerlerine karşılık gelen $\hat{y}_i$ kestirilmiş değerlerinden oluşan vektör ise,

$$\begin{aligned} \hat{\boldsymbol{Y}} &= X\hat{\boldsymbol{\beta}} \\ &= X\left(X'X\right)^{-1} X'\boldsymbol{Y} \end{aligned} \tag{1.5}$$

biçiminde elde edilir. Böylece $\boldsymbol{e} = \boldsymbol{Y} - \hat{\boldsymbol{Y}}$ vektörü de artık vektörü olarak adlandırılır. Bu kesimde verilen varsayımların incelenebilmesi için de artıklar kullanılır.

1.2. Gözlem Uzaklıkları Matrisi

(1.5) eşitliğinin sağ yanı $\hat{\boldsymbol{Y}} = H\boldsymbol{Y}$ ile gösterilirse,

$$H = X(X'X)^{-1}X' \tag{1.6}$$

matrisi, (n x n) boyutlu gözlem uzaklıkları matrisi olarak adlandırılır. Bununla birlikte H matrisi, gözlenen değerlerden oluşan $\boldsymbol{Y}$ vektörünü onun en küçük kareler kestiricisi olan $\hat{\boldsymbol{Y}}$ vektörüne dönüştürdüğü için şapka matrisi ya da dönüşüm matrisi olarak da tanımlanmaktadır (Rousseeuw and Leroy, 2003).

H matrisi, simetrik $(H' = H)$ ve eşgüçlüdür (idempotent) $(HH = H)$; matrisin izi (trace) ve rankı,

$$\begin{aligned} iz(H) &= \sum_{i=1}^{n} h_{ii} = k+1, \\ rank(H) &= k+1. \end{aligned} \tag{1.7}$$

biçiminde verilir. Burada h_{ii}, dönüşüm matrisinin i'inci köşegen öğesini göstermektedir. Bu değer, $1/n$ ve 1 ile sınırlıdır $(1/n \le h_{ii} \le 1)$; aritmetik ortalaması ise (1.7)'den de görülebileceği gibi $(k+1)/n$'dir.

$\boldsymbol{x}_i' = (1 \;\; x_{i1} \;\; x_{i2} \;\; \ldots \;\; x_{ik})$ gözlem vektörü iken, h_{ii} değeri (1.6)'dan

$$h_{ii} = \boldsymbol{x}_i'(X'X)^{-1}\boldsymbol{x}_i \tag{1.8}$$

ile hesaplanır.

Dönüşüm matrisinin eşgüçlü ve simetrik olma özellikleri kullanılarak h_{ii} aşağıdaki gibi de yazılabilir:

$$h_{ii} = (HH)_{ii} = \sum_{j=1}^{n} h_{ij} h_{ji} = \sum_{j=1}^{n} h_{ij} h_{ij}$$
$$h_{ii} = \sum_{j=1}^{n} h_{ij}^2, \qquad \textit{bütün } i\textit{'ler için} \tag{1.9}$$

Sonuç olarak, dönüşüm matrisinin köşegen öğeleri (h_{ii}) herhangi bir $\boldsymbol{x}_i'$ gözlem vektörünün $\boldsymbol{X}_j$'lerin aritmetik ortalamalarından oluşan merkeze standart uzaklığını belirtir (Draper and John, 1981).

1.3. Artıklar

ε_i hata teriminin kestiricisi, $e_i = y_i - \hat{y}_i$ eşitliği ile tanımlanan artık terimidir. $\boldsymbol{e}$, (n x 1) boyutlu bir artık vektörüdür ve

$$\begin{aligned} \boldsymbol{e} &= \boldsymbol{Y} - X\hat{\boldsymbol{\beta}} \\ &= \boldsymbol{Y} - H\boldsymbol{Y} \\ &= (I - H)\boldsymbol{Y} \end{aligned} \tag{1.10}$$

ilişkilerine sahiptir. Burada I, (n x n) boyutlu bir birim matristir. Artıkların varyans-kovaryans matrisi $V(\boldsymbol{e})$ de genelde $V(\varepsilon) = \sigma^2 I$'nın bir kestiricisi olarak kullanılır. σ^2'nin kestiricisi Artık Kareler Ortalaması (AKO) da

$$AKO = \frac{\sum_{i=1}^{n} e_i^2}{serbestlik \quad derecesi} = \frac{AKT}{n-k-1} \tag{1.11}$$

ile verilir. Burada AKT (Artık Kareler Toplamı),

$$
\begin{aligned}
AKT &= \sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2 = \boldsymbol{e}'\boldsymbol{e} = \left(\boldsymbol{Y}-X\hat{\beta}\right)'\left(\boldsymbol{Y}-X\hat{\beta}\right) \\
&= \boldsymbol{Y}'\boldsymbol{Y}-\hat{\beta}'X'\boldsymbol{Y}-\boldsymbol{Y}'X\hat{\beta}+\hat{\beta}'X'X\hat{\beta} \\
&= \boldsymbol{Y}'\boldsymbol{Y}-2\hat{\beta}'X'\boldsymbol{Y}+\hat{\beta}'X'X\hat{\beta} \\
&= \boldsymbol{Y}'\boldsymbol{Y}-\hat{\beta}'X'\boldsymbol{Y}
\end{aligned}
\tag{1.12}
$$

biçiminde hesaplanır (Ryan, 1997).

Artıkları çeşitli sınıflara ayırmak mümkündür:

Standartlaştırılmış artıklar, artık varyansının karekökü ile oranlanarak,

$$
d_i = \frac{e_i}{\sqrt{AKO}}, \qquad i = 1,2,...,n \tag{1.13}
$$

elde edilir. Genelde birim normal sapma olarak adlandırılır ve [-2, +2] aralığında bulunur.

Student türü artıklar, (1.13)'deki artıklarda pay ve payda bağımlı olduğundan her bir artığın kendi standart hatası olan $\sqrt{V(e_i)} = \sqrt{AKO\left(1-h_{ii}\right)}$ değeriyle oranlanarak,

$$
r_i = \frac{e_i}{\sqrt{AKO\left(1-h_{ii}\right)}}, \qquad i = 1,2,...,n \tag{1.14}
$$

ile verilir; içsel artık olarak da adlandırılır. İçsel denmesinin nedeni, AKO'nun n gözlemli kestirilmiş modelden elde edilmiş olmasıdır. Büyük artığa ve büyük h_{ii}'ye sahip bir gözlemin en küçük kareler kestirimi üzerinde etkili olacağı düşünüldüğünde student türü artıkların kullanılması öngörülür. Aynı zamanda r_i değerleri [-3, +3] aralığında yer alır (Montgomery and Peck, 1992).

Çıkartılmış artıklar, i. gözlem çıkartıldıktan sonra geriye kalan n-1 gözlem üzerine regresyon modeli kurulup bu model yardımıyla i'inci gözlemin kestirim değerinin $\left(\hat{y}_{i(i)}\right)$ elde edilmesiyle i'inci gözlem için

$$e_{(i)} = y_i - \hat{y}_{i(i)} \tag{1.15}$$

biçiminde hesaplanır. Bu artık *i*'inci PRESS artığı olarak da adlandırılır. Bu işlem, her bir gözlem için $i = 1, 2, ..., n$ tekrarlanarak $e_{(1)}, e_{(2)}, ..., e_{(n)}$ elde edilir. Bununla birlikte, PRESS artıkları kolayca,

$$e_{(i)} = \frac{e_i}{1-h_{ii}} \tag{1.16}$$

yardımıyla hesaplanabilir. Bir artık ile PRESS artığı arasındaki olası büyük fark, bu gözlem olmaksızın modelin zayıfladığını gösterir (Montgomery and Peck, 1992).
i'inci PRESS artığının varyansı aşağıda verildiği gibi bulunur:

$$V\left[e_{(i)}\right] = V\left[\frac{e_i}{1-h_{ii}}\right] = \frac{1}{\left(1-h_{ii}\right)^2}\left[\sigma^2\left(1-h_{ii}\right)\right] = \frac{\sigma^2}{1-h_{ii}}. \tag{1.17}$$

Standartlaştırılmış PRESS artıkları ise,

$$\frac{e_{(i)}}{\sqrt{V\left[e_{(i)}\right]}} = \frac{e_i/\left(1-h_{ii}\right)}{\sqrt{\sigma^2/\left(1-h_{ii}\right)}} = \frac{e_i}{\sqrt{\sigma^2\left(1-h_{ii}\right)}} \tag{1.18}$$

olarak elde edilir. σ^2'nin kestirimi için AKO'nun kullanılması durumunda, standartlaştırılmış PRESS artıkları student türü artıklar haline dönüşür (Montgomery and Peck, 1992).

R-Student türü artıklar, (1.14) eşitliğinde de pay ve payda tam olarak bağımsız olmadığından $V(e_i)$ yerine yukarıda tanımlanan çıkartılmış artıklara ilişkin varyans kullanılarak,

$$t_i = \frac{e_i}{\sqrt{s_{(i)}^2(1-h_{ii})}}, \qquad i = 1,2,...,n \tag{1.19}$$

ile t-dağılımına sahip biçimde bulunur. i'inci gözlemin çıkartıldığı ve σ^2'nin kestirimi olarak $s_{(i)}^2$'nin kullanıldığı bu artıklar dışsal artıklar olarak da adlandırılır. Burada $s_{(i)}^2$,

$$s_{(i)}^2 = \frac{(n-k-1)AKO - \left(e_i^2/(1-h_{ii})\right)}{n-k-2} \tag{1.20}$$

biçiminde hesaplanır. (1.20)'deki eşitliğin payında bulunan $e_i^2/(1-h_{ii})$ dikkate alınırsa, a_i düzeltilmiş artık olmak üzere $a_i^2 = e_i^2/(1-h_{ii})$ eşitliği bir başka biçimde,

$$a_i^2 = AKT - AKT_{(i)} \tag{1.21}$$

olarak yazılabilir. (1.20) ve (1.21) eşitlikleri yardımıyla da $s_{(i)}^2$,

$$s_{(i)}^2 = \frac{(n-k-1)AKO - AKT + AKT_{(i)}}{n-k-2} = \frac{AKT_{(i)}}{n-k-2} \tag{1.22}$$

biçiminde de hesaplanabilir (Hadi and Simonoff, 1993). $s_{(i)}^2$'nin çıkarımı Ek-A.1'de verilmiştir. R-student türü artıkları, Atkinson 1981'de "jackknife" ve Cook ve Weisberg de 1982'de "dışsal student türü" artıklar olarak adlandırmışlardır. R-student türü artıkların student türü artıklara tercih edilmesinin nedenleri şunlardır (Chatterjee and Hadi, 1986):

- t_i'nin t dağılımından (t_{n-k-2}) gelmesi,
- büyük sapmaları daha açık yansıtması,

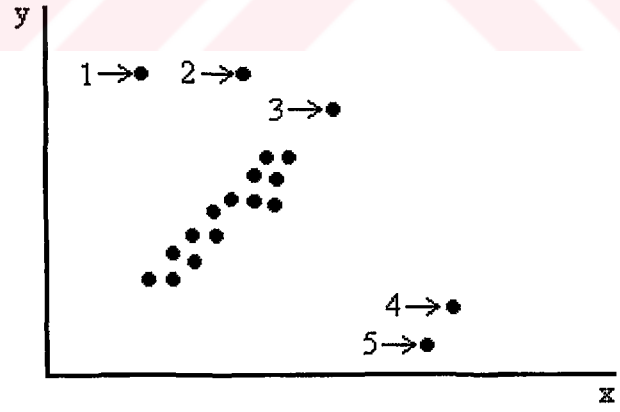
- $s^2_{(i)}$'nin i'inci gözlemdeki büyük hataları düzeltmede sağlam bir kestirici olması.

1.4. Kuşkulu Gözlemler ve Aykırı Değerler

Regresyon çözümlemelerindeki gözlemler incelendiğinde bazı ya da tüm regresyon sonuçlarının üç tür gözlemden etkilendiği görülmüştür. Bu gözlemler,

- bağımlı değişken yönünde kuşkulu olan gözlemler,
- bağımsız değişkenler yönünde kuşkulu olan gözlemler,
- hem bağımlı hem de bağımsız değişkenler yönünde kuşkulu olan gözlemlerdir.

Çeşitli durumlardaki kuşkulu gözlemler Şekil 1.1'de gösterilmiştir. Burada, 1 ve 2 ile numaralanmış gözlemler y değerlerine göre kuşkulu, 3, 4 ve 5 ile numaralanmış gözlemler x değerlerine göre kuşkulu ve aynı zamanda 4 ve 5 ile numaralanmış gözlemler hem de y değerlerine göre kuşkulu gözlemlerdir (Neter ve diğerleri, 1996).



Şekil 1.1. Farklı Kuşkulu Gözlemlerin Görünümü

Hawkins (1980)'e göre "*bir aykırı değer, diğer gözlemlerden oldukça sapan ve başka bir mekanizma tarafından yaratıldığı konusunda kuşku uyandıran bir gözlemdir.*" Bir gözlem,

$$\left(\overline{x} - ls, \overline{x} + ls\right) \tag{1.23}$$

aralığının dışında kalıyorsa aykırı değer olarak belirtilir. Burada l, genellikle iki ya da üç olarak alınır. Ortalama $(\bar{x})$ ve standart sapma (s) aykırı değerlerden oldukça etkilenen kestiricilerdir (Hawkins, 1980).

Regresyonda, y-ekseni yönünde kuşkulu olan gözlemlere aykırı değer denir. y-ekseni yönünde uzakta olan gözlemlerin student türü artıklarının (r_i ya da t_i) mutlak değeri de diğer gözlemlerinkilerle karşılaştırıldığında oldukça büyüktür.

Bir veri kümesindeki aykırı değerlerin varlık nedenleri; hatalı ölçüm, veri girişindeki ya da veri kaydındaki hatalar, ölçüm yapan aygıttaki sorunlar ve nadir olaylardır (örneğin, Ocak ayında bir gün İstanbul'daki yüksek hava sıcaklığı). Eğer bu nedenlerin aykırı değerleri yarattığı biliniyor ise, aykırı değerler düzeltilebilir ya da veri kümesinden çıkartılabilir.

Aykırı değerler, en küçük kareler yönteminde AKT'nin en küçüklenmesi ilkesinden dolayı kestirilmiş doğruyu kendilerine doğru çekerler. Bazı durumlarda kestirilmiş denklemi düzeltmek için bu değerlerin silinmesi kestirim konusunda araştırmacıları yanlış bir yöne çekebilir. Bunun nedeni, bu değerlerin bazen veriyi yorumlamada yararlı bilgilere sahip olmaları ve onların çıkarılmasının verinin içerdiği anlamı olumsuz yönde etkileyecek olmasıdır. Bu nedenle, bir aykırı değeri veriden uzaklaştırma kararı vermeden önce söz konusu değer çok dikkatli bir şekilde incelenmelidir.

Genel anlamda, aykırı değerlerin regresyon çözümlemesinde yarattığı sorunlar aşağıda verilmiştir (High, 2004):

1. Örneklem ortalamasına ve varyansına bağlı istatistiksel testleri saptırırlar.
2. Regresyon katsayılarının değerleri, t ve F değerleri, R^2 ve AKO aykırı değerlerin varlığından etkilenir.
3. Regresyondaki AKT'yi en küçükleme ilkesi aykırı değerlerin varlığından oldukça etkilenir.
4. Kestiricilerin yanlılığına neden olurlar.
5. İstatistiksel anlamlılıkta p-değerlerini saptırırlar.
6. Yanlış karar vermeye yol açarlar.

1.5. Uç Değerler

x-uzayında veri kümesinden uzakta bulunan noktalar uç değerler olarak adlandırılır. Bir başka deyişle, uç değerler x-ekseni yönündeki kuşkulu gözlemlerdir. Yani, bir gözlemin uç değer olması demek gözlenen x_i verilerinden uzakta bulunması demektir. Bütün bu tanımlardan da anlaşıldığı gibi, regresyonda uç değerler tamamıyla bağımsız değişkenlerle ilgilidir.

Rousseeuw ve Van Zomeren (1990) daha genel olarak, x_i değerlerinin yanısıra y_i değerlerini de bu tanımın içine alıp aşağıdaki noktaları tanımlamıştır:

- İyi uç değerler, (x_i, y_i) noktalarının regresyon düzlemine yakın olanlarıdır. Bir başka deyişle, regresyon katsayılarının doğruluğunu (precision) arttıran noktalar iyi uç değerlerdir.
- Kötü uç değerler, regresyon düzlemine uzak olan (x_i, y_i) noktalarıdır. Bir başka deyişle, regresyon katsayılarının doğruluğunu azaltan noktalar kötü uç değerlerdir.

Bu tanımlar göz önünde bulundurulduğunda, Şekil 1.1'deki 3 numaralı gözlemin iyi uç değer; 4 ve 5 numaralı gözlemlerin de kötü uç değerler olduğu söylenebilir.

Bir yüksek uç değer, veri kümesindeki diğer gözlemlerle karşılaştırıldığında büyük h_{ii}'ye sahip olan gözlem olarak ifade edilir. Hoaglin ve Welsch 1978'de, i gözlemi için,

$$h_{ii} > 2k'/n \tag{1.24}$$

olması durumunda bu gözlemin yüksek uç değer (high leverage) olarak adlandırılabileceğini belirtmiştir. Yüksek uç değerler,

$$h_{ii} + \frac{e_i^2}{AKT} \leq 1 \tag{1.25}$$

eşitsizliğinden görülebileceği gibi küçük artığa sahip olma eğilimindedir (Chatterjee ve diğerleri, 2000).

1.6. Etkili Gözlemler

Etkili gözlemler (influential observations), veri kümesindeki diğer gözlemlerle karşılaştırıldığında tek tek ya da toplu olarak kestirilmiş regresyon denklemine etki eden gözlemler olarak tanımlanır. Bir başka deyişle, çözümlemeden çıkartıldığında hesaplanan çeşitli değerleri (katsayılar, standart hatalar, t-değerleri vb.) önemli ölçüde değiştiren gözlemlere etkili gözlemler denir. Bununla birlikte, bir gözlem bütün regresyon çıktıları üzerinde aynı etkiye sahip olmayabilir. Bu nedenle, etkili gözlem araştırmalarında "Ne üzerine etki?" sorusunun önem kazandığı durumlar olmaktadır.

Etkili gözlemler, hem y-ekseni hem de x-ekseni yönünde kuşkulu gözlemler olabilmektedir. Şekil 1.1'de, aykırı değerler olan 1 ve 2 numaralı gözlemler karşılaştırıldığında 1 numaralı gözlemin yarattığı etki oldukça büyüktür; 2 numaralı gözlemin ise etkili bir gözlem olduğu tartışılır. Yüksek uç değerler olan 3, 4, 5 numaralı gözlemlerden de (aynı zamanda 4 ve 5 numaralı gözlemler aykırı değerdir) 4 ve 5 oldukça etkilidir. Çünkü bu gözlemlerin hem x değerleri hem de y değerleri regresyondaki ilişkiyle uyuşmamaktadır. 3 numaralı gözlem ise y değerinin regresyondaki ilişkiyle uyumlu olması dolayısıyla çok etkili bir gözlem değildir.

Bazı durumlarda ise 4 ve 5 numaralı gözlemlere benzer gözlemlerden biri çözümlemeden kaldırıldığı zaman kestirimde bir değişiklik olmadığı görülür. Bunun nedeni, kalan gözlemin çıkartılanın etkisini gizlemiş olmasıdır.

Aykırı değerler, yüksek uç değerler ve etkili gözlemler arasındaki ilişkiler aşağıda verildiği gibidir (Chatterjee and Hadi, 1988):

- Aykırı değerler etkili gözlem olmak zorunda değildir.
- Yüksek uç değerler etkili gözlem olmak zorunda değildir.
- Etkili gözlemler genellikle aykırı değerler ya da yüksek uç değerlerdir.

İKİNCİ BÖLÜM
ÇOKLU DOĞRUSAL REGRESYONDA TEK KUŞKULU GÖZLEMLERİN VARLIĞI ve İNCELENMESİ

Yapılan araştırmalar, kuşkulu gözlemlerin veri kümelerinde sanılanın aksine daha yaygın olduğunu göstermektedir. Bu değer, ortalama %5 ile %10 arasında değişmektedir. Bu da bir veri kümesinde, kuşkulu gözlemlerin tek tek ya da çoklu gruplar halinde bulunabileceğinin bir belirtisidir (Wisnowski, 1999). Tek kuşkulu gözlemler iki yolla saptanabilir:

- Çeşitli grafikler yardımıyla,
- Çeşitli istatistikler yardımıyla.

2.1. Tek Kuşkulu Gözlemlerin Grafikler Yardımıyla Saptanması

2.1.1. Eklenmiş Değişken Grafikleri

Mosteller ve Tukey'in 1977'de verdiği kısmi regresyon grafikleri (uyarlanmış değişken grafikleri) olarak da bilinen "eklenmiş değişken grafikleri", eklenen değişkenin regresyondaki durumu ile ilgili bilgi verir (Chatterjee ve diğerleri, 2000). Bir X_j bağımsız değişkeni için çizilecek grafik iki farklı artık kümesinden oluşur. Biri Y'nin X_j dışındaki bağımsız değişkenler üzerine modellenmesi sonucu elde edilen artık kümesi, ikincisi ise X_j'nin öteki bağımsız değişkenler üzerine modellenmesi sonucu elde edilen artık kümesidir. (1.2) eşitliği,

$$\boldsymbol{Y} = X_{(j)}\beta_{(j)} + \boldsymbol{X}_j\beta_j + \boldsymbol{\varepsilon} \tag{2.1}$$

biçiminde yazılıp her iki tarafı $I - H_{(j)}$ ile çarpılırsa,

$$\left(I - H_{(j)}\right)\boldsymbol{Y} = \left(I - H_{(j)}\right)X_{(j)}\beta_{(j)} + \beta_j\left(I - H_{(j)}\right)\boldsymbol{X}_j + \left(I - H_{(j)}\right)\boldsymbol{\varepsilon} \tag{2.2}$$

elde edilir. Burada $H_{(j)}$, j'inci bağımsız değişken çıkartılarak elde edilen (n x n) boyutlu gözlem uzaklıkları matrisidir. $\left(I-H_{(j)}\right)\boldsymbol{X}_{(j)}=0$ olduğundan (2.2) eşitliği,

$$\left(I-H_{(j)}\right)\boldsymbol{Y}=\left(I-H_{(j)}\right)\boldsymbol{X}_j\beta_j+\left(I-H_{(j)}\right)\varepsilon \tag{2.3}$$

biçiminde de yazılabilir. Eklenmiş değişken grafikleri (2.3) eşitliğindeki,

$$\left\{\left(I-H_{(j)}\right)\boldsymbol{X}_j,\left(I-H_{(j)}\right)\boldsymbol{Y}\right\} \tag{2.4}$$

değerlerine göre çizilir. k sayıda bağımsız değişken için k farklı eklenmiş değişken grafiği elde edilir. Eklenmiş değişken grafiklerinden elde edilen bilgiler aşağıda verildiği gibidir (Myers, 1986):

- Grafikten, hangi bağımsız değişkenin açıklayıcı gücünün iyi olduğu (doğrusal ilişkinin varlığı) görülür.
- Grafik, bir eğriyi andırıyorsa dönüşüm uygulanmasının yararlı olacağı düşünülür.
- Grafik, kuşkulu gözlemlere ve hangi regresyon katsayılarının daha fazla etkilendiğine karar vermede yardımcı olur.

2.1.2. Bileşen Artı Artık (C+R) Grafikleri

Wood'un 1973'de belirttiği bileşen artı artık (C+R) grafikleri, regresyon çözümlemesinde kullanılan en eski grafik tiplerinden biridir. Kısmi artık grafikleri olarak da bilinir (Montgomery and Peck, 1992). $\boldsymbol{X}_j$ bağımsız değişkeni için i'inci kısmi artık,

$$e_{ij}^{*} = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \ldots - \hat{\beta}_{j-1} x_{i,j-1} - \hat{\beta}_{j+1} x_{i,j+1} - \ldots - \hat{\beta}_k x_{ik}$$
$$= e_i + \hat{\beta}_j x_{ij} \qquad i = 1,2,\ldots,n \tag{2.5}$$

biçiminde tanımlanırsa bu grafikler,

$$\left\{ X_j, \left(e + \hat{\beta}_j X_j \right) \right\} \tag{2.6}$$

değerlerine göre çizilir. Burada e, Y'nin tüm bağımsız değişkenler üzerine modellenmesiyle elde edilen artıklar ve $\hat{\beta}_j$ da X_j bağımsız değişkeninin katsayısının kestirimidir. Bu grafiklerden, Y ve X_j arasındaki bağıntının doğrusallığı ile ilgili bilgi alındığı gibi kuşkulu gözlemler de saptanabilmektedir (Chatterjee ve diğerleri, 2000).

2.1.3. Uç Değer-Artık (L-R) Grafikleri

McCulloch ve Meeter'ın 1983'de ve Gray'in 1986'da ele aldığı uç değer-artık (L-R) grafikleri,

$$\left\{ h_{ii}, d_i^2 \right\} \tag{2.7}$$

değerlerine göre çizilir. Burada h_{ii}, gözlem uzaklıkları matrisinin i'inci köşegen öğesi ve d_i, i'inci standartlaştırılmış artıktır. Her türlü kuşkulu gözlemin (yüksek uç değer, aykırı değer ya da ikisinin birlikte bulunduğu durumlar) saptanmasında da kullanılırlar (Chatterjee ve diğerleri, 2000).

2.1.4. Tek Kuşkulu Gözlemlerin Saptanmasında Kullanılabilen Diğer Grafikler

Tek kuşkulu gözlemlerin saptanmasında kullanılabilen diğer grafikler, kestirilmiş değerlere $(\hat{y}_i)$ karşı artık (e_i), student türü artık (r_i) ya da r-student türü artık (t_i) değerlerinin oluşturduğu klasik grafiklerdir. Bunun yanında, çeşitli etki istatistikleriyle

gözlemlerin indeks numaralarının grafiklenmesi de etkili gözlemlerin saptanmasında kullanılabilir.

Kullanılan bir diğer grafik tipi matris biçimindeki saçılım (scatter) grafikleridir. Bu grafikler, bütün değişkenleri aynı anda grafikleyip aralarında doğrusal bağıntı bulunup bulunmadığını gösterir. Ayrıca, bu grafiklerden kuşkulu gözlemler de belirlenebilmektedir. Bu grafiklerin son satırı, Y ile her bir X değişkeni arasındaki bağıntıyı tek tek diğer bağımsız değişkenlerin yokluğunda gösterdiğinden kısmi cevap (response) grafiklerini verir (Cook and Weisberg, 1994).

2.2. Tek Kuşkulu Gözlemlerin İstatistikler Yardımıyla Saptanması

Tek kuşkulu gözlemlerin saptanmasında kullanılan istatistikler üç başlık altında incelenebilir:

- Aykırı değerlerin saptanmasında kullanılan istatistikler
- Uç değerlerin saptanmasında kullanılan istatistikler
- Etkili gözlemlerin saptanmasında kullanılan istatistikler

2.2.1. Aykırı Değerlerin Saptanmasında Kullanılan Testler

2.2.1.1. Ortalama Değişim (Mean-Shift) Aykırı Değer Modellemesi

v'inci gözlemin aykırı değer olup olmadığının araştırılmasında $\boldsymbol{U}$ ile gösterilen yeni bir bağımsız değişken tanımlanmıştır. $\boldsymbol{U}$'nun i'inci elemanı $i \neq v$ iken $u_i = 0$ ve v'inci elemanı $u_v = 1$ olarak belirtilir. X'in v'inci satırı çıkartıldığında,

$$y_i = \boldsymbol{x}_i'\beta + \varepsilon_i, \qquad i \neq v \tag{2.8}$$

denkleminin kestirimi ve tüm gözlemler için X ve $\boldsymbol{U}$ üzerinden,

$$y_v = \boldsymbol{x}_v'\beta + \delta + \varepsilon_v, \qquad i = v \tag{2.9}$$

denkleminin kestirimi elde edilebilir. y_v'nin beklenen değeri $x_v'\beta$'dan δ kadar farklıdır. Burada δ, U'nun katsayısı olup ortalama değişim olarak adlandırılmaktadır. v'inci gözlemin aykırı değer olup olmadığına karar verebilmek için $H_0 : \delta = 0$ yokluk hipotezinin $H_a : \delta \neq 0$ alternatif hipotezine karşı test edilmesi gerekir. Eğer hata terimleri normal dağılımdan geliyorsa, test istatistiği $n - k - 1$ serbestlik derecesi ile t-dağılımıdır. Yokluk hipotezinin reddi, v'inci gözlemin aykırı değer olduğunun göstergesidir (Weisberg, 1985).

2.2.1.2. Aykırı Değerleri Saptamada Kullanılan Bir Test

Aykırı değerleri saptamada, (1.19)'da gösterilen r-student türü artıklar kullanılır. Bu artıkların yönleriyle ilgilenilmediğinden mutlak değerleri göz önünde bulundurulur. Hangi gözlemin $| t_i |$ değerinin büyük olduğu bilinmediğinden her bir gözlem için bir test içeren n sayıda test uygulanır ve bu test Bonferroni testi olarak adlandırılır. t_i'nin $n - k - 2$ serbestlik dereceli t-dağılımından geldiği göz önünde bulundurulduğunda bu testler için uygun Bonferroni kritik değeri $t_{n-k-2;(\alpha/2n)}$ olur. Sonuç olarak, $| t_i |$ değerleri Bonferroni kritik değerinden büyük olan gözlemler aykırı değer olarak düşünülebilir (Montgomery and Peck, 1992).

2.2.2. Uç Değerlerin Saptanmasında Kullanılan İstatistikler

2.2.2.1. Yüksek Uç Değerleri Belirlemede Kullanılan Bir Test

Hoaglin ve Welsch 1978'de, (1.24)'de belirtilen durumu sağlayan gözlemlerin yakından incelenmesi gerektiğini belirtmişlerdir. Eğer kurulan regresyon modeli doğruysa, bir örneklem için uç değer kümesi ki-kare yoğunluğuna benzer bir frekansa sahiptir. Eğer her bir bağımsız değişken de normal dağılımdan geliyorsa herhangi bir h_{ii} değeri için F_i aşağıda gösterildiği biçimde hesaplanır:

$$F_i = \frac{\left[h_{ii} - (1/n)\right]/k}{(1 - h_{ii})/(n - k - 1)} \tag{2.10}$$

Burada belirtilen F_i değeri, sırasıyla k ve $n-k-1$ serbestlik dereceli F dağılımına yönelir. Bu nedenle, yüksek uç değerlerin tespiti için F_i ile $F_{k,(n-k-1);(\alpha/n)}$ tablo değeri karşılaştırılır. $F_i > F_{k,(n-k-1);(\alpha/n)}$ olduğu durumlara sahip gözlemler yüksek uç değer olarak kabul edilir (Kleinbaum ve diğerleri, 1988).

2.2.2.2. Ağırlıklı Uzaklık Kareler Toplamı (AUKT)

Daniel ve Wood 1980'de, x-uzayında uzaktaki noktaları tespit etmek için bir başka ölçüt olan ağırlıklı uzaklık kareler toplamını (AUKT) kullanmayı öngörmüşlerdir:

$$AUKT_i = \sum_{j=1}^{k} \frac{\left[\hat{\beta}_j \left(x_{ij} - \bar{x}_j\right)\right]^2}{AKO}, \qquad i = 1,2,...,n \tag{2.11}$$

biçiminde hesaplanan değerler artan sıraya göre dizilip değerler arasında ani bir sıçrama (artma) olup olmadığı kontrol edilir. Böyle bir durum söz konusu ise bir ya da birden fazla yüksek uç değerin varlığından söz edilir. Eğer, $AUKT_i$ değerleri küçükten büyüğe doğru düzgün bir şekilde artıyorsa x-uzayında çok uzakta bir nokta bulunmuyor demektir. Bu nedenle, özellikle $AUKT_i$ değerleri yüksek olan gözlemler dikkatle incelenmelidir (Montgomery and Peck, 1992).

2.2.2.3. Gözlem Uzaklıkları Matrisinin Köşegen Öğesiyle Uç Değer Saptanması

(1.24)'de verilen durum, Huber (1981) tarafından en büyük h_{ii} üzerine belirginleştirilerek aşağıdaki durumlar sunulmuştur:

$$\begin{aligned} enb(h_{ii}) \le 0.2 &\Rightarrow i'inci \;\; gözlem \;\; güvenilirdir \\ 0.2 < enb(h_{ii}) \le 0.5 &\Rightarrow i'inci \;\; gözlem \;\; risklidir \\ enb(h_{ii}) > 0.5 &\Rightarrow uygunsa \;\; i'inci \;\; gözlem \;\; uzaklaştırılabilir \end{aligned} \tag{2.12}$$

0.2'den büyük olan h_{ii} değerlerine sahip gözlemler risklidir diye belirtilmiştir. Nedeni, *Y-X* ilişkisi hakkında çok fazla bilginin tek bir gözlemden sağlanıyor olmasındandır. 1'e yaklaşan h_{ii} değerlerine sahip gözlemler ise neredeyse regresyonu tamamen kontrol ediyor demektir. Bu nedenle, bu gözlemler uygunsa veri kümesinden çıkartılabilir.

2.2.2.4. Mahalanobis Uzaklığı (MU)

Uç değerleri tespit etmede kullanılan klasik bir ölçüt de Mahalanobis Uzaklığı'dır. Bu uzaklık, x_i'lerin oluşturduğu çok değişkenli bir veri kümesinde bir gözlemin veri kümesinin merkezine olan uzaklığını belirtir. x_i' gözlem vektörü,

$$x_i' = \begin{pmatrix} 1 & x_{i1} & x_{i2} & \dots & x_{ik} \end{pmatrix} = \begin{pmatrix} 1 & z_i \end{pmatrix} \tag{2.13}$$

biçiminde tanımlanırsa, sırasıyla z_i'nin ortalama vektörü ve kovaryans matrisi aşağıda verildiği gibi bulunur:

$$\bar{z} = \frac{1}{n}\sum_{i=1}^{n} z_i \tag{2.14}$$

$$C = \frac{1}{n-1}\sum_{i=1}^{n} \left(z_i - \bar{z}\right)' \left(z_i - \bar{z}\right) \tag{2.15}$$

(2.14) ve (2.15) kullanılarak *i*'inci gözlem için Mahalanobis Uzaklığı,

$$MU_i = \sqrt{\left(z_i - \bar{z}\right) C^{-1} \left(z_i - \bar{z}\right)'}, \qquad i = 1,2,\dots,n \tag{2.16}$$

biçiminde hesaplanır. Bununla birlikte, veri kümesi normal dağılımdan geliyorsa i'inci gözlemin bir uç değer olup olmadığını belirlemek için MU_i^2 değerleri k serbestlik dereceli ve 0.95 güvenilirlikteki ki-kare değeriyle karşılaştırılabilir.

Rousseeuw ve Van Zomeren (1990), gözlem uzaklıkları matrisinin köşegen öğesi ile Mahalanobis Uzaklığı arasında monoton bir ilişkinin varlığından bahsetmiştir. Her bir j için $(1/n)\sum_{i=1}^{n} x_{ij} = 0$, dolayısıyla $\bar{z} = 0$ alınıp yaklaşık olarak bir kayıp olmadığı varsayılırsa $X'X$ matrisi,

$$X'X = \begin{bmatrix} n & 0 & \dots & 0 \\ 0 & & & \\ \vdots & & (n-1)C & \\ 0 & & & \end{bmatrix} \tag{2.17}$$

biçiminde yazılabilir. (1.8)'de verilen eşitlik kullanılarak,

$$\begin{aligned} h_{ii} &= \begin{pmatrix} 1 & z_i \end{pmatrix} \begin{bmatrix} 1/n & 0 \\ 0 & \left(1/(n-1)\right)C^{-1} \end{bmatrix} \begin{bmatrix} 1 \\ z_i \end{bmatrix} \\ &= \frac{MU_i^2}{n-1} + \frac{1}{n} \end{aligned} \tag{2.18}$$

elde edilir (Rousseeuw and Leroy, 2003).

2.2.3. Etkili Gözlemlerin Saptanmasında Kullanılan İstatistikler

Bir gözlemin etkisini incelemek için "gözlemi çıkartma" ilkesi uygulanır. Bu ilke, regresyon modelini oluşturan veri kümesinden her bir gözlemin çıkartılması yoluyla elde edilen modellerdeki çeşitli değerlerin veri kümesinin oluşturduğu modeldeki değerlerle karşılaştırılması esasına dayanır. Hangi gözlem ya da gözlemler için bu değerlerdeki değişim büyük olursa, o gözlem ya da gözlemlerin modeli büyük ölçüde etkilediği belirtilir.

Etkili gözlemlerin saptanması için regresyon modeliyle ilgili çeşitli değerlerin karşılaştırılmasında kullanılan istatistikler aşağıda ele alınmıştır:

2.2.3.1. DFFITS İstatistiği

Welsch ve Kuh'un 1977'de belirttiği ve adını kestirimlerdeki farktan (difference in fits) alan bu istatistik, veri kümesinden elde edilen *i*'inci kestirilmiş değerle *i*'inci gözlem çıkartılıp elde edilen *i*'inci kestirilmiş değer arasındaki farka dayanır:

$$DFFITS_i = \frac{\hat{y}_i - \hat{y}_{i(i)}}{s_{(i)}\sqrt{h_{ii}}}, \qquad i = 1,2,...,n \tag{2.19}$$

$s_{(i)}$, eşitlik (1.22)'de verildiği gibidir. $\hat{y}_{i(i)}$, *i*'inci gözlem çıkartıldıktan sonra elde edilen modeldeki *i*'inci kestirim değeridir. DFFITS İstatistiği'nin mutlak değeri $2\sqrt{k'/n}$ ile karşılaştırılır. Bu değerden büyük olan $|DFFITS_i|$ değerine sahip gözlemlerin $\hat{y}$ uyum kestirimleri üzerinde etkili oldukları düşünülür. DFFITS İstatistiği aşağıda verildiği gibi de hesaplanabilir:

$$DFFITS_i = \sqrt{\frac{h_{ii}}{1-h_{ii}}}\ t_i, \qquad i = 1,2,...,n \tag{2.20}$$

Yukarıdaki eşitlikten de görülebileceği gibi bu istatistik, hem gözlem uzaklıkları matrisinin köşegen öğesinden hem de r-student türü artıklardan etkilenmektedir. Bir başka deyişle bu istatistik, yüksek uç değerlilik ve aykırı değerlilikle yakından ilgilidir (Belsley ve diğerleri, 1980). DFFITS'in çıkarımı Ek-A.2'de verilmiştir.

2.2.3.2. DFBETAS İstatistiği

Belsley, Welsch ve Kuh'un 1980'de belirttiği ve adını kestirilmiş regresyon katsayılarındaki farktan (difference in betas) alan bu istatistik, *i*'inci gözlemin

çıkartılmasıyla sadece j'inci kestirilmiş regresyon katsayısı üzerindeki değişimi gösterir (Belsley ve diğerleri, 1980). DFBETAS İstatistiği,

$$DFBETAS_{j(i)} = \frac{\hat{\beta}_j - \hat{\beta}_{j(i)}}{s_{(i)}\sqrt{C_{jj}}}, \qquad i = 1,2,...,n \tag{2.21}$$

biçiminde hesaplanır. Burada $\hat{\beta}_{j(i)}$, i'inci gözlemin çıkartılmasıyla elde edilen j'inci kestirilmiş regresyon katsayısı ve C_{jj}, β_0 katsayısını içeren regresyon modellerinde $(X'X)^{-1}$ matrisinin (j+1)'inci köşegen öğesidir. DFBETAS İstatistiği'nin mutlak değeri $2/\sqrt{n}$ ile karşılaştırılır. Bu değerden büyük olan $|DFBETAS_{j(i)}|$ değerine sahip gözlemlerin, j'inci regresyon katsayısı üzerinde etkili oldukları kabul edilir. DFBETAS İstatistiği'ni bir başka biçimde hesaplamak için önce (k' x n) boyutlu,

$$R = (X'X)^{-1} X' \tag{2.22}$$

matrisini tanımlamak gerekir. Daha sonra DFBETAS İstatistiği,

$$DFBETAS_{j(i)} = \frac{r_{ji}}{\sqrt{\mathbf{r}_j'\mathbf{r}_j}} \frac{t_i}{\sqrt{1-h_{ii}}}, \qquad i = 1,2,...,n \tag{2.23}$$

biçiminde de hesaplanabilir. Burada $\mathbf{r}_j'$, R matrisinin j'inci satırının oluşturduğu bir vektördür. Dikkat edilmesi gereken durum ise bu istatistiğin de yüksek uç değerlilik ve aykırı değerlilikle yakın ilgisinin bulunmasıdır (Montgomery and Peck, 1992). DFBETAS'ın çıkarımı Ek-A.3'de verilmiştir.

2.2.3.3. Cook Uzaklığı (D)

Cook (1977)'un belirttiği ve Cook Uzaklığı'nın kareleri olarak kullanılan bu istatistik, bütün model üzerindeki etkiyi içerir. Bir başka deyişle Cook Uzaklığı, etkiyi hem y_i'nin kestiricisinin $\hat{y}_i$'ya olan uyumu hem de x_i'nin geri kalan gözlemlerden uzaklığı anlamında ele alır. Cook Uzaklığı,

$$D_i = \frac{\left(\hat{\beta} - \hat{\beta}_{(i)}\right)' (X'X)\left(\hat{\beta} - \hat{\beta}_{(i)}\right)}{k'(AKO)}, \qquad i = 1,2,...,n \tag{2.24}$$

biçiminde verilir. Her bir D_i değeri, $F_{k',(n-k');0.5}$ kritik değeriyle karşılaştırılabilir. D_i'nin kritik değerden büyük olduğu durumlar için i'inci gözlemin etkili gözlem olduğu kabul edilir. Bununla birlikte, Cook ve Weisberg ile Montgomery ve Peck 1982'de D_i'nin yaklaşık 1.0 olduğu değerler için i'inci gözlemin etkili olduğunu belirtmişlerdir (Rousseeuw and Leroy, 2003). Cook Uzaklığı,

$$D_i = \left(\frac{r_i^2}{k'}\right)\left(\frac{h_{ii}}{1-h_{ii}}\right), \qquad i = 1,2,...,n \tag{2.25}$$

biçiminde hesaplanabilir. Yukarıdaki eşitlikten de görülebileceği gibi Cook Uzaklığı, regresyon modelindeki parametre sayısının dışında hem gözlem uzaklıkları matrisinin köşegen öğesinden hem de student türü artıklardan etkilenmektedir (Montgomery and Peck, 1992). D_i'nin çıkarımı Ek-A.4'de verilmiştir.

2.2.3.4. Düzeltilmiş Cook Uzaklığı (D^*)

Atkinson'ın 1981'de verdiği ve Cook Uzaklığı'nın bir başka uyarlaması olan bu istatistik,

$$D_i^* = |t_i| \left[\left(\frac{n-k'}{k'} \right) \left(\frac{h_{ii}}{1-h_{ii}} \right) \right]^{1/2} = |DFFITS_i| \left(\frac{n-k'}{k'} \right)^{1/2}, \quad i = 1,2,...,n \tag{2.26}$$

biçiminde hesaplanır. Her bir D_i^* değeri, $2\left[(n-k')/n\right]^{1/2}$ ile karşılaştırılır. Bu değerden büyük olan D_i^* değerine sahip gözlemler etkili gözlem olarak düşünülebilir.

Düzeltilmiş Cook Uzaklığı'nın Cook Uzaklığı'na avantajları, Chatterjee ve Hadi tarafından,

- Düzeltilmiş Cook Uzaklığı'nın kuşkulu değerlere daha fazla önem vermesi,
- D_i^* değerlerinin grafiksel gösterim için daha uygun olması.

olarak belirtilmiştir (Chatterjee and Hadi, 1986).

2.2.3.5. COVRATIO İstatistiği

Belsley, Welsch ve Kuh'un 1980'de belirttiği ve adını varyans-kovaryans matrisinin oranlanmasından (covariance ratio) alan bu istatistik, *i*'inci gözlemin parametre kestiricilerinin doğruluğu üzerine etkisini içerir. Bir başka deyişle COVRATIO İstatistiği, genel varyansın bir gözlemin çıkartılmasıyla nasıl etkilendiğini gösterir:

$$COVRATIO_i = \frac{det\left\{ s_{(i)}^2 \left(X'_{(i)} X_{(i)} \right)^{-1} \right\}}{det\left\{ AKO\left(X'X \right)^{-1} \right\}}, \quad i = 1,2,...,n \tag{2.27}$$

biçiminde hesaplanan COVRATIO İstatistiği'nin $1 \pm (3k'/n)$ aralığının dışında kalan değerlerine sahip gözlemler varyanslar üzerinde etkili gözlem olarak kabul edilebilir.

COVRATIO İstatistiği aşağıda verildiği gibi de hesaplanabilir:

$$COVRATIO_i = \left(\frac{n-k'-r_i^2}{n-k'-1} \right)^{k'} \Big/ (1-h_{ii}), \qquad i=1,2,...,n \tag{2.28}$$

Yukarıdaki eşitlikten de görülebileceği gibi COVRATIO İstatistiği, yüksek uç değerlilik ve aykırı değerlilikle yakından ilgidir. $COVRATIO_i$'nin 1.0'dan büyük olması, *i*'inci gözlemin parametre kestiricilerinin doğruluğunu arttırdığı; 1.0'dan küçük olması ise *i*'inci gözlemin parametre kestiricilerinin doğruluğunu azalttığı şeklinde yorumlanır (Rawlings ve diğerleri, 1998).

COVRATIO İstatistiği'ndeki determinantlar bazen ürkütücü olabildiğinden, COVRATIO yerine aynı özelliklere sahip aşağıdaki istatistik de kullanılabilir.

$$FVARATIO_i = \frac{var\left(\hat{y}_{i(i)}\right)}{var\left(\hat{y}_i\right)} = \frac{s_{(i)}^2}{AKO(1-h_{ii})}, \qquad i=1,2,...,n \tag{2.29}$$

2.2.3.6. DFTSTAT İstatistiği

Belsley, Welsch ve Kuh'un 1980'de belirttiği ve adını *t*-istatistik değerlerindeki farktan (difference in *t*-statistics) alan bu istatistik, *i*'inci gözlemin çıkartılmasıyla *j*'inci regresyon katsayısının sıfıra eşit olup olmadığını test ederken elde edilen *t*-istatistik değerindeki değişimi gösterir (Rousseeuw and Leroy, 2003). DFTSTAT İstatistiği,

$$DFTSTAT_{j(i)} = \frac{\hat{\beta}_j}{\sqrt{AKO\left(C_{jj}\right)}} - \frac{\hat{\beta}_{j(i)}}{\sqrt{s_{(i)}^2\left(\left(X_{(i)}'X_{(i)}\right)^{-1}\right)_{jj}}}, \qquad i=1,2,...,n \tag{2.30}$$

biçiminde hesaplanır. Büyük $|DFTSTAT_{j(i)}|$ değerleri için *i*'inci gözlemin etkili olduğu düşünülebilir.

2.2.3.7. Welsch Uzaklığı (WU)

Welsch'in 1982'de verdiği bu istatistik,

$$WU_i = |DFFITS_i| \sqrt{\frac{n-1}{1-h_{ii}}}, \qquad i = 1,2,...,n \tag{2.31}$$

biçiminde hesaplanır. $n > 15$ olan regresyon modellerinde kullanılan bu istatistik değerleri, $3\sqrt{k'}$ ile karşılaştırılır. Bu değerden büyük olan WU_i değerine sahip gözlemlerin etkili olduğu düşünülebilir (Chatterjee and Hadi, 1986).

2.2.3.8. PRESS İstatistiği

Allen'ın 1971'de belirttiği açık adı önkestirim hata kareler toplamı (prediction error sum of squares) olan bu istatistik, denklemlerin geçerlilik incelemesinde kullanıldığı gibi zaman zaman gözlemlerin artıklar üzerindeki etkilerini görmek için de kullanılabilir. (1.15) ve (1.16)'da belirtilen eşitlikler kullanılarak PRESS İstatistiği,

$$PRESS = \sum_{i=1}^{n} e_{(i)}^2 = \sum_{i=1}^{n} \left(y_i - \hat{y}_{i(i)}\right)^2 = \sum_{i=1}^{n} \left(\frac{e_i}{1-h_{ii}}\right)^2 \tag{2.32}$$

biçiminde hesaplanır. Özellikle, PRESS İstatistiği'nin AKT'den çok büyük olması durumunda etkili gözlemlerin varlığından söz edilir. Bununla birlikte bu istatistik, bir regresyon çözümlemesinde etkili gözlemlerin önemli bir faktör olup olmadığına karar vermede yardımcı olabilmektedir (Freund and Wilson, 1998). PRESS İstatistiği'nin çıkarımı Ek-A.5'te verilmiştir.

2.2.3.9. Andrews-Pregibon İstatistiği (AP)

Andrews ve Pregibon'un 1978'de verdiği ve

$$AP_i = \frac{det\{W'_{(i)}W_{(i)}\}}{det\{W'W\}}, \qquad i = 1,2,...,n \tag{2.33}$$

biçiminde hesaplanan bu istatistik, *i*'inci gözlemin $X'X$ matrisinin determinantı ya da AKT üzerindeki etkisini içerir. (2.33)'de belirtilen W matrisi,

$$W = \begin{bmatrix} 1 & x_{11} & x_{12} & . & . & . & x_{1k} & y_1 \\ 1 & x_{12} & x_{22} & . & . & . & x_{2k} & y_2 \\ . & . & . & & & & . & . \\ . & . & . & & & & . & . \\ . & . & . & & & & . & . \\ 1 & x_{n1} & x_{n2} & . & . & . & x_{nk} & y_n \end{bmatrix} \tag{2.34}$$

biçiminde tanımlanmaktadır. $W_{(i)}$ ise, (2.34)'de tanımlanan matrisin *i.* satırının çıkartılmasıyla elde edilen matristir. Diğer gözlemlerle karşılaştırıldığında küçük AP_i değerine sahip gözlemlerin etkili olabileceği düşünülür (Chatterjee and Hadi, 1986). Andrews-Pregibon İstatistiği,

$$AP_i = 1 - h_{ii} - \frac{e_i^2}{e'e}, \qquad i = 1,2,...,n \tag{2.35}$$

biçiminde de hesaplanabilir. Buradan da görülebileceği gibi AP_i değerlerinin, yüksek uç değerlerle ve aykırı değerlerle yakın ilgisi bulunmaktadır.

2.2.3.10. Etkili Gözlemlerin Saptanmasında Kullanılabilen Diğer İstatistikler

Andrews-Pregibon İstatistiği'ne bir alternatif olarak, Tatlıdil'in 1981'de doğrusal regresyon modelleri için kullanmanın daha uygun olduğunu önerdiği ve

$$det\{W'W\} = AKT.\left(det\{X'X\}\right) \tag{2.36}$$

eşitliği göz önünde bulundurularak elde edilen

$$R_i = \frac{AKT_{(i)}}{AKT}, \qquad i = 1,2,...,n \tag{2.37}$$

istatistiği etkili gözlemlerin saptanmasında kullanılabilir (Tatlıdil, 1981). Bu istatistik, diğer gözlemlerle karşılaştırıldığında küçük R_i değerine sahip gözlemlerin etkili olabileceğini belirtir.

Bunun yanında, Cook ve Weisberg'in 1980'de belirttiği ve

$$CW_i = -\frac{1}{2} log\left(COVRATIO_i\right) + \frac{k'}{2} log\left\{\frac{F_{k',n-k';\alpha}}{F_{k',n-k'-1;\alpha}}\right\}, \qquad i = 1,2,...,n \tag{2.38}$$

biçiminde hesaplanan istatistik de etkili gözlemlerin saptanmasında kullanılabilir. Bu istatistik, güven elipsi hacimlerinin oranı esasına dayanır. (2.38)'deki eşitliğin sağ tarafı göz önünde bulundurulursa, sabit olan kısım çıkartıldığında Cook-Weisberg İstatistiği'nin COVRATIO İstatistiği ile birebir ilişki içinde olduğu görülür. Diğer gözlemlerle karşılaştırıldığında çok büyük ya da çok küçük CW_i değerine sahip gözlemler etkili gözlem olarak düşünülür (Chatterjee and Hadi, 1986).

Çizelge 2.1. Tek Etkili Gözlemlerin Saptanmasında En Çok Kullanılan İstatistikler

ETKİ İSTATİSTİKLERİ	GÖSTERİM	FORMÜL	*i* GÖZLEMİNİN ETKİLİ OLABİLME DURUMU
DFFITS	$DFFITS_i$	$\frac{\hat{y}_i - \hat{y}_{i(i)}}{s_{(i)}\sqrt{h_{ii}}}$	$\lvert DFFITS_i \rvert > 2\sqrt{k'/n}$
DFBETAS	$DFBETAS_{j(i)}$	$\frac{\hat{\beta}_j - \hat{\beta}_{j(i)}}{s_{(i)}\sqrt{C_{jj}}}$	$\lvert DFBETAS_{j(i)} \rvert > 2/\sqrt{n}$
COOK UZAKLIĞI	D_i	$\frac{\left(\hat{\beta} - \hat{\beta}_{(i)}\right)'(X'X)\left(\hat{\beta} - \hat{\beta}_{(i)}\right)}{k'(AKO)}$	$D_i > F_{k',n-k';0.5}$
DÜZELTİLMİŞ COOK UZAKLIĞI	D_i^*	$\lvert DFFITS_i \rvert \left(\frac{n-k'}{k'}\right)^{1/2}$	$D_i^* > 2\left[(n-k')/n\right]^{1/2}$
COVRATIO	$COVRATIO_i$	$\frac{det\left\{s_{(i)}^2 \left(X'_{(i)}X_{(i)}\right)^{-1}\right\}}{det\left\{AKO\left(X'X\right)^{-1}\right\}}$	$COVRATIO_i \begin{cases} < 1 - 3k'/n \\ > 1 + 3k'/n \end{cases}$
WELSCH UZAKLIĞI	WU_i	$\lvert DFFITS_i \rvert \sqrt{\frac{n-1}{1-h_{ii}}}$	$WU_i > 3\sqrt{k'}$

ÜÇÜNCÜ BÖLÜM
ÇOKLU DOĞRUSAL REGRESYONDA ÇOKLU KUŞKULU GÖZLEMLERİN VARLIĞI ve İNCELENMESİ

Tek bir kuşkulu gözlemin saptanması için kullanılan çeşitli tekniklerin hemen hemen hepsi bu gözlemin veri kümesinden ayrılması temeline dayanmaktadır. Fakat bazı durumlarda bir kuşkulu gözlem bir diğerini çeşitli biçimlerde etkileyebilmektedir. Bu etkiler gizleme ve sürükleme etkileri olarak da bilinmektedir.

Gizleme (masking) etkisi, kuşkulu gözlemlerin başka kuşkulu gözlemler tarafından gizlenmesi sonucu tespit edilememesidir. Bir başka deyişle, iki kuşkulu gözlemli bir durum için veri kümesinden bir kuşkulu gözlem çıkarıldığında diğerinin kuşkulu gözlem olarak belirebilmesi demektir.

Sürükleme (swamping) etkisi ise, kuşkulu gözlemlerin regresyon doğrusunu kendilerine doğru çekmesiyle diğer bazı gözlemlerin uzaklaşması sonucu, kuşkulu gözlem olarak ele alınmalarıdır. Bir başka deyişle, iki kuşkulu gözlemli bir durum için birinin veri kümesinden çıkarılması sonucu diğerinin iyi bir gözlem durumunu almasıdır.

Yukarıda belirtilen bu iki etki nedeniyle, gözlem gruplarının potansiyel etkilerini ve dolayısıyla kuşkulu gözlem olup olmadıklarını incelemek için ikinci bölümde gösterilen bazı istatistikler geliştirilerek kullanılabilmektedir.

3.1. Çoklu Gözlemlerin Çıkarılmasına Dayalı İstatistikler

3.1.1. MDFFITS İstatistiği

Bu istatistiğin DFFITS İstatistiği'nden farkı, tek bir gözlemin değil de kuşku duyulan birden fazla gözlemin veri kümesinden çıkartılıp elde edilen kestirilmiş değerle, tüm veriler kullanıldığında bulunan kestirilmiş değer arasındaki farklılığın incelenmesidir:

$$MDFFITS_{(D_m)} = \left[\hat{\beta} - \hat{\beta}_{(D_m)}\right]' X'_{(D_m)} X_{(D_m)} \left[\hat{\beta} - \hat{\beta}_{(D_m)}\right] \qquad (3.1)$$

Burada D_m, m sayıda çıkartılacak gözlemlerden oluşan kümeyi; $\hat{\beta}_{(D_m)}$ ise m sayıdaki gözlem çıkartıldıktan sonra elde edilen parametre kestiricilerinin oluşturduğu sütun vektörünü göstermektedir (Belsley ve diğerleri, 1980).

MDFFITS İstatistiği, diğer gözlem kümeleri ile karşılaştırıldığında büyük değerler veren gözlem kümelerinin etkili gözlemlere sahip olduğunu belirtir. Bununla birlikte m, yaklaşık 20 gözlemi aştığında MDFFITS değerini hesaplamak zorlaşmaktadır.

3.1.2. Çoklu COVRATIO İstatistiği

Bu istatistik, (2.27)'de belirtilen COVRATIO İstatistiği'nin, birden çok gözlemin aynı anda çıkartılmasıyla elde edilmiş biçimidir:

$$COVRATIO_{(D_m)} = \frac{det\left\{s^2_{(D_m)}\left(X'_{(D_m)}X_{(D_m)}\right)^{-1}\right\}}{det\left\{AKO\left(X'X\right)^{-1}\right\}} \tag{3.2}$$

Burada $s^2_{(D_m)}$, m sayıdaki gözlem çıkartıldıktan sonra elde edilen varyans kestiricisidir. (3.2)'deki istatistik, diğer gözlem kümeleri ile karşılaştırıldığında çok büyük ya da çok küçük değerler veren gözlem kümelerinin varyans-kovaryans matrisi üzerinde büyük bir etkiye sahip olduğunu belirtir. Bununla birlikte, (2.29)'da belirtilen FVARATIO İstatistiği de birden çok gözlem için ele alınıp kullanılabilir (Belsley ve diğerleri, 1980).

3.1.3. Çoklu Andrews-Pregibon İstatistiği

Bu istatistik, Andrews ve Pregibon'un 1978'de belirttiği ve (2.33)'de verilen AP İstatistiği'nin (2.34)'de gösterilen W matrisinden birden çok satırın çıkartılmasıyla elde edilmiş biçimi olarak sunulur. Satır çıkartılması ile kastedilen durum, çoklu kuşkulu gözlemlerin saptanabilmesi için etkili olduğundan kuşku duyulan gözlemlerin aynı anda veri kümesinden çıkartılmasıdır:

$$AP_{(D_m)} = \frac{det\left\{W'_{(D_m)}W_{(D_m)}\right\}}{det\left\{W'W\right\}} \tag{3.3}$$

Bu istatistik, gözlem gruplarının uzaklık ölçütü olarak da yorumlanabilir. Bu yorum,

$$1-\sqrt{AP_{(D_m)}} \tag{3.4}$$

biçiminde tanımlanan ve $W'W$'nun oluşturduğu bir elipsin hacmindeki göreli değişimle ilişkilidir (Draper and John, 1981). (3.3)'deki determinant değerleri özünde hacim değerleri olarak düşünülebilir. Bu nedenle, diğer gözlem kümeleri ile karşılaştırıldığında küçük $AP_{(D_m)}$ değerlerini veren gözlem kümeleri etkili gözlemlere sahiptir denilebilir (Rousseeuw and Leroy, 2003).

Çoklu Andrews-Pregibon İstatistiği'nin hesaplanmasındaki uzun süreç göz önüne alındığında ikinci bölümde bahsedilen ve bilgisayar programlarından daha kolayca elde edilebilen (2.37)'deki Tatlıdil İstatistiği geliştirilip çoklu kuşkulu gözlemlerin saptanmasında kullanılabilir (Tatlıdil, 1981).

$$R_{(D_m)} = \frac{AKT_{(D_m)}}{AKT} \tag{3.5}$$

biçiminde hesaplanan bu istatistik, diğer gözlem kümeleri ile karşılaştırıldığında küçük $R_{(D_m)}$ değerlerini veren gözlem kümelerinin etkili gözlemlere sahip olduğunu belirtir.

Çizelge 3.1. Çoklu Etkili Gözlemlerin Saptanmasında Kullanılabilen İstatistikler

ETKİ İSTATİSTİKLERİ	GÖSTERİM	FORMÜL	D_m GÖZLEM KÜMESİNİN ETKİLİ OLABİLME DURUMU
MDFFITS	$MDFFITS_{(D_m)}$	$\left[\hat{\beta}-\hat{\beta}_{(D_m)}\right]' X'_{(D_m)}X_{(D_m)}\left[\hat{\beta}-\hat{\beta}_{(D_m)}\right]$	Diğer gözlem kümelerine göre büyük değerler için
ÇOKLU COVRATIO	$COVRATIO_{(D_m)}$	$\dfrac{det\left\{s^2_{(D_m)}\left(X'_{(D_m)}X_{(D_m)}\right)^{-1}\right\}}{det\left\{AKO\left(X'X\right)^{-1}\right\}}$	Diğer gözlem kümelerine göre çok küçük veya çok büyük değerler için
ÇOKLU ANDREWS-PREGIBON	$AP_{(D_m)}$	$\dfrac{det\left\{W'_{(D_m)}W_{(D_m)}\right\}}{det\left\{W'W\right\}}$	Diğer gözlem kümelerine göre küçük değerler için
ÇOKLU TATLIDİL	$R_{(D_m)}$	$\dfrac{AKT_{(D_m)}}{AKT}$	Diğer gözlem kümelerine göre küçük değerler için

3.2. İleri Araştırma Yöntemi

İleri araştırma yöntemi, bir veri kümesinde çoklu kuşkulu gözlemler tarafından oluşturulan gizleme ve sürükleme problemlerinden kaçınmak için geliştirilen bir yöntemdir (Chambers ve diğerleri, 2004). Hawkins 1983'te, bütün olası kuşkulu gözlemlerin çıkartılıp daha sonra bu gözlemlerin sırayla modele dahil edilip incelenmesi gerektiğini savunmuştur. Fakat, gizleme etkisi nedeniyle ne kadar gözlemin ve hangi gözlemlerin çıkartılacağının belli olmaması bu işlemin doğruluğunu azaltmıştır. Bu nedenle geliştirilen ileri araştırma yöntemi, çoklu gizlenmiş kuşkulu gözlemlerin saptanmasında ve onların kestirilmiş model üzerine etkilerini incelemede kullanılan sağlam bir yöntemdir (Atkinson and Riani, 2004).

İleri araştırma yöntemi için iki farklı uyarlama tanımlanmıştır. Bunlar, Hadi ve Simonoff (1993)'un önerdiği uyarlama ile Atkinson ve Riani (2000)'nin önerdiği uyarlamalardır (Chambers ve diğerleri, 2004). Bu uyarlamalar, aşağıda tek tek incelenmektedir:

3.2.1. Hadi ve Simonoff'un Önerdiği Uyarlama

İleri araştırma yönteminin, ikiden fazla bağımsız değişkenli çoklu regresyonda kullanılması gerekliliği göz önünde bulundurulup n gözlemli veri için regresyon modeli kurulur ve $|e_i|$ değerleri hesaplanır. Bu değerlerden en küçük k + 1 (modeldeki parametre sayısı) tanesine sahip gözlemler ile "temel" (basic) altküme oluşturulur. B, temel altkümeyi göstermek üzere $\hat{\beta}_B$, temel altkümedeki gözlemlerin oluşturduğu regresyon modelindeki kestirilmiş katsayıları; X_B de temel altkümedeki gözlemlerin oluşturduğu tam ranklı matrisi belirtir. X_B matrisinin tam ranklı olmaması durumunda ise $|e_i|$ değerleri göz önünde bulundurularak matris tam ranklı oluncaya kadar gözlem eklenir. $h = \lfloor (n+k-1)/2 \rfloor$ sayıda gözlemden oluşan aykırı değerlerden yoksun M altkümesinin belirlenebilmesi için aşağıdaki adımlar sırayla uygulanır. Burada $\lfloor . \rfloor$ sembolü, içindeki değerin tamsayı kısmını belirtir.

- B altkümesindeki gözlemlerle bir regresyon modeli oluşturulur ve

$$\frac{\left|y_i - \mathbf{x}_i'\hat{\beta}_B\right|}{\sqrt{1-\mathbf{x}_i'\left(X_B'X_B\right)^{-1}\mathbf{x}_i}}, \quad i \in B$$

$$\frac{\left|y_i - \mathbf{x}_i'\hat{\beta}_B\right|}{\sqrt{1+\mathbf{x}_i'\left(X_B'X_B\right)^{-1}\mathbf{x}_i}}, \quad i \notin B \tag{3.6}$$

değerleri hesaplanarak elde edilen sonuçlar yardımıyla gözlemler küçükten büyüğe doğru sıralanır.

- s temel altküme büyüklüğü olmak üzere, eğer $s = h$ ise ilk h gözlem M altkümesini oluşturur. Eğer $s < h$ ise $s + 1$ sıralı gözlem ile yeni bir temel altküme oluşturularak bu iki adım tekrar uygulanır.

İleri araştırma yöntemi, aykırı değerlerden yoksun s büyüklüğündeki M altkümesi ile aşağıdaki adımları takip ederek sürdürülür:

- $$u_i = \frac{y_i - \mathbf{x}_i'\hat{\beta}_M}{\hat{\sigma}_M\sqrt{1-\mathbf{x}_i'\left(X_M'X_M\right)^{-1}\mathbf{x}_i}}, \quad i \in M$$

$$= \frac{y_i - \mathbf{x}_i'\hat{\beta}_M}{\hat{\sigma}_M\sqrt{1+\mathbf{x}_i'\left(X_M'X_M\right)^{-1}\mathbf{x}_i}}, \quad i \notin M \tag{3.7}$$

değerleri hesaplanır. Burada $\hat{\beta}_M$, M altkümesindeki gözlemlerin oluşturduğu regresyon modelindeki kestirilmiş katsayıları; $\hat{\sigma}_M$, M altkümesindeki gözlemlerin oluşturduğu regresyon modeli için standart sapma kestiricisini; X_M de yine bu altkümedeki gözlemlerin oluşturduğu tam ranklı matrisi belirtir.

- Gözlemler, $|u_i|$ değerlerine göre küçükten büyüğe doğru sıralanır ve $u_{(s+1)}$ $(s+1)$'inci sıralı $|u_i|$ değeri olmak üzere, eğer $u_{(s+1)} \geq t_{(\alpha/2(s+1),s-k)}$ ise

$|u_i| \geq t_{(\alpha/2(s+1),s-k)}$ durumunu sağlayan bütün gözlemler aykırı değer olarak belirtilip ileri araştırma yöntemi sonlanır. Aksi takdirde, sıralı s + 1 gözlem alınarak yeni bir M altkümesi oluşturulur ve bu adımlar tekrar uygulanır. Eğer $n = s + 1$ durumu gerçekleşirse veride aykırı değer olmadığı belirtilip ileri araştırma yöntemi sonlanır (Hadi and Simonoff, 1993).

3.2.2. Atkinson ve Riani'nin Önerdiği Uyarlama

Atkinson ve Riani (2000), ileri araştırma yöntemine başlarken $m = k'$ olarak alınmasını ve çok sayıda altküme oluşturulmasını önermiştir. Burada m, altkümelerdeki gözlem sayısını belirtmektedir. Oluşturulabilecek bütün altkümelerin sayısı $\binom{n}{k'}$ biçiminde hesaplanır. Fakat, bu sayı çok büyük olduğunda altkümelerin sayısı genellikle 1000 olarak alınır. Başlangıç altkümesi, kesinlikle kuşkulu gözlem içermeyecek şekilde "en küçük ortanca artık kareye" (least median squared residual) sahip altküme olarak seçilir.

(2.34)'deki n x $(k'+1)$ boyutlu W matrisi için,

$$S^{(m)}_{i_1,\ldots,i_m} \equiv \left\{ w_{i_1},\ldots,w_{i_m} \right\} \tag{3.8}$$

biçiminde tanımlanan sıralanmış farklı m sayıda gözlemden oluşan kümeler olsun. Burada, $1 \leq i_1,\ldots,i_m \leq n$ ve $i_1,\ldots,i_m$ de bu kümelerdeki i'inci gözlem olmak üzere w'_{i_1}, W matrisinin i_1'inci satırıdır. $\iota' = [i_1,\ldots,i_m]$ ve $e_{i,S^{(m)}_\iota}$, $S^{(m)}_\iota$'deki i'inci gözlem için bulunan artık değer olursa $S^{(k')}_*$ biçiminde gösterilen başlangıç altkümesi aşağıdaki gibi seçilir:

$$e^2_{[Ortc.],S^{(k')}_*} = enk_\iota \left[e^2_{[Ortc.],S^{(k')}_\iota} \right] \tag{3.9}$$

Burada $e^2_{[a],S_i^{(k)}}$, $e^2_{i,S_i^{(k)}}$ $i=1,2,...,n$ arasından seçilen a'ıncı sıralı artık karedir; *Ortc.* ise,

$$Ortc. = \lfloor (n+k'+1)/2 \rfloor \tag{3.10}$$

biçiminde tanımlanıp $(n+k'+1)/2$ değerinin tamsayı kısmını belirtmektedir.

Bu uyarlamada m gözlemli başlangıç altkümesi olan $S_*^{(m)}$'ye gözlem eklerken, bu altküme yardımıyla oluşturulan regresyon modelinden elde edilen bütün artık kareler $\left(e^2_{i,S_*^{(m)}},\ i=1,2,...,n\right)$ sıralanır ve en küçük artık kareye sahip gözlem altkümeye dahil edilerek yeni m+1 gözlemli altküme oluşturulur. Çoğu zaman altkümeye tek bir gözlem dahil edilmesine karşın bazen de duruma göre altkümeye iki ya da daha fazla gözlem dahil edilebilir. İleri araştırma yöntemi, bütün gözlemler altkümeye girene kadar devam eder (Atkinson and Riani, 2002).

Bu yöntem, sağlam bir kestirim ile en küçük kareler kestiricilerinin birleşimi olarak ele alındığı için veride q sayıda kuşkulu gözlem varsa ileri araştırma yöntemi son q adımda bu gözlemleri altkümeye alacaktır. O zamana kadar da artık grafikleri ve parametre kestiricileri yaklaşık aynı kalacaktır. Parametre kestiricileri ve artıklar sabit olmasına karşın, ileri araştırma sürecinde gözlemlerden küçük artıklılar teker teker seçildiğinden σ^2'nin kestiricisi sabit kalmaz. Bu nedenle, kuşkulu gözlemlerin yokluğunda bile artık kareler ortalama kestiricisi $m < n$ için $s^2_{S_*^{(m)}} < s^2_{S_*^{(n)}} = s^2$ olur. $s^2_{S_*^{(m)}}$'deki azalış sonucu parametreler için hesaplanan t-değerleri artacağından m gözlem sayısı arttığında da t-değerleri azalır.

Bu yöntemde aykırı değer saptanması için kullanılabilen iki grafik, altkümede bulunmayan gözlemler için hesaplanan en küçük r-student türü artıklar ve altkümede bulunan gözlemler için hesaplanan en büyük student türü artıkların altküme büyüklüklerine göre belirlenmesiyle oluşur. En küçük r-student türü artıklar ve en büyük student türü artıklar sırasıyla,

$$t_{[m+1]} = enk\left|t_{i,S_*^{(m)}}\right| \qquad i \notin S_*^{(m)}, \quad m = k+1,\ldots,n-1 \tag{3.11}$$

ve

$$r_{[m]} = enb\left|r_{i,S_*^{(m)}}\right| \qquad i \in S_*^{(m)}, \quad m = k+1,\ldots,n \tag{3.12}$$

biçiminde hesaplanır. $\left\{m, t_{[m+1]}\right\}$ değerlerine göre oluşturulan grafiklerde oluşan sivri uç nokta ilk aykırı değer girişinden bir önceki adımı gösterir. Bununla birlikte, $\left\{m, r_{[m]}\right\}$ değerlerine göre oluşturulan grafiklerde ise keskin bir artış olduğu adım, ilk aykırı değerin altkümeye girdiği adımdır. Bazı durumlarda her iki grafikte de başka aykırı değerlerin altkümeye girişi sonrası gizleme etkisi sebebiyle azalışlar meydana gelebilir (Atkinson and Riani, 2000).

3.2.3. İleri Araştırma Yöntemi İçin Sarkıt (Stalactite) Grafiği

Woodruff ve Rocke 1994'de, bir problemin büyüklüğünün artmasıyla çoklu kuşkulu gözlemleri saptamanın zorlaşacağını belirtmiştir. Bu nedenle, amaç kuşkulu gözlemlerin saptanması olduğu için az sayıdaki ileri araştırma yeterli olabilmektedir. Az sayıdaki ileri araştırmanın yeterli olabilmesi için birkaç farklı başlangıç altkümesi belirlenerek sarkıt grafiği çizilebilir. Bu grafikte, satırlar altkümedeki gözlem sayısını; sütunlar ise gözlem numaralarını göstermektedir. Büyük mutlak değerli artıklara (2 ve 3 değerlerinden büyük artıklar) sahip gözlemler grafik üzerinde sembollerle işaretlenir. Bu grafikten, yaklaşık bütün altküme büyüklükleri için kuşkulu gözlemler görülebilir. Farklı birkaç başlangıç altkümesine dayalı çizilen sarkıt grafiklerinden genelleme yapılarak kuşkulu gözlemler saptanabilir. Bununla birlikte, sarkıt grafiklerinde bazı durumlarda altkümedeki gözlem sayısı m'nin küçük değerleri gösterilmeyebilir (Atkinson, 1994).

3.3. Bazı Sağlam Regresyon Kestirimleri

İstatistikte, hemen hemen en basit bir durumda bile varsayımların varlığından söz edilebilir. Matematiksel modellerde ortaya çıkan önemsiz sayılabilecek bir hatanın karar aşamasında da sadece küçük bir hataya neden olduğu bir ilke olarak kabul edilir. Ne yazıkki, bu durum her zaman düşünüldüğü gibi gerçekleşmez. Araştırmalar normal dağılım altında yaygın olan istatistiksel yöntemlerin çoğu varsayımlardan önemsiz sapmalara karşı oldukça duyarlı olduğunu göstermiştir. Bu nedenle, bu tür sapmalara karşı duyarsızlığı gösteren "sağlamlık" (robustness) kavramı ortaya çıkmıştır.

Sağlam yöntemler aşağıdaki özelliklere sahiptir (Huber, 1981):

- Varsayılan modelde yüksek bir verim sağlarlar.
- Model varsayımlarından küçük sapmaların performans üzerindeki etkisinin çok az olmasını amaçlarlar.
- Genelde, sağlam yöntemlerin kullanıldığı modellerdeki büyük sapmalar sonuçlar üzerinde büyük bir felakete (farklılığa) neden olmaz.
- Sağlam yöntemler, klasik parametrik oluşumlara parametrik olmayan ya da serbest dağılımlı oluşumlara oranla daha fazla uygulanır.

Çoğu istatistikçiler, sağlamlık kavramının kuşkulu gözlemleri ihmal etmek amacına sahip olduğunu düşünürler. Fakat tam tersine amaç, en küçük kareler kestirimiyle elde edilen artıklarla saptanamayan kuşkulu gözlemleri sağlam regresyon kestirimleri kullanılarak elde edilen artıklardan ortaya çıkarabilmektir.

Aslında, ikinci bölümde gösterilen çeşitli istatistikler ve sağlam regresyon yöntemleri aynı amacı ters sırayla takip etmektedir. Bu istatistikler, kuşkulu gözlemleri saptama amacıyla ayırıp kalan uygun veriyi en küçük kareler yardımıyla kestirme yolu izlerken, sağlam regresyon yöntemleri veriyle regresyon ilişkisi kurup daha sonra bu sağlam kestirimden elde edilen büyük artıklara sahip kuşkulu gözlemleri saptama yolunu izler.

En küçük kareler kestiriminin, bağımlı ve bağımsız değişkenlerdeki kuşkulu gözlemlere duyarlı olmasından dolayı kullanılan bazı sağlam regresyon kestirimleri aşağıda verildiği gibidir:

1. En Küçük Ortanca Kareler (EKOK) Kestirimi
2. En Küçük Budanmış (Trimmed) Kareler (EKBK) Kestirimi

Bu kestirimleri tek tek incelemeden önce bu kestirimlerle bağlantılı birkaç açıklama yapılacaktır.

Temel Küme (Elemental Set) Yaklaşımı:

Bu yaklaşım, genellikle doğrusal regresyonda kuşkulu gözlemlerin yer aldığı veri kümeleri için uygun parametre kestiricilerini bulmak amacıyla kullanılır (Smyth and Hawkins, 2000). Verilen veri kümesiyle bağlantılı birden çok kestirim, altkümeler oluşturularak elde edilir. k' sayıda katsayı içeren doğrusal regresyon modeli için bu yaklaşım, k' sayıda durumdan oluşan farklı altkümeler yardımıyla farklı katsayı vektörleri verir. Bu vektörlerin her biri verinin bir parçasının özelliklerini içerir.

Temel küme yaklaşımı, düzgün olmayan ve parametre uzayının kestiriminde fazlaca yerel minimuma sahip veri kümelerinde faydalı sonuçlar verir. Bu yaklaşım, farklı altkümeler kullanılarak çokça kestirim ve her bir kestirim için de kestirilmiş parametreleri ve kestirilmiş artık kümesini içerir. Bu farklı kestirimlerdeki parametre kestiricilerinin ortalaması alınarak ya da bazı ölçüt (EKOK ve EKBK ölçütleri) fonksiyonlarını en küçükleyen farklı temel kümeler araştırılarak uygun bir regresyon modeli elde edilebilir (Smyth and Hawkins, 2000).

Yüksek Bozulma Noktası (High Breakdown Point):

Bozulma noktası, verideki kuşkulu gözlemlere karşı dirençli (resistant) kestiriciler elde etmek için geliştirilen bir durumdur. Donoho ve Huber 1983'de, rastgele seçilen n gözlemli bir örneklem $S_n = (x_1, ..., x_n)$ ve bir regresyon kestiricisi T_n olmak üzere T_n'nin S_n'deki bozulma noktasını,

$$\varepsilon_n^*(T_n, S_n) = \frac{1}{n} enb\left\{ m \mid \underset{i_1,...,i_m}{enb} \ \underset{y_1,...,y_m}{sup} \| T_n(z_1, ..., z_n) \| < +\infty \right\} \qquad (3.13)$$

biçiminde göstermişlerdir. Burada $(z_1,...,z_n)$ örneklemi, $x_{i1},...,x_{im}$ değerlerinin $y_1,...,y_m$ ile değiştirildiği S_n örnekleminden elde edilmiştir. Bozulma noktası, S_n'ye bağlı kalmamakla birlikte 1/2'ye eşittir. Kuşkulu gözlemlerin çok olduğu durumlar için de bozulma noktası, $\left(\lfloor (n-k)/2 \rfloor + 1\right)/n$ biçiminde hesaplanabilir. EKOK ve EKBK kestiricileri, yüksek bozulma durumuna dayalı kestiricilerdir ("Robust Regression", 2003).

3.3.1. En Küçük Ortanca Kareler (EKOK) Kestirimi

Rousseeuw, 1984'de hem doğrusal dönüşüm için uygun hem de %50'lik yüksek bir bozulma noktasına sahip olan en küçük ortanca kareler (EKOK) kestirimini sunmuştur. Bu kestirimin kullanılma nedeni, büyük bir inceleme noktasına sahip olarak (yaklaşık 1/2) parametrelerin sağlam kestiricilerini bulma ve artıkların etkilerini azaltma yoluyla çoklu kuşkulu gözlemleri saptamaya eğilimli bir yöntem olmasıdır (Rousseeuw and Leroy, 2003). Kestirim,

$$EKOK = \underset{\beta}{enk} \ \underset{i}{ortc}\left(y_i - \mathbf{x}_i'\beta\right)^2 \tag{3.14}$$

biçiminde tanımlanır. Bu kestirimin hesaplanmasında PROGRESS programı kullanılacaktır ("Robustness", 1999). Bu programın EKOK için kullandığı algoritma Ek-B.1'de verilmiştir.

3.3.2. En Küçük Budanmış (Trimmed) Kareler (EKBK) Kestirimi

EKBK kestirimi, Rousseeuw tarafından 1984'te

$$EKBK = \underset{\beta}{enk} \sum_{i=1}^{h} \left(\left(y_i - \mathbf{x}_i'\beta\right)^2\right)_{i:n} \tag{3.15}$$

biçiminde belirlenmiştir (Rousseeuw and Leroy, 2003). Burada $\left(\left(y_i - \mathbf{x}_i'\beta\right)^2\right)_{i:n}$, i'inci sıralı artık kareyi göstermekte olup kestirimdeki "budama" sözcüğü de kuşkulu

gözlemlerin değişkenler üzerindeki etkilerinin sınırlanması anlamında kullanılır. (3.15)'deki kestirim dikkatle incelenirse, en küçük artık kareler kestirimine olan benzerliği görülür. Her iki kestirim de artık kareler toplamıyla ilgilenmekte olup EKBK kestirimi büyük artık kareleri toplama dahil etmemektedir. EKOK kestirimi gibi EKBK kestirimi de yaklaşık %50'lik yüksek bir bozulma noktasına sahiptir. EKBK kestiriminin hesaplanmasında da PROGRESS programı ("Robustness", 1999) kullanılacak olup algoritması Ek-B.2'de verilmiştir.

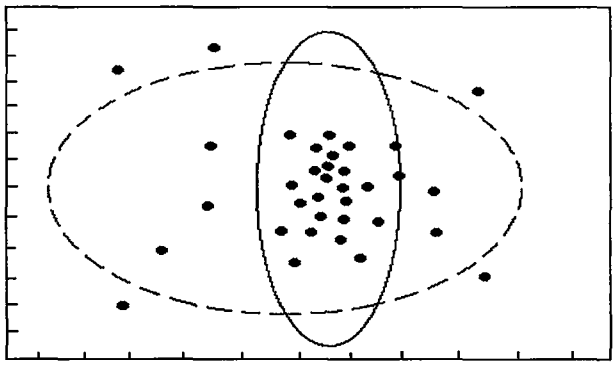
Yapılan incelemeler sonucu, Rousseeuw ve Van Driessen 1998'de yerel etkilere daha az duyarlı olması sebebiyle EKBK kestiriminin EKOK kestirimine oranla daha sağlam olduğunu belirtmiştir (Yaffee, 2002).

3.4. En Küçük Hacimli Elips (EKHE) Yöntemi

En küçük hacimli elips yöntemi, Rousseeuw'un 1985'de belirttiği ve çoklu kuşkulu gözlemlerin saptanmasında gizleme ve sürükleme etkilerini ortadan kaldırabilen sağlam bir yöntemdir (Rousseeuw and Van Zomeren, 1990). X, k boyutlu bir örneklem olmak üzere,

$$T(X) = X\text{'teki en az } h \text{ gözlemi içeren en küçük hacimli elipsin merkezi} \tag{3.16}$$

biçiminde tanımlanıp h, $\lfloor n/2 \rfloor + 1$ ya da $\lfloor (n+k+1)/2 \rfloor$'ye eşit olarak alınabilir. En küçük hacimli elipsin bir görünümü aşağıda verildiği gibidir:



Şekil 3.1. En Küçük Hacimli Elipsin Görünümü

Şekil 3.1'de, ortalama ve varyans-kovaryans matrisi yardımıyla elde edilmiş $(1-\alpha)$ güvenilirlikteki bir güven elipsi kesikli çizgiler ile gösterilmiştir. Güven elipsinin büyüklüğünü beş gözlem etkilemektedir. Güven elipsinin dışında kalan bu beş gözlem kuşkulu gözlemler olarak belirtilir. Bunun yanında, sürekli çizgiyle gösterilen elips en küçük hacimli elips olup gözlemlerin çoğunu içerisinde bulundurmaktadır.

Gözlemlerin en azından yarısını içine alacak biçimde belirlenen kestiricilere dayalı bir yöntem olan EKHE'nin, gözlem sayısı arttıkça hesaplanması zaman alıcı ve zahmetli olmaktadır. Bu yöntem ile kuşkulu gözlemlerin belirlenmesi aşağıda açıklandığı gibidir:

Öncelikle, k' sayıda gözlem rastgele seçilerek bir altküme oluşturulur. Oluşturulan bu altküme için ortalama ve varyans-kovaryans matrisi yardımıyla MU_i değerleri hesaplanır. Altkümedeki gözlem sayısı s ise MU_i değerlerinden en küçük s + 1 tanesiyle yeni alt küme oluşturulur. Altkümede n - h gözlem bulunana kadar bu işlemler tekrarlanır. Son adımda elde edilen altkümeye ait MU_i değerleri ile bu altküme için elips hacmi belirlenir. Bu işlem, $\binom{n}{k'}$ sayıda seçilen tüm altkümeler için tekrarlanıp içlerinden en küçük hacime sahip altküme belirlenir. Bu altkümedeki gözlemler iyi, altkümenin dışında kalan gözlemler ise kuşkulu olarak belirtilir (Kıral ve Billor, 2001). Bu işlemleri, el yordamıyla gerçekleştirmenin imkansız olması nedeniyle geliştirilen MINVOL programının ("Robustness", 1999) kullandığı algoritma Ek-B.3'de verilmiştir.

EKHE yönteminde, n/k değeri küçük olursa kuşkulu değerleri saptama zorlaşır. Bu durumda, az sayıdaki bazı veriler şans eseri çoklu doğrusal bağlantılı olabilirler. Bu nedenle, $n/k > 5$ olduğu durumlarda EKHE'nin kullanılması önerilir. Bununla birlikte bu yöntem, çok değişkenli normallik ve yaklaşık doğrusallık gibi varsayımların sağlanmaması durumlarında da kullanılabilir (Rousseeuw and Van Zomeren, 1990).

3.5. En Küçük Kovaryans Determinantı (EKKD) Yöntemi

En Küçük Kovaryans Determinantı Yöntemi, Rousseeuw'un 1985'de belirttiği ve çoklu kuşkulu gözlemlerin saptanmasında gizleme ve sürükleme etkilerini ortadan kaldırabilen bir başka sağlam yöntemdir. Bu yöntem, h gözlemli bir altkümeden yararlanılarak varyans-kovaryans matrisi determinantının en küçüklenmesi temeline dayanır.

$$T(X) = \text{Varyans - kovaryans matrisi determinantını en küçük yapan } h \text{ gözlemli } X\text{'teki bu } h \text{ gözlemin ortalaması} \tag{3.17}$$

biçiminde tanımlanıp h, $\lfloor (n+k+1)/2 \rfloor$'ye eşit olarak alınır. EKHE yöntemi gibi EKKD yöntemi de en uygun yarı veriyle hesaplandığından kuşkulu gözlemlerin bu yöntemin konum-ölçek (location-scale) kestiricilerini kaydırmayacağı belirtilir.

EKKD yönteminde amaç, n gözlem üzerinden klasik varyans-kovaryans matrisinin determinantını en küçük yapan h gözlemi bulmaktır. EKKD, yaklaşık olarak normal olması nedeniyle EKHE ile karşılaştırıldığında bir takım avantajlara sahiptir. Bunlar, EKKD yönteminin EKHE'ye göre istatistiksel olarak daha etkin olması ve EKKD'ye dayalı olarak hesaplanan sağlam uzaklıkların EKHE'ye dayalı olarak hesaplanan sağlam uzaklıklardan daha kesin olmasıdır. Bu nedenle de EKKD, çok değişkenli veri kümeleri içerisinde problemli gözlemleri belirlemeye daha uygun bir yöntemdir (Kıral ve Billor, 2001).

Bu yöntemin hesaplanmasında FAST-MCD programı kullanılacaktır ("Robustness", 1999). Bu program, Rousseeuw ve Van Driessen (1999) tarafından geliştirilmiş olup büyük veri kümelerine rahatlıkla uygulanabilir ve süreç bakımından daha hızlı sonuç verir. FAST-MCD programının kullandığı algoritma Ek-B.4'de verilmiştir.

3.6. Yüksek Bozulma Noktasına Sahip Yöntemlerle Kuşkulu Gözlem Saptaması

Yüksek bozulma noktasına sahip çeşitli yöntemlerin, yalnız çoklu aykırı değerleri, yalnız çoklu iyi ve kötü uç değerleri ya da her iki gözlem tipini de saptadığı ölçütler vardır.

Çizelge 3.2. Kestiricilerin Bozulma Noktaları (Rousseeuw and Leroy, 2003)

Kestirici	Bozulma Noktası (%)
Ortalama	0
Ortanca	50
En Küçük Ortanca Kareler	50
En Küçük Budanmış Kareler	50

Aşağıda çeşitli sağlam yöntemler ile çoklu kuşkulu gözlem saptama ele alınmıştır:

3.6.1. Sağlam Kestirimlerden Elde Edilen Artıklar Yardımıyla Saptama

Sağlam kestirimlerden elde edilen artıklar ile sadece aykırı değerler belirlenebilir. Bunun için önce standartlaştırılmış artıklar hesaplanır. Standartlaştırılmış artıkları hesaplarken standart sapma kestiricisinin $(\hat{\sigma})$ kullanılması gerekir. EKOK kestiriminde standart sapma kestiricisi,

$$\hat{\sigma} = C_1 \sqrt{ortc_i\, e_i^2} \tag{3.18}$$

biçiminde hesaplanır. Burada e_i, i'inci gözlem için EKOK kestiriminden elde edilen artığı, C_1 de artıkların Gauss Dağılımı'na uygunluğu için kullanılan düzeltme sabitini göstermektedir.

EKBK kestiriminde ise standart sapma kestiricisi,

$$\hat{\sigma} = C_2 \sqrt{\frac{1}{n} \sum_{i=1}^{h} \left(e^2\right)_{i:n}} \tag{3.19}$$

biçiminde hesaplanır. Burada $\left(e^2\right)_{i:n}$, EKBK kestiriminden elde edilen sıralı artık kareleri, C_2 de bir başka düzeltme sabitini belirtmektedir.

Elde edilen standartlaştırılmış artıklar yardımıyla,

$$\left\{\hat{y}_i,(e_i/\hat{\sigma})\right\} \text{ ya da } \left\{\text{gözlem no},(e_i/\hat{\sigma})\right\} \tag{3.20}$$

değerleri ile grafikler çizilerek çoklu aykırı değerler saptanabilir.
EKOK ve EKBK kestirimlerinden elde edilen çözümleri daha kuvvetli hale getirmek için ağırlıklı en küçük kareler kestirimi de kullanılabilir. Bu kestirim için ağırlık,

$$w_i = \begin{cases} 1, & |e_i/\hat{\sigma}| \le 2.5 \\ 0, & |e_i/\hat{\sigma}| > 2.5 \end{cases} \tag{3.21}$$

biçiminde tanımlanır. Burada, mutlak standartlaştırılmış artık sınırının 2.5 olmasının nedeni, etkili olabilecek aykırı değerlerin saptanması içindir. 2.5 değerini aşan mutlak standartlaştırılmış artıklara sahip gözlemler aykırı değer olarak saptanabilir. Ağırlıklı en küçük kareler kestirimi,

$$\underset{\beta}{enk} \sum_{i=1}^{n} w_i \left(y_i - \mathbf{x}_i'\beta\right)^2 \tag{3.22}$$

biçiminde belirtilir. Bu kestirimde kullanılan ağırlıklarla ya da bu kestirimden elde edilen artıklar yardımıyla çoklu aykırı değerler saptanabilir (Rousseeuw and Leroy, 2003). Bu kestirim gibi, (3.22)'ye benzer farklı bir fonksiyonla kurulan ve ağırlıklar üzerine dayalı başka sağlam kestirimlerle de çoklu aykırı değerler saptanabilir. Huber'ın 1981'de belirttiği kestirim için ağırlık

$$w_i = \begin{cases} 1, & |e_i/\hat{\sigma}| \le t \\ \dfrac{t}{|e_i/\hat{\sigma}|}, & |e_i/\hat{\sigma}| > t \end{cases} \tag{3.23}$$

biçimindedir. Burada $\hat{\sigma}$, artıkların standart sapmasını belirtmektedir. t değeri ise sabit bir değer olup Montgomery ve Peck tarafından 1982'de t = 2.0 alınması önerilmiştir (Lawrence and Arthur, 1990).

Andrews de 1972'de ağırlığı,

$$w_i = \begin{cases} \sin\left((e_i/\hat{\sigma})/d\right), & |e_i/\hat{\sigma}| \le d\pi \\ 0, & |e_i/\hat{\sigma}| > d\pi \end{cases} \tag{3.24}$$

biçiminde tanımlayıp bir başka kestirim sunmuştur. $\hat{\sigma}$, artıkların standart sapmasını belirtmekle birlikte $d = 2.1$ alınması önerilmiştir (Barnett, 1994).

3.6.2. EKHE ve EKKD Yöntemleri Kullanılarak Saptama

EKHE ve EKKD yöntemleri, sadece bağımsız değişkenleri ele aldığı için çoklu uç değerlerin saptanmasında kullanılan sağlam yöntemlerdendir. Bu yöntemlerden elde edilen her bir gözlem için sağlam uzaklıklar (SU_i), $\sqrt{\chi^2_{k;0.975}}$ değeriyle karşılaştırılır. Sağlam uzaklıkları, bu değerden büyük olan gözlemler uç değerler olarak belirtilir. Bununla birlikte, regresyondaki aykırı değerlerin ve uç değer olarak saptanan gözlemlerin iyi mi yoksa kötü mü uç değer oldukları belirlenebilir.

Rousseeuw ve Van Zomeren (1990), sağlam regresyon yöntemlerinden elde edilen standartlaştırılmış artıkların EKHE ya da EKKD yöntemlerinden elde edilmiş sağlam uzaklıklara karşı grafiğinin çizilmesini önermişlerdir.

$$\left\{SU_i, (e_i/\hat{\sigma})\right\} \tag{3.25}$$

değerleriyle çizilebilen grafiklerde, $y = -2.5$ ve $y = +2.5$ için çizilen iki yatay doğru küçük ve büyük standartlaştırılmış artıkları belirlemede ve $x = \sqrt{\chi^2_{k;0.975}}$ için çizilen bir dikey doğru da küçük ve büyük uzaklıkları belirlemede yardımcı olur. $y = -2.5$ ve $y = +2.5$ yatay doğrularının belirlediği bölgenin dışında kalan gözlemler aykırı

değerler olarak belirtilir. Bu değerlerden, $x=\sqrt{\chi^2_{k;0.975}}$ için çizilen dikey doğrudan büyük olan kısımda yer alanları kötü uç değerler olarak adlandırılır. Aykırı değer olmayıp $x=\sqrt{\chi^2_{k;0.975}}$ için çizilen dikey doğrudan büyük olan kısımda yer alan gözlemler de iyi uç değerler olarak belirtilir.

Çoklu uç değerlerin h_{ii} ve MU_i ile saptanmasının gizleme ve sürükleme etkileri nedeniyle doğru sonuçları vermeyeceği gerçeği, iyi ve kötü uç değerlerin belirlenmesinde y_i değerlerinin de kullanılması gerektiği, basit bir yöntemin Çizelge 3.3'de özetlenen durumların hepsini birarada veremeyeceğinin göstergesidir.

Çizelge 3.3. Farklı Durumlardaki Gözlemlerin Tanımlanması

Durum 1	Küçük SU_i ve küçük standartlaştırılmış artık → iyi gözlem
Durum 2	Küçük SU_i ve büyük standartlaştırılmış artık → aykırı değer
Durum 3	Büyük SU_i ve küçük standartlaştırılmış artık → iyi uç değer
Durum 4	Büyük SU_i ve büyük standartlaştırılmış artık → kötü uç değer

Rousseeuw ve Van Zomeren (1990)'in hangi yöntemlerden hesaplanan uzaklıklar ve standartlaştırılmış artıkların birlikte grafiklenmesi gerektiği önerisi Çizelge 3.4'de gösterilmiştir.

Çizelge 3.4. Uygun (Uzaklık, Standartlaştırılmış artık) Grafikleri

Uzaklık (x)	**Artık (y)**	**Sonuç**
(MU_i)	EKK $(e_i/\hat{\sigma})$'sı	Çoklu kuşkulu gözlemleri belirleyemez.
EKHE (SU_i)'si	EKOK $(e_i/\hat{\sigma})$'sı	Çoklu kuşkulu gözlemleri belirler.
EKKD (SU_i)'si	EKBK $(e_i/\hat{\sigma})$'sı	Çoklu kuşkulu gözlemleri belirler.

DÖRDÜNCÜ BÖLÜM
UYGULAMA

Bu bölümde, birinci bölümde açıklanan genel bilgiler doğrultusunda sırasıyla tek ve çoklu aykırı, uç, etkili gözlemler araştırılacak ve bu gözlemleri saptamada, iki gerçek veri kümesi için hangi yöntemin daha geçerli olduğu belirlenecektir.

4.1. Kira Verileri ve Tam Küme Çözümlemesi

Çözümlemede kullanılan veri, EVYAP GAYRİMENKUL LTD. ŞTİ.'den Gayrimenkul Danışmanı Sn. Kaan Göçük tarafından derlenmiştir. Veri, 2005 yılı Ağustos-Kasım ayları arasında Kadıköy-Merkez'deki 76 tane kiralık daire bilgisini içermektedir. Bu 76 gözlemin, rastgele seçilen ve içlerinde kuşkulu gözlem olmadığı varsayılan 20 tanesi geçerlilik çözümlemesi için ayrılmış ve kalan 56 gözlemli veri kümesinden çözümlemeye alınan değişkenler, dairenin kirası, büyüklüğü, kaçıncı katta olduğu, kiralarken evsahibine ödenecek olan deposite, dairenin ısınma durumu (soba, doğalgaz sobası, merkezi kalorifer, kombi), mutfak ve banyosunun durumu (yeni tip ve bakımlı, eski tip ve masraf isteyen) ve dairenin bulunduğu bölge (denize yakın veya uzak) olarak belirlenmiştir. Çözümlemede kullanılan bu değişkenler şu biçimde gösterilmiştir:

X1: Kiralık dairenin (m^2) cinsinden büyüklüğü
X2: Kiralık dairenin bulunduğu kat numarası
X3: Daire kiralanırken evsahibine ödenecek olan deposite miktarı (YTL)
S: Kiralık dairenin ısınması (s:soba, ds:doğalgaz sobası, mk:merkezi kalorifer, k:kombi)
M_B1: Kiralık dairenin mutfak ve banyosunun yeni tip ve bakımlı olup olmadığı
B1: Kiralık dairenin denize yakın olup olmadığı
Y: Dairenin kirası (YTL)

M_B1 ve B1 isimli değişkenler, 1 (olumlu durum) ve 0 (olumsuz durum) değerleri ile kodlanmış ve veri kümesi Çizelge 4.1'de verilmiştir.

Çizelge 4.1. Kiralık Dairelerle İlgili Veri

Gözlem No	Y	X1	X2	X3	S	M_B1	B1
1	425	60	1	390	ds	0	1
2	500	75	2	390	ds	0	1
3	550	70	2	390	k	1	1
4	550	90	0	650	k	1	0
5	600	85	3	650	k	1	0
6	650	110	2	650	mk	1	0
7	750	115	3	910	k	1	0
8	500	80	3	910	ds	1	0
9	650	90	3	1300	k	0	0
10	500	85	0	500	ds	1	0
11	400	60	0	300	s	0	0
12	750	130	2	650	k	0	0
13	500	85	2	500	ds	0	0
14	325	55	0	500	s	1	0
15	600	60	0	650	k	1	0
16	500	50	5	650	mk	1	0
17	800	135	0	1300	k	1	1
18	700	75	2	650	ds	1	0
19	650	90	2	650	k	1	0
20	500	70	3	780	ds	1	0
21	1200	100	3	1300	k	1	1
22	500	85	2	650	k	0	0
23	600	70	1	650	k	1	1
24	750	130	3	1300	k	0	0
25	500	85	3	650	s	0	0
26	550	75	3	650	ds	0	0
27	650	120	4	650	k	0	0
28	600	85	2	975	ds	0	1
29	1350	120	5	2600	k	1	1
30	1500	170	4	1600	k	1	1
31	650	85	2	1300	ds	1	0
32	500	65	3	650	ds	1	0
33	1000	100	2	1300	k	1	1
34	1000	110	3	1300	mk	1	1
35	1600	130	5	1300	mk	1	0
36	400	80	0	500	s	0	1
37	500	75	2	650	ds	0	0
38	550	70	2	800	ds	1	0
39	750	110	3	1300	k	1	1
40	600	85	2	800	k	0	0
41	500	80	4	500	s	0	1
42	500	75	0	650	mk	0	1
43	750	110	1	1120	k	1	0
44	600	85	1	650	ds	1	1
45	750	65	2	1300	k	1	1
46	750	100	3	1300	k	1	1
47	550	85	1	650	ds	1	1
48	500	65	0	650	mk	1	1
49	850	110	4	1300	mk	1	0
50	650	85	2	650	k	1	1
51	500	70	0	650	mk	1	0
52	500	70	1	650	s	1	0
53	650	65	5	650	ds	1	0
54	500	70	0	500	ds	0	1
55	1250	135	3	1600	k	1	1
56	650	85	3	650	ds	1	0

Kiralık dairenin ısınma durumu, çözümlemede kullanılmak üzere göstermelik (dummy) değişkenler yardımıyla, aşağıdaki gibi tanımlanmıştır:

Çizelge 4.2. Göstermelik Değişkenler

Isınma Durumu (S)	**I1**	**I2**	**I3**
Doğalgaz Sobası	1	0	0
Kombi	0	1	0
Merkezi Kalorifer	0	0	1
Soba	0	0	0

Veri, SPSS10.0 programı yardımıyla incelenmiş ve varsayımların, Y değişkenine logaritmik dönüşüm uygulandığında daha iyi sağlandığı görülmüştür. Ayrıca, R^2 değerindeki artış da dikkate alınarak bağımlı değişken olarak $\log y$ kullanılmış ve çözümlemeden elde edilen sonuçlar Çizelge 4.3'de verilmiştir.

Çizelge 4.3. Tam Küme Çözümleme Sonuçları

R	R^2	Düzeltilmiş R^2	Kestirimin Standart Hatası	Durbin-Watson
,928	,861	,837	,1318	2,253

	Kareler Toplamı	sd	Kareler Ortalaması	F	p
Regresyon	5,041	8	,630	36,289	,000
Artık	,816	47	1,737E-02		
Toplam	5,857	55			

	Katsayılar		Standartlaştırılmış Katsayılar		
	B	Standart Hata	Beta	t	p
Sabit	5,407	,085		63,520	,000
X1	5,8E-03	,001	,430	5,551	,000
X2	4,7E-02	,015	,207	3,177	,003
X3	2,4E-04	,000	,301	3,523	,001
I1	,104	,064	,151	1,638	,108
I2	,149	,069	,228	2,174	,035
I3	,169	,077	,183	2,203	,033
M_B1	,113	,042	,164	2,705	,009
B1	7,5E-02	,039	,114	1,940	,058

Çizelge 4.3'de Durbin-Watson İstatistiği ≈ 2.0 olduğu için gözlemlerin bağımsızlığından söz edilebilir. Yine bu çizelgedeki F değerine göre model anlamlı çıkmış ve değişkenler tek tek incelendiğinde I1 ile B1 değişkenleri 0.05'lik anlam düzeyinde anlamsız bulunmuştur.

4.2. Tek Kuşkulu Gözlemlerin İncelenmesi

4.2.1. Aykırı, Uç Değer ve Etkili Gözlem İstatistikleri

Kuşkulu gözlemlerin belirlenmesi sırasında kullanılan istatistikler için kritik değerler birinci ve ikinci bölümlerde belirtildiği gibi olup Çizelge 4.4'de verilmiştir. Bunun yanında, kritik değerlerin uygun olmadığı durumlarda gözlemler birbirleriyle karşılaştırılarak yorumlanmıştır.

Çizelge 4.4. Bazı İstatistikler İçin Kritik Değerler

İstatistikler	Kritik Değerler
h	0,321
MU^2	15,507
DFFITS	0,802
DFBETAS	0,267
D	0,940
D^*	1,832
COVRATIO	[0,518; 1,482]
WU	9,0

SPSS10.0 Paketi yardımıyla Çizelge 4.5, Çizelge 4.6, Çizelge 4.7 oluşturulmuştur. Çizelge 4.5'deki sonuçlar incelendiğinde, h_{ii} ve MU^2 değerlerine göre 16., 29. ve 41. gözlemlerin yüksek uç değer oldukları; r_i, $e_{(i)}$ ve t_i değerlerine göre de 21. ve 35. gözlemlerin aykırı değer oldukları görülmüştür. Çizelge 4.6'ya göre, 16. ve 35. gözlemlerin etkili oldukları söylenebilir. Bununla birlikte, varyans üzerindeki etkiyi gösteren COVRATIO ve FVARATIO İstatistikleri bu gözlemlerin yanısıra 21., 29., 30., 41. ve 42. gözlemlerin de etkili olduklarını göstermektedir. DFBETAS İstatistiği'ne sıfıra yakın çok küçük değerler alması nedeniyle Çizelge 4.6'da yer verilmemiştir. Çizelge 4.7'ye göre Welsch Uzaklığı, Andrews-Pregibon, Tatlıdil ve Cook-Weisberg İstatistikleri incelenerek 16., 21., 29., 35. ve 41. gözlemler etkili olarak bulunmuştur.

Çizelge 4.5'deki artıklar incelenip aykırı değer olduklarından kuşkulanılan 15., 18., 21. ve 35. gözlemler ikinci bölümün 2.2.1.1 nolu kesiminde anlatılan ortalama değişim aykırı değer modellemesi yardımıyla test edilmiş ve Çizelge 4.8, Çizelge 4.9, Çizelge 4.10 ve Çizelge 4.11'deki $\boldsymbol{U}$ değişkeninin anlamlılığı 0.05 anlam düzeyinde incelenerek 21. ve 35. gözlemlerin aykırı değer olduklarına karar verilmiştir.

Bütün bu istatistiklerin yanısıra, aykırı değerlerin saptanması için Bonferroni Testi yapılmış, Bonferroni kritik değeri = $t_{46;0.000446} = 3.553$ ile $| t_i |$ değerleri karşılaştırılmış ve bu test ile sadece 35. gözlemin aykırı değer olduğu görülmüştür. Yüksek uç değerleri belirlemede kullanılan test yardımıyla da $F_{8,47;0.000893} = 4.108$ değerini aşan 29. gözlemin ($F_{29} = 4.393$) yüksek uç değer olduğu anlaşılmıştır. Buna paralel olarak, $enb(h_{ii}) = 0.2 < 0.438 \leq 0.5$ olduğundan 29. gözlemin riskli olduğu söylenebilir. Ayrıca, PRESS İstatistiği hesaplanmış ve PRESS = 1.174 bulunmuştur. Bu değer, Çizelge 4.3'de verilen AKT'den büyük olduğu için etkili gözlemlerin varlığından söz edilebilir.

Çizelge 4.5. Uç ve Aykırı Değerlerle İlgili İstatistikler

i	h_{ii}	MU^2	e_i	d_i	r_i	$e_{(i)}$	t_i
1	,142	6,802	-,020	-,152	-,164	-,023	-,162
2	,145	6,994	,009	,072	,077	,011	,077
3	,193	9,648	-,025	-,186	-,207	-,030	-,205
4	,127	6,022	-,034	-,257	-,275	-,039	-,272
5	,108	4,943	-,058	-,437	-,462	-,064	-,458
6	,194	9,705	-,095	-,720	-,802	-,118	-,799
7	,093	4,151	-,069	-,527	-,554	-,077	-,549
8	,086	3,766	-,228	-1,731	-1,811	-,250	-1,857
9	,192	9,594	-,048	-,363	-,404	-,059	-,400
10	,148	7,182	-,020	-,150	-,163	-,023	-,161
11	,208	10,450	,167	1,265	1,422	,210	1,438
12	,190	9,453	,066	,498	,554	,081	,550
13	,103	4,688	,000	,004	,004	,001	,004
14	,245	12,479	-,173	-1,313	-1,510	-,229	-1,532
15	,166	8,125	,226	1,717	1,880	,271	1,934
16	,326	16,938	-,150	-1,141	-1,390	-,223	-1,404
17	,197	9,877	-,149	-1,128	-1,259	-,185	-1,267
18	,079	3,345	,246	1,864	1,942	,267	2,004
19	,089	3,897	,040	,305	,319	,044	,316
20	,088	3,839	-,139	-1,058	-1,108	-,153	-1,110
21	,078	3,293	,319	2,423	2,523	,346	2,685
22	,124	5,862	-,080	-,607	-,649	-,091	-,645
23	,123	5,788	,047	,357	,382	,054	,378
24	,156	7,621	-,136	-1,029	-1,120	-,161	-1,123
25	,209	10,503	,023	,172	,194	,029	,192
26	,108	4,946	,071	,541	,573	,080	,569
27	,191	9,521	-,113	-,855	-,951	-,139	-,950
28	,140	6,738	-,005	-,040	-,043	-,006	-,042
29	,438	23,111	-,081	-,614	-,819	-,144	-,816
30	,271	13,946	,020	,155	,182	,028	,180
31	,162	7,943	-,041	-,310	-,339	-,049	-,336
32	,093	4,135	-,080	-,604	-,634	-,088	-,630
33	,073	3,022	,184	1,393	1,446	,198	1,464
34	,174	8,609	,060	,453	,499	,072	,495
35	,233	11,834	,396	3,007	3,433	,517	3,924
36	,226	11,471	-,071	-,541	-,615	-,092	-,611
37	,105	4,775	,023	,171	,181	,025	,179
38	,085	3,694	-,002	-,018	-,019	-,003	-,018
39	,074	3,087	-,208	-1,582	-1,644	-,225	-1,675
40	,128	6,074	,067	,505	,541	,076	,537
41	,328	17,084	-,034	-,260	-,317	-,051	-,314
42	,270	13,865	-,024	-,180	-,211	-,033	-,209
43	,109	5,009	,002	,019	,020	,003	,019
44	,125	5,903	,005	,041	,043	,006	,043
45	,178	8,787	,098	,743	,819	,119	,816
46	,078	3,293	-,151	-1,144	-1,191	-,163	-1,196
47	,125	5,903	-,082	-,620	-,663	-,093	-,659
48	,209	10,499	-,079	-,602	-,677	-,100	-,673
49	,173	8,520	-,074	-,564	-,620	-,090	-,616
50	,120	5,636	-,006	-,045	-,048	-,007	-,048
51	,203	10,203	-,033	-,252	-,283	-,042	-,280
52	,221	11,197	,089	,675	,766	,114	,762
53	,189	9,391	,090	,681	,756	,111	,752
54	,151	7,337	,105	,798	,866	,124	,864
55	,114	5,279	,087	,658	,699	,098	,695
56	,095	4,224	,067	,511	,537	,074	,533

Çizelge 4.6. Tek Etkili Gözlemlerle İlgili İstatistikler I

i	DFFITS	D	D*	COVRATIO	FVARATIO
1	-,003	,000	,008	1,406	1,189
2	,002	,000	,004	1,418	1,195
3	-,006	,001	,013	1,492	1,265
4	-,005	,001	,011	1,371	1,169
5	-,007	,003	,016	1,305	1,140
6	-,023	,017	,052	1,331	1,250
7	-,007	,004	,016	1,262	1,119
8	-,022	,034	,049	,693	1,039
9	-,011	,004	,026	1,456	1,260
10	-,003	,001	,008	1,418	1,199
11	,044	,059	,100	1,031	1,233
12	,015	,008	,035	1,412	1,253
13	,000	,000	,000	1,353	1,139
14	-,056	,082	,128	1,027	1,286
15	,045	,078	,103	,720	1,130
16	-,073	,104	,166	1,234	1,452
17	-,037	,043	,084	1,110	1,229
18	,021	,036	,048	,620	1,018
19	,004	,001	,009	1,306	1,119
20	-,013	,013	,031	1,048	1,090
21	,027	,060	,062	,355	,955
22	-,011	,007	,026	1,278	1,156
23	,007	,002	,015	1,346	1,161
24	-,025	,026	,057	1,128	1,178
25	,006	,001	,014	1,523	1,290
26	,009	,004	,020	1,277	1,137
27	-,027	,024	,061	1,259	1,238
28	-,001	,000	,002	1,411	1,189
29	-,063	,058	,144	1,898	1,792
30	,008	,001	,017	1,655	1,401
31	-,008	,002	,018	1,417	1,217
32	-,008	,005	,019	1,238	1,117
33	,014	,018	,033	,869	1,052
34	,013	,006	,029	1,401	1,231
35	,120	,398	,275	,118	,991
36	-,021	,012	,048	1,459	1,310
37	,003	,000	,006	1,347	1,140
38	,000	,000	,000	1,326	1,117
39	-,017	,024	,038	,769	1,039
40	,010	,005	,022	1,316	1,165
41	-,017	,005	,038	1,773	1,518
42	-,009	,002	,020	1,648	1,398
43	,000	,000	,001	1,362	1,147
44	,001	,000	,002	1,387	1,168
45	,021	,016	,048	1,297	1,224
46	-,013	,013	,029	,999	1,074
47	-,012	,007	,027	1,275	1,157
48	-,021	,013	,048	1,404	1,278
49	-,016	,009	,035	1,362	1,225
50	-,001	,000	,002	1,379	1,161
51	-,008	,002	,019	1,500	1,280
52	,025	,019	,058	1,392	1,296
53	,021	,015	,048	1,340	1,244
54	,019	,015	,043	1,237	1,184
55	,011	,007	,025	1,247	1,141
56	,007	,003	,016	1,268	1,122

Çizelge 4.7. Tek Etkili Gözlemlerle İlgili İstatistikler II

i	WU	AP	R	CW
1	,026	,858	1,000	-,079
2	,013	,855	1,000	-,081
3	,049	,806	,999	-,092
4	,039	,871	,999	-,073
5	,055	,888	,995	-,063
6	,189	,795	,987	-,067
7	,056	,901	,994	-,056
8	,167	,850	,930	,075
9	,094	,805	,996	-,087
10	,028	,851	1,000	-,081
11	,365	,758	,957	-,012
12	,127	,805	,994	-,080
13	,000	,897	1,000	-,071
14	,478	,719	,952	-,011
15	,365	,772	,925	,066
16	,657	,646	,960	-,051
17	,303	,775	,967	-,028
18	,162	,847	,920	,099
19	,030	,909	,998	-,063
20	,104	,889	,974	-,015
21	,208	,797	,865	,220
22	,090	,868	,991	-,058
23	,052	,874	,998	-,069
24	,203	,821	,973	-,031
25	,050	,791	1,000	-,096
26	,068	,886	,993	-,058
27	,219	,793	,980	-,055
28	,007	,860	1,000	-,080
29	,624	,554	,987	-,144
30	,066	,728	1,000	-,114
31	,064	,836	,998	-,081
32	,064	,899	,991	-,051
33	,111	,886	,956	,026
34	,103	,821	,995	-,078
35	1,019	,575	,749	,460
36	,176	,767	,993	-,087
37	,021	,895	1,000	-,070
38	,002	,915	1,000	-,066
39	,128	,873	,942	,052
40	,078	,866	,994	-,065
41	,151	,670	,998	-,129
42	,076	,729	,999	-,114
43	,002	,891	1,000	-,072
44	,006	,875	1,000	-,076
45	,173	,811	,987	-,061
46	,098	,894	,971	-,005
47	,093	,867	,991	-,058
48	,175	,784	,990	-,079
49	,127	,820	,991	-,072
50	,006	,880	1,000	-,075
51	,071	,795	,999	-,093
52	,213	,769	,988	-,077
53	,172	,802	,988	-,069
54	,151	,835	,984	-,051
55	,088	,877	,990	-,053
56	,055	,900	,994	-,057

Çizelge 4.8. 15. Gözlem İçin Ortalama Değişim Aykırı Değer Modellemesi

	Katsayılar		Standartlaştırılmış Katsayılar	t	p
	B	Standart Hata	Beta		
Sabit	5,371	,085		63,365	,000
X1	6,2E-03	,001	,461	5,992	,000
X2	5,2E-02	,014	,230	3,572	,001
X3	2,3E-04	,000	,289	3,473	,001
I1	,101	,062	,146	1,634	,109
I2	,125	,068	,191	1,833	,073
I3	,161	,075	,174	2,154	,036
M_B1	,109	,041	,158	2,679	,010
B1	8,8E-02	,038	,134	2,302	,026
U	,271	,140	,111	1,934	,059

Çizelge 4.9. 18. Gözlem İçin Ortalama Değişim Aykırı Değer Modellemesi

	Katsayılar		Standartlaştırılmış Katsayılar	t	p
	B	Standart Hata	Beta		
Sabit	5,409	,083		65,543	,000
X1	5,7E-03	,001	,427	5,691	,000
X2	4,7E-02	,014	,209	3,309	,002
X3	2,4E-04	,000	,304	3,670	,001
I1	9,2E-02	,062	,133	1,479	,146
I2	,152	,067	,232	2,278	,027
I3	,173	,074	,187	2,331	,024
M_B1	,103	,041	,149	2,525	,015
B1	8,2E-02	,038	,124	2,171	,035
U	,267	,133	,109	2,004	,051

Çizelge 4.10. 21. Gözlem İçin Ortalama Değişim Aykırı Değer Modellemesi

	Katsayılar		Standartlaştırılmış Katsayılar	t	p
	B	Standart Hata	Beta		
Sabit	5,409	,080		67,602	,000
X1	5,9E-03	,001	,442	6,070	,000
X2	4,4E-02	,014	,194	3,153	,003
X3	2,3E-04	,000	,293	3,646	,001
I1	,109	,060	,157	1,811	,077
I2	,141	,065	,215	2,178	,035
I3	,175	,072	,189	2,429	,019
M_B1	,107	,039	,155	2,719	,009
B1	6,1E-02	,037	,093	1,667	,102
U	,346	,129	,142	2,685	,010

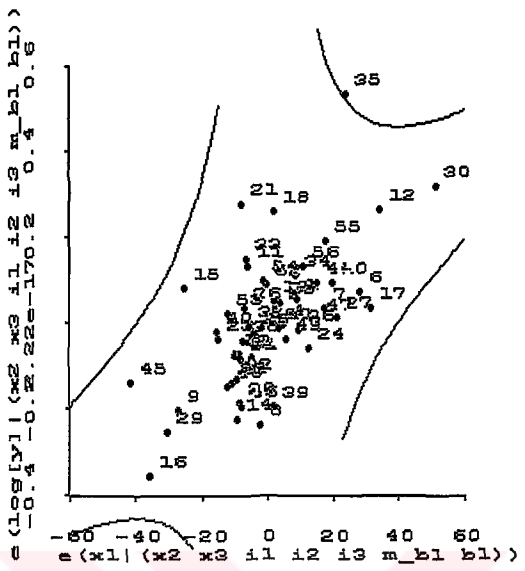
Çizelge 4.11. 35. Gözlem İçin Ortalama Değişim Aykırı Değer Modellemesi

	Katsayılar		Standartlaştırılmış Katsayılar		
	B	Standart Hata	Beta	t	p
Sabit	5,469	,076		71,846	,000
X1	5,0E-03	,001	,372	5,366	,000
X2	3,8E-02	,013	,168	2,908	,006
X3	2,5E-04	,000	,312	4,171	,000
I1	,115	,056	,166	2,065	,045
I2	,178	,061	,273	2,943	,005
I3	,128	,068	,139	1,891	,065
M_B1	,105	,037	,152	2,859	,006
B1	8,7E-02	,034	,132	2,551	,014
U	,517	,132	,212	3,924	,000

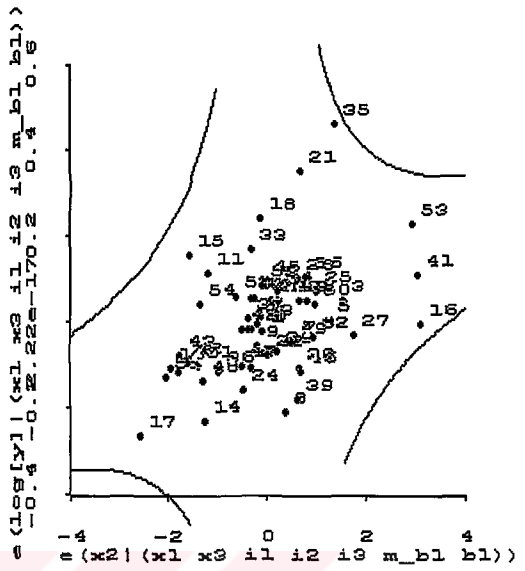
4.2.2. Tek Kuşkulu Gözlemlerin Grafiklerle İncelenmesi

Bütün bu istatistik incelemelerinden sonra, R-CODE Programı yardımıyla Eklenmiş Değişken Grafikleri ve Bileşen Artı Artık Grafikleri; SPSS10.0 Programı yardımıyla da Uç Değer-Artık Grafiği, artıklarla ve Cook Uzaklıklarıyla ilgili grafikler çizilip incelenmiştir.

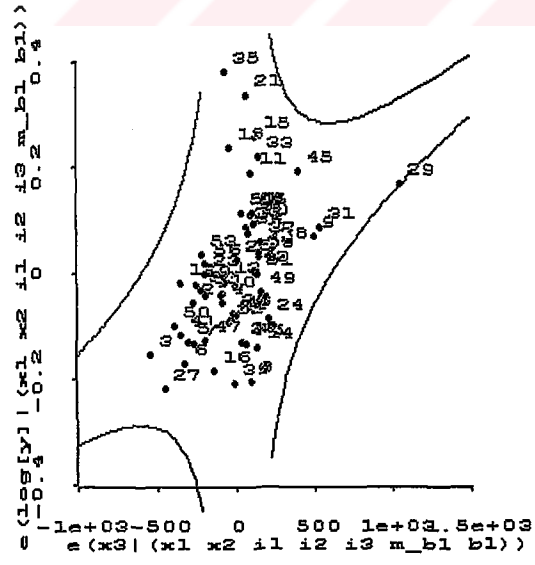
Şekil 4.1-Şekil 4.8, Eklenmiş Değişken Grafikleri olup bu grafiklerde genellikle 11., 14., 21., 29. ve 35. gözlemler etki sınırına yaklaştıkları ve veri kümesinden uzakta göründükleri için kuşku duyulan gözlemlerdir. Şekil 4.9-Şekil 4.16, Bileşen Artı Artık Grafikleri olup bu grafiklerde de özellikle 21. ve 35. gözlemler diğer gözlemlere göre kuşkulanılan gözlem konumundadır. Şekil 4.17, Uç Değer-Artık Grafiği'ni göstermekte olup bu grafikten 16., 29. ve 41. gözlemlerin yüksek uç değer oldukları; 21. ve 35. gözlemlerin de aykırı değer oldukları görülmektedir. Şekil 4.18 ve Şekil 4.19'dan da 35. gözlemin etkisinin diğer gözlemlerle karşılaştırıldığında ne kadar fazla olduğu görülebilir.



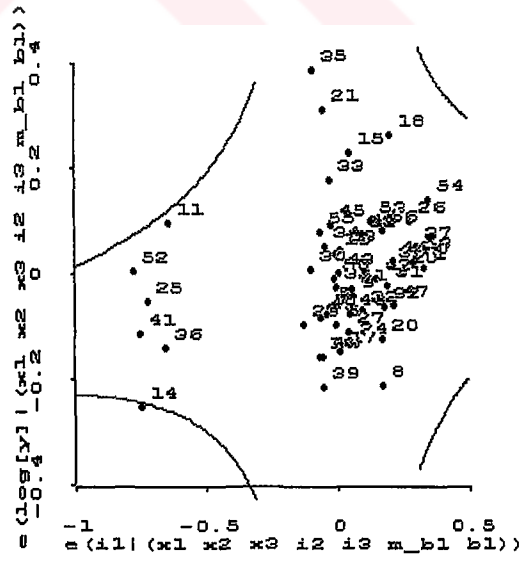
Şekil 4.1. X1 İçin Eklenmiş Değişken Grafiği



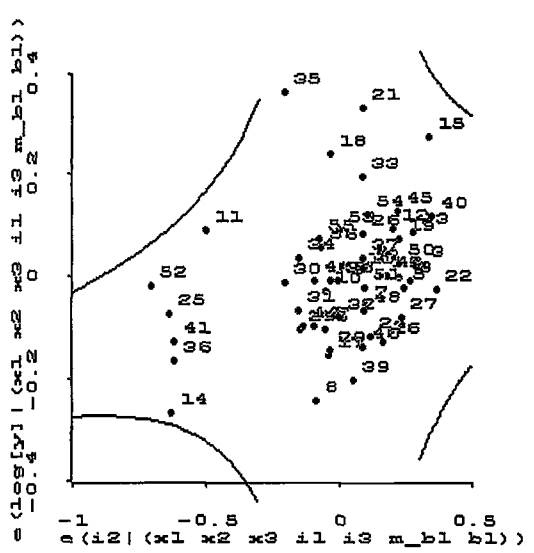
Şekil 4.2. X2 İçin Eklenmiş Değişken Grafiği



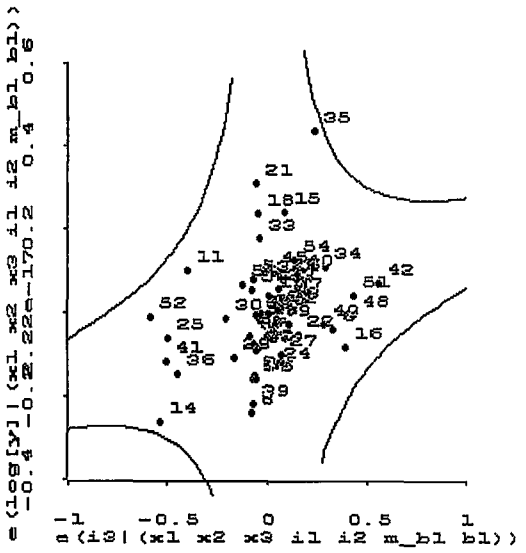
Şekil 4.3. X3 İçin Eklenmiş Değişken Grafiği



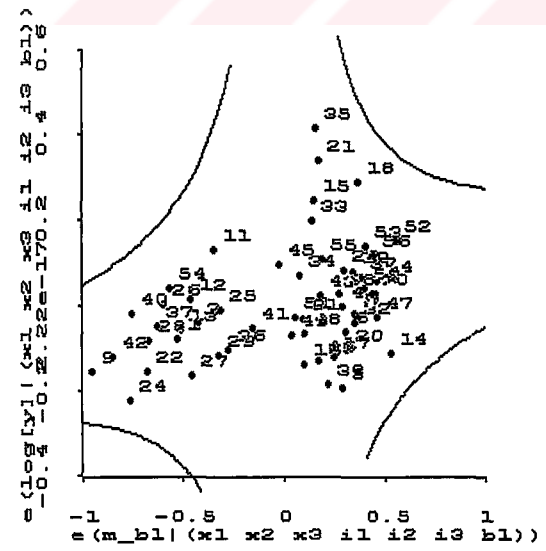
Şekil 4.4. I1 İçin Eklenmiş Değişken Grafiği



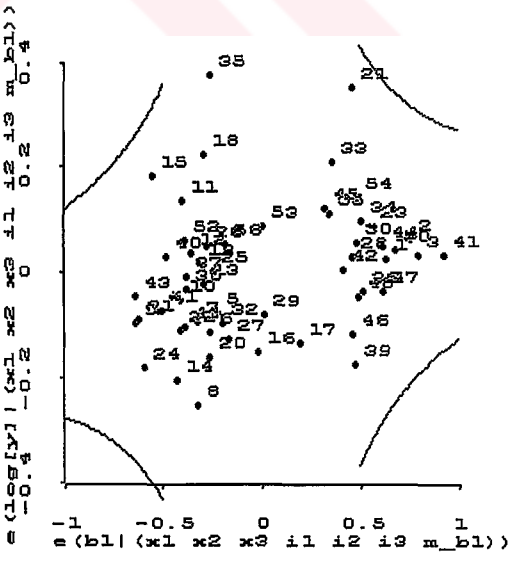
Şekil 4.5. I2 İçin Eklenmiş Değişken Grafiği



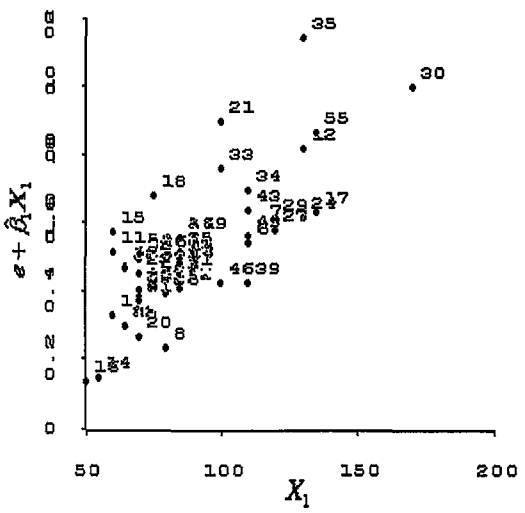
Şekil 4.6. I3 İçin Eklenmiş Değişken Grafiği



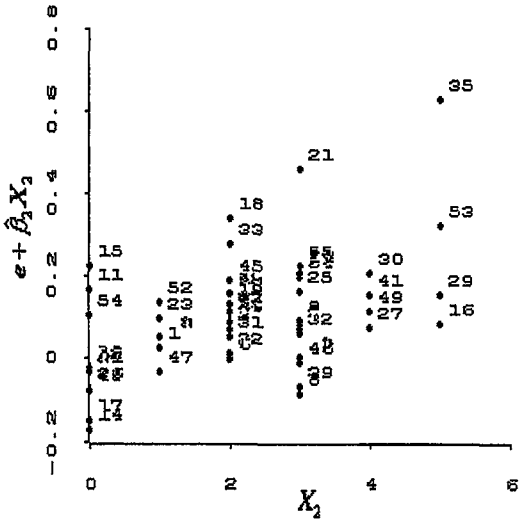
Şekil 4.7. M_B1 İçin Eklenmiş Değişken Grafiği



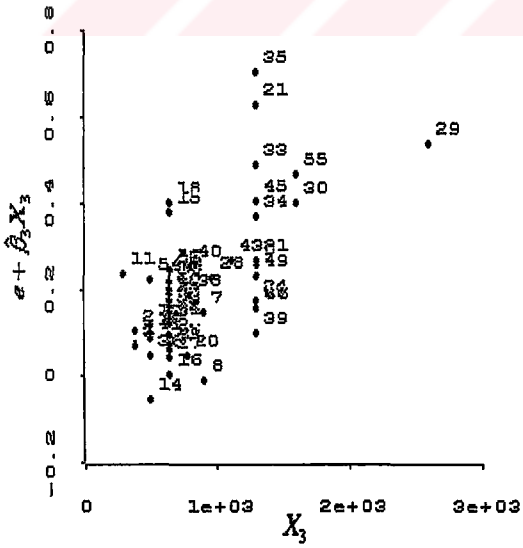
Şekil 4.8. B1 İçin Eklenmiş Değişken Grafiği



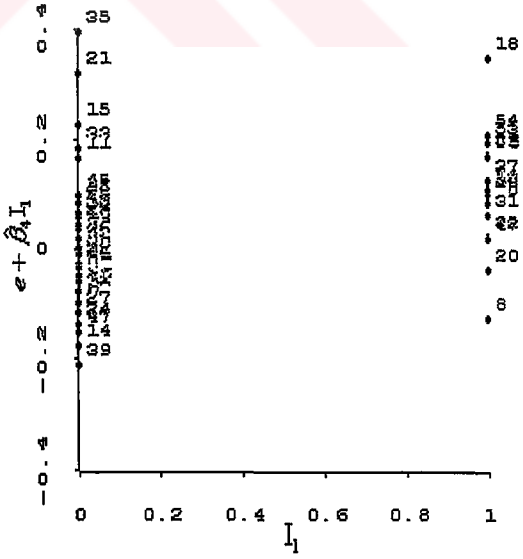
Şekil 4.9. X1 İçin Bileşen Artı Artık Grafiği



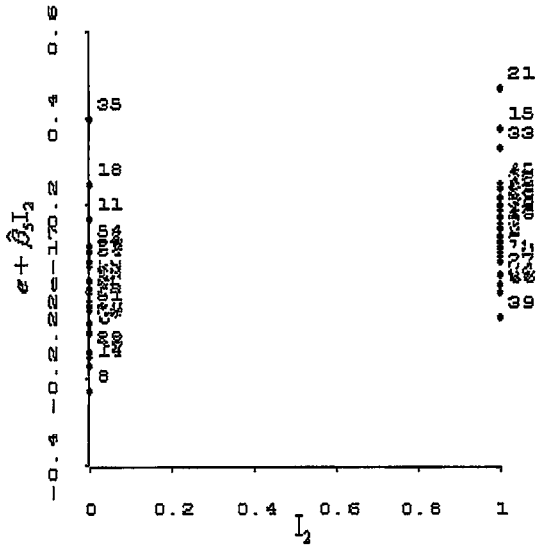
Şekil 4.10. X2 İçin Bileşen Artı Artık Grafiği



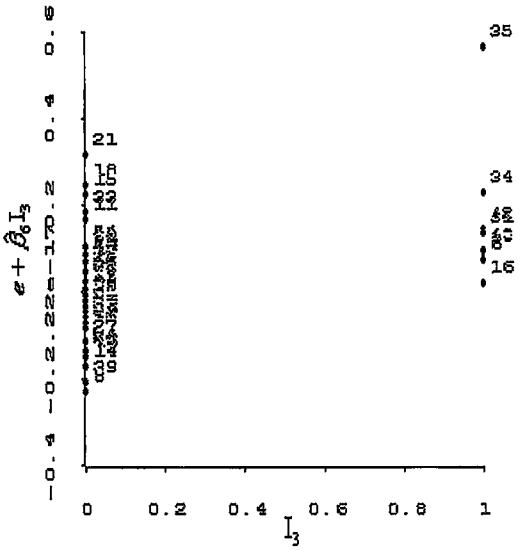
Şekil 4.11. X3 İçin Bileşen Artı Artık Grafiği



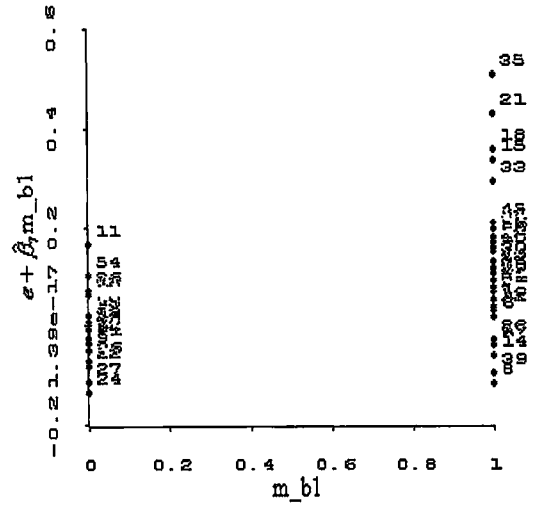
Şekil 4.12. I1 İçin Bileşen Artı Artık Grafiği



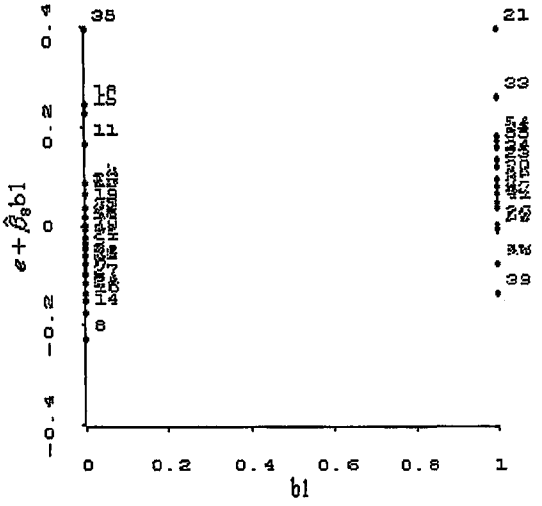
Şekil 4.13. I2 İçin Bileşen Artı Artık Grafiği



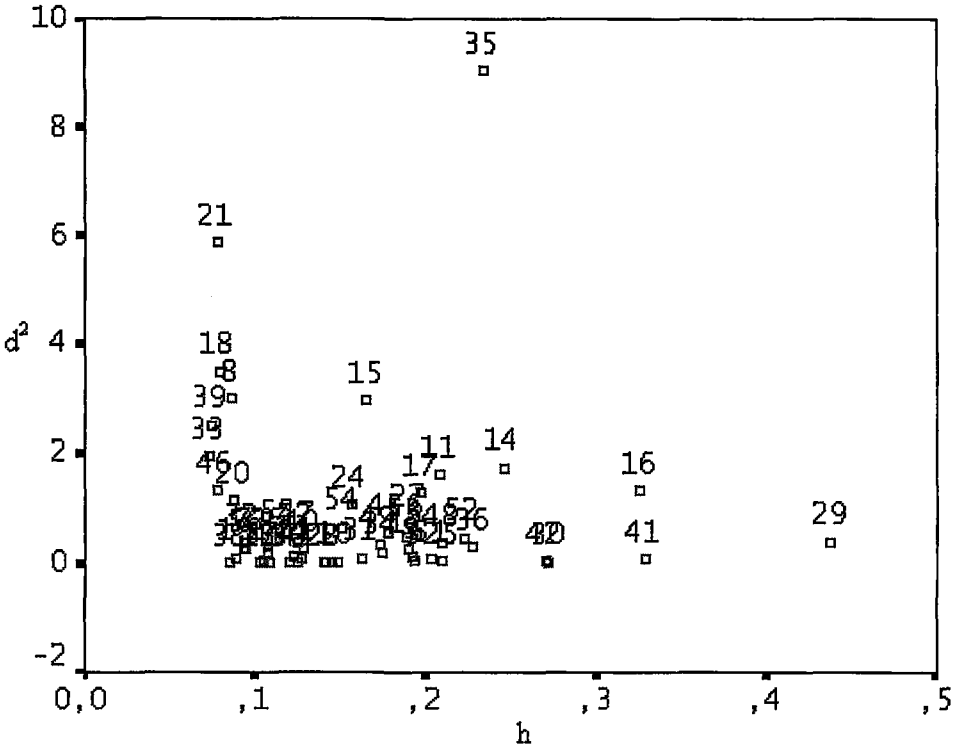
Şekil 4.14. I3 İçin Bileşen Artı Artık Grafiği



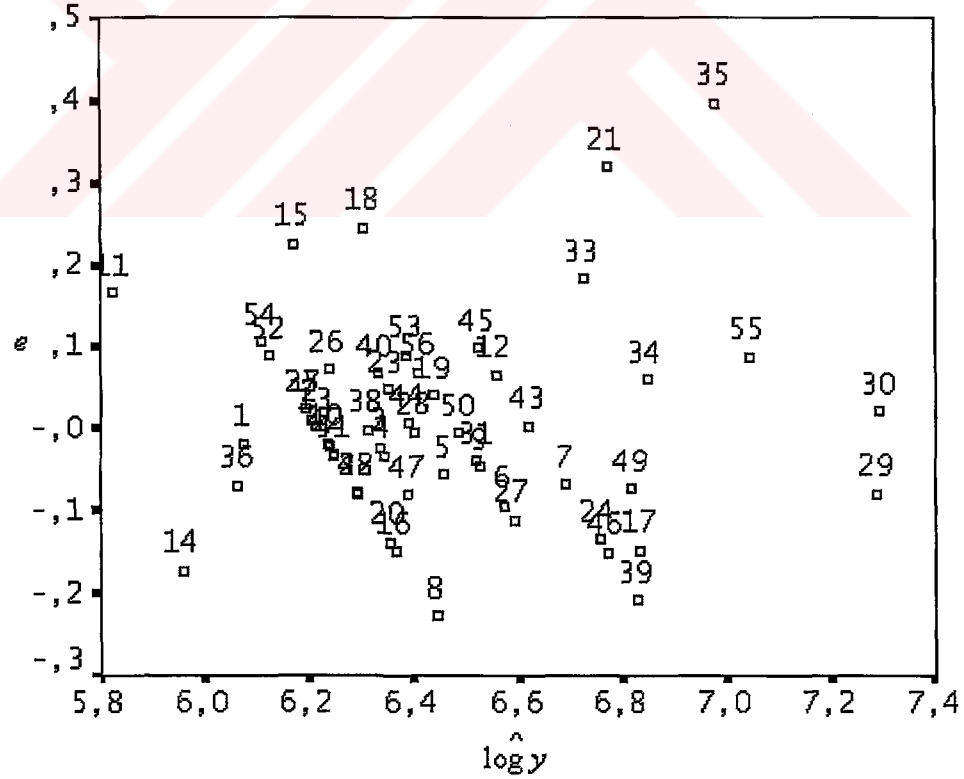
Şekil 4.15. M_B1 İçin Bileşen Artı Artık Grafiği



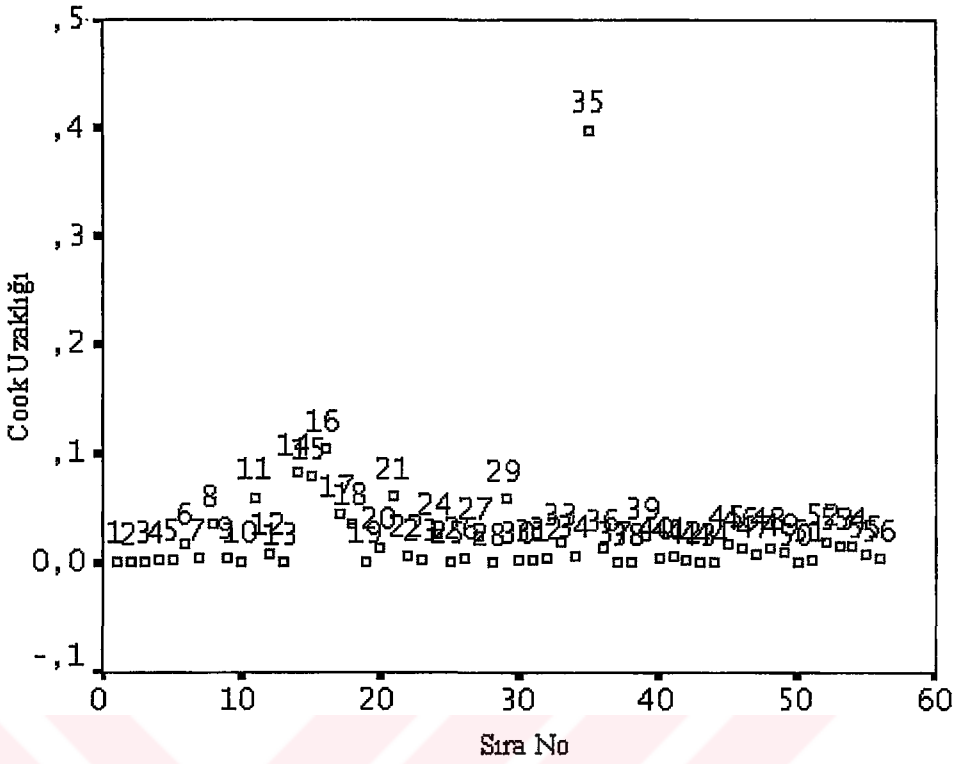
Şekil 4.16. B1 İçin Bileşen Artı Artık Grafiği



Şekil 4.17. Uç Değer-Artık Grafiği



Şekil 4.18. Kestirilmiş Değer-Artık Grafiği



Şekil 4.19. Gözlemlerin Cook Uzaklıkları Grafiği

4.2.3. Tek Kuşkulu Gözlemler İçin Genel Sonuç

Çizelge 4.12. Tek Kuşkulu Gözlemler İçin İnceleme

i	h_{ii}	MU^2	r_i	$e_{(i)}$	t_i	Bonf.	DFT	D	D^*	CVR	FVR	WU	AP	R	CW
16	+	+					+	+	+		+	+	+		
21			+	+	+					*	+			+	+
25										+					
29	*	*					+			*	*	+	*		*
30										+	+				
35			*	*	*	+	*	*	*	*	+	*	*	*	*
41	+	+								*	+		+		
42										+	+				

* : Etkili + : Daha az etkili ; DFT : DFFITS CVR : COVRATIO FVR : FVARATIO

Çizelge 4.12'de, tek kuşkulu gözlemlerin saptanabilmesi için kullanılan istatistikler ve bu istatistiklerin işaret ettiği gözlemler verilmiştir. "+" sembolü, Çizelge 4.4'de verilen kritik değerlerden çok fazla uzaklaşmayan veya gözlemlerin kendi aralarında karşılaştırılmaları sonucu çoğunluktan aşırı biçimde kopmayan gözlemleri belirtmektedir. "*" sembolü de Çizelge 4.4'de verilen kritik değerlerden fazlaca uzaklaşan veya gözlemlerin kendi aralarında karşılaştırılmaları sonucu çoğunluktan

aşırı biçimde kopan gözlemleri belirtmektedir.

Çizelge 4.12'ye göre 16., 29. ve 41. gözlemler tek uç değerler ve tek etkili gözlemler olarak saptanmış, 29. gözlem diğerlerine oranla daha etkili bulunmuştur. Bununla birlikte, 21. ve 35. gözlemler de tek aykırı değerler ve tek etkili gözlemler olarak saptanmış, 35. gözlem diğerine oranla daha etkili bulunmuştur. Bütün bu gözlemlerin dışında, 25., 30. ve 42. gözlemlerin de COVRATIO ve FVARATIO İstatistikleri yardımıyla varyans üzerinde etkili oldukları saptanmıştır.

4.3. Çoklu Kuşkulu Gözlemlerin İncelenmesi

4.3.1. Çoklu Kuşkulu Gözlemlerin İstatistiklerle İncelenmesi

Bazı gözlemlerin, gizleme ve sürükleme etkilerinden dolayı tek kuşkulu gözlemlerin saptanması sırasında yanlış konumda bulunması ve bu durum sonucunda da kuşkulu gözlemlerin tam olarak saptanamaması problemi çoklu kuşkulu gözlemlerin incelenmesiyle aşılabilmektedir. Öncelikle, tek etkili gözlemlerin saptanması sırasında kullanılan DFFITS, COVRATIO, AP ve R değerleri dikkate alınarak bu değerlerin işaret ettiği gözlemlerin önemli görülen bazı kombinasyonları çıkartılmış üçüncü bölümdeki (3.1), (3.2), (3.3) ve (3.5) eşitlikleri doğrultusunda EXCEL Paketi kullanılıp incelenmiştir. Buradan, Çizelge 4.13, Çizelge 4.14, Çizelge 4.15 ve Çizelge 4.16 elde edilmiştir. Bu çizelgelerde, 35. gözlemin yanısıra Çizelge 4.13'de 16. ve 29., Çizelge 4.14'de 8. ve 21., 18. ve 21., Çizelge 4.16'daki Tatlıdil İstatistiği için de 16. ve 21. gözlemler birlikte kestirimler üzerindeki etkiyi daha da arttırmıştır. Sadece Çizelge 4.15'de 29., 30. ve 41. gözlemlerin etkisi büyük COVRATIO durumu için görülür ve bu gözlemlerin veri kümesinde bulunması parametre kestiricilerinin doğruluğunu arttırdığı şeklinde yorumlanır. Bir başka dikkat edilmesi gereken durum, Çizelge 4.16'daki çoklu AP ve R değerlerinin uyumsuzluğudur. Bu uyumsuzluğun nedeni, gizleme ve sürükleme etkilerine duyarsızlığın çok da mümkün olmadığıdır. Sonuçlardaki çeşitlilik ve bütün gözlem kombinasyonlarının hesaplanmasının çok zaman aldığı düşünüldüğünde yukarıda belirtilen yöntemlerin fazla tercih edilmediği söylenebilir. Bu yöntemlerin aksine gizleme ve sürükleme etkilerine duyarsız daha sağlam olan ileri araştırma yöntemi, EKOK ve EKBK kestirimleri, EKHE ve EKKD yöntemleri kullanılabilir.

Çizelge 4.17, S-PLUS2000 Paketi'nin "fwd" isimli kütüphanesi ("Robust Diagnostic Regression Analysis: Software and Datasets", 2000) ile elde edilen ve aykırı değerleri belirlemede kullanılan ileri araştırma yönteminde gözlemlerin başlangıç altkümesine dahil olma basamaklarını göstermektedir. Çizelge 4.17'de bulunmayan 2., 3., 6., 13., 30., 36., 41., 43. ve 44. gözlemler başlangıç altkümesini oluşturan gözlemlerdir. Burada dikkatle incelenmesi gereken durum ise altkümeye son basamaklarda giren gözlemler ve dolayısıyla hangi basamaktan sonra altkümeye aykırı değer girişinin başladığıdır. Bu durumu belirlemek için Şekil 4.20-Şekil 4.26 incelenmiş ve şu sonuçlarla karşılaşılmıştır:

Çizelge 4.13. Çoklu Etkili Gözlemlerle İlgili İstatistikler I

Çıkartılan Gözlemler	MDFFITS
11 ve 15	0.116
11 ve 35	0,141
14 ve 15	0.051
14 ve 35	0.279
15 ve 35	0.293
16 ve 35	0,037
29 ve 35	0.157
11, 14 ve 15	0.071
11, 29 ve 35	0.135
14, 15 ve 35	0.119
16, 29 ve 35	0.325

Çizelge 4.14. Çoklu Etkili Gözlemlerle İlgili İstatistikler II

Çıkartılan Gözlemler	Küçük COVRATIO
8 ve 18	0.693
8 ve 21	0.207
8 ve 35	0.062
18 ve 21	0.205
18 ve 35	0.062
21 ve 35	0.028
8, 21 ve 35	0.013
18, 21 ve 35	0.013

Çizelge 4.15. Çoklu Etkili Gözlemlerle İlgili İstatistikler III

Çıkartılan Gözlemler	Büyük COVRATIO
29 ve 30	4.124
29 ve 41	4.433
29 ve 42	4.080
30 ve 41	3.448
41 ve 42	3.412
29, 30 ve 41	6.193
29, 30 ve 42	5.694

Çizelge 4.16. Çoklu Etkili Gözlemlerle İlgili İstatistikler IV

Çıkartılan Gözlemler	AP	R
16 ve 29	0.357	0.944
16 ve 35	0.369	0.735
21 ve 35	0.438	0.615
21 ve 41	0.535	0.868
29 ve 35	0.319	0.741
29 ve 41	0.371	0.984
35 ve 41	0.385	0.749
16, 29 ve 35	0.204	0.727
16, 21 ve 35	0.280	0.604

Çizelge 4.17. İleri Araştırma Yönteminde Gözlemlerin Kümeye Dahil Olması

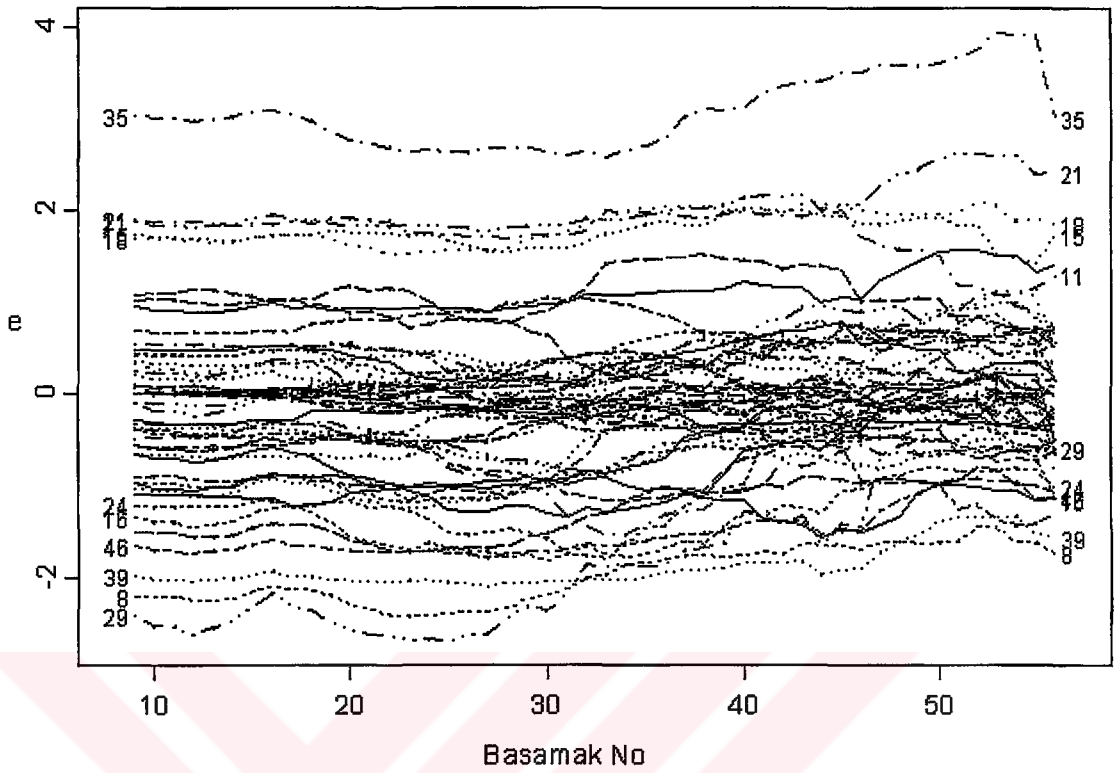
Basamak No(m)	Gözlem No	Basamak No(m)	Gözlem No
10	50	34	32
11	42	35	5
12	51	36	7
13	10	37	49
14	4	38	54
15	37	39	22
16	45	40	9
17	55	41	16
18	1	42	27
19	34	43	20
20	53	44	33
21	26	45	24
22	56	46	52
23	25	47	29
24	48	48	46
25	40	49	17
26	19	50	39
27	23	51	11
28	28	52	8
29	38	53	15
30	47	54	18
31	12	55	21
32	31	56	35
33	14		

- Şekil 4.20'ye göre, son basamaklara kadar büyük artıklara sahip 8., 11., 15., 18., 21., 35. ve 39. gözlemler aykırı değer olarak belirlenebilir.
- Şekil 4.21'e göre 16. ve 29. gözlemler uç değer olarak belirlenebilir.
- Şekil 4.22'ye göre değişkenlerin anlamsızlığı 50. basamak ve çevresinde başlamaktadır.
- Şekil 4.23'e göre Düzeltilmiş Cook Uzaklığı değerlerinde 50. basamak çevresinde keskin bir artış görülmektedir.
- Şekil 4.24'e göre s^2'deki artış ve Şekil 4.25'e göre de R^2'deki azalış 50. basamak çevresinde hızlanmaktadır.
- Şekil 4.26'daki sarkıt grafiğine göre de 8., 15., 18., 21., 29., 35. ve 39. gözlemler diğer gözlemlerle karşılaştırıldığında büyük artıklarından ötürü aykırı değer olarak belirlenebilir. Bu grafik, ileri araştırma yönteminde kullanılan başlangıç altkümesi temel alınarak oluşturulmuştur. Bundan dolayı, bu grafikten yaklaşık bütün altküme büyüklükleri için kuşkulu olabilecek gözlemler görülebilir.

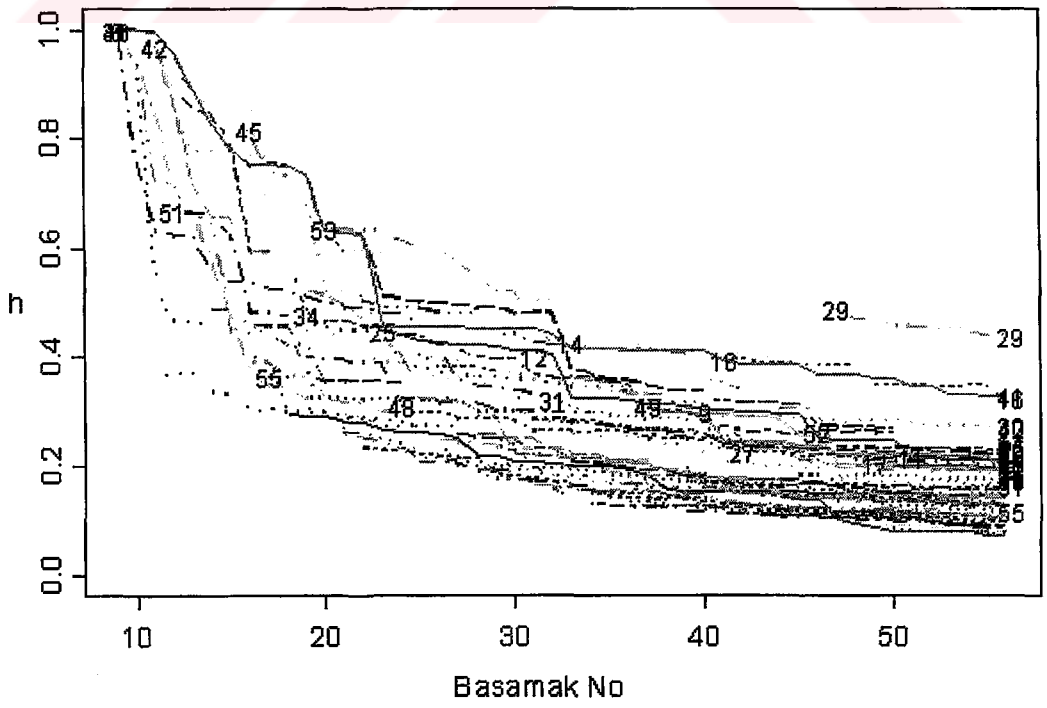
Sonuçta, aykırı değer girişinin yaklaşık 50. basamakta gerçekleştiği söylenebilir.

İleri araştırma yönteminde kullanılan, en küçük r-student türü artıkların ve en büyük student türü artıkların basamak numarasına karşı çizilmesiyle elde edilen grafiklere, sonuç bulmaya yönelik anlaşılır olmadığından yer verilmemiştir.

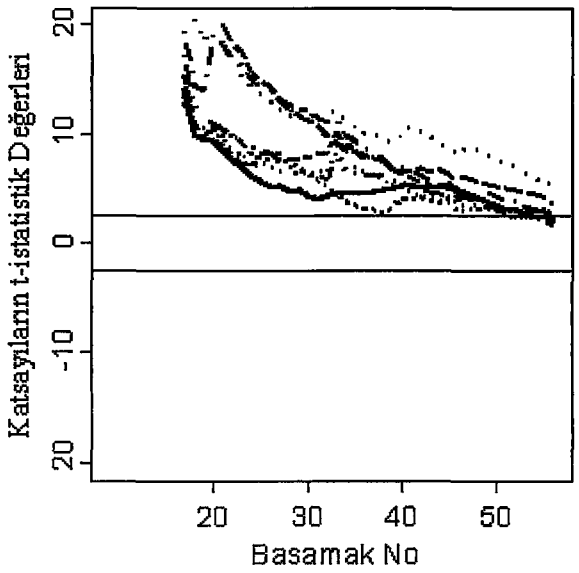
İleri araştırma yönteminin yanısıra, çoklu kuşkulu gözlemlerin saptanmasında kullanılan sağlam yöntemlerden EKOK ve EKBK kestirimleri FORTRAN tabanlı kod biçiminde bulunan PROGRESS Programı ile, EKHE ve EKKD yöntemleri de yine FORTRAN tabanlı kod biçiminde bulunan sırasıyla MINVOL ve FAST-MCD Programları ile elde edilmiştir. Çizelge 4.18'de bu yöntemlerden elde edilen çeşitli sonuçlar verilmiştir. Sadece, bağımsız değişkenlerdeki değerlerin çok fazla tekrar etmesinden dolayı seçilen en iyi altkümenin varyans-kovaryans matrisinin en az bir satır ve sütununun sıfıra çok yakın küçük değerlerden oluşması, FAST-MCD algoritmasına göre EKKD yönteminden bir sonuca ulaşılamamasını sağlamıştır. Çizelge 4.18'de verilen standartlaştırılmış artıklardan [-2.5, 2.5] aralığının dışında kalanları aykırı değer; sağlam uzaklıklardan da $\chi^2_{8;0.025} = 17.535 \Rightarrow \sqrt{17.535} = 4.187$ değerinden büyük olanları uç değer olarak belirtilir.



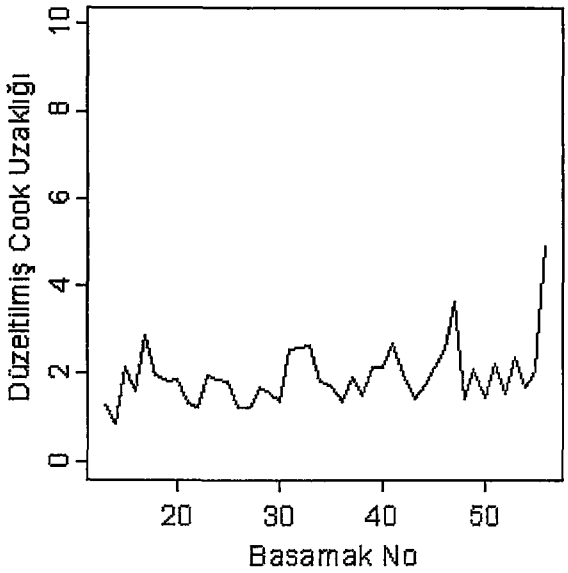
Şekil 4.20. Gözlem Girişiyle Birlikte Artıkların Durumunu Gösteren Grafik



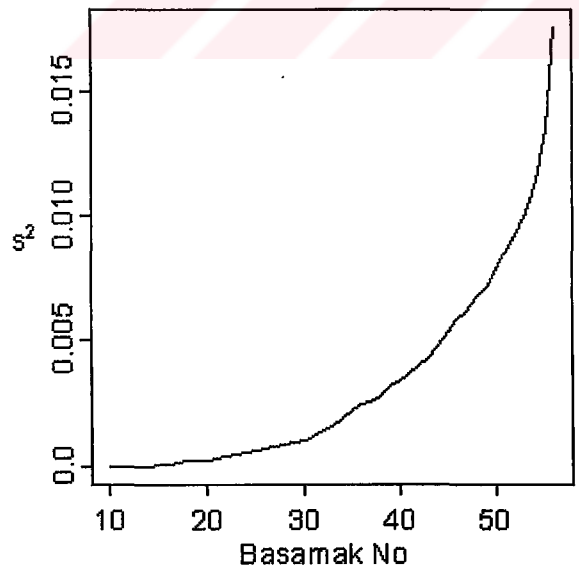
Şekil 4.21. Gözlem Girişiyle Birlikte Uç Değerliliği Gösteren Grafik



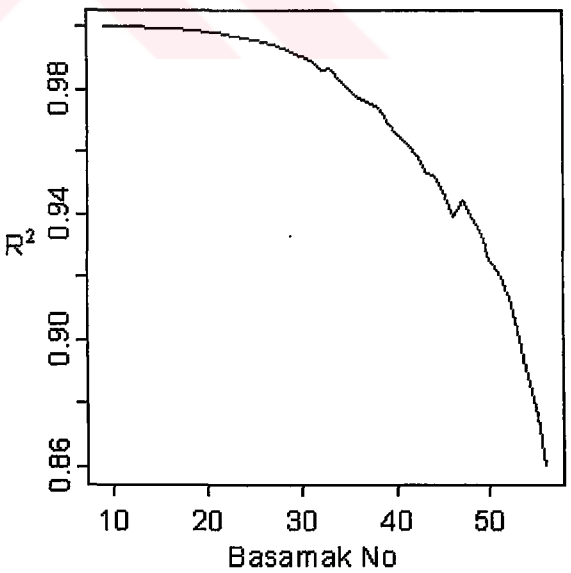
Şekil 4.22. Gözlem Girişiyle Birlikte Değişkenlerin Anlamlılığı



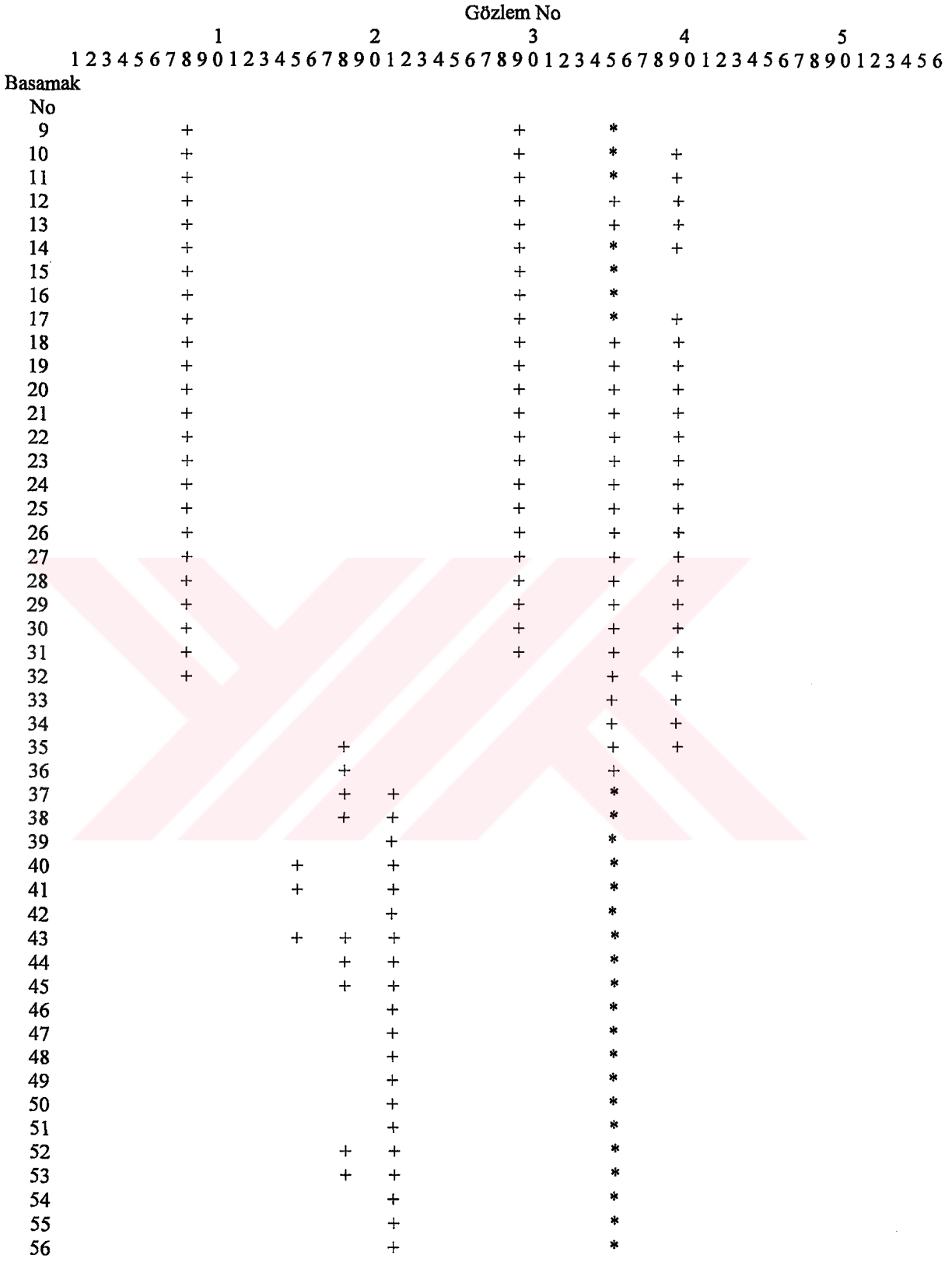
Şekil 4.23. Gözlem Girişiyle Birlikte D^* Değerlerindeki Değişim



Şekil 4.24. Gözlem Girişiyle Birlikte s^2 Değerlerindeki Değişim



Şekil 4.25. Gözlem Girişiyle Birlikte R^2 Değerlerindeki Değişin



+ : Mutlak değeri 2'ye eşit veya 2'den büyük artığa sahip gözlemleri belirtmektedir.
* : Mutlak değeri 3'e eşit veya 3'ten büyük artığa sahip gözlemleri belirtmektedir.

Şekil 4.26. İleri Araştırma Yöntemi İçin Sarkıt Grafiği

EKOK kestirimine göre, 14. ve 15. gözlemlerin sınırda olması sonucu 8., 17., 18., 21., 35., 39. ve 46. gözlemler aykırı değer olarak bulunmuştur. EKBK kestirimine göre ise 14., 18., 21., 30., 33., 34., 35. ve 55. gözlemler aykırı değer olarak bulunmuştur. EKHE yöntemine göre de 16., 29., 30. ve 35. gözlemler uç değer olarak saptanmıştır. Burada, gerek EKOK kestiriminden elde edilen sonuçların ileri araştırma yönteminden elde edilen sonuçlarla yaklaşık olarak uyuşması gerekse EKKD yönteminden sonuç alınamaması EKBK yönteminden elde edilen sonuçların tutarlılığının tartışılma nedenleridir.

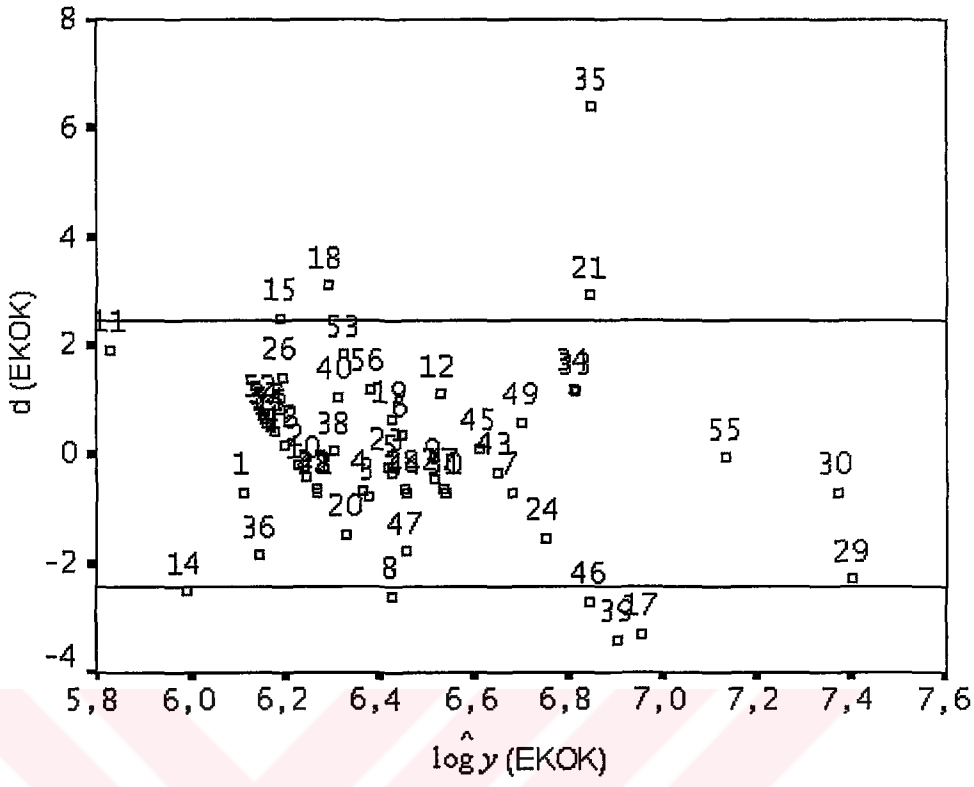
4.3.2. Çoklu Kuşkulu Gözlemlerin Grafiklerle İncelenmesi

EKOK ve EKBK kestirimlerinden elde edilen sonuçlarla ilgili grafikler Şekil 4.27-Şekil 4.30'da gösterilmiştir. Şekil 4.27 ve Şekil 4.28, EKOK kestiriminden elde edilen standartlaştırılmış artıkları temel alan grafikler olup y = -2.5 doğrusunun altında ve y = +2.5 doğrusunun üstünde kalan gözlemler aykırı değer olarak belirlenir. Bu nedenle, 8., 17., 18., 21., 35., 39. ve 46. gözlemler EKOK kestirimine göre aykırı değer olarak saptanmıştır. Bununla birlikte, Şekil 4.29 ve Şekil 4.30 da EKBK kestiriminden elde edilen standartlaştırılmış artıkları temel alan grafikler olup yine y = -2.5 doğrusunun altında ve y = +2.5 doğrusunun üstünde kalan gözlemler aykırı değer olarak belirlenir. Bu nedenle, 14., 18., 21., 30., 33., 34., 35. ve 55. gözlemler EKBK kestirimine göre aykırı değer olarak saptanmıştır.

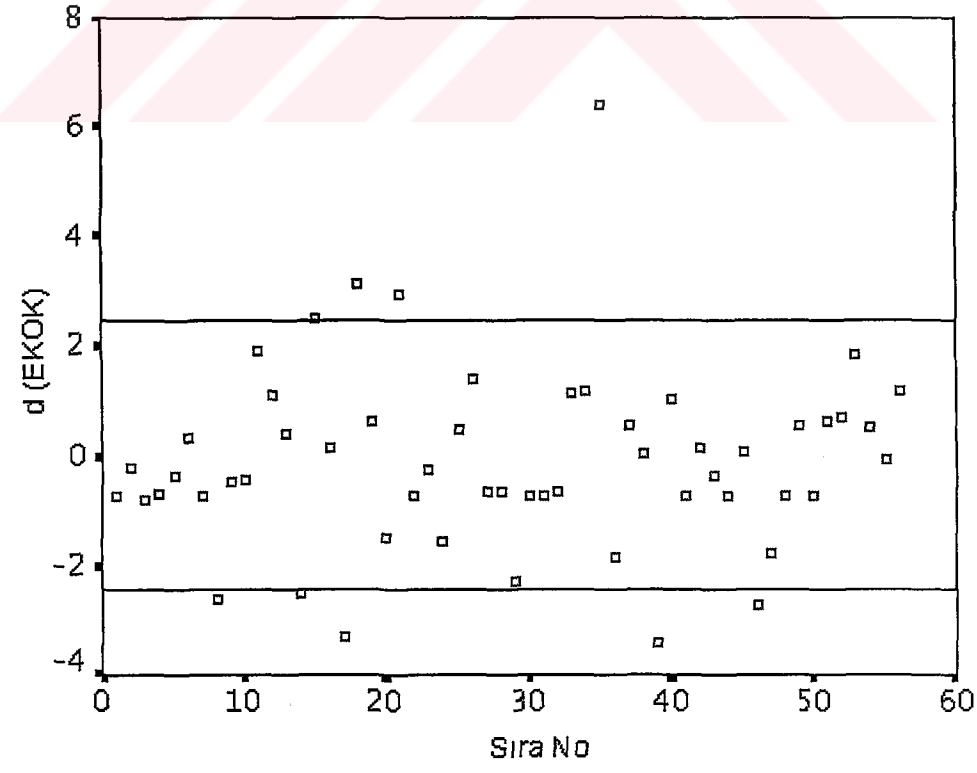
Sonuç olarak, çoklu kuşkulu gözlemlerin saptanmasında tercih edilen, aykırı değerlerin, iyi ve kötü uç değerlerin beraber görüldüğü Rousseeuw ve Leroy (2003) tarafından önerilen grafik Şekil 4.32'de gösterilmiştir. y = -2.5 ve y= +2.5 doğruları ile sağlam uzaklığın $\chi^2_{8;0.025} = 17.535 \Rightarrow \sqrt{17.535} = 4.187$ değeriyle sınırladığı bu grafiğe göre 8., 17., 18., 21., 39. ve 46. gözlemler aykırı değer, 16., 29., ve 30. gözlemler iyi uç değer ve 35. gözlem de kötü uç değer olarak belirtilir. Şekil 4.31'de de kuşkulu gözlemlerin gizleme ve sürüklemenin etkisinde kalmış durumları gösterilmiştir. y = +2.5 doğrusu ile Mahalanobis Uzaklığı'nın $\chi^2_{8;0.05} = 15.507 \Rightarrow \sqrt{15.507} = 3.938$ değeriyle sınırladığı bu grafik, çoklu kuşkulu gözlemlerin saptanmasında tercih edilmemekle beraber tek kuşkulu gözlemlerin belirlenmesinde faydalı olabilmektedir.

Çizelge 4.18. EKOK, EKBK ve EKHE Yöntemlerinden Elde Edilen Sonuçlar

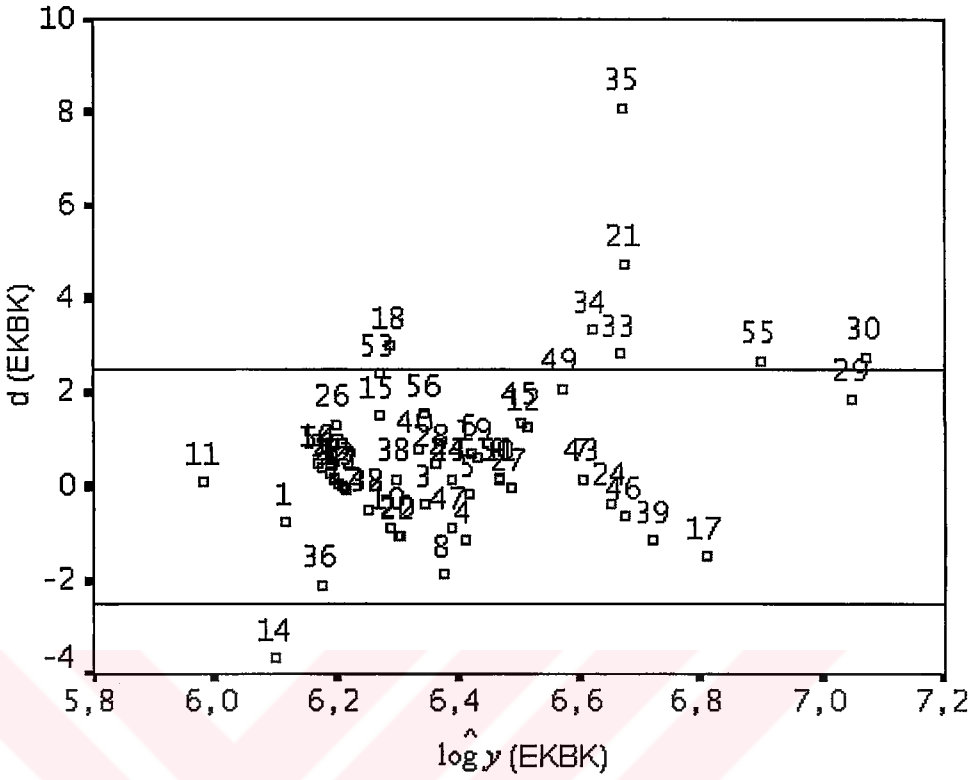
i	d (EKOK)	$\hat{\log y}$ (EKOK)	d (EKBK)	$\hat{\log y}$ (EKBK)	SU (EKHE)
1	-,710	6,109	-,770	6,117	2,612
2	-,200	6,227	,160	6,196	2,570
3	-,800	6,377	-,400	6,345	3,119
4	-,670	6,365	-1,150	6,411	2,432
5	-,360	6,430	-,180	6,416	2,278
6	,350	6,451	,680	6,421	3,491
7	-,710	6,679	,140	6,607	2,161
8	-2,610	6,428	-1,890	6,375	1,910
9	-,460	6,518	,140	6,467	3,069
10	-,410	6,244	-,890	6,288	2,728
11	1,920	5,831	,080	5,983	3,204
12	1,090	6,529	1,230	6,513	3,209
13	,400	6,176	,060	6,204	2,103
14	-2,520	5,990	-3,680	6,102	3,468
15	2,500	6,192	1,470	6,271	3,001
16	,150	6,197	,450	6,171	5,269
17	-3,300	6,955	-1,490	6,810	3,688
18	3,100	6,292	2,970	6,290	1,768
19	,630	6,428	,580	6,430	1,930
20	-1,470	6,332	-1,060	6,303	1,968
21	2,930	6,846	4,740	6,676	1,823
22	-,710	6,269	-1,080	6,304	2,405
23	-,250	6,421	,140	6,387	2,426
24	-1,550	6,749	-,380	6,653	2,829
25	,480	6,170	,140	6,197	3,189
26	1,400	6,193	1,290	6,198	2,198
27	-,640	6,533	-,060	6,486	3,154
28	-,660	6,455	,460	6,360	2,550
29	-2,300	7,402	1,850	7,049	4,716
30	-,710	7,369	2,740	7,071	4,611
31	-,710	6,539	,140	6,467	2,753
32	-,660	6,265	-,500	6,254	2,082
33	1,140	6,815	2,790	6,667	1,739
34	1,180	6,811	3,310	6,621	3,332
35	6,380	6,849	8,100	6,673	4,255
36	-1,860	6,145	-2,120	6,176	3,383
37	,570	6,162	,250	6,188	2,142
38	,040	6,307	,140	6,297	1,874
39	-3,410	6,904	-1,170	6,722	1,866
40	1,050	6,313	,750	6,334	2,457
41	-,710	6,269	-,040	6,214	4,128
42	,160	6,197	,140	6,197	3,725
43	-,350	6,649	,140	6,607	2,285
44	-,710	6,459	,140	6,387	2,482
45	,090	6,612	1,330	6,504	3,052
46	-2,710	6,846	-,640	6,676	1,823
47	-1,790	6,459	-,890	6,387	2,482
48	-,710	6,269	-,490	6,253	3,338
49	,570	6,702	2,060	6,571	3,545
50	-,710	6,539	,160	6,466	2,350
51	,630	6,158	-,070	6,216	3,349
52	,710	6,151	-,010	6,211	3,247
53	1,830	6,327	2,380	6,272	3,293
54	,510	6,168	,390	6,176	2,771
55	-,070	7,135	2,650	6,899	2,713
56	1,190	6,381	1,530	6,346	2,064



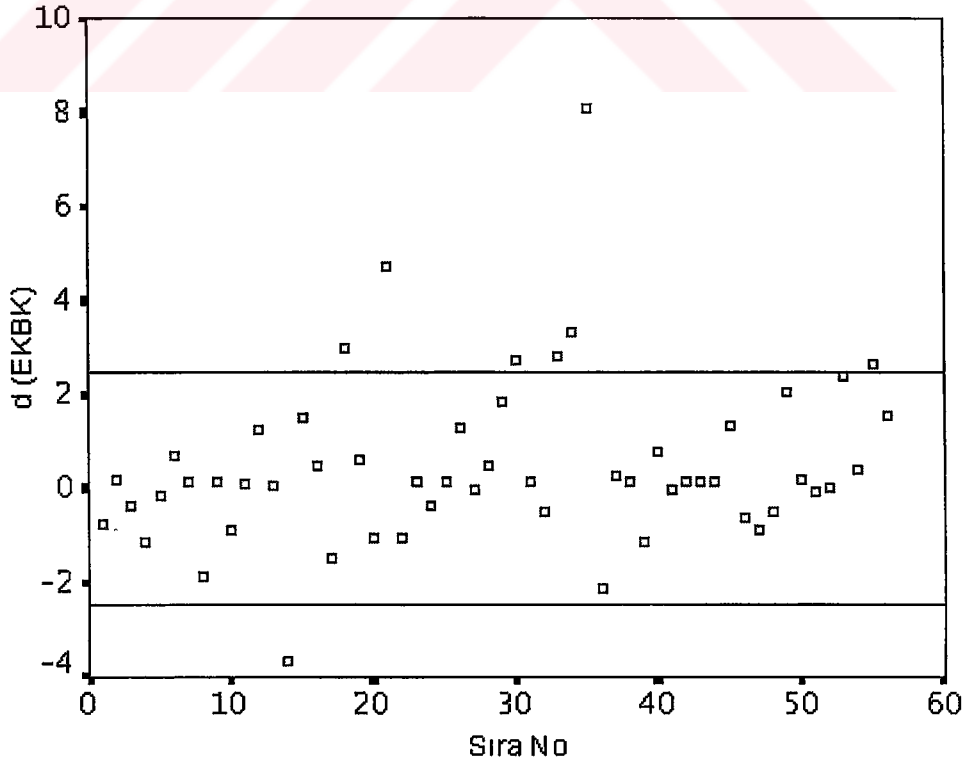
Şekil 4.27. EKOK Kestiriminden Elde Edilen Kestirilmiş Değer-Artık Grafiği



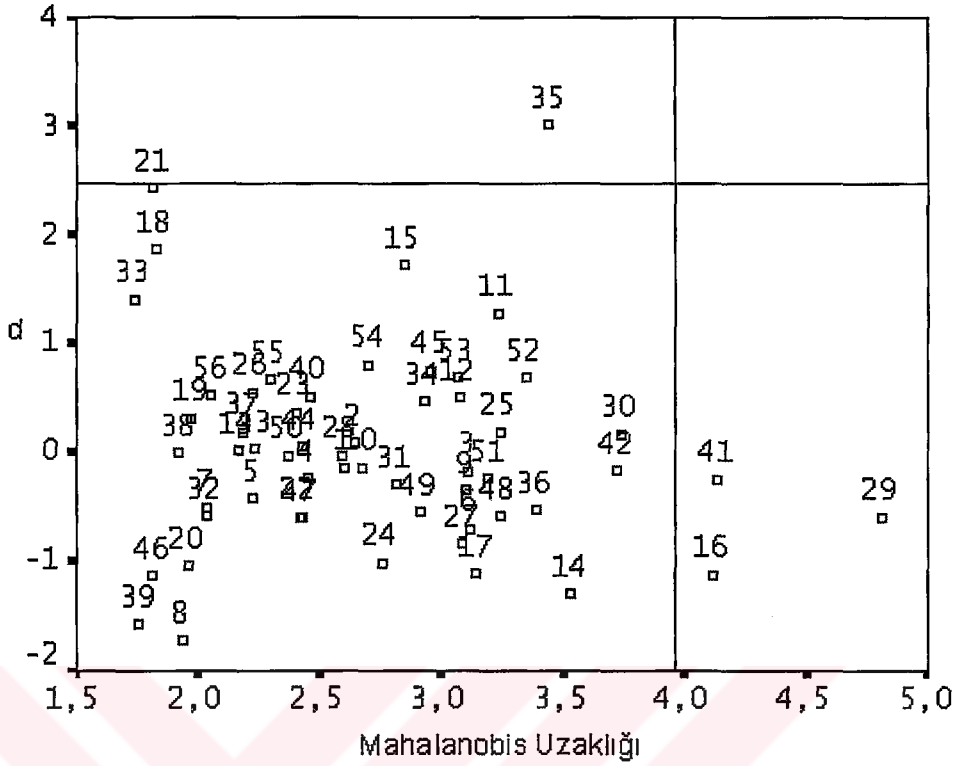
Şekil 4.28. EKOK Kestiriminden Elde Edilen Gözlem Artıklarının Grafiği



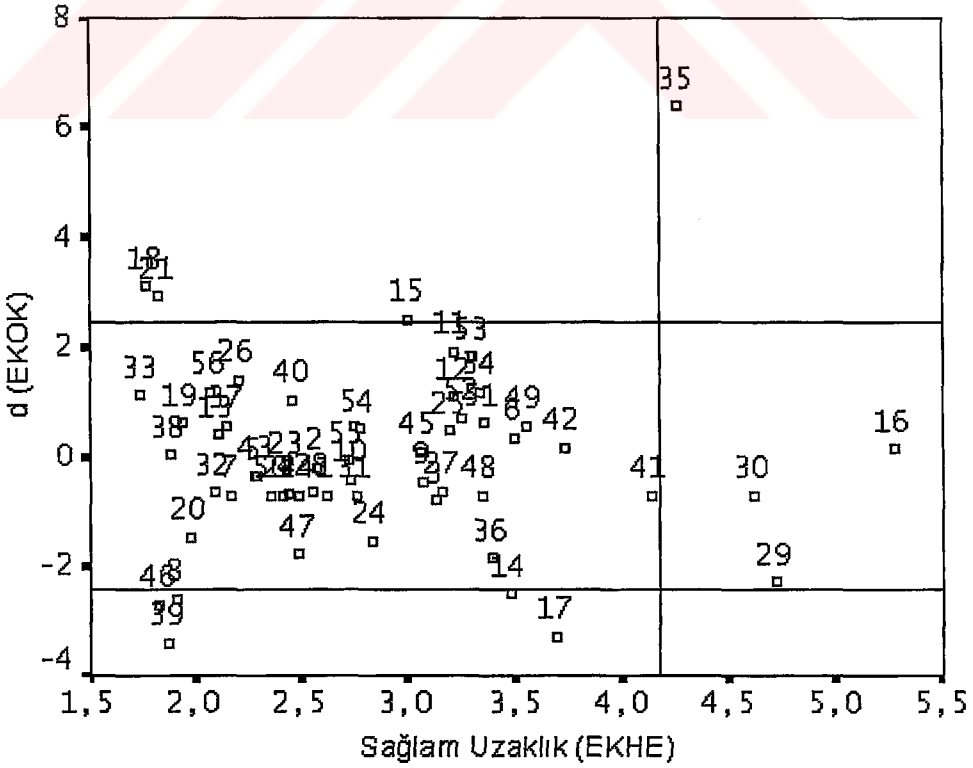
Şekil 4.29. EKBK Kestiriminden Elde Edilen Kestirilmiş Değer-Artık Grafiği



Şekil 4.30. EKBK Kestiriminden Elde Edilen Gözlem Artıklarının Grafiği



Şekil 4.31. Mahalanobis Uzaklığı-EKK Kestiriminden Elde Edilen Artık Grafiği



Şekil 4.32. EKHE'den Elde Edilen Sağlam Uzaklık-EKOK Kestiriminden Elde Edilen Artık Grafiği

4.3.3. Çoklu Kuşkulu Gözlemler İçin Genel Sonuç

Çizelge 4.19. Çoklu Kuşkulu Gözlemler İçin İnceleme

i	MDFFITS	COVRATIO	AP	R	İAY	EKOK	EKBK	EKHE
8		+			+	+		
14	+						+	
15	+				+			
16	+		+	+				*
17						+		
18		+			+	+	+	
21		+	+	*	+	+	*	
29	+	*	*		+			*
30		+					+	*
33							+	
34							+	
35	*	*	*	*	*	*	*	+
39					+	+		
41		+	+					
42		+						
46						+		
55							+	

* : Etkili + : Daha az etkili

Çizelge 4.19'da, çoklu kuşkulu gözlemlerin saptanabilmesi için kullanılan istatistikler ve bu istatistiklerin işaret ettiği gözlemler verilmiştir. "+" sembolü, çoklu gözlemlerin kendi aralarında karşılaştırılmaları sonucu çoğunluktan aşırı biçimde kopmayan veya Şekil 4.27-Şekil 4.30 ve Şekil 4.32'deki kritik değerlerden çok fazla uzaklaşmayan gözlemleri belirtmektedir. "*" sembolü de çoklu gözlemlerin kendi aralarında karşılaştırılmaları sonucu çoğunluktan aşırı biçimde kopan veya Şekil 4.27-Şekil 4.30 ve Şekil 4.32'deki kritik değerlerden fazlaca uzaklaşan gözlemleri belirtmektedir.

Çizelge 4.19'a göre 8., 18., 21., 35. ve 39. gözlemler, sağlam ve tercih edilen yöntemler olan ileri araştırma yöntemi (İAY) ve EKOK kestiriminin aykırı değer olarak belirlediği ortak gözlemlerdir. Bunun yanında, 15., 17., 29. ve 46. gözlemlerin bu iki yöntemden sadece birinin aykırı değer olarak belirlediği gözlemler olduğu görülebilir. 16., 29. ve 30. gözlemler ise EKHE yöntemine göre uç değer olup 35. gözlem de bütün yöntemlerin işaret ettiği hem uç ve aykırı hem de etkili olan tek gözlemdir. Bunların yanısıra, 8., 14., 15., 16., 18., 21., 29., 30., 41. ve 42. gözlemler çeşitli etki istatistiklerine göre etkili görünen gözlemlerdir.

4.4. Geçerlilik Çözümlemesi ve Model Karşılaştırmaları

Tam küme, EKK (En Küçük Kareler), EKOK (En Küçük Ortanca Kareler), Huber ve Andrews kestirimleri ile; 4.2 ve 4.3 nolu kesimlerdeki yöntemlerle bulunan tek kuşkulu ve çoklu kuşkulu gözlemlerin çeşitli kombinasyonları da tam kümeden çıkartılıp EKK kestirimi ile çözümlenmiş ve bulunan değerler Çizelge 4.20'de verilmiştir. Bu kestirimlerden, Huber ve Andrews kestirimleri sağlam kestiriciler olup MATLAB6.5 Programı ile hesaplanmıştır. Çizelge 4.20'de, R^2 belirtme katsayısını, $\hat{\sigma}$ kestirimin standart hatasını, $p_1, \ldots, p_8$ değerleri sırasıyla bağımsız değişkenlerin anlamlılıklarını belirtmektedir. $PRESS^*$ ise eşitlik (2.32)'de çıkartılmış artıklara dayanan PRESS'ten farklı olarak geçerlilik çözümlemesi için kullanılan istatistiktir. Bu istatistik, bu bölümün başında veri kümesinden ayrılan 20 tane gözlemin Çizelge 4.20'deki her bir modelde teker teker yerlerine konulması sonucu elde edilen önkestirim ($\tilde{y}$) değerlerine dayanmakta olup,

$$PRESS^* = \sum_{i=57}^{76} (y_i - \tilde{y}_i)^2 \qquad (4.1)$$

biçiminde hesaplanmaktadır. Tam kümeden çıkartılıp geçerliliği araştırılacak olan tek ve çoklu kuşkulu gözlem kombinasyonları, tek kuşkulu gözlemler kendi arasında olmak üzere çoklu kuşkulu gözlemler de sırasıyla EKOK, İAY'den elde edilenler ve her iki yöntemde ortak bulunan kuşkulu gözlemler olarak aşağıda tanımlanmıştır:

Tek Kuşkulu Gözlem Araştırması
- (A) 21. gözlem çıktı
- (B) 35. gözlem çıktı
- (C) 21. ve 35. gözlemler çıktı

Çizelge 4.20. Kiralik Veri Kümesi İçin Geçerlilik Çözümlemesi

MODELLER		R^2	$\hat{\sigma}$	p_1	p_2	p_3	p_4	p_5	p_6	p_7	p_8	PRESS*
TAM KÜME EKK		0,861	0,132	0,000	0,003	0,001	0,108	0,035	0,033	0,009	0,058	0,189
TEK KUŞKULU	A	0,870	0,124	0,000	0,003	0,001	0,077	0,035	0,019	0,009	0,102	0,197
	B	0,877	0,115	0,000	0,006	0,000	0,045	0,005	0,065	0,006	0,014	0,182
	C	0,890	0,106	0,000	0,005	0,000	0,024	0,004	0,036	0,005	0,025	0,180
ÇOKLU KUŞKULU	A	0,916	0,094	0,000	0,008	0,000	0,028	0,001	0,025	0,002	0,002	0,165
	B	0,929	0,088	0,000	0,002	0,000	0,014	0,001	0,026	0,000	0,001	0,183
	C	0,892	0,099	0,000	0,002	0,000	0,015	0,003	0,051	0,002	0,016	0,206
	D	0,920	0,088	0,000	0,000	0,000	0,018	0,004	0,030	0,002	0,001	0,194
	E	0,906	0,099	0,000	0,002	0,000	0,012	0,003	0,038	0,001	0,016	0,379
	F	0,930	0,087	0,000	0,000	0,000	0,014	0,003	0,021	0,001	0,001	0,191
	G	0,896	0,103	0,000	0,003	0,000	0,024	0,003	0,038	0,003	0,012	0,183
	H	0,919	0,093	0,000	0,001	0,000	0,017	0,001	0,023	0,002	0,006	0,182
EKOK		0,810	0,083	0,000	0,005	0,000	0,114	0,006	0,157	0,000	0,002	0,195
HUBER		0,886	0,119	0,000	0,005	0,001	0,092	0,025	0,064	0,012	0,043	0,180
ANDREWS		0,892	0,116	0,000	0,011	0,000	0,075	0,014	0,097	0,013	0,034	0,181

Çoklu Kuşkulu Gözlem Araştırması

- EKOK
  - (A) 17.,18.,21.,35. ve 39. gözlemler çıktı
  - (B) 8.,17.,18.,21.,35.,39. ve 46. gözlemler çıktı
- İAY
  - (C) 8.,21.,29.,35. ve 39. gözlemler çıktı
  - (D) 8.,15.,18.,21.,29.,35. ve 39. gözlemler çıktı
  - (E) 8.,21.,35. ve 39. gözlemler çıktı
  - (F) 8.,15.,18.,21.,35. ve 39. gözlemler çıktı
- Ortak
  - (G) 21.,35. ve 39. gözlemler çıktı
  - (H) 8.,18.,21.,35. ve 39. gözlemler çıktı

Çizelge 4.20'de, R^2, $\hat{\sigma}$ ve $p_1, \ldots, p_8$ değerleri (F) ve (B) kombinasyonlarını işaret etmektedir. Bununla birlikte, geçerlilikte güvenilir bir istatistik olan PRESS* değerlerinin en küçüğü dikkate alınırsa (A) kombinasyonunun geçerli olduğu söylenebilir. O halde, EKOK kestirim yöntemi aykırı değerlerin saptanmasında diğerlerine oranla daha iyi yöntemdir denilebilir. Bunun dışında, sağlam kestirimler olan Huber ve Andrews'in tam küme EKK kestirimine göre R^2, $\hat{\sigma}$ ve PRESS* değerleri dikkate alındığında daha iyi sonuç verdiği görülebilir.

4.5. Hava Kirliliği Verileri ile Çözümleme

Yukarıdaki tüm incelemeler, bir kez de oldukça farklı bir veri kümesi olan ve 6 bağımsız değişkenden oluşan 50 gözlemli (20 gözlem daha önce geçerlilik çözümlemesi için ayrılmış) Candan (1995)'da verilen hava kirliliği ile ilgili veri kümesinde de yapılmış ve Çizelge 4.21 elde edilmiştir. Bu veri kümesinde, havadaki SO_2 (kükürtdioksit) miktarının, basınca (mb), rüzgar hızına (m/sn), neme (%), minimum sıcaklığa (C°), yağış miktarına (mm) ve aya (Aralık, Ocak, Şubat) göre değişim gösterip göstermediği incelenmektedir. Ay değişkeni, bu uygulamada iki göstermelik değişkenle tanımlanmıştır. Bu uygulama için önceki kesimlerde kullanılan çizelgeler ve şekiller burada verilmemiştir. Çizelge 4.21'de belirtilen kuşkulu gözlem kombinasyonları aşağıda tanımlandığı gibidir:

Tek Kuşkulu Gözlem Araştırması
- (A) 27. gözlem çıktı
- (B) 42. gözlem çıktı
- (C) 27. ve 42. gözlemler çıktı

Çoklu Kuşkulu Gözlem Araştırması
- EKOK
  - (A) 2.,8.,9.,24.,26. ve 29. gözlemler çıktı
  - (B) 2.,8.,9.,24.,26.,29. ve 33. gözlemler çıktı
- İAY
  - (C) 26.,28.,29.,33.,34. ve 42. gözlemler çıktı
  - (D) 25.,26.,28.,29.,33.,34. ve 42. gözlemler çıktı
  - (E) 4.,24.,25.,26.,28.,29.,33.,34. ve 42. gözlemler çıktı
  - (F) 4.,24.,25.,26.,28.,29.,33.,34.,37. ve 42. gözlemler çıktı
- Ortak
  - (G) 26.,29. ve 33. gözlemler çıktı
  - (H) 24.,26.,29. ve 33. gözlemler çıktı

Çizelge 4.21'de, R^2 ve $\hat{\sigma}$ (F) kombinasyonunu işaret etmektedir. Bunun yanında, $p_1, \ldots, p_7$ değerlerindeki büyük farklılıklar kombinasyonlar üzerine birşey söylemeyi zorlaştırmaktadır. PRESS* değerlerinin en küçüğü dikkate alındığında da tam küme üzerine uygulanan Huber ve Andrews sağlam kestiricilerinin yanısıra (B) kombinasyonu dikkat çekmektedir. O halde, sadece 42. gözlemin veri kümesinden uzaklaştırılması modelin geçerliliği için yeterli olabilmektedir.

Sonuç olarak, kuşkulu gözlemlerin saptanmasında hangi yöntemin daha iyi sonuç verdiği çalışılan veri kümesine bağlı olarak değişmektedir denilebilir.

Çizelge 4.21. Hava Kirliliği Veri Kümesi İçin Geçerlilik Çözümlemesi

MODELLER		R^2	$\hat{\sigma}$	p_1	p_2	p_3	p_4	p_5	p_6	p_7	PRESS*
TAM KÜME EKK		0,410	0,207	0,059	0,950	0,365	0,020	0,262	0,016	0,206	0,456
TEK KUŞKULU	A	0,452	0,198	0,019	0,944	0,425	0,023	0,129	0,018	0,093	1,190
	B	0,408	0,198	0,046	0,812	0,206	0,023	0,406	0,029	0,395	0,404
	C	0,448	0,189	0,014	0,915	0,243	0,026	0,216	0,032	0,203	0,461
ÇOKLU KUŞKULU	A	0,583	0,179	0,004	0,960	0,230	0,011	0,110	0,003	0,542	0,640
	B	0,631	0,168	0,021	0,868	0,089	0,003	0,036	0,001	0,939	0,611
	C	0,596	0,165	0,485	0,248	0,224	0,004	0,044	0,044	0,218	0,456
	D	0,634	0,157	0,481	0,234	0,499	0,004	0,034	0,035	0,123	0,532
	E	0,690	0,137	0,767	0,288	0,288	0,002	0,004	0,039	0,016	0,609
	F	0,719	0,132	0,795	0,054	0,434	0,001	0,001	0,084	0,008	0,753
	G	0,541	0,180	0,108	0,685	0,385	0,004	0,095	0,012	0,862	0,443
	H	0,572	0,176	0,134	0,960	0,224	0,002	0,047	0,007	0,886	0,501
EKOK		0,550	0,175	0,022	0,851	0,092	0,003	0,035	0,001	0,924	0,607
HUBER		0,339	0,219	0,057	0,882	0,431	0,034	0,257	0,034	0,328	0,400
ANDREWS		0,363	0,215	0,054	0,860	0,419	0,032	0,267	0,032	0,295	0,396

BEŞİNCİ BÖLÜM
SONUÇ ve TARTIŞMA

Çalışmanın birinci bölümünde, çoklu doğrusal regresyon ile ilgili genel bilgiler verilerek aykırı, uç değer ve etkili gözlem kavramlarının tanımları üzerinde durulmuş ve bunların arasındaki ilişkiler belirtilmiştir.

Çalışmanın ikinci bölümünde, tek kuşkulu gözlemlerin çoklu doğrusal regresyonda varlığı çeşitli grafiklerle ve istatistiklerle incelenmiş ve hangi istatistiğin hangi kuşkulu gözlem tipini saptamada kullanıldığı verilmiştir.

Çalışmanın üçüncü bölümünde, gizleme ve sürükleme etkileri açıklanmış, bunların varlığında ikinci bölümde anlatılan yöntemlerin özellikle çoklu kuşkulu gözlemleri saptamada başarılı olamayacağı üzerinde durulmuştur. Bu nedenle, çoklu kuşkulu gözlemlerin saptanması için ikinci bölümde verilen istatistiklerin çoklu durumlara uygun biçimde geliştirilmesinin yanısıra kullanılan sağlam yöntemler ve bu yöntemlerden elde edilen çeşitli sonuçların grafiklerle sunulması açıklanmıştır.

Çalışmanın dördüncü bölümünde, önceki bölümlerde anlatılan bilgiler doğrultusunda uygulama verileri ile çoklu doğrusal regresyon modelleri oluşturulmuş ve çeşitli programlar yardımıyla tek ve çoklu kuşkulu gözlemler saptanmaya çalışılmıştır.

Genel olarak uygulamanın sonunda şu sonuçlar elde edilmiştir:

- Bütün veri kümeleri için yalnızca tek kuşkulu ve yalnızca çoklu kuşkulu gözlemlerin değil her iki durum için de kuşkulu gözlemlerin incelenmesi gerektiği,

- Özellikle, tek kuşkulu gözlemlerin saptanmasında kullanılan bazı istatistikler için belirlenen kritik değerlerin kuşkulu gözlemleri saptamada uygun olmaması durumunda gözlemlerin kendi aralarında karşılaştırılıp bir sonuca varılabileceği,

- Gizleme ve sürükleme etkilerinden bahsedilen durumlarda kuşkulu gözlemleri saptamak için kesinlikle İAY, EKOK, EKHE vb. sağlam yöntemlerin kullanılması gerektiği,

- Sağlam bir yöntem olan EKKD yönteminin, veri kümesindeki bağımsız değişkenlerin değerlerinin çok fazla tekrar etmesiyle sonuç vermediği ve bu yöntemin vereceği sonuçlarla birlikte grafiklenen EKBK yönteminin de hassas bir kestirim olmasından dolayı tek başına değerlendirilmemesi gerektiği,

- Geçerlilik çözümlemesi yardımıyla, kuşkulu gözlemleri saptamada hangi yöntem ya da yöntemlerin daha iyi sonuç verdiği ve bu sonucun bütün veri kümeleri için değilde sadece çalışılan veri kümesi için geçerli olduğu belirlenmiştir.

KAYNAKLAR

Atkinson A. C. (December 1994), *Fast Very Robust Methods for the Detection of Multiple Outliers*, Journal of the American Statistical Association, Theory and Methods, Volume 89, No. 428, pg. 1329-1339

Atkinson A. C. and Riani M. (2000), *Robust Diagnostic Regression Analysis*, Springer, New York

Atkinson A. C. and Riani M. (2002), *Forward Search Added Variable t Tests and the Effect of Masked Outliers on Model Selection and Transformation* (http://www.lse.ac.uk/collections/statistics/documents/researchreport73.pdf)

Atkinson A. C. and Riani M. (2004), *Distribution Theory and Simulations for Tests of Outliers in Regression* (http://www.lse.ac.uk/collections/statistics/documents/researchreport104.pdf)

Barnett V. and Lewis T. (1994), *Outliers In Statistical Data*, Third Edition, John Wiley&Sons Ltd., Chichester

Belsley D. A., Kuh E. and Welsch R. E. (1980), *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*, John Wiley&Sons, Inc.

Candan, M. (1995), Doğrusal Regresyon Çözümlemesinde Sağlam Kestiriciler, Yayınlanmamış Yüksek Lisans Tezi, Hacettepe Üniv., Ankara

Chambers R., Hentges A. and Zhao X. (May 2004), *Robust Automatic Methods for Outlier and Error Detection*, Journal of the Royal Statistical Society, Series A, Statistics In Society, Blackwell Publishing, Volume 167, Part 2, pg. 323-339

Chatterjee S. and Hadi A. S. (1986), *Influential Observations, High Leverage Points and Outliers In Linear Regression*, Statistical Science, Institute of Mathematical Statistics, Volume 1, No. 3, pg. 379-416

Chatterjee S. and Hadi A. S. (1988), *Sensitivity Analysis In Linear Regression*, John Wiley&Sons, Inc.

Chatterjee S., Hadi A. S. and Price B. (2000), *Regression Analysis by Example*, Third Edition, John Wiley&Sons, Inc.

Cook R. D. (February 1977), *Detection of Influential Observation In Linear Regression*, Technometrics, Volume 19, No. 1, pg. 15-18

Cook R. D. and Weisberg S. (1994), *An Introduction To Regression Graphics*, John Wiley&Sons, Inc.

Draper N. R. and Smith H. (1966), *Applied Regression Analysis*, John Wiley&Sons, Inc.

Draper N. R. and John J. A. (February 1981), *Influential Observations and Outliers In Regression*, Technometrics, Volume 23, No. 1, pg. 21-26

Freund R. J. and Wilson W. J. (1998), *Regression Analysis: Statistical Modeling of A Response Variable*, Academic Press

Hadi A. S. and Simonoff J. S. (December 1993), *Procedures for the Identification of Multiple Outliers In Linear Models*, Journal of the American Statistical Association, Volume 88, No. 424, pg. 1264-1272

Hardin J. and Rocke D. M. (2004), *Outlier Detection In the Multiple Cluster Setting Using the Minimum Covariance Determinant Estimator*, Computational Statistics and Data Analysis, Volume 44, pg. 625-638

Hawkins D. M. (1980), *Identification of Outliers*, Chapman and Hall Ltd.

High R. (2004), *Outlier.....s*, University of Oregon (http://darkwing.uoregon.edu/~robinh/outl.txt)

Huber P. J. (1981), *Robust Statistics*, John Wiley&Sons, Inc.

Kıral G. ve Billor N. (2001), *Bacon Temel Bileşenler Analizi İle Sapan Değerlerin Belirlenmesi*, Çukurova Üniv., Adana (http://idari.cu.edu.tr/sempozyum/bil27.htm)

Kleinbaum D. G., Kupper L. L. and Muller K. E. (1988), *Applied Regression Analysis and Other Multivariable Methods*, Second Edition, PWS-KENT Publishing Company, a division of Wadsworth, Inc.

Lawrence D. K. and Arthur J. L. (1990), *Robust Regression; Analysis and Applications*, Marcel Dekker, Inc., New York and Basel

Montgomery D. C. and Peck E. A. (1992), *Introduction To Linear Regression Analysis*, Second Edition, John Wiley&Sons, Inc.

Myers R. H. (1986), *Classical and Modern Regression With Applications*, PWS Publishers

Neter J., Kutner M. H., Nachtsheim C. J. and Wasserman W. (1996), *Applied Linear Statistical Models*, Fourth Edition, The McGraw-Hill Companies, Inc.

Orhunbilge N. (2002), *Uygulamalı Regresyon ve Korelasyon Analizi*, Gözden Geçirilmiş 2. Baskı, İ. Ü. İşletme Fakültesi, İstanbul

Rawlings J. O., Pantula S. G. and Dickey D. A. (1998), *Applied Regression Analysis; A Research Tool*, Second Edition, Springer-Verlag New York, Inc.

Robustness, (1999) (http://www.agoras.ua.ac.be/Robustn.htm)

Robust Diagnostic Regression Analysis: Software and Datasets, (2000) (http://www.riani.it/ar/Software.html)

Robust Regression, (2003) (http://www.quantlet.com/mdstat/scripts/xag/html/xaghtmlnode11.html)

Rousseeuw P. J. and Van Zomeren B. C. (September 1990), *Unmasking Multivariate Outliers and Leverage Points*, Journal of the American Statistical Association, Volume 85, No. 411, pg. 633-639

Rousseeuw P. J. and Van Driessen K. (August 1999), *A Fast Algorithm For The Minimum Covariance Determinant Estimator*, Technometrics, Volume 41, No. 3 pg. 212-223

Rousseeuw P. J. and Leroy A. M. (2003), *Robust Regression and Outlier Detection*, John Wiley&Sons, Inc.

Ryan T. P. (1997), *Modern Regression Methods*, John Wiley&Sons, Inc.

Smyth G. K. and Hawkins D. M. (2000), *Robust Frequency Estimation Using Elemental Sets*, Journal of Computational and Graphical Statistics, Volume 9, pg. 196-214

Tatlıdil H. (1981), *Doğrusal Regresyonda ve Çok Değişkenli Verilerde Kuşkulu Gözlemlerin Testi*, Doktora Tezi, Hacettepe Üniv., Ankara

Weisberg S. (1985), *Applied Linear Regression*, Second Edition, John Wiley&Sons, Inc.

Wisnowski J. W. (1999), *Multiple Outliers in Linear Regression: Advances in Detection Methods, Robust Estimation and Variable Selection*, Arizona State Univ. (http://www.stormingmedia.us/33/3367/A336763.html)

Yaffee R. A. (2002), *Robust Regression Analysis: Some Popular Statistical Package Options* (http://www.nyu.edu/its/socsci/Docs/RobustReg2.pdf)

EK-A

ÇIKARIMLAR

A.1. $s^2_{(i)}$'nin Çıkarımı

$s^2_{(i)}$'nin çıkarımına geçmeden önce, $X'X$ matrisi ve X matrisinin i'inci satırı olarak $\boldsymbol{x}_i'$ vektörü ele alınırsa $\left(X'X-\boldsymbol{x}_i\boldsymbol{x}_i'\right)$ matrisi, $X'X$ matrisinin veri kümesinden i'inci gözlem çıkartılıp hesaplanmış halidir.

$$\left(X'X-\boldsymbol{x}_i\boldsymbol{x}_i'\right)^{-1}=\left(X'X\right)^{-1}+\frac{\left(X'X\right)^{-1}\boldsymbol{x}_i\boldsymbol{x}_i'\left(X'X\right)^{-1}}{1-\boldsymbol{x}_i'\left(X'X\right)^{-1}\boldsymbol{x}_i} \tag{A.1}$$

eşitliği ispatlanacak olursa, her iki taraf $\left(X'X-\boldsymbol{x}_i\boldsymbol{x}_i'\right)$ matrisiyle çarpılıp,

$$I=I+\frac{\left(X'X\right)^{-1}\boldsymbol{x}_i\boldsymbol{x}_i'}{1-\boldsymbol{x}_i'\left(X'X\right)^{-1}\boldsymbol{x}_i}-\left(X'X\right)^{-1}\boldsymbol{x}_i\boldsymbol{x}_i'-\frac{\left(X'X\right)^{-1}\boldsymbol{x}_i\boldsymbol{x}_i'}{1-\boldsymbol{x}_i'\left(X'X\right)^{-1}\boldsymbol{x}_i}\left(X'X\right)^{-1}\boldsymbol{x}_i\boldsymbol{x}_i'$$

$$I=I+\frac{\left(X'X\right)^{-1}\boldsymbol{x}_i\boldsymbol{x}_i'-\left(X'X\right)^{-1}\boldsymbol{x}_i\boldsymbol{x}_i'\left(1-\boldsymbol{x}_i'\left(X'X\right)^{-1}\boldsymbol{x}_i\right)-\left(X'X\right)^{-1}\boldsymbol{x}_i\left(\boldsymbol{x}_i'\left(X'X\right)^{-1}\boldsymbol{x}_i\right)\boldsymbol{x}_i'}{1-\boldsymbol{x}_i'\left(X'X\right)^{-1}\boldsymbol{x}_i}$$

$$I=I+\frac{\left(X'X\right)^{-1}\boldsymbol{x}_i\boldsymbol{x}_i'-\left(X'X\right)^{-1}\boldsymbol{x}_i\boldsymbol{x}_i'+\left(X'X\right)^{-1}\boldsymbol{x}_i\boldsymbol{x}_i'\left(\boldsymbol{x}_i'\left(X'X\right)^{-1}\boldsymbol{x}_i\right)-\left(X'X\right)^{-1}\boldsymbol{x}_i\boldsymbol{x}_i'\left(\boldsymbol{x}_i'\left(X'X\right)^{-1}\boldsymbol{x}_i\right)}{1-\boldsymbol{x}_i'\left(X'X\right)^{-1}\boldsymbol{x}_i}$$

$I=I$ olduğu görülür.

Bununla birlikte, (A.1)'deki eşitlik (1.8)'den de yararlanılarak,

$$\left[X_{(i)}'X_{(i)}\right]^{-1}=\left(X'X\right)^{-1}+\frac{\left(X'X\right)^{-1}\boldsymbol{x}_i\boldsymbol{x}_i'\left(X'X\right)^{-1}}{1-h_{ii}} \tag{A.2}$$

biçiminde de yazılabilir. Bu eşitliğin her iki tarafı $\left(X'\boldsymbol{Y}-\boldsymbol{x}_i y_i\right)$ ile çarpılırsa,

$$\hat{\boldsymbol{\beta}}_{(i)}=\hat{\boldsymbol{\beta}}-\left(X'X\right)^{-1}\boldsymbol{x}_i y_i+\frac{\left(X'X\right)^{-1}\boldsymbol{x}_i\boldsymbol{x}_i'\left(X'X\right)^{-1}\left(X'\boldsymbol{Y}-\boldsymbol{x}_i y_i\right)}{1-h_{ii}}$$

$$\hat{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}_{(i)}=\frac{\left(X'X\right)^{-1}\boldsymbol{x}_i y_i-\left(X'X\right)^{-1}\boldsymbol{x}_i y_i h_{ii}-\left(X'X\right)^{-1}\boldsymbol{x}_i\boldsymbol{x}_i'\hat{\boldsymbol{\beta}}+\left(X'X\right)^{-1}\boldsymbol{x}_i h_{ii} y_i}{1-h_{ii}}$$

$$\begin{aligned}\hat{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}_{(i)}&=\frac{\left(X'X\right)^{-1}\boldsymbol{x}_i\left[y_i-y_i h_{ii}-\boldsymbol{x}_i'\hat{\boldsymbol{\beta}}+h_{ii}y_i\right]}{1-h_{ii}}=\frac{\left(X'X\right)^{-1}\boldsymbol{x}_i\left[y_i-\hat{y}_i\right]}{1-h_{ii}}\\&=\frac{\left(X'X\right)^{-1}\boldsymbol{x}_i e_i}{1-h_{ii}}\end{aligned}\qquad\text{(A.3)}$$

elde edilir.

$$\left(n-k-2\right)s_{(i)}^2=\sum_{j\neq i}\left(y_j-\boldsymbol{x}_j'\hat{\boldsymbol{\beta}}_{(i)}\right)^2\qquad\text{(A.4)}$$

eşitliğinin sağ tarafı ele alınırsa, (A.3)'deki eşitlik yardımıyla,

$$\begin{aligned}\sum_{j\neq i}\left(y_j-\boldsymbol{x}_j'\hat{\boldsymbol{\beta}}_{(i)}\right)^2&=\sum_{j=1}^{n}\left(y_j-\boldsymbol{x}_j'\hat{\boldsymbol{\beta}}+\frac{\boldsymbol{x}_j'\left(X'X\right)^{-1}\boldsymbol{x}_i e_i}{1-h_{ii}}\right)^2-\left(y_i-\boldsymbol{x}_i'\hat{\boldsymbol{\beta}}+\frac{h_{ii}e_i}{1-h_{ii}}\right)^2\\&=\sum_{j=1}^{n}\left(e_j+\frac{h_{ij}e_i}{1-h_{ii}}\right)^2-\frac{e_i^2}{\left(1-h_{ii}\right)^2}\end{aligned}\qquad\text{(A.5)}$$

elde edilir. Bu eşitliğin sağ tarafındaki ilk terim açılıp,

$$\sum_{j=1}^{n}\left(e_j+\frac{h_{ij}e_i}{1-h_{ii}}\right)^2=\sum_{j=1}^{n}e_j^2+\frac{2e_i}{1-h_{ii}}\sum_{j=1}^{n}e_jh_{ij}+\frac{e_i^2}{(1-h_{ii})^2}\sum_{j=1}^{n}h_{ij}^2 \qquad (A.6)$$

$HY=H\hat{Y}$ olduğundan $\sum_{j=1}^{n}e_jh_{ij}=0$ ve (1.9)'daki eşitlik göz önünde bulundurulursa (A.4)'deki eşitlik,

$$\begin{aligned}(n-k-2)s_{(i)}^2&=\sum_{j=1}^{n}e_j^2+\frac{h_{ii}e_i^2}{(1-h_{ii})^2}-\frac{e_i^2}{(1-h_{ii})^2}\\&=\sum_{j=1}^{n}e_j^2-\frac{e_i^2}{1-h_{ii}}\\&=(n-k-1)AKO-\frac{e_i^2}{1-h_{ii}}\end{aligned} \qquad (A.7)$$

biçiminde yazılabilir. Bu eşitliğin, her iki tarafı $(n-k-2)$'ye bölünerek (1.20)'deki eşitlik elde edilir.

A.2. DFFITS İstatistiği'nin Çıkarımı

(A.3)'deki eşitliğin her iki tarafı x_i' ile çarpılırsa,

$$\hat{y}_i-\hat{y}_{i(i)}=\frac{h_{ii}e_i}{1-h_{ii}} \qquad (A.8)$$

elde edilir. Bu eşitliğin de her iki tarafı $s_{(i)}\sqrt{h_{ii}}$'ye bölünürse,

$$DFFITS_i=\frac{\hat{y}_i-\hat{y}_{i(i)}}{s_{(i)}\sqrt{h_{ii}}}=\frac{h_{ii}e_i}{1-h_{ii}}\left(\frac{1}{s_{(i)}\sqrt{h_{ii}}}\right)=\frac{e_i}{s_{(i)}\sqrt{1-h_{ii}}}\left(\frac{h_{ii}}{1-h_{ii}}\right)^{1/2}=\sqrt{\frac{h_{ii}}{1-h_{ii}}}\;\;t_i$$

biçiminde (2.19) ve (2.20)'deki eşitlikler elde edilir.

A.3. DFBETAS İstatistiği'nin Çıkarımı

(A.3)'deki eşitlik,

$$\hat{\beta}_j - \hat{\beta}_{j(i)} = \frac{r_{ji} e_i}{1 - h_{ii}} \tag{A.9}$$

biçiminde yazılıp (2.22)'deki R matrisi ele alınırsa,

$$(RR')' = \left[(X'X)^{-1} X'X (X'X)^{-1} \right]' = (X'X)^{-1} = C = RR' \tag{A.10}$$

eşitliği yardımıyla $C_{jj} = \mathbf{r}_j' \mathbf{r}_j$ biçiminde yazılabilir. (2.21)'deki eşitlik yardımıyla,

$$DFBETAS_{j(i)} = \frac{\hat{\beta}_j - \hat{\beta}_{j(i)}}{s_{(i)} \sqrt{C_{jj}}} = \left(\frac{r_{ji} e_i}{1 - h_{ii}} \right) \frac{1}{s_{(i)} \sqrt{\mathbf{r}_j' \mathbf{r}_j}} = \frac{r_{ji}}{\sqrt{\mathbf{r}_j' \mathbf{r}_j}} \frac{t_i}{\sqrt{1 - h_{ii}}}$$

biçiminde (2.23)'deki eşitlik elde edilir.

A.4. Cook Uzaklığı'nın Çıkarımı

(2.24) ve (A.3)'deki eşitlikler birarada kullanılırsa,

$$D_i = \frac{\mathbf{x}_i' (X'X)^{-1} (X'X) (X'X)^{-1} \mathbf{x}_i e_i^2}{k' (AKO) (1 - h_{ii})^2} = \left(\frac{h_{ii}}{k' (AKO)} \right) \left(\frac{e_i}{1 - h_{ii}} \right)^2 = \left(\frac{e_i^2}{k' (AKO) (1 - h_{ii})} \right) \left(\frac{h_{ii}}{1 - h_{ii}} \right)$$

biçiminde elde edilen bu eşitlik (1.14) yardımıyla (2.25)'deki gibi yazılabilir.

A.5. PRESS İstatistiği'nin Çıkarımı

(1.15)'deki eşitlik,

$$e_{(i)} = y_i - \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}_{(i)} = y_i - \boldsymbol{x}_i'\left(X_{(i)}'X_{(i)}\right)^{-1} X_{(i)}'\boldsymbol{Y}_{(i)} \tag{A.11}$$

biçiminde yazılıp (A.2)'deki eşitlik kullanılarak,

$$\begin{aligned} e_{(i)} &= y_i - \boldsymbol{x}_i'\left[\left(X'X\right)^{-1} + \frac{\left(X'X\right)^{-1}\boldsymbol{x}_i\boldsymbol{x}_i'\left(X'X\right)^{-1}}{1-h_{ii}}\right] X_{(i)}'\boldsymbol{Y}_{(i)} \\ &= y_i - \boldsymbol{x}_i'\left(X'X\right)^{-1} X_{(i)}'\boldsymbol{Y}_{(i)} - \frac{\boldsymbol{x}_i'\left(X'X\right)^{-1}\boldsymbol{x}_i\boldsymbol{x}_i'\left(X'X\right)^{-1} X_{(i)}'\boldsymbol{Y}_{(i)}}{1-h_{ii}} \\ &= \frac{\left(1-h_{ii}\right)y_i - \left(1-h_{ii}\right)\boldsymbol{x}_i'\left(X'X\right)^{-1} X_{(i)}'\boldsymbol{Y}_{(i)} - h_{ii}\boldsymbol{x}_i'\left(X'X\right)^{-1} X_{(i)}'\boldsymbol{Y}_{(i)}}{1-h_{ii}} \\ &= \frac{\left(1-h_{ii}\right)y_i - \boldsymbol{x}_i'\left(X'X\right)^{-1} X_{(i)}'\boldsymbol{Y}_{(i)}}{1-h_{ii}} \end{aligned} \tag{A.12}$$

elde edilir. Burada, $X_{(i)}'\boldsymbol{Y}_{(i)} = X'\boldsymbol{Y} - \boldsymbol{x}_i y_i$ biçiminde gösterilirse,

$$\begin{aligned} e_{(i)} &= \frac{\left(1-h_{ii}\right)y_i - \boldsymbol{x}_i'\left(X'X\right)^{-1}\left(X'\boldsymbol{Y} - \boldsymbol{x}_i y_i\right)}{1-h_{ii}} \\ &= \frac{\left(1-h_{ii}\right)y_i - \boldsymbol{x}_i'\left(X'X\right)^{-1} X'\boldsymbol{Y} + \boldsymbol{x}_i'\left(X'X\right)^{-1}\boldsymbol{x}_i y_i}{1-h_{ii}} \\ &= \frac{\left(1-h_{ii}\right)y_i - \boldsymbol{x}_i'\hat{\boldsymbol{\beta}} + h_{ii}y_i}{1-h_{ii}} = \frac{y_i - \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}}{1-h_{ii}} = \frac{e_i}{1-h_{ii}} \end{aligned} \tag{A.13}$$

elde edilir. Bu eşitlik kullanılarak da (2.32)'deki PRESS İstatistiği hesaplanabilir.

EK-B
PROGRAM ALGORİTMALARI

B.1. PROGRESS Programı'nda Kullanılan EKOK Kestirim Algoritması

EKOK kestirimi ile regresyon katsayılarının hesaplanması tam olarak net olmadığından bu kestirim için hesaplama formülü üretmek neredeyse imkansızdır. Bu durum diğer sağlam kestirimler için de geçerlidir. PROGRESS'in tanımladığı algoritma, k' farklı gözlemli altkümeler oluşturur. Oluşturulan bir altküme $J=\{i_1,...,i_{k'}\}$ biçiminde gösterilirse EKOK kestirimi bu altküme için,

$$\underset{i=1,...,n}{ortc}\left(y_i-\mathbf{x}_i'\beta_J\right)^2 \tag{B.1}$$

biçiminde hesaplanır. Amaç, altküme sayısını arttırıp en iyi sonuca ulaşılmasını sağlamaktır. Oluşturulabilecek toplam altküme sayısı $\binom{n}{k'}$ kadardır. Fakat bazı durumlarda, bu hesaplama biçimi n ve k''ne bağlı olarak büyük bir artış gösterir. Bu nedenle, rastgele seçilen belli sayıdaki altkümeden yola çıkılarak da bir çözüm elde edilebilir. Bir altkümenin iyi olabilmesi, içinde bulunduracağı k' iyi gözleme ve Ω oranıyla bulundurabileceği kuşkulu gözleme dayanır. m rastgele seçilen altküme sayısını göstermek üzere iyi bir altküme seçilebilme olasılığı,

$$1-\left(1-\left(1-\Omega\right)^{k'}\right)^m \tag{B.2}$$

biçiminde hesaplanabilir. %50'ye yakın kuşkulu gözlemli bir veri kümesi için seçilen büyük k' gözlemli altkümelerin içinde herhangi iyi altküme bulunma olasılığı düşmektedir. PROGRESS ise küçük k' değerleri için büyük m değerleri alarak beklenen iyi altküme sayısını arttırmayı hedefler.

B.2. PROGRESS Programı'nda Kullanılan EKBK Kestirim Algoritması

EKBK kestirimi, oluşturulan bir altküme $J = \{i_1, ..., i_{k'}\}$ için,

$$\sum_{i=1}^{h} \left((y_i - \boldsymbol{x}_i' \boldsymbol{\beta}_J)^2 \right)_{i:n} \tag{B.3}$$

biçiminde hesaplanır. Amaç, altküme sayısını arttırıp (B.3)'ü en küçüklemektir. Toplam altküme sayısını belirlerken kullanılan prosedür, programın EKOK kestirimi için kullandığının aynıdır. Bununla birlikte, EKBK kestirimi artık kareleri sıralama işlemi içerdiğinden hesaplama zamanı bakımından örneklem büyüklüğü (n) arttıkça EKOK kestirimine oranla daha uzun süre içerir.

B.3. EKHE İçin MINVOL Programı'nın Kullandığı Algoritma

En küçük hacimli elipsin bulunmasında verinin yarısıyla ilgilenildiğinden ve çeşitli biçimlerde bütün yarıların düşünülmesi çok zor olduğundan birçok yaklaşık algoritma üretilmiştir. Rousseeuw ve Leroy'un 1987'de belirttiği algoritmaya göre, içinde k' farklı gözlem bulunan bir altküme oluşturulur ve bu altküme $J = \{i_1, ..., i_{k'}\}$ biçiminde gösterilir. Bu altküme için ortalama vektörü ve varyans-kovaryans matrisi, sırasıyla,

$$\begin{aligned} \boldsymbol{T}_J &= \frac{1}{k'} \sum_{i \in J} \boldsymbol{x}_i', \\ C_J &= \frac{1}{k} \sum_{i \in J} (\boldsymbol{x}_i' - \boldsymbol{T}_J)' (\boldsymbol{x}_i' - \boldsymbol{T}_J) \end{aligned} \tag{B.4}$$

biçiminde hesaplanır. Burada C_J, tersi alınabilir bir matristir.

$$m_J^2 = \underset{i=1,...,n}{ortc} (\boldsymbol{x}_i' - \boldsymbol{T}_J) C_J^{-1} (\boldsymbol{x}_i' - \boldsymbol{T}_J)' \tag{B.5}$$

biçiminde hesaplanmak üzere, elips hacmi aşağıdaki eşitlikle orantılıdır:

$$\left(det\left(m_J^{2k}C_J\right)\right)^{1/2} = \left(det\left(C_J\right)\right)^{1/2}\left(m_J\right)^k \tag{B.6}$$

(B.6) eşitliğinin en küçük değeri elde edilene kadar bu algoritma çok sayıdaki J altkümesi için tekrarlanır ve sonunda en küçük değeri veren J için,

$$\begin{aligned} T(X) &= \boldsymbol{T}_J, \\ C(X) &= \left(\chi^2_{k;0.5}\right)^{-1} m_J^2 C_J \end{aligned} \tag{B.7}$$

biçiminde hesaplanır.

EKHE yöntemiyle çözüm elde etmek için kullanılabilen bu algoritma, ağırlık kavramı kullanılarak biraz daha sağlam hale getirilebilir. Ağırlıklar w_i ile gösterilmek üzere,

$$w_i = \begin{cases} 1, & \left(\boldsymbol{x}_i' - T(X)\right)\left(C(X)\right)^{-1}\left(\boldsymbol{x}_i' - T(X)\right)' \le c^2 \\ 0, & \text{diğer durumlarda} \end{cases} \tag{B.8}$$

biçiminde tanımlanır. Burada, c^2 sabit bir değer olup veri çoğunluğunun normal dağılımdan gelmesi durumunda $\chi^2_{k;0.975}$'e eşit olarak alınabilir. Böylece,

$$\begin{aligned} T_w(X) &= \frac{\sum_{i=1}^{n} w_i \boldsymbol{x}_i'}{\sum_{i=1}^{n} w_i}, \\ C_w(X) &= \frac{\sum_{i=1}^{n} w_i \left(\boldsymbol{x}_i' - T_w(X)\right)'\left(\boldsymbol{x}_i' - T_w(X)\right)}{\sum_{i=1}^{n} w_i - 1} \end{aligned} \tag{B.9}$$

biçiminde hesaplanır.

Sonuç olarak, (B.7)'den ya da (B.9)'dan elde edilen ortalama vektörü ve varyans-kovaryans matrisi yardımıyla en küçük hacimli elips için Şekil 3.1'dekine benzer bir güven bölgesi oluşturulabileceği gibi her bir *i* gözlemi için sağlam uzaklıklar da

$$SU_i = \sqrt{\left(\boldsymbol{x}_i' - T(X)\right)\left(C(X)\right)^{-1}\left(\boldsymbol{x}_i' - T(X)\right)'} \tag{B.10}$$

biçiminde hesaplanır (Rousseeuw and Van Zomeren, 1990).

B.4. EKKD İçin FAST-MCD Programı'nın Kullandığı Algoritma

EKKD yönteminden çözüm elde etmek için de birçok yaklaşık algoritma üretilmiştir. Rousseeuw ve Van Driessen'in 1999'da belirttiği algoritma, "C-adım" olarak adlandırılan bir teorem üzerine kuruludur. Bu teoreme göre, rastgele seçilmiş *h* gözlemli bir J_1 altkümesi için,

$$\begin{aligned} \boldsymbol{T}_1 &= \frac{1}{h}\sum_{i\in J_1}\boldsymbol{x}_i', \\ C_1 &= \frac{1}{h}\sum_{i\in J_1}\left(\boldsymbol{x}_i' - \boldsymbol{T}_1\right)'\left(\boldsymbol{x}_i' - \boldsymbol{T}_1\right) \end{aligned} \tag{B.11}$$

biçiminde hesaplanan (Hardin and Rocke, 2004) $(\boldsymbol{T}_1, C_1)$ kestiricileri yardımıyla elde edilen Mahalanobis Uzaklıkları doğrultusunda, $det(C_2) \le det(C_1)$ olacak biçimde yeni bir *h* gözlemli J_2 altkümesi düşünülür. "C-adım" teoremindeki "C" harfi, konsantrasyon (concentration) kelimesinden gelmekte olup en küçük uzaklığa sahip *h* gözleme ve $det(C_2) \le det(C_1)$ durumuna dikkat edilmesi anlamında kullanılır (Rousseeuw and Van Driessen, 1999). "C-adım" teoreminin algoritması aşağıdaki gibi belirtilebilir:

- Rastgele seçilen h gözlemli bir J_1 altkümesi için hesaplanan $(\boldsymbol{T}_1, C_1)$ ile her bir i gözlemi için Mahalanobis Uzaklıkları hesaplanır.
- Bu Mahalanobis Uzaklıkları küçükten büyüğe doğru sıralanır.
- Bu uzaklıklardan en küçük h tanesine sahip gözlemler J_2 altkümesini oluşturur.
- J_2 altkümesi için de $(\boldsymbol{T}_2, C_2)$ kestiricileri hesaplanır.

Yukarıdaki işlemler sonucunda, eğer $det(C_2) = 0$ veya $det(C_2) = det(C_1)$ durumu olursa algoritma sonlanır. Aksi takdirde, algoritma bir J_3 altkümesi ve diğer başka altkümeler için tekrarlanır. Bu durumu bir iterasyon süreci takip eder. Algoritma, $det(C_1) \geq det(C_2) \geq det(C_3) \geq \cdots$ durumu yakınsayana kadar sürdürülür.

"C-adım" teoremi göz önüne alındığında, $h < n \leq 600$ ve $k \geq 2$ olan bir veri kümesi için FAST-MCD Programı'nın kullandığı algoritma ve verdiği sonuçlar aşağıda belirtildiği gibidir (Rousseeuw and Van Driessen, 1999):

- "C-adım" teoreminin algoritmasındaki rastgele seçim yerine h gözlemli bir J_1 altkümesi oluşturulur. Bunun için, rastgele seçilen k+1 gözlemli bir altkümeden elde edilen $(\boldsymbol{T}_0, C_0)$ kestiricileri yardımıyla her bir gözleme ait Mahalanobis Uzaklıkları hesaplanır. Bu uzaklıklar küçükten büyüğe doğru sıralanıp içlerinden en küçük h tanesi ile J_1 altkümesi oluşturulur. Bu işlem tekrarlanarak genellikle 500 tane altküme oluşturulur. h gözlemli altkümeyi rastgele seçmeyip de rastgele seçilen k+1 gözlemli altkümeden oluşturmanın nedeni, h sayıdaki gözlemin k+1 sayıdaki gözleme oranla kuşkulu gözlemleri içerme olasılığının yüksek olmasıdır. Sonuçta, içinde kuşkulu gözlemlerin bulunduğu altküme ile iterasyonlar sonucu yakınsama elde edilemez.
- Oluşturulan 500 altküme için "C-adım" teoreminin algoritmasında belirtilen ileri iki basamak bulunur.
- En küçük $det(C_3)$'e sahip 10 altkümeye yakınsayana kadar tam iterasyon uygulanır.

- Sonunda, en küçük determinantlı varyans-kovaryans matrisine sahip küme yardımıyla en iyi h gözlem ve bu h gözlemin oluşturduğu küme için $(\boldsymbol{T}, C)$ kestiricileri verilir.
- Bununla birlikte, $(\boldsymbol{T}, C)$ kestiricileri yardımıyla MINVOL Programı'ndakine benzer şekilde ağırlıklı kestiriciler de verilebilir.
- Sonuç olarak, çoklu kuşkulu gözlemlerin saptanmasında kullanılacak sağlam uzaklıklar verilir.

EK-C

R-CODE

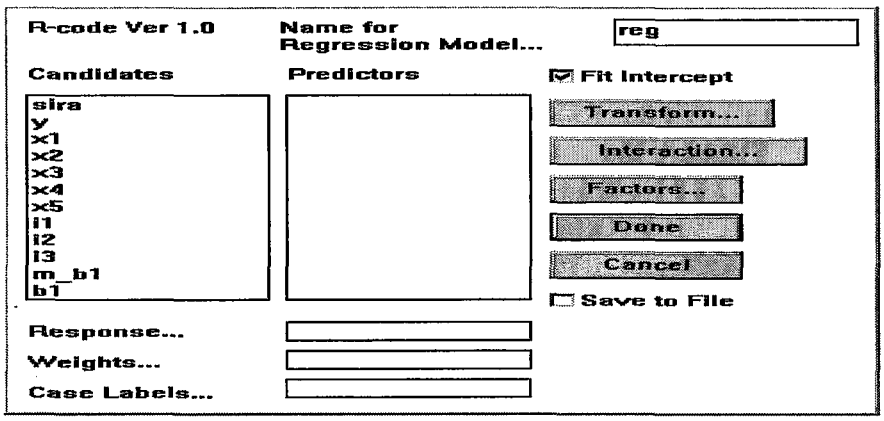
R-code (regresyon için yazılmış kod), Luke Tierney tarafından 1990'da açıklanan "Xlisp-Stat" dilinde yazılmış bir bilgisayar programıdır. R-code programını kullanmak için doğrusal regresyon çözümlemeleri konusunda bilgi sahibi olmak gerekir (Cook and Weisberg, 1994).

R-code programı, Windows İşletim Sistemi'nde kurulduktan sonra iki dosya belirir. Bunlar, "Xlisp-Stat" programını çalıştırıp R-code'u yükleyecek "Launch R-code" ve yeni bir veri kümesi hazırlamada kullanılan "Lspedit" 'tir. R-code, ya "Lspedit" 'te hazırlanmış verileri ya da .lsp uzantılı dosyalardaki verileri algılar.

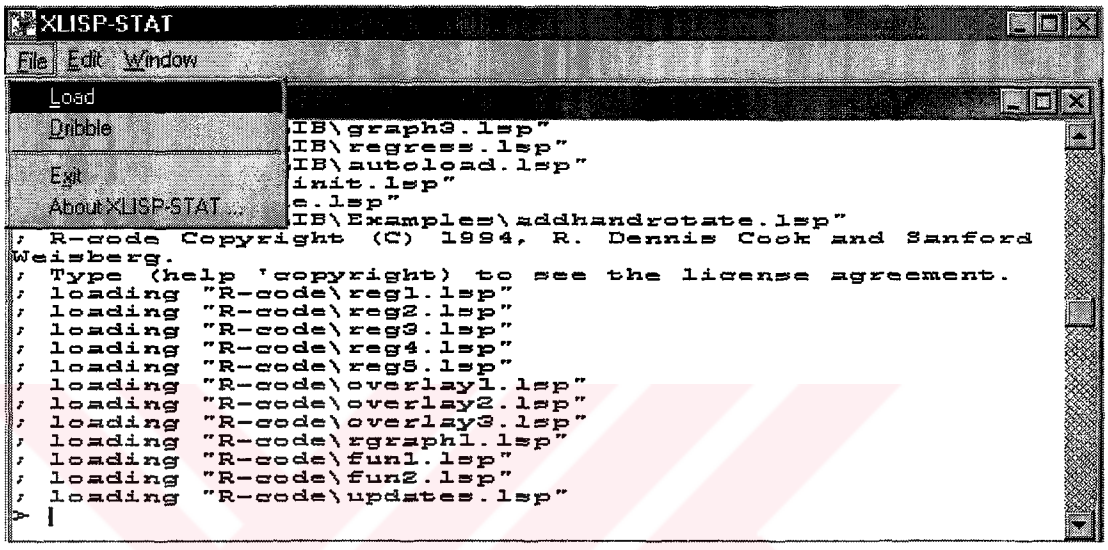
"Lspedit" 'te veri hazırlamak için önce "Lspedit" çalıştırılır ve herbir değişkenin verileri sütunlar halinde girilir. Burada önemli olan, herbir satırdaki değer sayılarının aynı olması ve herbir değerin arasında birer boşluk bırakılmasıdır. Eksik değerli veriler R-code'da çalışmaz. "Lspedit" 'te veri girildikten sonra bu pencere kapatılır ve bu işlem sırasında veri istenilen isimde .lsp uzantılı dosya olarak kaydedilmiş olur. Daha sonra R-code çalıştırılıp çıkan ekrana

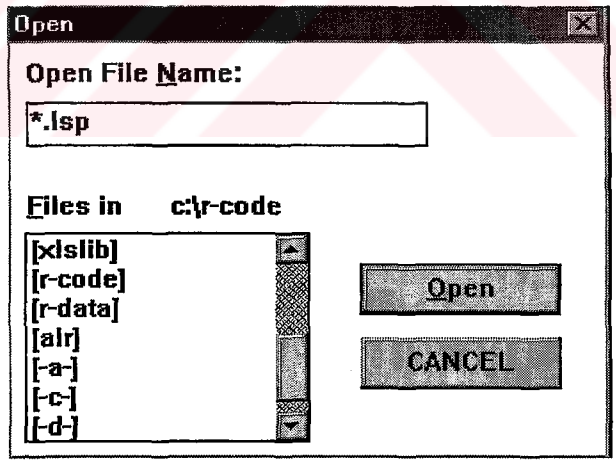
> (r-code)

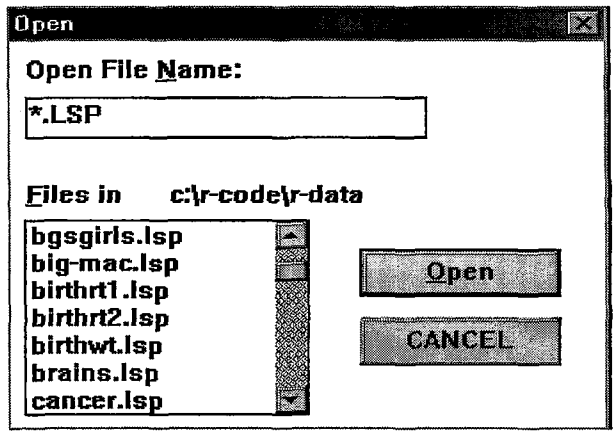
komutu girilip açılan pencerede kaydedilen .lsp uzantılı dosya seçilir ve herbir sütun için değişken ismi belirlenir. Bu aşamadan sonra, çıkan regresyon penceresinde "Save to File" kutusu işaretlenip veri kümesi yeni haliyle kaydedilir. Daha sonra da bağımlı ve bağımsız değişkenler ilgili kutularda belirtilerek "Done" seçeneğiyle regresyon çözümlemesine geçilir. Aşağıda bu aşamadaki regresyon penceresinin görünümü verilmiştir:



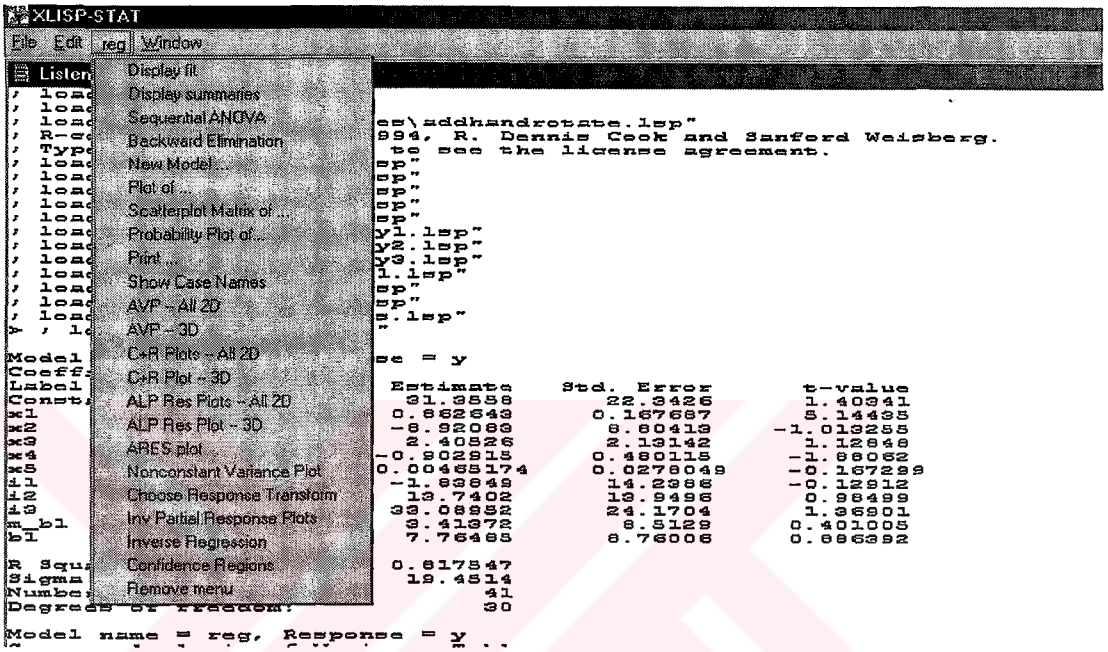
Programdaki hazır veri kümelerinden faydalanmak için de "File" menüsünden "Load" işaretlenip çıkan pencerede "R-data" bölümüne girilip istenilen hazır bir veri kümesi ile regresyon çözümlemesi yapılabilir.



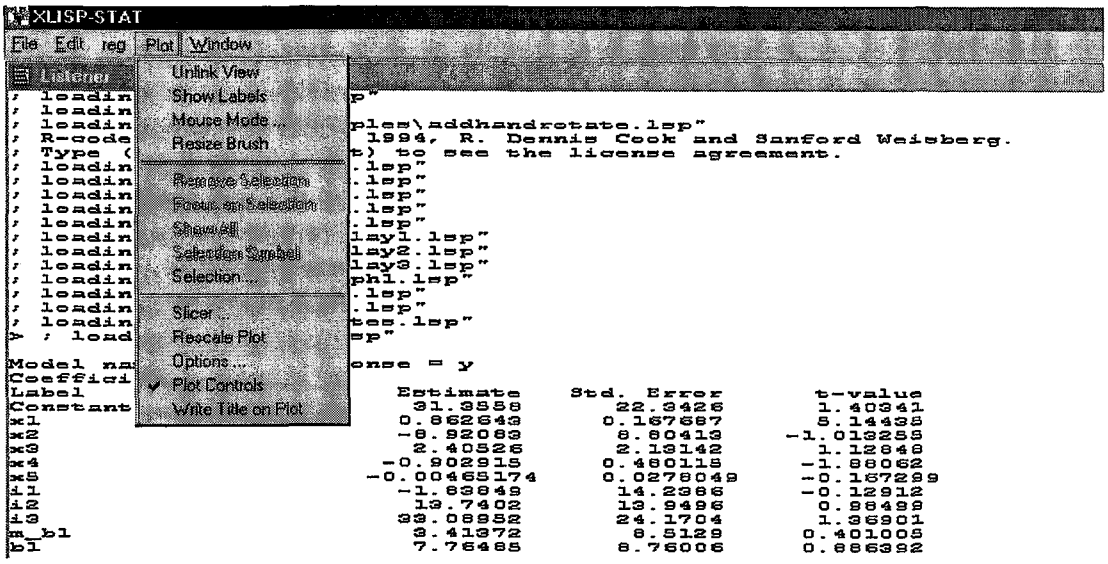




"Done" seçeneğiyle regresyon çözümlemesine geçildiğinde, R-code programında çeşitli çözümlemelerin yapıldığı ve iki veya üç boyutlu çeşitli grafiklerin elde edildiği "reg" isimli yeni bir menü belirir.



Yukarıda, "reg" menüsündeki seçenekler gösterilmiştir. Eğer herhangi bir grafik çizdirilirse de bu grafik ve ayarları üzerinde çeşitli incelemelerin yapıldığı "plot" menüsü belirir.



"Xlisp-Stat" dilindeki çeşitli komutlarla da R-code programındaki çeşitli işlemler yapılabilir. Bu komutlardan bazıları ve yaptıkları işlemler aşağıda verildiği gibidir:

- > (def varname (list numbers))

Verilerin komutla girilmesi durumunda kullanılır. Burada, "varname" değişken isminin, "numbers" da verilerin birer boşluk bırakılarak girildiği kısımdır.

- > (length varname)

Bu komut, "varname" isimli değişkendeki gözlem sayısını verir.

- > (def xbar (mean varname))

Bu komut, "varname" isimli değişkenin aritmetik ortalamasını "xbar" 'a eşitleyerek verir.

- > (def sdx (standard-deviation varname))

Bu komut, "varname" isimli değişkenin standart sapmasını "sdx" 'e eşitleyerek verir.

- > (histogram varname)

Bu komut, "varname" isimli değişkenin histogramını verir.

- > (def p (plot varname1 varname2))

Bu komut, "varname1" ve "varname2" isimli değişkenlerin oluşturduğu serpme grafiğini "p" isminde tanımlayarak verir.

- > (exit)

Bu komut, programdan çıkmayı sağlar.

R-code programının ürettiği bazı hata mesajları ve anlamları da aşağıda verildiği gibidir:

- error: misplaced right paren

Bu hata mesajı, girilen komutta sağ kısma bir parantez daha konulması gerektiğini belirtir.

- error: can't assign to a constant

"Xlisp-Stat" dilinde, "pi", "e" ve "t" gibi isimler tanımlıdır. Bu hata mesajı, bunların değişken ismi olarak kullanılamayacağını belirtir.

- error: not a number – ABC

Bu hata mesajı, sayı kullanılması gerekli bir yerde ABC karakterlerinin kullanıldığını belirtir.

- error: unbound variable – ABC

Bu hata mesajı, ABC isimli değişkenin tanımlı olmadığını belirtir.

- error: sequences of different lengths

Bu hata mesajı, çeşitli operasyonlarda ortaya çıkabilecek gözlem sayısı uyumsuzluğunu belirtir.

- error: illegal zero argument

Bu hata mesajı, sıfıra bölme teşebbüsünü belirtir.

- error: too few arguments veya error: too many arguments

Bu hata mesajları, farklı sayıdaki belirtece sahip fonksiyonu belirtir.

- error: not a list – NIL

Bu hata mesajı, bir fonksiyonun veya mesajın listelenemediğini belirtir.

- error: dimensions do not match

Bu hata mesajı, uyumsuz boyutlara sahip matrislerle aritmetik işlemler yapılmaya çalışıldığını belirtir.

- error: arguments not all the same length

Bu hata mesajı, bir fonksiyondaki bir belirtecin farklı uzunlukta olduğunu belirtir.

- error: not enough memory to allocated to color port

Bu hata mesajı, grafiklerde renk kullanmak için bilgisayarın hafızasının yetersiz olduğunu belirtir.

- error: insufficient node space

Bu hata mesajı, program hafızasının yeterli olmadığını belirtir.

ÖZGEÇMİŞ

Adı Soyadı : Barış AŞIKGİL
Doğum Yeri : Ankara
Doğum Yılı : 1980
Medeni Hali : Bekar

Eğitim Durumu :
İlkokul : Çankaya İlkokulu (1986-1991)
Ortaokul-Lise : Ankara Anadolu Lisesi (1991-1998)
Lisans : Orta Doğu Teknik Üniversitesi, İstatistik Bölümü (1998-2003)

Yabancı Dil : İngilizce, Almanca

İş Tecrübesi :
2004- : Mimar Sinan G. S. Üniversitesi, İstatistik Bölümü Araştırma Görevlisi