ÇUKUROVA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

YÜKSEK LİSANS TEZİ

Yekta Sitara KOÇ

ROBUST TAHMİN EDİCİLERİ VE ÖZELLİKLERİ

İSTATİSTİK ANABİLİM DALI

ADANA, 2007

ÇUKUROVA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

ROBUST TAHMİN EDİCİLERİ VE ÖZELLİKLERİ

Yekta Sitara KOÇ

YÜKSEK LİSANS TEZİ

İSTATİSTİK ANABİLİM DALI

Bu tez 25/ 09 / 2007 Tarihinde Aşağıdaki Jüri Üyeleri Tarafından Oybirliği/ Oyçokluğu İle Kabul Edilmiştir.

İmza……	İmza……..	İmza……
Prof.Dr. Fikri AKDENİZ	**Prof.Dr. Hamza EROL**	**Prof.Dr. H.AltanÇABUK**
DANIŞMAN	ÜYE	ÜYE

Bu tez Enstitümüz İstatistik Anabilim Dalı'nda hazırlanmıştır.

Kod No:

Prof.Dr. Aziz ERTUNÇ
Enstitü Müdürü
İmza ve Mühür

ÖZ
YÜKSEK LİSANS TEZİ

ROBUST TAHMİN EDİCİLERİ VE ÖZELLİKLERİ

Yekta Sitara KOÇ

ÇUKUROVA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ
İSTATİSTİK ANABİLİM DALI

Danışman: Prof. Dr. Fikri AKDENİZ

Yıl: 2007, **Sayfa: 110**

Jüri: Prof.Dr. Fikri AKDENİZ
Prof.Dr. Hamza EROL
Prof.Dr. H. Altan ÇABUK

Robust tahmin ediciler, veri kümesinde güvenli gözlemlerin homojen dağılmaması durumunda güvenli sonuçlar bulmak ve sapan değerlerin etkisini azaltmak amacıyla kullanılır.

Tezin temel amacı; klasik regresyon analizinde sapan gözlemlerin varlığı nedeniyle standart varsayımların sağlanmaması durumunda en küçük kareler yöntemine alternatif olarak sunulan robust regresyon yöntemlerinin incelenmesidir.

Bu çalışmada önce sapan değerler, kırılma noktası ve etki fonksiyonu kavramları ele alınacak, sonra robust basit regresyon ve çoklu regresyondaki tahmin ediciler incelenecektir. Örneklerle bu tahmin ediciler, en küçük kareler tahmin edicisiyle karşılaştırılacaktır.

Anahtar kelimeler: En küçük kareler tahmin edici, En küçük medyan kareler tahmin edici, Kırılma noktası, Robust tahmin edici, Sapan değer

ABSTRACT

MSc.THESIS

ROBUST TAHMİN EDİCİLERİ VE ÖZELLİKLERİ

Yekta Sitara KOÇ

DEPARTMENT OF STATISTICS
INSTITUTE OF NATURAL AND APPLIED SCIENCES
UNIVERSITY OF ÇUKUROVA

Supervisor: Prof. Dr. Fikri AKDENİZ
Year: 2007, **Sayfa: 110**
Jury: Prof.Dr. Fikri AKDENİZ
Prof.Dr. Hamza EROL
Prof.Dr. H.Altan ÇABUK

Robust estimators are used for reducing the effects(weights) of outlying observations in the data set to get more reliable and stable estimators.

The aim of this thesis is to propose robust regression procedures as an alternative method to Least Squares procedure which is widely used in classical regression analysis and very sensitive to outlying observations.

In this thesis, firstly outlier and breaking point concepts will be introduced, secondly a general overview of estimators for robust simple and multiple regression will be given and finally these estimators will be compared with classical Least Squares estimators and examples will be provided.

Keywords: Breaking Point, , Least median squares estimator, Least squares estimator, Outlier, Robust estimator,

TEŞEKKÜR

Bu tezin hazırlanması esnasında değerli bilgilerini ve kaynaklarını benimle paylaşan danışmanım sayın Prof.Dr. Fikri AKDENİZ'e; yardımlarını esirgemeyen İstatistik bölümü öğretim elemanlarına teşekkürlerimi sunarım.

Ayrıca eğitim ve öğretim hayatım boyunca maddi ve manevi katkılarını esirgemeyen aileme teşekkürü bir borç bilirim.

İÇİNDEKİLER **SAYFA**

TABLOLAR DİZİNİ

ŞEKİLLER DİZİNİ

1.GİRİŞ

Regresyon analizinde amaç; gözlenen değerlere uyan en iyi denklemi oluşturmaktır. Birçok regresyon tekniği olmasına karşın bunlardan en kullanışlı; olanı klasikliği ve hesaplama kolaylığından dolayı en küçük kareler (EKK) tahmin edicisidir. Fakat bu tahmin edici sapan değerlere karşı çok hassastır. Bu probleme çözüm bulmak amacıyla sapanlar değerlerden çok etkilenmeyen yeni istatistiksel teknikler geliştirildi. Böylece robust (dayanıklı) tahmin ediciler ortaya çıktı. Bu yöntem verinin çoğunluğuna uygun bir model tasarlamaya çalışır. Yani, veri kümesinin küçük bir bölümü sapan değerlerden oluşsa bile kalan büyük bölüm güvenilir sonuçlar verir. Bu çalışmada basit regresyon ve çoklu regresyon olmak üzere iki ana başlık altında robust tahmin edicileri ele alınmış ve bunların sonuçları örnekler aracılığıyla klasik EKK analiziyle karşılaştırılmıştır. Konunun daha iyi anlaşılması için bazı tanımlar bu kısımda verilmiştir.

1.1. Kırılma Noktası (Breakdown Point)

n tane veri noktasından oluşan herhangi bir örneklemi alalım:

$$Z = \{(x_{11},..., x_{1p}, y_1),..., (x_{n1},..., x_{np}, y_n)\} \tag{1.1}$$

T bir regresyon tahmin edicisi olsun. Yani Z örnekleminde T 'ye başvurularak $\hat{\theta}$ regresyon katsayılarının bir vektörü elde edilir.

$$\mathrm{T(Z)}=\hat{\theta} \tag{1.2}$$

dir. Orijinal veri noktalarının herhangi m tanesi keyfi değerlerle değiştirilsin. Böylece Z örneklemi Z' ne dönüştürülür. Bu durumda maksimum yanlılık bias (m;T,Z) ile ifade edilir ve

$$\text{bias (m;T,Z)} = \sup_{Z'} \| T(Z') - T(Z) \| \tag{1.3}$$

olarak tanımlanır. Burada bias (m;T,Z), sonsuz ise yani m sapan değer T üzerinde geniş bir etkiye sahipse bu tahmin edicinin kırılması olarak açıklanır. Bu nedenle, örnekteki T tahmin edicisinin kırılma noktası sonlu örneklem için

$$\varepsilon_n^*(T,Z) = \min\left\{\frac{m}{n}; bias(m;T,Z) sonsuz\right\} \tag{1.4}$$

şeklinde tanımlanır. Diğer bir ifadeyle bu T (Z) den uzaktaki keyfi değerler için T tahmin edicisinin neden olduğu bozulmanın en küçük kırılmasıdır. Dikkat edilirse bu tanım, dağılımların olasılığını içermez!

EKK 'ya göre bütün sınırlar üzerinde T 'yi taşımak için bir sapan değerin yeterli olduğu görülür. Bu nedenle kırılma noktası,

$$\varepsilon_n^*(T,Z) = \frac{1}{n}$$

dir. Burada n örneklem genişliği arttıkça $\varepsilon_n^*(T,Z)$ ifadesi 0'a yaklaşır. Bu nedenle EKK, % 0 'lık kırılma noktasına sahiptir denilir. Bu EKK yönteminin sapan değerlere karşı aşırı hassaslığını da gösterir.

1.2. Etki Fonksiyonu (Influence function=IF)

Kırılma noktası, güvenirliğin evrensel bir ölçüsüyken; etki fonksiyonu, tahmin edici üzerindeki son derece küçük etkileri ölçer. F dağılımında T tahmin edicisinin etki fonksiyonu,

$$\mathrm{IF}(z;T,F) = \lim_{\varepsilon\to 0} \frac{T(\text{◀}-\varepsilon\,\widetilde{.F} + \varepsilon.\Delta z\,\widetilde{)} + T(F)}{\varepsilon}$$

dir. Örnek uzayın tüm z noktalarında limit vardır. Δz terimi, z noktasındaki bütün yığınının olduğu olasılık dağılımıdır.

Etki fonksiyonu, veri kümesinde sapan değerlerin etkilendiği yanlılığı ifade eder. Homojen olmayan verilerin dağılımında bu ifade, tahmin edicinin,

$$T\left((1-\varepsilon)F+\varepsilon\Delta_z\right)\approx T(F)+\varepsilon IF(z;T,F)$$

şeklindeki bir yaklaşımını verir. Sonlu örneklemlerle çalışıldığında zaman etki fonksiyonunun kesikli tipi kullanabilir. Buna hassas eğri (sensitivity curve=SC) denir. Hassas eğri,

$$\text{SC(z)} = \frac{T\left(\frac{n-1}{n}.F_n+\frac{1}{n}\Delta_z\right)-T(F_n)}{\frac{1}{n}} \approx IF(z;T,F)$$

ile tanımlanır. Burada F_n ,

$$Z = \{z_1,...,z_n\} = \{(x_{11},...,x_{1p},y_1),...,(x_{n1},...,x_{np},y_n)\}$$

ile oluşturulmuş deneysel bir dağılımdır. Eğer bir gözlem ile z sapan değerinin yeri değiştirilirse

$$T(z_1,\ldots,z_{n-1},z)=T(F_n)+\frac{1}{n}SC(z)$$

ifadesi bulunur. Hatta gözlemlerin $\varepsilon = m/n$ küçük bir kırılmasıyla yer değiştirmesi olasıdır. Bu durumda da

$$T(z_1,\ldots,z_{n-1},z)=T(F_n)+\frac{m}{n}IF(z;T,F)$$

ifadesi elde edilir. Bütün tahmin ediciler bir kırılma noktasına sahiptir. Bununla birlikte mutlaka bir etki noktasına sahip olması gerekmez.

2.ÖNCEKİ ÇALIŞMALAR VE REGRESYON ANALİZİNDE SAPAN DEĞERLER

Regresyon analizinde amaç; gözlenen değerlere en iyi uyan denklemleri oluşturmaktır. Klasik lineer model,

$$y_i = x_{i1}\,\theta_1 + \ldots x_{ip}\theta_p + e_i \qquad i=1,,\ldots,n \qquad (2.1)$$

dir. Burada n örneklem genişliği, x_{i1}, … , x_{ip} açıklayıcı değişkenler, y_i yanıt değişkendir. e_i hatalarının ise 0 ortalamalı ve bilinmeyen varyanslı normal dağılıma sahip olduğu varsayılır. Bilinmeyen parametre vektörü yani

$$\theta = \begin{bmatrix} \theta_1 \\ . \\ . \\ \theta_p \end{bmatrix} \quad , \qquad (2.2)$$

veriden tahmin edilir. Veri için

değişkenler

$$\text{olaylar} \begin{bmatrix} x_{11}\ldots\ldots x_{1p} & y_1 \\ . & . \\ x_{i1}\ldots\ldots x_{ip} & y_i \\ . & . \\ x_{n1}\ldots\ldots x_{np} & y_n \end{bmatrix} \qquad (2.3)$$

matris gösterimi kullanılsın. Böyle bir veri kümesine regresyon tahmin edicisi uygulandığından

$$\hat{\theta} = \begin{bmatrix} \hat{\theta}_1 \\ . \\ \hat{\theta}_p \end{bmatrix} \qquad (2.4)$$

elde edilir. Burada $\hat{\theta}_j$ tahminleri, regresyon katsayıları olarak adlandırılır. Gerçek θ_j bilinmemesine rağmen $\hat{\theta}_j$ tahmin edicileri ile açıklayıcı değişkenler çarpılarak

$$\hat{y}_i = x_{i1}\hat{\theta}_1 + ... + x_{1p}\hat{\theta}_p \tag{2.5}$$

tahmin edilen değerleri elde edilir. Bu durumda i-inci olayın r_i rezidüsü, y_i gözlenmiş değerleri ile $\hat{y}_i$ tahmin edilen değerinin farkı olarak tanımlanır. Yani,

$$r_i = y_i - \hat{y}_i \tag{2.6}$$

dır. En popüler regresyon tahmin, edicisi Gauss ve Legendre tarafından bulundu. Bu tahmin edici,

$$\min_{\hat{\theta}} \sum_{i=1}^{n} r_i^2 \tag{2.7}$$

ifadesine karşılık gelir. Bu tahmin edicinin amacı (2.7) ile verilen ifadeyi minimum yaparak modeli en iyi duruma getirmektir. Bu yöntem çok iyi bilinen en küçük kareler yöntemidir (EKK). Bu yöntem, istatistiğin önemli bir köşe taşıdır. Popüler olmasının nedeni ise anlaşılmasının kolaylığıdır. 1800 'lü yıllarda bulunduğunda bilgisayarlar yoktu ve EKK tahmin edicisi veriden belirli bir matris cebiri ile kolayca hesaplanabilirdi. Günümüzde bile birçok istatistiksel paket program hala gelenekselliği ve hesaplanma hızından dolayı aynı tekniği kullanmaktadır. Yine de tek boyutlu durumda (2.7) 'de verilen EKK kriteri gözlemlerin aritmetik ortalamasını verir ki o zaman da en uygun konum tahmin edicisi olduğu görülür. Sonradan Gauss EKK 'yı en iyi duruma getiren hata dağılımı olarak normal dağılımı (Gaussian Dağılımı) önerdi. Böylece güzel bir matematik teorisi oldu. Ondan sonra Gaussian varsayımları ve EKK 'nın kombinasyonu, istatistik tekniklerinin üretilmesi için standart bir mekanizma oldu. (Örneğin Çoklu yer, Varyans Analizi, Minimum varyans sınıflaması gibi)

Daha yakın zamanlarda birçok araştırmacı, gerçek verilerin klasik varsayımları tamamen sağlamadığını farketmeye başladı. (Student 1927, Pearson 1931, Box 1953 ve Tukey 1960) .

Basit doğrusal regresyon modelinde sapan değerlerin etkilerini inceleyelim.

$$y_i = \theta_1 x_i + \theta_2 + e_i \tag{2.8}$$

basit doğrusal regresyon modelinde θ_1 eğimi ve θ_2 sabiti tahmin edilmelidir. Bu (2.1) ‘ deki ifadede p = 2 özel durumudur. Çünkü

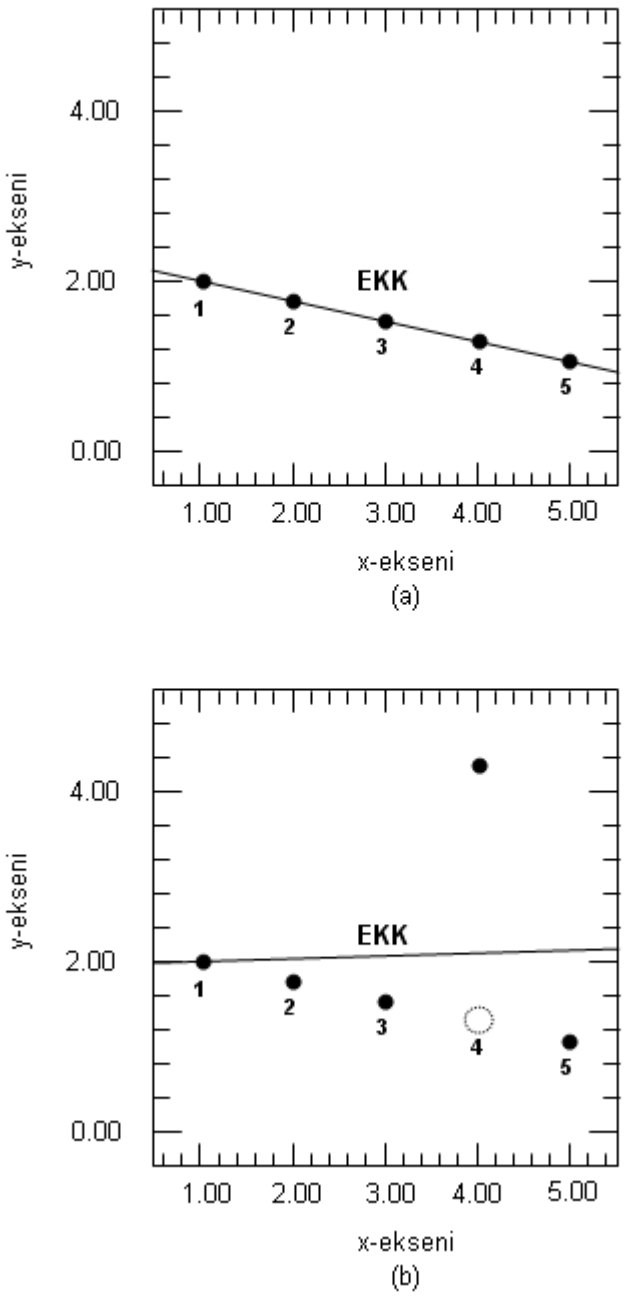
$x_{i1} = x_i$ ve $x_{i2} = 1$ tüm i = 1,…..n için

olur. Genelde açılayıcı değişkenin 1 olması sabit terimli regresyonu elde etmek için kullanılan standart bir yoldur. Basit doğrusal regresyon modelinde saçılım grafiği olarak adlandırılan (x_i , y_i) ‘nin grafiği oluşturulur. p ‘nin genel hali için çoklu regresyon modelinde bu mümkün olmaz. İşlemleri görsel olarak açıklamak amacıyla basit doğrusal regresyon modelini kullanmak daha iyidir.

Şekil 2.1(a) (x_1, y_1), … , (x_5, y_5) noktalarının saçılım grafiği olsun. Burada bu noktalar doğrusal olarak uzanıyorlar. Böylece grafikte,

$$\hat{y} = \hat{\theta}_1 x + \hat{\theta}_2$$

olur. EKK doğrusundan görülebileceği gibi EKK çözümü veriye çok iyi uyar. Fakat kopyalarken ya da taşırken y_4 değerinin örneğin ondalıklı kısmının hatalı alındığını varsayalım. Bu durumda (x_4, y_4) değeri ideal doğrudan uzaklaşacaktır. Şekil2.1(b), 4.noktanın.(kesikli çizgiyle yuvarlak içine alınan nokta) orijinal durumundan uzaklaştığını ve yukarı doğru hareket ettiğini gösteriyor. Bu nokta y-yönünde sapan değer olarak adlandırılır. Bu ise şekil2.1(a) ‘daki EKK den oldukça farklı olan EKK doğrusu üzerinde oldukça geniş bir etkiye sahiptir. Bu olayın literatüre önemli katkısı



Şekil 2.1. (a) beş noktalı orijinal veri kümesi ve EKK regresyon doğrusu. (b) ise (a)'daki veri ile aynı fakat y-yönünde sapan var.

olmuştur. Çünkü genelde y_i gözlemler olarak, $x_{i1},..., x_{ip}$ iler de sabit sayılar olarak düşünülür. Böyle sapan değerler geniş pozitif rezidülere ya da geniş negatif rezidülere sahip olur. Gerçekten bu örnekte 4.nokta doğrudan en uzak konumdadır.

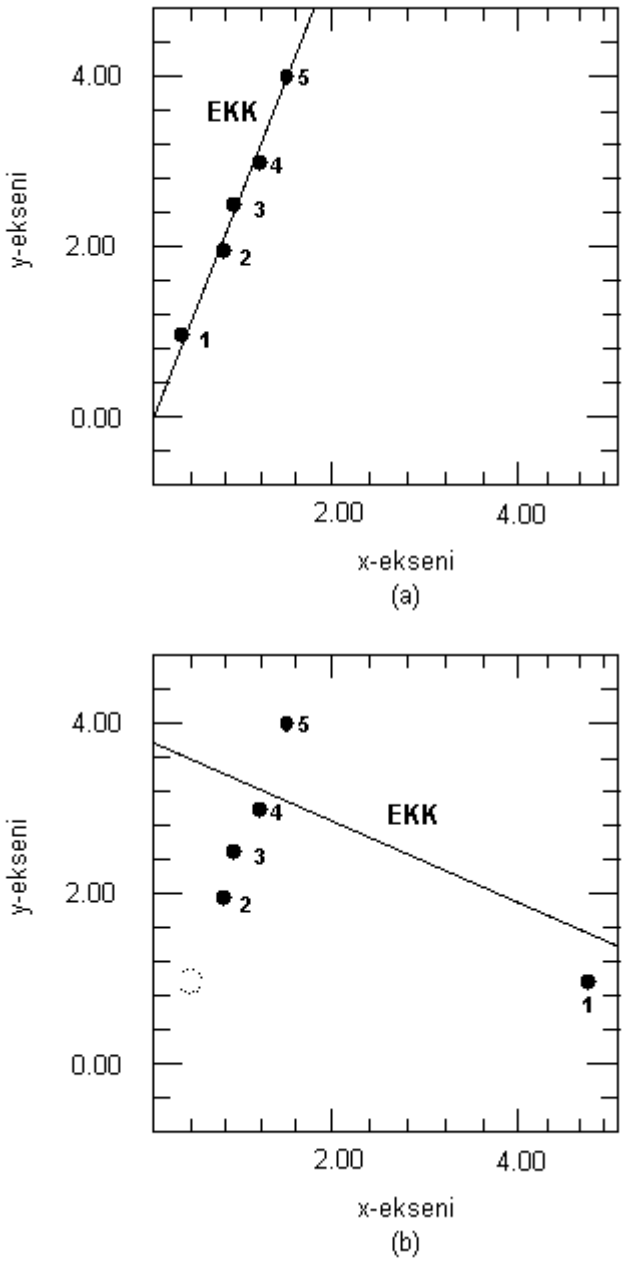
Bu nedenle (2.6) ile verilen r_i si şüpheli bir biçimde geniştir. p genişliğindeki (2.1) çoklu genel regresyonda bile ki bu durumda veriyi gözümüzde canlandıramayız böyle sapanlar rezidülerin listesinden veya rezidü grafiklerinden bulunabilir.

Fakat genelde $x_{i1},..., x_{ip}$ açıklayıcı değişkenler rasgele değişkenliğe bağlı olarak gözlenmiş niceliklerdir. Gerçekten de birçok uygulamada bir yanıt değişken ve bazı açıklayıcı değişkenlerden seçilmek zorunda olan değişkenlerin bir listesi alınır. Bu nedenle y_i yanıt değişkende sadece kötü noktaların (gross errors) neden meydana geleceğinin bir sebebi yoktur. Bazı durumlarda p genel olarak 1 den büyük olduğu için $x_{i1},..., x_{ip}$ açıklayıcı değişkenlerden birinde sapan olması bile daha muhtemeldir ve bu nedenle bir şeye yanlış gitmek için daha fazla fırsat vardır. Böyle bir sapanın etkisi için Şekil2.2 deki basit regresyonun bir örneğini inceleyelim:

Şekil2.2a, EKK doğrusuna iyi uyan (x_1, y_1), … , (x_5 , y_5) noktalarını içeriyor.x_1 i hatalı girersek Şekil2.2b 'yi elde ederiz.O zaman bu noktaya x yönünde sapan deriz.Bu nokta EKK doğrusunu eğdiği için EKK doğrusu üzerindeki etkisi çok geniştir.Bu nedenle şekil2b 'deki (x_1, y_1) noktası **kaldıraç noktası** olarak adlandırılır.

EKK tahmin edicisi üzerindeki bu çekim şu şekilde açıklanabilir: x_1 orijinal doğrudan uzaklaştığı için orijinal doğrudaki r_1 rezidüsü çok büyük bir değer alır.$\sum_{i=1}^{5} r_i^2$ katkısıyla çok daha büyük olur. Bu nedenle orijinal doğru EKK yönünden seçilemez ve gerçekten şekil2b doğrusu $r_2^2,..., r_5^2$ terimleri azar azar artsa bile r_1^2 genişliğini azaltarak eğdiği için en küçük $\sum_{i=1}^{5} r_i^2$ ye sahip olur.

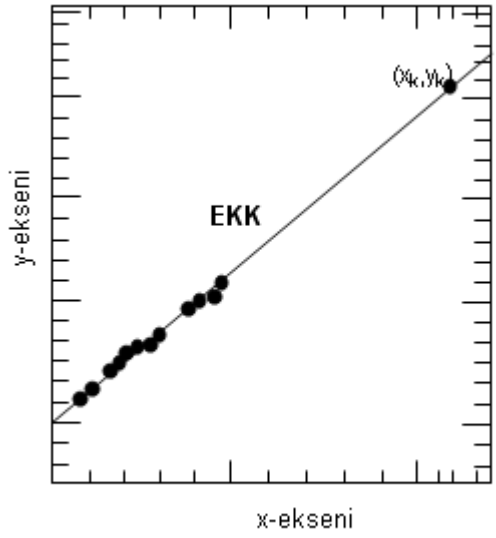
Genelde (x_k, y_k) gözlemini her ne zaman x_k örneklemdeki gözlenmiş x_i yığınından uzaklaşırsa kaldıraç noktası olarak adlandırılır. y_k bu durumda hesaba alınmaz. Bu nedenle (x_k, y_k) noktası muhakkak bir **regresyon sapanı** olmak



Şekil 2.2. (a) Beş noktalı orijinal veri kümesi ve EKK regresyon doğrusu. (b) ise (a) 'daki veri ile aynı fakat x-yönünde sapandır ("kaldıraç noktası").

zorunda değildir. (x_k, y_k) verinin çoğunluğu tarafından tanımlanan regresyon doğrusuna yaklaştığında Şekil2.3 te görüldüğü gibi iyi bir kaldıraç noktası olarak düşünülebilir. Bu nedenle (x_k, y_k) nın bir kaldıraç noktası olduğunu söylemek

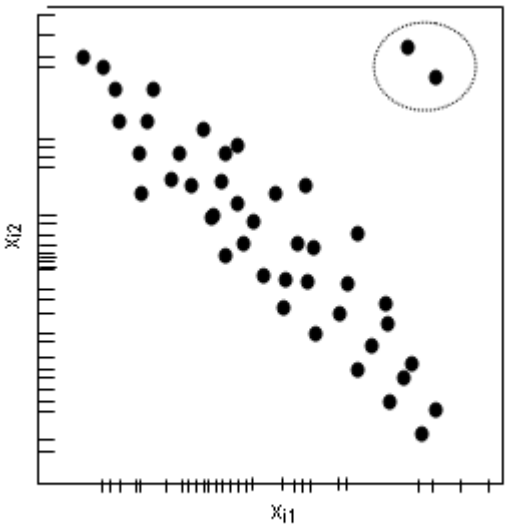
uzaktaki x_k bileşenine göre sadece güçlü bir biçimde etkili $\hat{\theta}_1$ ve $\hat{\theta}_2$ regresyon katsayılarına göre gücünü söylemek demektir. Fakat bu durum (x_k, y_k) gerçekten $\hat{\theta}_1$ ve $\hat{\theta}_2$ üzerinde çok geniş bir etkiye sahip olması anlamına gelmek zorunda değildir. Çünkü (x_k, y_k) diğer verilere göre eğim kümesine tamamen uyabilir (Böyle bir durumda kaldıraç noktası bazı güven bölgelerini daraltacağından oldukça yararlı bile olabilir.)



Şekil 2.3. (x_k,y_k) noktası x_k sapan olduğu için bir kaldıraç noktasıdır. Fakat . (x_k,y_k) bir regresyon sapanı değildir çünkü diğer veri kümeleriyle birlikte modele uyuyor.

Çoklu regresyonda ($x_{i1},..., x_{ip}$) p – boyutlu bir uzayda uzanır. (Bazen faktör uzayı olarak adlandırılır.) O zaman kaldıraç noktası bir veri kümesinde ($x_{i1},..., x_{ip}$) ‘e göre uzakta kalan ($x_{k1},..., x_{kp}$) için ($x_{k1},..., x_{kp}$, y_k) olarak yine tanımlanır. Önceden, böyle kaldıraç noktaları esas y_k değerine bağlı olarak EKK regresyon katsayıları üzerinde güçlü geniş bir etkiye sahipti fakat bu durumda kaldıraç noktalarını tanımlamak daha da zordur. Çünkü boyut daha büyüktür. Gerçekten 10 tane açıklayıcı değişken olduğunda böyle bir noktayı bulmak çok zor olabilir. Bunu gözümüzde canlandıramayız.

Bu problemin basit bir örneği Şekil3.4 'te veriliyor. Burada x_{i1}, x_{i2} ye karşı belirli bir veri kümesine göre veriliyor. Bu grafikte 2 tane kaldıraç noktasını kolayca görürüz. Fakat x_{i1} ve x_{i2} ayrı düşünüldüğünde bu noktalar görülemez. Gerçekten $\{x_{11}, x_{21}, ..., x_{n1}\}$ tek boyutlu örneklemi sapanlar içermez ve $\{x_{12}, x_{22}, ..., x_{n2}\}$ de içermez. Genel olarak her değişkene ayrı ayrı bakmak ya da bütün değişken çiftlerine bakmak bile yeterli değildir. Uzaktaki $(x_{i1}, ..., x_{ip})$ yi belirlemek oldukça zor bir problemdir. Fakat biz çoğunlukla regresyon sapanlarından bahsedeceğiz. Yani $(x_{i1}, ..., x_{ip}, y_i)$ eş zamanlı olarak hem yanıt hem de açıklayıcı değişkenleri göz önüne alarak verinin çoğunluğuna göre lineer ilişkiden ayrılan durumları takip eder.



Şekil 2.4. Regresyon veri kümesinin (x_{i1},x_{i2}) açıklayıcı değişkenlerinin grafiği. İki kaldıraç noktası var(kesikli yuvarlak içinde). Burada koordinatların ikisi de sapan değil.

Birçok insan EKK rezidülerine bakarak regresyon sapanlarının keşfedilebildiğini tartışır. Maalesef sapanlar kaldıraç noktaları olduğunda bu doğru olamaz. Örneğin tekrar Şekil2.2(b) ye dikkat edelim. Kaldıraç noktası olan durum1 o doğruya oldukça yakın olsun diye EKK doğrusunu eğer. Sonuç olarak $r_i = y_i - \hat{y}_i$ rezidüsü negatif küçük bir sayıdır. Diğer taraftan r_2 *ve* r_5 rezidüleri iyi noktalara karşılık gelmesine rağmen bu rezidüler daha geniş mutlak değerlere sahiptir. Eğer en

geniş EKK rezidülü noktaları silme gibi bir kurula başvurulacaksa iyi noktalar ilk başta silinmelidir. Tabi ki böyle tek değişkenli bir veri kümesine bakılabildiği için böyle değişkenli bir veri kümesinde gerçekten tümüyle problemsizdir. Fakat EKK rezidülerinin dikkatli analizlerine rağmen bile sapanların görülemediği birçok çok değişkenli veri kümeleri vardır.

Sonuç olarak regresyon sapanlar (hem x de hem y de) standart EKK analizlerinde ciddi bir risk oluşturur. Esas olarak bu problemin 2 çıkış yolu vardır. İlki ve muhtemelen en çok bilineni regresyon diagnostiklerini yapmak için bir yaklaşımdır. Diagnostikler etki noktalarını belirlemek amacıyla veriden hesaplanan belli niceliklerdir. Bu sapanlar kaldırıldıktan ya da düzeltildikten sonra kalan olaylara EKK analizi yapılır. Tek bir sapan olduğunda bu yöntemlerden bazıları belli bir zaman diliminde silinen bir noktanın etkisine bakılarak oldukça iyi çalışır. Maalesef diagnostik sapanlarından birkaç tane olduğunda sapanları teşhis etmek çok daha zor olur ve böyle çoklu sapanlar için diagnostikler oldukça gereklidir ve sıklıkla geniş hesaplamaların artışını sağlar. (bütün mümkün alt kümelerin sayısı oldukça büyüktür.)

Bir diğer yaklaşım robust regresyonudur. Bu yaklaşım sapanlar tarafından çok güçlü bir biçimde etkilenmeyen tahmin ediciler tasarlamaya çalışır. Belli belirsiz robustlık fikrine sahip birçok istatistikçi robustlığın amacının basit bir biçimde sapanları ihmal etmek olduğuna inandı fakat bu doğru değildir. Aksine sapanların tanımlanabildiği robust regresyonunda rezidülere bakılır ki burada genelde EKK rezidüleri yoluyla yapılamaz. Bu nedenle diagnostikler ve robust regresyonu gerçekten de sadece şu basamaklarda aynı amaca sahiptir: Diagnostik araçları kullanıdığında ilk olarak sapanların silinmesine çalışılır ve o zaman EKK ile " iyi " veriye uygulanır. Aksine robust analizi ilk olarak verinin çoğunluğunu bir regresyona uydurmak ister ve o zaman robust çözümünden geniş rezidülere sahip noktalar olarak sapanları keşfeder.

Sonraki basamak ortaya çıkarılan yapı hakkında düşünmektir. Mesela orijinal veri kümesine geri dönülebilir ve konu sebep bilgisi sapanlar çalışması için kullanılır ve orijinleri açıklanır. Yine de eğer sapmalar modelin başarısızlığı için belirti olup

olmadığı araştırılması gerekir. Örneğin belirli bir dönüşüm yapmak ya da ikinci dereceli (kuadratik)bir terim ekleyerek tahmin edilebilir.

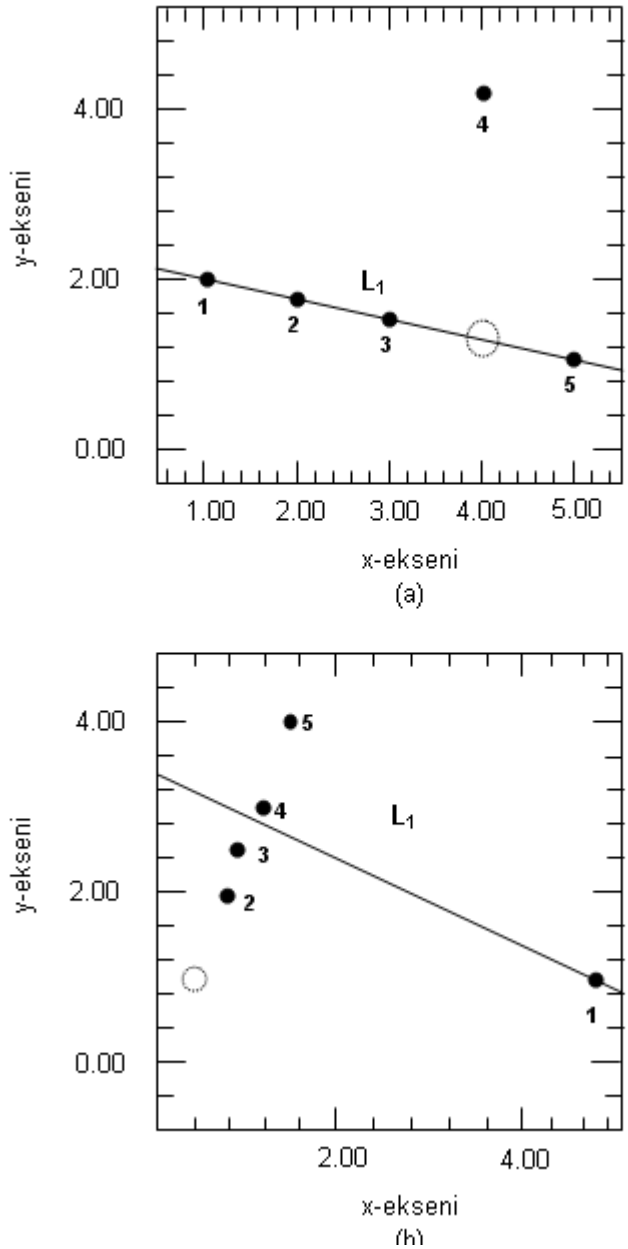
Diagnostikler kadar robust tahmin ediciler de vardır ve ikisini birbirinden ayırt etmek amacıyla ne kadar etkili olduklarını ölçmek gerekir. İkinci kısımda bazı robust metotlarında sapanların sayısı hesaplanarak kıyaslanacaktır. Alt bölümlerde temiz veri kümeleri kadar bozuk veri kümelerinin de analiz edileceği birçok robust yöntemine başvurulacaktır.

Daha robust bir regresyon tahmin edicisine doğru ilk adım 1887 de Edgeworth den geldi. Edgeworth (2.7) de r_i rezidülerinin karesi alındığı için sapanların EKK üzerinde çok geniş bir etkiye sahip olduğunu inceledi. Bu nedenle en küçük mutlak değerler regresyon tahmin edicisini ileri sürdü. Bu tahmin edici şu şekilde tanımlanır:

$$\min_{\hat{\theta}} \sum_{i=1}^{n} | r_i | \qquad (2.9)$$

Bu teknik genelde L_1 regresyonu olarak bilinir, EKK ise L_2 regresyonu olarak bilinir. Ondan önce Laplace basit medyanı elde etmek için (2.9) kriterini kullanmıştı. Medyan gibi L_1 regresyon tahmin edicisi tamamen tek değildir. (Harter 1977 ve Gentle 1977) tek değişkenli medyan‘ ın kırılma noktası % 50 kadar yüksektir. Maalesef L_1 regresyonun kırılma noktası hala % 0 dan daha iyi değildir. Nedenini görmek için Şekil2.5 ‘e bakalım:

Şekil2.5 L_1 regresyonun da sapanların etkisinin şematik bir özetini verir. Şekil2.5(a), y-yönündeki sapanların etkisini gösteriyor. EKK ‘nın aksine L_1 regresyon doğrusu böyle bir sapana karşı robusttır. Yaklaşık olarak L_1 regresyonu 4. gözlem hala düzgün kaldığında ve hala kalan noktalara iyi bir biçimde uyarken aynı şekilde kalır. Bu nedenle L_1 uzaktaki y_i ye karşı bizi korur ve bu hususta EKK ‘dan dolayı



Şekil 2.5. (a) L_1 regresyonunun y-yönündeki sapmaya karşı robustlığı (dayanıklılığı). (b) L_1 regresyonunun x-yönündeki sapana karşı hassaslığı("kaldıraç noktası").

tercih edilebilir. Son yıllarda L_1 yaklaşımının istatistiğe belirli bir yer kazandırdığı görülür. (Bloomfield ve Steiger 1980,1983; Narula ve Wellington 1982; Devroye ve Gyorfi 1984) Fakat L_1 regresyonu uzaktaki x e karşı korumaz. Bu durum şekil 5b den görülebilir. Burada kaldıraç noktasının etkisi şekildeki EKK doğrusu üzerindekinden bile daha güçlüdür. Bu kaldıraç noktası yeterince uzağa gittiğinde

ortaya çıkar. L_1 doğrusu onun arasından sağa geçer. Bu nedenle basit hatalı bir gözlem tamamen L_1 tahmin edicisinin yerine geçebilir. Bu nedenle sonlu örnekli kırılma noktası hala 1/n dir.

Bu yönde sonraki adım Huber (1973 s.800,1981) in bulduğu M tahmin edicilerinin kullanımıdır. Bunlar r_i rezidülerinin başka bir fonksiyonuyla (2.7) de ki r_i^2 karelenmiş rezidülerin yer değiştirmesi fikrine dayanıyor.

$$\min_{\hat{\theta}} \sum_{i=1}^{n} p(r_i) \tag{2.10}$$

Burada p simetrik bir fonksiyondur [p(-t)=p(t), $\forall$ t için] ve 0 da tek minimuma sahiptir. $\hat{\theta}_i$ regresyon katsayılarına göre bu ifadenin farkı

$$\sum_{i=1}^{n} \psi(r_i) x_i = 0 \tag{2.11}$$

verir. Burada ψ, p nin türevi ve x_i i-inci durumdaki açıklayıcı değişkenin satır vektörüdür:

$$x_i = (x_{i1},..., x_{ip})$$
$$0 = (0,\ldots,0) \tag{2.12}$$

Bu nedenle (2.11) gerçekten p tane denklemin bir sistemidir. Çözümünü bulmak her zaman çok kolay değildir. Pratik olarak yeniden ağırlaklandırılmış EKK (Holland ve Welsch 1977) ya da H-algoritması diye adlandırılan (Huber ve Dutter 1974, Marazi 1980) yöntemlerine dayanan planlar tekrar kullanılır. (2.7) ya da (2.9) un aksine (2.11) in çözümü y ekseninin büyütülmesine nazaran uygun bir şekle dönüştürülemez. Bu nedenle σ nın belirli bir tahmini ile rezidüleri standartlaştırmak zorundayız. Yani

$$\sum_{i=1}^{n} \psi(r_i / \hat{\sigma}) x_i = 0 \qquad (2.13)$$

yazılmalıdır. Burada $\hat{\sigma}$ eş zamanlı tahmin edilmelidir.

Minimax asimtotik varyans tartışmalarına neden olduğu için Huber şu fonksiyonu kullanmayı ileri sürdü:

$$\psi(t) = \min(c, \max(t, -c)) \qquad (2.14)$$

(2.14) 'lü M-tahmin edicileri istatistiksel açıdan L_1 regresyonundan daha etkilidir. (Normal dağılımlı bir modelde) Aynı zamanda uzaktaki y_i lere göre hala daha robusttır. Fakat kırılma noktaları uzaktaki x_i nin etkisi yüzünden hala 1/n dir.

Kaldıraç noktalarının bu yaralanabilirliği yüzünden genelleştirilmiş M-tahmin edicileri ileri sürüldü. .Mallows (1975) sunduğu bu tahmin edici belirli bir ağırlık fonksiyonu aracılığıyla uzaktaki x_i sapanlarının etkisini sınırlandırmak esas amacıdır. Mallows (2.13) yerine şu eşitliği kullandı:

$$\sum_{i=1}^{n} w(x_i) \psi(r_i / \hat{\sigma}) x_i = 0 \qquad (2.15)$$

Schwepp (Hill 1977 e bakınız.), Mallows un aksine şunu kullandı:

$$\sum_{i=1}^{n} w(x_i) \psi(r_i / w(x_i) \hat{\sigma}) x_i = 0 \qquad (2.16)$$

Bu tahmin ediciler bir tek sapan gözlemin etkisini sınırlandırmak umuduyla yapıldı. Bunların etkisi **etki fonksiyonu** (Hampel 1974) diye adlandırılan fonksiyonlar aracılığıyla ölçülebilir.

Böyle bir kritere dayanarak ψ ve w nun en iyi seçimlerini 1978 de Hampel, 1980 de Krasker, 1982 de Krasker ve Welsch, 1985 te Ronchetti ve Rousseeuw ve yine 1985 de Samarrow, 1986 da Hampel in son araştırması yapıldı. Bu nedenle karşılık gelen GM (Genelleştirilmiş M) tahmin edicileri genel olarak şu an sınırlandırılmış etki tahmin edicileri olarak adlandırılacaktır. Buradan şu sonuç çıkar:

Bütün GM tahmin edicilerinin kırılma noktası p nin bir fonksiyonu gibi azalan bir değerden daha iyi olmayabilir. Burada p yine regresyon katsayılarının sayısıdır. Bu çok yeterli değildir. Çünkü bu artan boyutla birlikte kırılma noktasının azalması anlamına gelir. Burada sapanların meydana gelmesi için daha fazla fırsat vardır. Ayrıca Maronna – Bustos – Yohai üst sınırına ulaşıp ulaşılmadığı açık değildir. Eğer ulaşılırsa GM tahmin edicisinin bu amacı başarmış olabileceği konusunda açık değildir. Bölüm 3 ün 6. kısmında basit regresyon durumunda (p=2) küçük bir kıyaslama çalışması gösterilecektir. Bu çalışma aynı kırılma noktasına ulaşmış bütün GM tahmin edicilerini göstermeyecektir. Fakat elbette ki esas problem daha yüksek boyutlulardadır.

Çeşitli başka tahmin ediciler önerildi. Mesela Wald (1940) ın metodu, Nair ve Shrivastava (1942), Bartlett (1949), ve Brown ve Mood (1951) ın metotları; eş yönlü eğimlerin medyanı (Theil 1950, Adichie 1967, Sen 1968); dayanıklı doğru (Tukey 1970/1971, Velleman ve Hooglin 1981, Johnstone ve Velleman 1985b); R tahmin edicileri (Jureckova 1971, Jaeckel 1972); L tahmin edicileri (Bickel 1973, Koenker ve Bassett 1978; Andrews 1974 ün yöntemi). Maalesef basit regresyonda bu noktaların hiçbiri %30 luk kırılma noktasına ulaşamadı. Üstelik onların çoğu p>2 için bile tanımlanamadı. Bütün bunlar yüksek kırılma noktalı robust regresyonunun tamamen mümkün olup olmadığı hakkındaki soruları arttırdı. Bunun doğrulayıcı cevabı Siegel(1982) tarafından verildi. Siegel %50 kırılma noktalı tekrarlı medyan tahmin edicisini ileri sürdüler. Gerçekten %50 beklenebilecek en iyi kırılma noktasıdır. (Bozulmanın daha geniş miktarları için örneğin iyi ve kötü kısmını ayırt etmek için mümkün olmaz. Siegel'in tahmin edicisi aşağıdaki gibi tanımlanabilir:
Herhangi p gözlem için $(x_{i1}, y_{i1}),..., (x_{ip}, y_{ip})$ olsun. Bu noktalara tam olarak uyan parametre vektörü hesaplanır. Bu vektörün j-inci koordinatı

$$\hat{\theta}_j = \underset{i_1}{med}(...(\underset{i_{p-1}}{med}(\underset{i_p}{med}\theta_j(i_1,...,i_p)))...) \qquad (2.17)$$

dır. Bu tahmin edici kolayca hesaplanabilir fakat p gözlemin bütün altkümelerinin sayısı istenir. Bu ise çok fazla zaman alır. Küçük p-li problemlere başarılı bir

biçimde uygulanır. Fakat diğer regresyon tahmin edicilerinin aksine koordinat yönündeki yapısı sayesinde tekrarlı medyan x_i nin lineer dönüşümleri için eşdeğişkenli (equivariant) değildir. Yüksek kırılmalı regresyon yöntemlerini ve equivariantları düşünelim. Bunları tanımak için (1.7) ye dönelim. EKK yönteminin tam ismi karelerin en küçük toplamıdır.Fakat açık olarak birkaç kişi n pozitif sayı ile yapılacak tek makul şeyin onları eklemek olacakmış gibi toplam kelimesinin silinmesine karşılar.Belki de tarihsel isminin bir sonucu olarak birkaç kişi bu tahmin edicinin kare ya da başka bir şey yerine robust yapmaya çalışıyor.Medyan ile toplamın yerini değiştirerek daha robust yapabiliriz. Yani;

$$\min_{\hat{\theta}} med_i r_i^2 \tag{2.18}$$

dır. Buna en küçük kareler medyanı (the least median of squares) tahmin edicisi denir. Bunu Rousseeuw 1984 te buldu. Bu öneri aslında Hampel ın(1975,s.380) fikrine dayanıyordu. Bu tahmin edicinin x sapanlarına olduğu kadar y deki sapanlara karşı da çok sağlam olduğu ortaya çıkar. LMS (2.15) rezidüleri kullandığı için açıklayıcı değişkenler üzerinde lineer dönüşümlere karşı equivarianttır. Maalesef LMS asimtotik etkinin görünüşündeki noktadan yetersiz bir biçimde uygulanır.

Bunu tekrar etmek için Rousseeuw(1983,1984) en küçük budanmış kareler(the least trimmed squares) yöntemini sundu. Bu durum;

$$\min_{\hat{\theta}} \sum_{i=1}^{n} (r^2)_{i:n} \tag{2.19}$$

eşitliğiyle verilir. Burada

$$(r^2)_{1:n} \leq \ldots \leq (r^2)_{n:n}$$

karesi alınarak sıralanmış rezidülerdir. (2.16) formülü EKK ya çok benzer. Tek farkı sapanlardan uzakta kalması için modele uymasına izin vermesi suretiyle en geniş karesi alınmış rezidülerin toplamda kullanılmamasıdır. LMS gibi bu tahmin edici de x_i nin lineer dönüşümleri için uygun dönüşümlüdür ve izdüşüm takibine bağlıdır. En

sağlam oranlar $h \to n/2$ olduğunda başarıldı.Bu oran da kırılma noktasının %50 ye ulaşması durumudur.

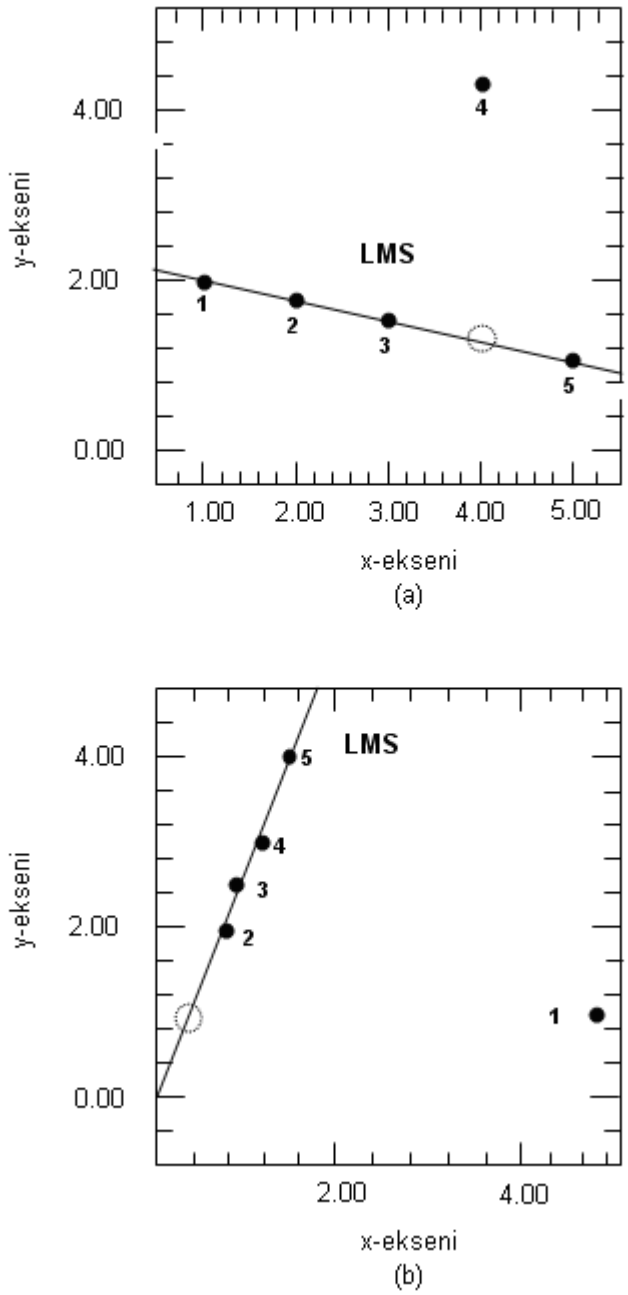
LMS ve LTS nin her ikisi de rezidülerin saçılımının robust ölçümünü minimize ederek tanımlanır. Bunu genelleştirirsek Rousseeuw ve Yohai (1984) S-tahmin edicisi diye adlandırılan

$$\min_{\hat{\theta}} S(\theta) \tag{2.20}$$

ile ifade edilen tahmin ediciyi buldular. Burada $S(\theta)$ ifadesi, $r_1(\theta),..., r_n(\theta)$ rezidülerinin ölçümünün robust M-tahmininin belli bir çeşididir. Burada karışık sabitlerin uygun bir seçimiyle kırılma noktası %50 ye de ulaşabilir. Bundan başka S-tahmin edicilerinin aslında M-tahmin edicileri gibi aynı asimtotik performansa sahip olduğu ortaya çıkar.

Şekil2.6 bu yeni tahmin edicilerin y de ve x deki sapanlara karşı sağlamlıklarını gösteriyor. Yüksek kırılma noktaları yüzünden bu tahmin ediciler aynı zamanda birkaç sapanla bile başa çıkabilir. Bu dayanıklılık açıklayıcı değişken sayısı olan p den de bağımsızdır. Bu nedenle LMS ve LTS çok değişkenli durumlarda regresyon sapanlarını keşfetmek için kullanılabilen analitik araçların verisinin güvenilir olması gerekir. LMS ve LTS 'nin esas kuralı robust modelinden uzaklaşan noktalar olarak sapanları belirleyebildikten yani pozitif ya da negatif rezidü durumlarını belirleyebildikten sonra verinin çoğunluğunun modellenmesidir. Şekil2.6a da 4.durum önemli bir rezidüye sahiptir. Hatta Şekil2.6b nin birinci durumunda daha da açık görünüyor.

Fakat genelde y_i ve dolayısıyla rezidüler ölçümün herhangi bir biriminde olabilir. Bu nedenle r_i rezidüsünün geniş olup olmadığına karar vermek amacıyla hata ölçeğinin $\hat{\sigma}$ tahmini ile onu kıyaslamamız gerekir. Elbette ki bu $\hat{\sigma}$ ölçüm tahmini kendi içinde sağlam olmalıdır. Bu nedenle sadece iyi bir veriye bağlıdır ve sapan değer(değerler) tarafından yukarıya çıkarılmaz. LMS regresyonu için



Şekil 2.6. LMS regresyonunun y-yönünde (a) 'da y-yönündeki sapana ve (b) 'de x-yönündeki sapana karşı dayanıklılığı.

$$\hat{\sigma} = C_1 \sqrt{\underset{i}{med}\, r_i^2} \tag{2.21}$$

tahmin edicisini kullanabiliriz. Burada r_i LMS modeline göre i-inci durumun rezidüsüdür. C_1 sabiti sadece normal dağılımda uygunluğa ulaşmak için kullanılan bir faktördür. Bunlar küçük örnekler için bir düzeltme yapar. LTS için

$$\hat{\sigma} = C_2 \sqrt{\frac{1}{n}\sum_{i=1}^{h} (r^2)_{i:n}} \tag{2.22}$$

gibi bir kural kullanabiliriz. C_2 ise başka bir düzeltme faktörüdür. Her durumda i-inci durumu ancak ve ancak $|r_i / \hat{\sigma}|$ geniş ise bir sapan olarak tanımlanır. (Şunu belirtelim ki bu oran orijinal ölçüm birimine bağlı değildir.) Bu bizi başka bir fikre getirir. İşlenmemiş LMS ve LTS çözümlerini geliştirmek amacıyla ve t – değerleri, güven aralıkları, … vb standart nicelikleri hesaplamak amacıyla sapanların belirlenmesine dayanan ağırlıklandırılmış en küçük kareler analizine başvurabiliriz. Örneğin ;

$$w_i = \begin{cases} 1, & |r_i / \hat{\sigma}| \le 2.5 \\ 0 & |r_i / \hat{\sigma}| \ge 2.5 \end{cases} \tag{2.23}$$

ağırlığını kullanabiliriz. Yani i – inci durumun LMS rezidüsü orta küçüklükte ise ağırlıklandırılmış EKK da tutulacaktır fakat eğer bir sapansa ihmal edilecektir. 2.5 sınırı, tabiî ki keyfi olarak, 2.5 $\hat{\sigma}$ den daha geniş birkaç rezidünün olacağı normal dağılım durumunda oldukça makuldur. (2.20.) deki gibi sapanların reddedilmesinin zorluğu yerine direkt redde de başvurulabilir. Örneğin çift bölgeli izin verildiği suretiyle $|r_i / \hat{\sigma}|$ nın sürekli fonksiyonu kullanılarak $2 \le |r_i / \hat{\sigma}| \le 3$ li noktaların 0 ile 1 arasında ağırlıkları verilebilir.

Nasıl olursa olsun aşağıdaki gibi tanımlanan ağırlıklandırılmış en küçük karelere başvuracağız:

$$\min_{\hat{\theta}} \sum_{i=1}^{n} w_i r_i^2 \tag{2.24}$$

Bu ifade hızlı bir biçimde hesaplanabilir. Sonuçlanmış tahmin edici normal dağılım altında hala yüksek kırılma noktasına sahiptir fakat istatistiksel anlamda daha etkilidir ve regresyon katsayılarının varyans-kovaryans matrisi ve belirleyicilik katsayısı R^2 gibi en küçük karelerle çalışıldığında aşikar olunan bütün standart çıktıyı verir.

Sonraki bölüm basit regresyondaki robust yöntemlerinin uygulaması işlenecektir. EKK, LMS ve yeniden ağırlıklandırılmış EKK açıklanacaktır

3. BASİT REGRESYON

3.1. Giriş

Basit regresyon modeli

$$y_i = \theta_1 x_i + \theta_2 + e_i \qquad \text{i=1,…,n} \qquad (3.1)$$

İki boyutlu veri kümesini bölüm1'de bazı problemlerde gördük. Bazı saçılım grafikleri yardımıyla ekk üzerinde x ve y yönünde sapanların etkilerini gösterdik. Bu bölümde bu problemlerle başa çıkabileceğimiz yüksek kırılmalı regresyon tekniklerine başvuracağız.

Basit regresyon ifadesi bazen de sabit terim olmadan yani;

$$y_i = \theta_1 x_i + e_i \qquad \text{i=1,…,n} \qquad (3.2)$$

olarak da kullanılıyor. Bu model açıklayıcı değişken sıfırken yanıt değişkenin sıfır olması gerektiği varsayımının olduğu sıradan uygulamalarda kullanılabilir. Grafiği çizilecek olursa orjinden geçen bir doğrudur.

Şimdi sağlam regresyon tekniğine olan ihtiyacı göstermek amacıyla Tablo3.1'de verilen Pilot-Plant (Daniel ve Wood 1971) verisini inceleyelim:

Burada yanıt değişken titrasyonla belirlenen asit miktarı, açıklayıcı değişken ise çekme ve tartma ile belirlenen organik asit miktarıdır. Yale ve Farsythe (1976) da bu veri kümesini analiz ettiler ve saçılım grafiğinin açıklayıcı değişken ve yanıt değişken arasında çok güçlü bir istatistiksel ilişki olduğunu gördüler.

Verinin EKK ile çözümünden

$$\hat{y} = 0{,}322x + 35{,}458$$

bulunur. En küçük kareler meydanından (LMS) ise

$$\hat{y} = 0{,}311x + 36{,}519$$

bulunur.

Tablo 3.1. Pilot-Plant Verisi

Gözlem (i)	Çekme (x_1)	Titrasyon (y_i)
1	123	76
2	109	70
3	62	55
4	104	71
5	57	55
6	37	48
7	44	50
8	100	66
9	16	41
10	28	43
11	138	82
12	105	68
13	159	88
14	75	58
15	88	64
16	164	88
17	169	89
18	167	88
19	149	84
20	167	88

Kaynak: Daniel ve Wood(1971).

dır. Grafikte görüldüğü gibi örnekte sapan değer yoktur. Böyle bir olayda umulacağı gibi sadece EKK ve sağlam tahmin ediciler arasında marjinal farklar olduğudur

Kabul edelim ki bir değer yanlış girilsin örneğin 6. x-gözleminde 37 yerine 370 alınsın. Bu hata x-yönünde bir sapana neden olur. O zaman modelleri tekrar oluşturduğumuzda ekk modeli;

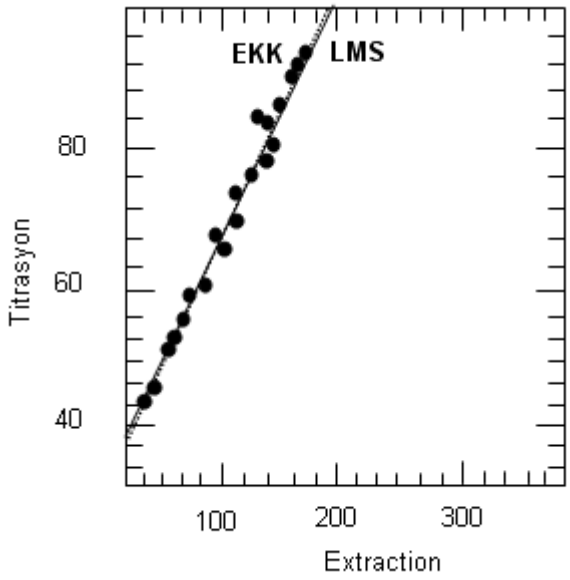
$$\hat{y} = 0{,}081x + 58{,}939$$

olur.En küçük kareler medyanı ise;

$$\hat{y} = 0{,}314x + 36{,}343$$

biçiminde elde edilir. Buradan da görüldüğü gibi EKK yöntemi tek bir sapandan oldukça etkilenirken sağlam model olan LMS, verinin büyük çoğunluğuna iyi bir

model olarak sapandan uzak durmayı başardı. Ayrıca LMS doğrusu, bozulmamış veriye



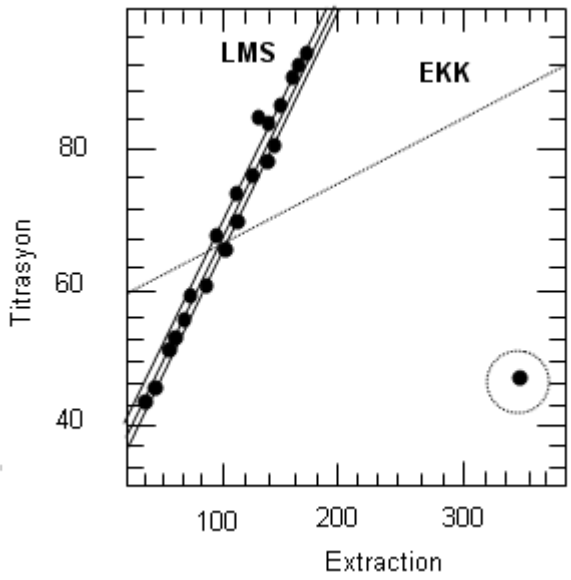
Şekil 3.1. EKK (kesikli doğru) ve LMS (düz doğru) modeliyle Pilot-Plant verisi

EKK tahminini uyguladığındaki doğruya yaklaştı. Robust teknikleri sapanları ihmal eder deyimi yanlış olduğunu ortaya koydu. Aksine LMS böyle noktaların varlığını ortaya çıkarır.

Kesenli basit regresyonun LMS çözümü ;

$$\min_{\hat{\theta}_1,\hat{\theta}_2} \underset{i}{med}(y_i - \hat{\theta}_1 x_i - \hat{\theta}_2)^2 \tag{3.3}$$

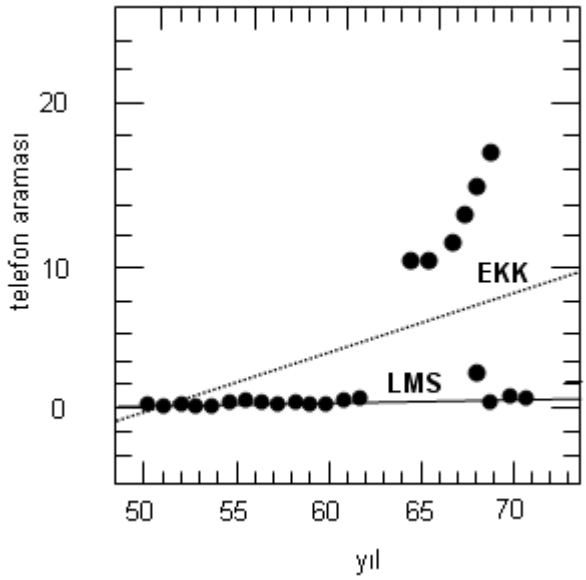
şeklinde olur. Geometrik olarak bu gözlemlerin yarısını içine alan en dar şeridi bulmayı ifade eder. (Buradaki yarısı kavramından kasıt [n/2]+1'in tam kısmıdır. Bununla beraber



Şekil 3.2. Şekil 1 'deki ile aynı veri fakat tek sapanı var. Kesikli doğru EKK modeline karşılık gelir. Düz çizgili doğru ise noktaların yarısını içeren en dar şeritle çevrilir.

bu şeridin kalınlığı düşey yönde ölçülebilir.) LMS doğrusu bu bandın tam ortasında uzanır. Şunu ifade etmeliyiz ki bu notasyonu insanlara anlatmak EKK'nin tanımından daha kolaydır aslında. Bozulmuş Pilot-Plant verisi için bu şerit Şekil 3.2 de çizilmiştir.

Bu örnekteki sapan yapaydır fakat bu tür hatalarla gerçek verilerde sık sık karşılaştığımızı fark etmek önemlidir. Veri noktalarının sapması; gözlemleri kaydederkenki hatalar, aktarmada ya da tanımlamadaki hatalar ya da incelenen olayda hariç tutulan olaylar yüzünden örneklemde mevcut olabilir. 2 boyutlu durumda(yukarıdaki örnekte olduğu gibi sadece gözlemleri grafikte çizerek farklı(alışılmamış) değerleri ortaya çıkarmak oldukça kolaydır. Bu görüşü izlemek daha yüksek boyutlulara gelindiğinde artık mümkün olmaz. Bu nedenle pratikte sapanların etkisini azaltabilen bir yönteme ihtiyaç vardır ve dolayısıyla rezidü grafiklerinde sapanları ortaya çıkar. Buna ek olarak sapan meydana gelmediği zaman alternatif yöntemin sonucu EKK çözümünden nadiren farklı olur. Buradan LMS regresyonunun bunların gerekliliğini karşılayacağı ortaya çıkar.



Şekil 3.3. Belçika'da 1950-1973 yılları arasındaki uluslar arası telefon aramaları sayısının EKK (kesikli doğru) ve LMS modeli (düz doğru)

Şimdi de sapan değerin olduğu bazı gerçek veri örneklerine bakalım. Ekonomi Bakanlığı tarafından yayınlanan Belçika istatistiksel incelemesinde uluslar arası yapılan telefon görüşmelerinin toplam sayısını içeren bir veri kümesi bulduk.

Grafik yıllar arttıkça arama sayısının arttığını göstermektedir. Fakat bu zaman serileri 1964 den 1969'a yoğun bir kirlenme içermektedir. Araştırma üzerine 1964 ten 1969 a kadar arama dakikalarının toplam sayısını veren başka bir kayıt sistemi kullanıldığı ortaya çıktı.(1963 ve 1970 yıllarında da kısmen etkilenildi çünkü geçişler yeni yılda tam olarak olmadı.Bu nedenle bazı aylardaki arama sayısı kalan aylardaki dakika sayısına eklendi.)Bu ise y- yönünde sapanların geniş bir kırılmasına sebep oldu.Bu veri tablo 2 de listelenmiş ve şekil 3 'te gösterilmiştir.

Bu veri için EKK çözümü

$$\hat{y} = 0.504x - 26.01$$

dır. Bu modelin grafiği ise Şekil 3.3 te kesikli noktalarla gösterilmiştir. Bu kesikli doğru 1964-1969 yılları arasındaki y değerlerinden çok fazla etkilenmiş. Sonuç

Tablo 3.2. Belçika'dan Yapılan Uluslararası Arama Sayısı

Yıl (x_i)	Arama Sayısı (y_i)
50	0.44
51	0.47
52	0.47
53	0.59
54	0.66
55	0.73
56	0.81
57	0.88
58	1.06
59	1.20
60	1.35
61	1.49
62	1.61
63	2.12
64	11.90
65	12.40
66	14.20
67	15.90
68	18.20
69	21.20
70	4.30
71	2.40
72	2.70
73	2.90

Milyon yerine on sayısı kullanılmış.

olarak EKK doğrusu geniş bir eğime sahiptir ve iyi ya da kötü noktalara uymuyor. Bu ise verilere ciddi bir şekilde bakmayarak ve rutin bir şekilde EKK metoduna başvurarak elde edilebilecek bir yoldur.Aslında iyi gözlemlerin bazıları(1972 gibi) kötü değerlerin bazılarının EKK rezidülerinden daha geniş olur. Şimdi de LMS yöntemini uygulayalım:O zaman model;

$$\hat{y} = 0.115x - 5.610$$

olur. (Şekil 3.3 te tam çizgiyle gösterilmiştir.) Görüldüğü gibi sapanlardan uzak durmuştur. Basitçe grafikteki veri noktalarına bakıldığında model ortaya çıktığı görülen örneğe denk gelir. Açıkça bu doğru verinin çoğunluğuna uyar.(Bu bir lineer

modelin ister istemez en iyi model olacağını ima etmez. Çünkü daha fazla veri toplamak ilişkinin daha karmaşık türünü ortaya çıkarabilir.)

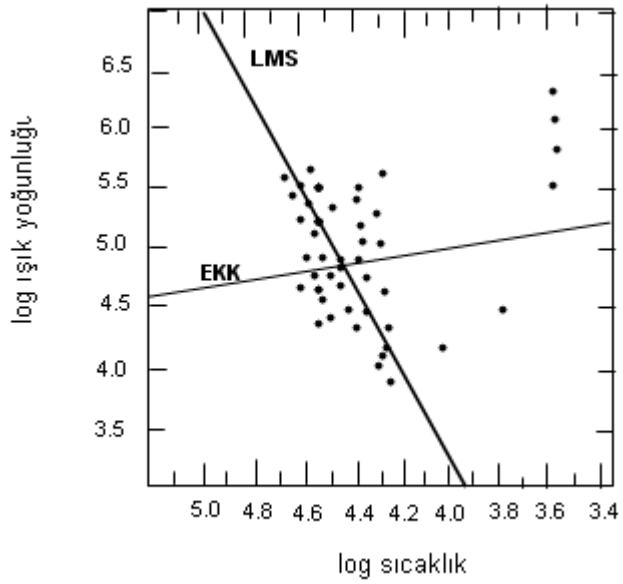
Tablo 3.3. CYG OB1 Yıldız Kümesinin Hertzprung – Russell Diyagramının Verileri

Yıldız İndeksi	$\log T_e$ (x_i)	$\log \lvert L/L_0 \rvert$ (y_i)	Yıldız İndeksi	$\log T_e$ (x_i)	$\log \lvert L/L_0 \rvert$ (y_i)
1.	4.37	5.23	25	4,38	5,02
2.	4.56	5.74	26	4,42	4,66
3.	4.26	4.93	27	4,29	4,66
4.	4.56	5.74	28	4,39	4,90
5.	4.30	5.19	29	4,22	4,39
6.	4.46	5.46	30	3,48	6,05
7.	3.84	4.65	31	4,38	4,42
8.	4.57	5.27	32	4,56	5,10
9.	4.26	5.57	33	4,45	5,22
10.	4.37	5.12	34	3,49	6,29
11.	3.49	5.73	35	4,23	4,34
12.	4.43	5.45	36	4,62	5,62
13.	4.48	5.42	37	4,53	5,10
14.	4.01	4.05	38	4,45	5,22
15.	4.29	4.26	39	4,53	5,18
16.	4.42	4.58	40	4,43	5,57
17.	4.23	3.94	41	4,38	4,62
18.	4.42	4.18	42	4,45	5,06
19.	4.23	4.18	43	4,50	5,34
20.	3.49	5.89	44	4,45	5,34
21.	4.29	4.38	45	4,55	5,54
22.	4.29	4.22	46	4,45	4,98
23.	4.42	4.42	47	4,42	4,50
24.	4.49	4.85			

Diğer bir örnek astronomiden. CYG OB1 yıldız sınıfının Hertzsprung-Russell diyagramındaki veri Tablo 3.3'te veriliyor. Bu yıldız sınıfı Cygnus yönünde 47 yıldız içeriyor. Burada x yıldız yüzeyindeki sıcaklık etkisinin (T_e) logaritması, y ise yıldızın ışık şiddetinin(L/L_o)logaritmasıdır. Bu sayılar bize Humphreys(1978)den

ham veri alanı ve Vansina ve De Greve(1982)'ye göre işleyen. C.Doom(kişisel konuşmayla) tarafından verildi.

Hertzsprung-Russell diyagramı Şekil 3.4'te gösterilmiştir. Sağdan sola doğru grafiklenmiş x logaritma sıcaklık değerinin olduğu bu noktaların dağılımıdır.Grafikte noktaların iki grubu görülüyor.Verinin çoğunluğu iniş aşağı bir yokuş oluşturarak kümelenmiş,4 yıldız ise yukarda sağ köşede kalmış.Diyagramın bu kısımları astronomide iyi bilinir: 43 yıldız esas sırada uzandığı söylenir bunun aksine kalan 4 yıldız ise devler olarak adlandırılır.(Devler 11,20,30 ve 34 indeksli noktalardır.)



Şekil 3.4. CYG OB1 yıldız kümesinin Hertsprung-Russell diyagramının EKK (kesikli doğru) ve LMS (düz doğru) modeli.

Bu veri kümesiyle LMS tahmin edicisine başvurmak esas diziye iyi bir biçimde uyan

$$\hat{y} = 3.898x - 12.2989$$

modelini verir(Düz çizgiyle gösterilmiştir.)Diğer taraftan EKK çözümü

$$\hat{y} = -0.409x + 6.78$$

olur. Bu da 4 dev yıldızla çekilen(bu 4 yıldız modele iyi uymaz) şekil4'teki kesikli çizgiyle gösterilen doğruya tekabül eder. Bu sapanlar kaldıraç noktasıdır ama onlar hata değildir. Daha özel bir ifadeyle açıklarsak eğer veriler iki farklı popülasyondan geliyor. Bu iki grup LMS rezidüleri temelinde kolayca ayrılabilir.(Geniş rezidüler dev yıldızlara tekabül eder.)Aksine EKK rezidüleri oldukça homojen ve devleri ana diziden ayırmamız için bize izin vermiyor.

3.2. En Küçük Medyan Kareler Doğrusunun Hesaplanması

Bilgisayar desteği olmaksızın yüksek kırılmalı regresyon tahminlerini hesaplamak asla mümkün olmaz. Gerçekten LMS, EKK'de kullanıldığı gibi açık bir formüle sahip değil. Yüksek kırılmalı regresyonun neden ucuz bir biçimde hesaplanamadığının derin sebepleri görülür.

Yukarıdaki örnekte, veri 2 boyutluydu bu nedenle grafiklendirilebildi. Saçılım grafiğindeki bir nokta bir rakamla gösterilir. Bu rakam, yaklaşık olarak aynı koordinatlara sahip noktaların sayısına denk gelir. 9 dan fazla rakam çakıştığında, (*) işareti yazar. Basit regresyonda, EKK tahmin edicileri üzerinde oldukça güçlü bir etkiye sahip olabilen noktalar böyle bir grafikte hemen ortaya çıkar.

O zaman her j-inci değişkenin m_i medyanı hesaplanır. Yukarıdaki Pilot-Plant çıktısında gösterildiği gibi medyan extraction değeri 107 ve medyan titrasyon değeri 69'a eşittir. Diğer / sonraki doğruda her değişkenin standart sapmasının bir robust versiyonu gibi düşünülebilen s_i değerini verir. Standartlaştırılmış gözlemlerin bir listesini alalım. Yani her x_{ij} ile

$$\frac{x_{ij} - m_j}{s_j} \tag{3.4}$$

nin yerlerini değiştilerecek (yanıt değişken aynı yöntemle standartlaştırılacak.) Sonuç tablosunun sütunlarının her biri 1 varyansına ve 0 medyanına sahiptir. Bu, herhangi tek bir değişkende sapanları ortaya çıkarmamızı mümkün kılar. Gerçekten, standartlaştırılmış gözlemin mutlak değeri genişse (yani 2.5 dan fazla ise) o zaman bu değişken için bir sapandır. Pilot-Plant çıktısında biz bu şekilde olay 6 nın geniş

exraction ölçümüne sahip olduğunu keşfettik (satandartlaştırılmış değeri 3,73 tü. Bu da sıradan bir durum değil. Bazen standartlaştırılmış değerler, bir sütundaki dağılımın çarpıklığını söyleyebilir Şimdi, parametrik olmayan (non parametrik) sperman korelasyon katsayıları gibi değişkenler arasındaki Pearson (product moment) korelasyon katsayılarını verdi.

Robust tahminlerini vermeden önce EKK sonuçlarıyla başlayalım. Standart hataların yer aldığı tahmin edilmiş katsayılarla bir tablo yazılsın. J-inci EKK regresyon katsayısının standart hatası $\sigma^2(X'X)^{-1}$ ile bulunan EKK regresyon katsayılarının varyans kovaryans matrisin j-inci diagonal elementinin kareköküdür. Burada X matrisi

$$X=\begin{bmatrix} x_{11}.....x_{1p} \\ . \quad\quad . \\ x_{i1}.....x_{ip} \\ . \quad\quad . \\ x_{n1}.....x_{np} \end{bmatrix}$$

X'in i-inci satırı, i-inci gözlemdeki p tane açıklayıcı değişkenden oluşan x_i vektörüne eşittir. Bilinmeyen σ^2,

$$S^2=\frac{1}{n-p}\sum_{i=1}^{n} r_i^2 \tag{3.5}$$

ile tahmin edilir. Burada r_i i-inci rezidüdür. Tahmin edilmiş $S^2(X'X)^{-1}$ varyans-kovaryans matrisi de hesaplanabilir. Basit regresyon için kesen terim olmadığında p=1,diğer durumlarda p=2 dir. Ölçek tahminiyle gösterilen miktar $S=\sqrt{S^2}$ dir.

θ_j parametreleri için güven aralıkları oluşturmak için hata terimlerinin 0 ortalamalı σ^2 varyanslı bağımsız normal dağılıma sahip olduğu varsayılmalıdır. Bu koşullar altında n-p serbestlik dereceli Student dağılımına sahip

$$\frac{\hat{\theta}_j - \theta_j}{\sqrt{S^2((X'X)^{-1})_{JJ}}} \quad \text{j=1,…,p} \tag{3.6}$$

elde edilir.Miktarların her biri iyi bilinir. $t_{n-p,1-n/2}$ ile bu dağılımın 1-$\alpha/2$ miktarını ifade edelim. O zaman θ_j gibi herhangi bir regresyon katsayısının önemlilik testi için de kullanılabilir.

$$\hat{\theta}_j - t_{n-p,1-\alpha/2}\sqrt{s^2((x'x)^{-1})_{jj}}, \hat{\theta}_j + t_{n-p,1-\alpha/2}\sqrt{s^2((x'x)^{-1})_{jj}} \tag{3.7}$$

Aynı sonuç, $\hat{\theta}_j$ gibi herhangi bir regresyon katsayısının önemlilik (significence) testi için de kullanılabilir.

$H_0 : \theta_j = 0$ (sıfır hipotezi)

$H_1 : \theta_j \neq 0$ (alternatif hipotez) (3.8)

Böyle bir test j – inci değişkenin modelden silinip silinemeyeceğini açıklamaya yardımcı olabilir. Eğer α nın belirli bir değeri için (3.5) teki sıfır hipotezi kabul edilebilire bu durum j – inci yanıt değişkenin açıklayıcı değişkene çok katkısı olmadığını gösterir. Bu hipotez için test istatistiği

$$\frac{\hat{\theta}_j}{\sqrt{s^2((x'x)^{-1})_{jj}}}, \qquad \text{j = 1,…,p} \tag{3.9}$$

dır. Bu oran çıktıda t değerine karşılık gelir. Sıfır hipotezi t nin mutlak değeri $t_{n-p,1-\alpha/2}$ den büyük olduğunda α seviyesinde reddedilir. Bu durumda j-inci regresyon katsayısının önemli bir biçimde 0 dan farklı olduğu söylenir.

Şimdiki örnekte n=20 ve p=2 dir. α = %5 olduğu için n-p=18 serbestlik dereceli student dağılımını %97,5 i 2,101 e eşittir. Bu yüzden EKK eğimi ortak t değeri 1.719 olduğu için önemli bir farkı olmadığı sonucu çıkarılabilir. Diğer taraftan kesen t değeri 8,911 olup oldukça önemlidir. p değeri gerçekte elde edilen t

değerinden daha büyük olan n-p serbestlik dereceli student dağılımlı rastgele değişkeninin olasılığıdır. p değeri 0,05 ten küçük olduğunda karşılık gelen regresyon katsayısı %5 anlamlılık seviyesinde önemlidir. Yukarıdaki çıktıda EKK eğiminin p değeri 0,10274 e eşittir. Bu nedenle önemli değildir. Bunun aksine EKK keseni % 0,1 anlamlılık seviyesinde bile 0,000 olup modeli çok anlamlı yapar. Yazılmış p değerlerinin hipotez testlerini yapmak çok kolaydır. Çünkü olasılık tabloları artık gerekli değildir. Bununla birlikte (3.6) tamamen robust olmadığı için dikkatli biçimde kullanılmalıdır.

Bu istatistiğin dağılım teoremi genelde sadece hataların takip edildiği ve pratikte nadiren karşılaşılan normal dağılımda korunur. Bu nedenle sapanların varlığının mümkün olduğunun farkında olduğumuz için robust modelinde ilk bakışta görülür. Veri kümesi etkili noktaları içerdiği görüldüğünde aşağıda verilecek olan RLS çözümünün t değerine gidilmesi gerekir.

Yanıt değişken ile açıklayıcı değişken(ler) arasındaki lineer ilişkinin dayanıklılığıyla ilgili bir fikir elde etmek için belirleyicilik katsayısı (R^2) çıktıda görülür. R^2 regresyon tarafından açıklanan toplam çeşitliliğin oranını ölçer. Tam bir formül için sabit terimli regresyon ile sabit terimsiz regresyonu ayırmamız gerek

EKK regresyonu için R^2 aşağıdaki gibi hesaplanır:

$$R^2 = 1 - \frac{SSE}{SST}$$, sabit terimsiz modelde

$$R^2 = 1 - \frac{SSE}{SST_m}$$, sabit terimli modelde

Burada;

SSE=Rezidü hatalarının kareler toplamı

$$= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

SST=Bütün kareler toplamı

$$=\sum_{i=1}^{n} y_i^2$$

ve

SST_m = Ortalamaya göre bütün düzeltilmiş kareler toplamı

$$=\sum_{i=1}^{n} (y_i - \bar{y})^2$$

Burada

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

dir. Sabit terimli basit regresyon durumunda belirleyicilik katsayı R^2 yi açıklayan x ve y arasındaki Pearson korelasyon katsayısının karesine eşittir.

Pilot-plant verisi için EKK ya karşılık gelen R^2 değeri 0,141 e eşittir. Bunun anlamı basit regresyon modelinde y nin sadece %14,1 i açıklandı. Buradan R^2 nin sıfıra eşit olması hipotezini test etmek de düşünülebilir. Formüle edersek;

$$H_0 : R^2 = 0 \qquad \text{(sıfır hipotezi)}$$

$$H_1 : R^2 \neq 0 \qquad \text{(alternatif hipotezi)} \qquad (3.10)$$

olur. Bu hipotezler sabit terimin olmadığı model hariç bütün regresyon katsayılarının vektörünün sıfıra eşit olup olmadığı testine denktir. Yani (3.10) aşağıdakine denktir:

H_0 = Bütün kesensiz θ_j lerin hepsi sıfırdır.

$H_1 = H_0$ doğru değildir. (3.11)

Burada katsayılar birlikte düşünüldüğünde tek bir regresyon katsayısı üzerinde yukarıdaki t testinden oldukça farklıdır. Eğer e_i normal dağılımlı ise o zaman aşağıdaki istatistikler F dağılımına sahiptir:

$$\frac{R^2/(p-1)}{(1-R^2)/(n-p)} = \frac{(SST_m - SSE)/(p-1)}{SSE/(n-p)} \sim F_{p-1,n-p},$$ sabit terimli regresyon için

$$\frac{R^2/p}{(1-R^2)/(n-p)} = \frac{(SST - SSE)/p}{SSE/(n-p)} \sim F_{p,n-p},$$ diğer durumlarda

Eğer uygun istatistiğin hesaplanmış değeri ortak F değerinin (1-α)-ıncı sıklık derecesinden daha az ise o zaman H_0 kabul edilebilir. Eğer değilse H_0 reddedilebilir. p değeri F değerini aşmış uygun serbestlik dereceli bir F dağılımında rasgele değişimin olasılığıdır.

Bozulmuş Pilot-Plant verisi için 1 ve 18 serbestlik dereceli F değerini 2,955 elde ettik. Uygun p değeri 0,103 dür ve bu nedenle %5 seviyesinde anlamlı değildir. Sonuç olarak H_0 reddedilemediği için anlamlılık yönünde açıklayıcı değişkenin yanıt değişkeni gerçekten açıklamadığı EKK tahmin edicilerinden görüldüğü söylenilebilir. t-testi, R^2 ve F-testinin yorumu bölüm4 te görülecek olan çok boyutlu veri kümeleri için hala doğru olacaktır. Şekil 3.2 incelendiğinde durum 6 EKK doğrusunu güçlü bir biçimde çektiği ve s yi de yukarı çektiği için standartlaştırılmış rezidüsü sadece -2.62 dir. Bu değer heyecan verici bir yönde olmadığı hemen hemen açıktır.

LMS regresyonunun sonuçları aşağıdaki gibidir:

İlk başta regresyon parametrelerinin tahminleri ve bununla birlikte karşılık gelen ölçüm tahmini verilir.

Bu ölçüm tahmini de robust şekilde tanımlanır. Bu amaçla ilk ölçüm tahmini s^0 hesaplanır. s^0,n ve p ye bağlı bir düzeltme çarpanının sonlu bir örneklemle çarpılmasından oluşan objektif fonksiyonunun bir değerine dayanır.

$$s^0 = 1{,}4826\left(1+\frac{5}{n-p}\right)\sqrt{\underset{i}{med}\, r_i^2}$$

Bu ölçek tahminiyle standartlaştırılmış r_i / s^0 rezidüleri hesaplanır ve aşağıdaki gibi i-inci gözlemin w_i ağırlığını tanımlamada kullanılır.

$$w_i = \begin{cases} 1 & ,\left|r_i / s^0\right| \le 2{,}5 \\ 0 & ,d.h \end{cases} \tag{3.12}$$

LMS regresyonu için ölçek tahmini o zaman şu şekilde verilir:

$$\sigma^* = \sqrt{\frac{\sum_{i=1}^{n} w_i r_i^2}{\sum_{i=1}^{n} w_i - p}} \tag{3.13}$$

σ^* ın %50lik kırılma noktasına sahip olduğuna dikkat edelim.Yani kirlenmenin %50 den azı için $\sigma^* \to \infty$ veya $\sigma^* \to 0$ olmaz.Şimdiki örnekte σ^*=1.33 dir. Bu değer başlangıçta hesaplanan EKK ölçümü olan s=15,6 dan daha küçüktür.

LMS aynı zamanda y deki gözlenmiş değişimin uygun modeli nasıl iyi açıkladığını tanımlamak için bir ölçüdür. Klasik yönteme benzer şekilde R^2 belirleyicilk katsayısına diyeceğiz. Sabit terimli regresyon durumunda;

$$R^2 = 1 - \left(\frac{med\,|\,r_i\,|}{mad(y_i)}\right)^2 \tag{3.14}$$

sabit (kesen) terim olmadığında ise;

$$R^2 = 1 - \left(\frac{med\,|\,r_i\,|}{med\,|\,y_i\,|}\right)^2 \tag{3.15}$$

dir. Burada "mad=medyanın mutlak sapması (median absolute deviation)" ın kısaltmasıdır ve

$$mad(y_i) = \left| y_i - \underset{i}{med}\, y_i \right| .$$

dir. Pilot-Plant çıktısında robust belirleyicilk katsayısı 0,996 dır.Yani verinin büyük çoğunluğuna oldukça iyi bir model olmuştur. LMS için de gözlenmiş y_i,tahmin edilmiş $\hat{y}_i$,r_i rezidüsü ve r_i / σ^* standartlaştırılmış rezidülü bir tablo verilir.Bu da açıkça gösteriyor ki r_6 / σ^* değeri -78,50 gibi çok büyük bir değere eşit olduğu için durum6 bir sapandır.

Bu çıktının en son kısmı yeniden ağırlıklandırılmış en küçük kareler regresyonu(RLS) ile ilgilidir. Rezidülerin kareleriyle w_i ağırlığının çarpımlarının toplamını minimum yapmaya denktir.

$$\min_{\hat{\theta}} \sum_{i=1}^{n} w_i r_i^2 \tag{3.16}$$

w_i ağırlıkları s^0 yerine σ^* son ölçüm tahminiyle birlikte (3.16) daki LMS çözümünden tanımlanır. Sadece 0 ve 1 değerlerini alabilen bu ağırlıkların etkisi w_i nin sıfıra eşit olduğu durumlardaki silmeye benzerdir. (3.16) ile ilgili ölçek tahmini

$$s = \sqrt{\frac{\sum_{i=1}^{n} w_i r_i^2}{\sum_{i=1}^{n} w_i - p}} \tag{3.17}$$

dir. (Bütün durumlar için w_i=1 yazıldığında sıradan EKK için ölçek tahmini tekrar bulunur.) Bu nedenle RLS sadece sıfırdan farklı bir değer alan öyle gözlemlerden oluşan azaltılmış veri kümeli klasik EKK gibi görülebilir. Bu azaltılmış veri kümesi artık regresyon sapanlarını içermediği için istatistikler ve sonuçlar bütün veri kümesindeki EKK lı birleşimlerden daha güvenilirdir. Ağırlıklar karmaşık bir biçimde veriye bağlı olduğu için vurgulanan dağılım teorisi tamamen gerekli

değildir. Fakat bu dağılım hala ne iyi bir yaklaşım gibi ne de bazı Monte-Carlo denemeleriyle doğrulanmış gibi kullanışlı değildir.

Şimdiki örnekte durum6 hariç bütün w_i ler 1 e eşittir. Burada durum6 gerçekten bir sapandır. RLS ile elde edilen regresyon katsayıları LMS nin ilk basamağında tanımlananlarla güçlü bir benzerlik gösterir. Dikkat edelim ki sapanlar olmadan eğim tahmini sıfırdan önemli derecede farklıdır.

RLS için belirleyicilik katsayısı EKK e benzer bir biçimde tanımlanır. Fakat bütün terimler w_i ağırlıklarıyla çarpılır. Bu örnek için bu durum son derece önemlidir. Çünkü F değeri çok büyüktür ve bu nedenle karşılık gelen p değeri sıfıra yakındır.

EKK ve RLS sonuçları gerçekte tamamen farklı olduğunda sapanları tanımlamak (RLS aracılığıyla) ve bunlar üzerinde çalışmak doğru olacaktır. Bazı insanlar bunun yerine EKK ve RLS arasından seçim yapmak düşüncesine yatkınlar ve genelde daha anlamlı t değerleri ve F değerlerinin tahminlerini tercih edecekler ve genelde yüksek olan R^2 yi en iyi regresyon varsayıyorlardı. Bu artık yapılmıyor. Çünkü EKK sonuçları sapanlara karşı çok hassastır. Burada R^2 iki şekilde etkilenebilir. Gerçekten de herhangi bir verinin EKK ya ait R^2 si bir ya da daha fazla kaldıraç noktasıyla 1 e yaklaşabilir. Genelde RLS çoğunluğun R^2 si tamamen çok yüksek olmadığı sonucuna uygun biçimde ve yüksek R^2 den sorumlu sapanları gösterir. (ya da bazı θ_i ler sıfırdan anlamlı bir şekilde farklı bulunabilir.

3.3. Örnekler

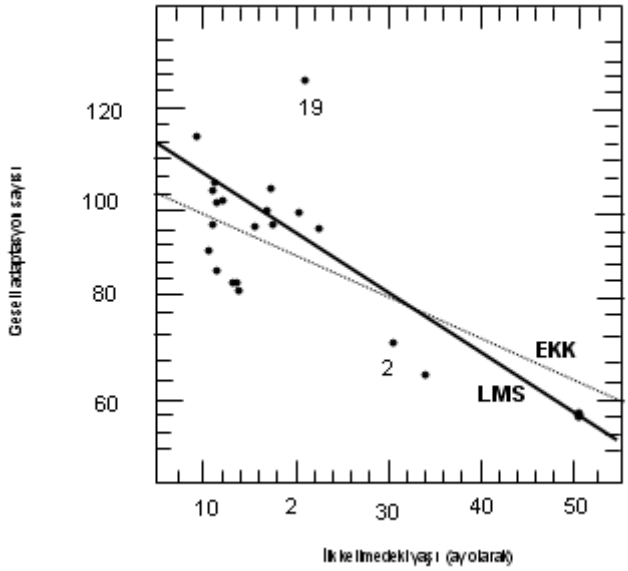
Burada gerçek veri örnekleriyle sağlanan sonuçları daha fazla açıklayacağız.

Örnek 3.3.1 : İlk kelime-Gesell adaptasyon skor verisi

Tablo 3.4. İlk Kelime-Gesell Adaptasyon Skor Verisi

Çocuk (i)	Ayca Yaşı (x_i)	Gesell Skor (y_i)
1	15	95
2	26	71
3	10	83
4	9	91
5	15	102
6	20	87
7	18	93
8	11	100
9	8	104
10	20	94
11	7	113
12	9	96
13	10	83
14	11	84
15	11	102
16	10	100
17	12	105
18	42	57
19	17	121
20	11	86
21	10	100

Kaynak : Mickey (1967)



Şekil 3.5. İlk kelimedeki yaşına göre Gesell adaptasyon skorunun saçılım grafiği

Bu iki boyutlu veri kümesi Mikey den gelir ve geniş bir biçimde gösterilir. Açıklayıcı değişken bir çocuğun ilk kelimesini söylediğindeki yaşı ve yanıt değişken

ise bunun Gesell adaptasyon değeridir. Bu veri 21 çocuk için Tablo 3.4 görülüyor ve şekil5 te grafiklenmiştir. Mickey 19 gözlemin bir sapan olduğuna basamak regresyonunda sapanları arama yaklaşımı ile karar verdi.Bu örnekte 19. indeks durumu sıfır ağırlığını aldı.Gerçekten de bu durum LMS modelinden geniş rezidülere sahip olduğu için sapan olarak tanımlanıyor.EKK ve RLS denklemleri bu veri kümesi için çok farklı değildir. 2 ve 18 nokta çifti EKK yı iyi bir yönde çekiyor.Bu noktalar iyi kaldıraç noktalarıdır ve küçük EKK rezidüleri gibi küçük LMS rezidülerine sahiptirler (Bu noktaların herhangi birinin ya da her ikisinin silinmesi güven aralıklarının genişliğinde önemli bir etkiye sahiptir.)Bu veri kümesinin 2 ve 18 noktalarının silinmesi x ve y arasındaki lineer ilişkiyi çok fazla bırakmayacağı için bu veri kümesinin lineer regresyonunun iyi bir örneği olmadığı söylenebilir.

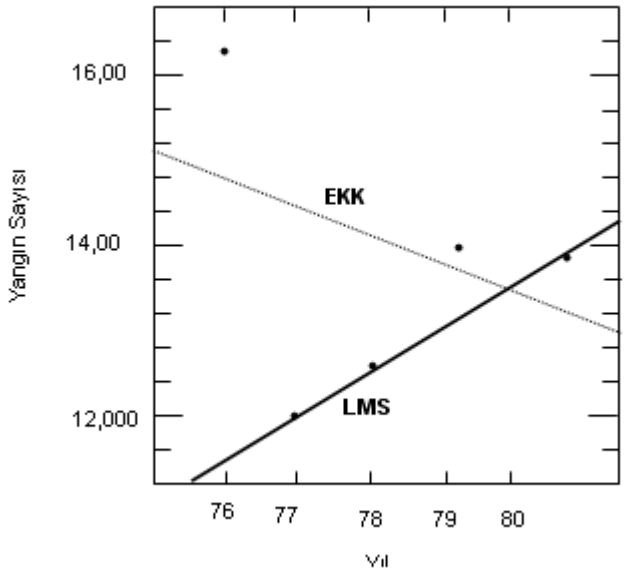
Örnek 3.3.2: 1976 - 1980 Yılları Arasındaki Yangın Sayısı

Tablo 3.5 te verilen bu veri kümesi Belçikalı yangın sigorta şirketlerine bildirilen yangın ihbarlarının 1980e kadar gidişatını göstermektedir.(Belçikalı sigorta şirketinin yıllık raporundan alınmıştır. Bu örnekte çok az nokta vardır)

Tablo 3.5. 1976-1980 arasında Belçika'da yangın ihbarlarının sayısı

$YIL(x_i)$	$YANGIN\ SAYISI$
76	16,694
77	12,271
78	12,904
79	14,036
80	13,874

Şekil 3.6 da saçılım grafiğine bakıldığında yıllar üzerinde hafif yukarı doğru bir eğilim olduğu dikkat çekiyor. Bununla birlikte 1976 daki sayı oldukça yüksektir. Bunun sebebi o yıl yaz dönemi oldukça sıcak ve kuruydu. (Belçika standartlarıyla kıyaslanırsa) Bu durum doğadan kaynaklı olarak ağaçların ve çalıların yanmasına sebep oldu. Bu durumda



Şekil 3.6. 1976- 1980 yıllarında Belçika 'daki yangın ihbarının sayısı

EKK için,

$$\hat{y} = -387{,}5x + 44180{,}8 \qquad \text{(kesikli çizgi)}$$

LMS için,

$$\hat{y} = 534{,}3x - 28823{,}3 \qquad \text{(düz çizgi)}$$

bulunur. Tahminlerinin çok farklı olduğu göze çarpıyor. EKK veriye azalan eğimle uyarken aksine LMS doğrusu artarak uyum sağlıyor. Sapan değer x – menzili dışında uzanır ve kötü bir biçimde yaramayan EKK' e sebep olur. Hatta bu sapan en geniş EKK rezidüsüne sahip olmaz. Bu örnek tekrar gösteriyor ki EKK rezidülerinin incelenmesi sapan(lar)ı ortaya çıkarmak için yeterli değidir.

Elbette ki, böyle küçük bir veri kümesinde herhangi bir güçlü istatistiksel sonuçlar çizemeyiz fakat EKK ve LMS nin esasen ne zaman farklı sonuçlar vereceğini veri kümesi hakkında dikkatli düşünülmesi gerektiğini gösterir.

Örnek 3.3.3: Çin de Fiyatların Artmasının Yıllık Oranı

Tablo 3.6, 1940 dan 1948 e (Simkin 1978) kadar Çin deki ana şehirler de ortalama fiyatların yıllık büyüme oranını içeriyor mesela 1940 da fiyatlar % 62 yükselmiş bir öncekine kıyasla. 1948 de çok fazla hükümet harcamaları sonucunda büyük bir sıçrama meydana gelmiş ve bu hiper enflasyon olarak adlandırılmış.

LMS regresyon denklemi,

$$\hat{y} = 0{,}102x - 2{,}468$$

dır. Buna karşılık EKK tahmini,

$$\hat{y} = 24{,}845\,x - 1049{,}468$$

dır. Bu iki modelde birbirinden tamamen farklıdır. EKK da sayılar daha büyük çıkmıştır.

Bu doğrulardan hangisinin daha iyi modeli vereceğini göstermek için tablo 6 her iki yöntemle tahmin edilmiş değerleri göstermektedir. Son iki yıl hariç LMS verinin çoğunluğuna doğru bir yaklaşım sağlamıştır. Diğer taraftan EKK modeli her yerde kötüdür. Burada tahmin edilmiş $\hat{y}_i$ ilk 3 yıl negatiftir, 1948 hariç sonraki yıllarda çok genişlediği ve kıyaslanamadığı görülür. EKK bütün sütun üzerine

Tablo 3.6. 1940-1948 Arası Serbest Çin deki Ana Şehirlerde Yıllık Ortalama Fiyat Artışı

Yıl	Büyüme Artışı	Tahmini Büyüme	
(x_i)	(y_i)	*LMS*	*EKK*
40	1,62	1,61	-55,67
41	1,63	1,71	-30,82
42	1,90	1,82	-5,98
43	2,64	1,92	18,87
44	2,05	2,02	43,71
45	2,13	2,12	68,56
46	1,94	2,22	93,40
47	15,50	2,33	118,25
48	364,00	2,43	143,09

Kaynak: Simkin (1978)

orijinal verinin doğrusal olmamasının etkisini karıştırırken buna karşı LMS verinin çoğunluğuna uyar. Çıktıda 8. ve 9. gözlemlerin birer sapan olduğuna dikkat edelim.

Bu durum standartlaştırılmış verinin 2. sütununda gösteriliyor. Yinede Spearman Rank korelasyonu, Pearson korelasyonundan daha yüksektir. Çünkü sapanlar değişkenler arasında neredeyse bir monotonluğa uyarken lineer şekilde dağılmıştır.

Rezidü grafiklerine bakarsak bunlar standartlaştırılmış rezidülere karşılık tahmin edilmiş yanıt değişkeni gösterir. Her iki grafikte de kesikli doğru 0 a doğru çiziliyor ve [-2.5, 2.5] aralığında düşey şerit gösteriyor. Bu doğrular sonuçların yorumunu kolaylaştırır. Gözlenmiş y_i değerleri, tahmin edilmiş $\hat{y}_i$ değerlerine eşit olduğunda o zaman sonuçlanmış rezidü 0 olur. Bu 0 doğrusunun komşuluğundaki noktalar modele en iyi uyanlardır. Rezidüler normal dağılımlı ise standartlaştırılmış rezidülerin % 98 i [-2.5, 2.5] aralığında uzanacağı beklenebilir. Bu nedenle düşey şeritten uzakta yerleşmiş olan standartlaştırılmış rezidülerin olduğu gözlemler uzakta olarak tanımlanabilir.

Yukarıdaki çıktıda ilk rezidü grafiğinin EKK modelinde kötü noktayı nasıl gizlediğini gösterir. EKK da sapanlar çekip koparılıyor ve ölçek tahminin iyi olmadığını gösteriyor. Sapanlarla ortak EKK rezidüsü bile düşey şerit içerisinde uzanır. Bu etki yüzünden EKK tahmin edicisine uygun bir rezidü grafiğinin yorumu tehlikelidir. LMS nin rezidü grafiğin de sapan şeritten çok uzaktadır. Robust tahmin edicilerine karşılık gelen rezidü grafikleri bölüm 4e de gösterileceği gibi bazı değişkenlerle olan problemlerde daha kullanışlı bile olabilir.

Bu grafiksel araçlar uzaktaki gözlemleri belirlemek için çok uygundur. EKK sonucu Robust ve Robust olmayan regresyon yöntemlerinin her ikisinin rezidü grafikleri yaklaşık olarak uyuşuyorsa güvenilirdir. Sapanları tanımlamadan başka rezidü grafiği modelin yeterliliğini ölçmek ve dönüşümleri bulmak için bir diagnostik araç sağlar. Yeterli bir model ve iyi davranışlı bir verinin rezidü grafiğinin en ideal örneği sabit düşey saçılımlı noktaların yatay gölgesidir. Rezidü örneğindeki anormallikler olayı bazı yönlere götürür örneğin grafik rezidülerinin varyansı, $\hat{y}$ veya diğer değişken artarken bu değer artar ya da azalır.

Bu heteroscedaticity olarak adlandırılır. Klasik modelin aksine homoscedaticity nin konuşulduğu durumda hatalar sabit varyansa sahiptir. Bu problemi hem açıklayıcı hem de yanıt değişkenin uygun bir dönüşümüne başvurarak yaklaşılabilir.

Örnek 3.3.4: Beyin Ve Ağırlık Verisi

Tablo3.7, 28 hayvanın vücut ağırlıklarını (kg) ve beyin ağırlıklarını (gr) veriyor. Bu örnek Weisberg 1980 ve Jerison 1973 te geniş veri kümelerinden alındı. Beyin ne kadar büyükse o kadar ağır bir vücudu yönetmesini gerekip gerekmediğini araştırıldı.

Bu ölçümlerin logaritmaları arasındaki ilişkinin temiz bir görüntüsü şekil 7 de gösteriliyor. Bu logaritmik dönüşüm ister daha küçük ister daha büyük ölçümlerin yerini tutmada yetersiz olan orijinal ölçümleri grafiklemek için gereklidir. Gerçekten de her iki orijinal değişken de birkaç önem sırasının üzerinden sıralanıyor. Bu dönüştürülmüş verinin lineer modeli beyin ağırlığı (y) ve vücut ağırlığı (x) arasındaki denklem

$$\hat{y} = \theta_2' \; x^{\hat{\theta}_1}$$

biçimindeki ilişkiye denk olacaktır. Şekil3.7 ye bakılırsa bunun dönüşümü daha lineer yaptığı görülür.

EKK modeli,

$$\log \hat{y} = 0{,}75092 \log x + 0{,}86914$$

ile verilir. Eğimle birleştirilmiş standart hata 0,0782 ye eşittir ve kesen terimi 0,1794 tür. Bölüm 4te bilinmeyen regresyon parametreleri için güven aralıklarını nasıl oluşturacağımızı açıklayacağız. Şimdiki örnek için n = 28 ve p = 2 dir. Bu nedenle 26 serbestlik dereceli t dağılımı kullanarak 2,0555 e eşittir. EKK sonuçlarını kullanarak eğim için % 95 lik güven aralığı [0.3353; 0.6567] olarak verilir. RLS şekil 7 de düz çizgi ile veriliyor. Burada model daha diktir.

Tablo 3.7. 28 Hayvanın Vücut ve Beyin Ağırlıkları

İndeks (i)	Çeşitler	Vücut Ağırlığı (kg) (x_i)	Beyin Ağırlığı (gr) (y_i)
1	Kunduz	1,350	8,100
2	İnek	465,00	423,00
3	Bozkurt	36,330	119,500
4	Keçi	27,660	115,00
5	Kobay	1,040	5,500
6	Diplodocus	11700,000	50,000
7	Asya Fili	2547,000	4603,000
8	Eşek	187,100	419,000
9	At	521,000	655,000
10	Potar Maymunu	10,000	115,000
11	Kedi	3,300	25,600
12	Zürafa	529,000	680,000
13	Goril	207,000	406,000
14	İnsan	62,000	1320,000
15	Afrika Fili	6654,000	5712,000
16	Triceratops	9400,000	70,000
17	Rhesus Maymunu	6,800	179,000
18	Kanguru	35,000	56,000
19	Hamster	0,120	1,000
20	Fare	0,023	0,400
21	Tavşan	2,500	12,100
22	Koyun	55,500	175,000
23	Jaguar	100,000	157,000
24	Şempanza	52,160	440,000
25	Brachiosaurus	87000	154,400
26	Sıçan	0,280	1,900
27	Köstebek	0,122	3,000
28	Domuz	192,000	180,000

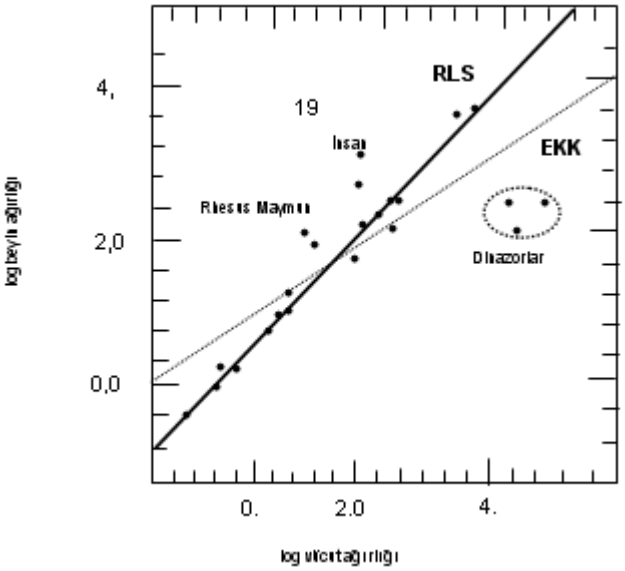
Kaynak: Weisberg (1980) ve Jerison (1973)

$$\log \hat{y} = 0{,}75092.\log x + 0{,}86914$$

RLS tekniği ile tahmin edilmiş eğimi EKK modeli ile ortak %95 güven aralığının dışında kalabilir. RLS deki regresyon katsayılarının standart hatası EKK ile kıyaslandığında dikkate değer şekilde azalmıştır. Yani eğim 0.0318, sabit terim ise 0.0618 dir. Bilinmeyen eğim için %95 lik güven aralığı şimdi [0.6848, 0.8171] olarak verilir. Buradaki aralık EKK' de verilen güven aralığından daha dardır. RLS regresyon katsayıları ile ortak t değerleri çok geniştir. Bu, eğim ve keseni 0 dan

önemli derecede farklı olmasını gerektirir. Bundan başka R^2 belirleyicilik katsayısı (modelin her yerde iyi olduğunun özet ölçümü) EKK için 0.608 den, RLS için 0.964 ten artar. Bu örnek ancak bütün EKK sonuçları ile sadece EKK regresyon katsayılarının sapanların varlığından şüpheye düşmediğini gösterir.

Tablo 3.8 standartlaştırılmış EKK ve RLS rezidülerini ve LMS nin temelinde tanımlanmış w_i yi listeliyor. RLS den dolayı sıradan olmayan gözlemleri ortaya çıkarması kolaydır ve onlara özel bir önem verir. Gerçekten de w_i nin 0 olan beş durumuna bakılırsa onların neden sapan olarak düşünülmek zorunda olduğu kolayca anlaşılır. En belirgin RLS rezidüleri (büyük olasılıkla negatif olanlar) durum 6, 16 ve 25 tir. Burada EKK modelinin düşük eğiminden bu hatalar sorumludur. 3 tane dinazor var ve bunların her biri ağır bir vücutla kıyaslanırsa küçük bir beyne sahiptir. Bu konuda veri de kalan diğer memelilerle kıyaslandı LMS regresyonu da durum 14 ve 17 için 0 ağırlığını gösterdi. Yani insan ve rhesus maymunu. Onlar için esas beyin ağırlığı lineer modeldeki beklenenden daha yüksektir. Dinazorların aksine rezidüleri bu nedenle pozitiftir. Sonuç olarak dinazorlar, insanlar ve rhesus maymunları verinin çoğunluğunu izlediği gibi aynı eğime uymadığı söylenilebilir.



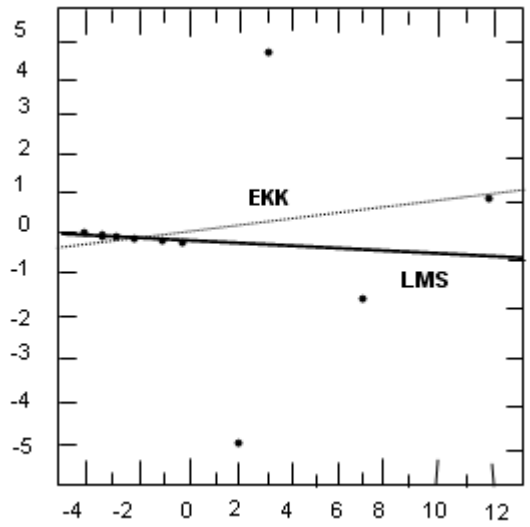
Şekil 3.7. 28 tane hayvan için logaritmik vücut ağırlıklarına göre logaritmik beyin ağırlıklarının EKK (kesikli doğru) ve RLS (düz doğru) modeli.

Tablo 3.8. Beyin ve Vücut Ağırlıkları Verisi İçin Standartlaştırılmış EKK ve RLS Rezidüleri

İndeks	*Türler*	*Standartlaşmış EKK Rezidüleri*	*Standartlaşmış RLS Rezidüleri*	w_1
1	Kunduz	-0,40	-0,27	1
2	İnek	0,29	-1,13	1
3	Bozkurt	0,29	0,17	1
4	Keçi	0,36	0,50	1
5	Kobay	-0,57	-0,65	1
6	Diplodocus	-2,15	-10,19	0
7	Asya Fili	1,30	1,08	1
8	Eşek	0,58	0,21	1
9	At	0,54	-0,43	1
10	Potar Maymunu	0,68	2,02	1
11	Kedi	0,06	0,68	1
12	Zürafa	0,56	-0,37	1
13	Goril	0,53	0,00	1
14	İnsan	1,69	4,15	0
15	Afrika Fili	1,13	0,08	1
16	Triceratops	-1,86	-9,20	0
17	Rhesus Maymunu	1,10	3,47	0
18	Kanguru	-0,19	-1,29	1
19	Hamster	-0,98	-0,81	1
20	Fare	-1,04	-0,17	1
21	Tavşan	-0,34	-0,39	1
22	Koyun	0,40	0,29	1
23	Jaguar	0,14	-0,80	1
24	Şempanza	1,02	2,22	1
25	Brachiosaurus	-2,06	-10,94	0
26	Sıçan	-0,84	-0,80	1
27	Köstebek	-0,27	1,35	1
28	Domuz	0,02	-1,50	1

3.4. Tam Modelin Özelliklerinin Bir Örneği

" Tam model " sözcüğü gözlemlerin geniş bir yüzdesini belli bir lineer denkleme tam olarak uydurma durumu anlamına gelir. Örneğin basit regresyonda bu verinin çoğunluğu düz bir doğru üzerinde tam uzanıyorsa olur. Böyle bir durumda bir robust regresyon yöntemi o doğruyu iyileştirebilir. Bu ise tam modelin özelliği olarak adlandırılır. Örneğin tekrarlı medyan bu özelliği sağlıyor. Gözlemlerin en azından $n - [\, n/2 \,] + 1$ 'i aynı doğru üzerinde uzandığında bu doğrunun denklemi LMS çözümü olacaktır.



Şekil 3.8. Tam modelin örneği. Siegel 'in veri kümesinin EKK (kesikli doğru) ve LMS (düz doğru) modelinin saçılım grafiği

Tablo 3.9 daki veri Emerson ve Hooglin (1983, s.139) dan geldi. Bu verilerin şekil 7 deki saçılım grafiğini inceleyelim Emerson ve Hooglin uygun gelecek özetin 0 eğimle doğru olacağını ileri sürdüler. Gerçekten 9 noktanın dışında 6 tanesi aslında 0 eğimli ve 0 kesenli doğrusu üzerinde uzanır. Tam model durumları gerçek veride de meydana gelir.

3.5. Orijinden Geçen Basit Regresyon

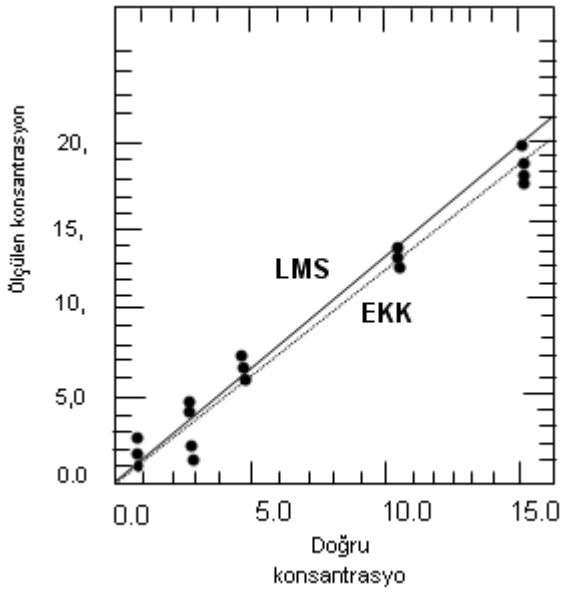
Kesen terimin 0 olduğu önceden bilindiğinde o zaman modeli buna zorlamak durumunda kalınır.

Afifi ve Azen (1979, s.125) Tablo 3.10 da ki örneği verdiler. Burada böyle bir model uygundur. Veri kandaki laktik asit miktarını ölçen bir aletin kalibrasyonu (doğru ölçüp ölçmediği) ile ilgilidir. x_i doğru konsantrasyon ile y_i miktarı kıyaslanıyor. Bu verideki açıklayıcı değişken düzenlenmiştir yani değeri önceden tespit edilmiştir. Sonuç olarak böyle bir veri kümesi kaldıraç noktalarını içermez. Şekil 3.9 da ki saçılım grafiği kabaca 2 değişken arasındaki lineer ilişkiyi gösterir. Gerçekten bu örnek için EKK ve LMS hemen hemen aynıdır. Buna rağmen literatür sapanlarla birlikte kalibrasyon verisini de içerir.

Tablo 3.9. Siegel' in Veri Kümesi

i	x_i	y_i
1	-4	0
2	-3	0
3	-2	0
4	-1	0
5	0	0
6	1	0
7	2	-5
8	3	5
9	12	1

Kaynak: Emerson ve Hoaglin (1983)

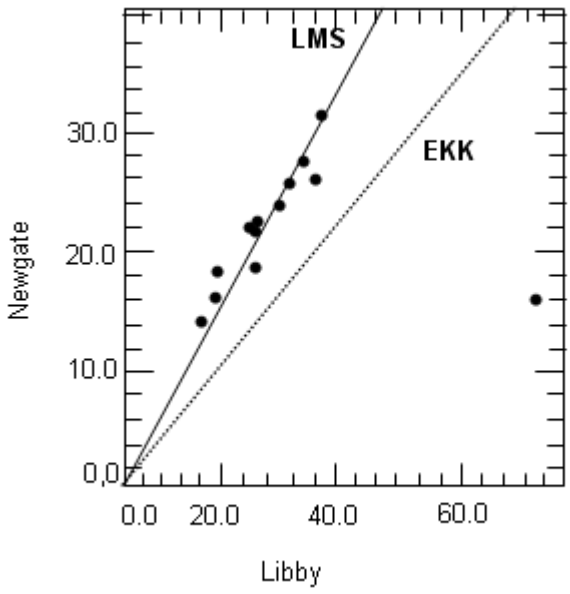


Şekil 3.9. Tablo 3.10 'daki verinin saçılım grafiği. Kesen terim olmadan bir model kullanıldı. Kesikli çizgiye karşılık gelen EKK tahmin edicisi ve düz doğruya karşılık gelen LMS modelidir.

Şimdi de Hampel' in kullandığı veriye bakalım. Burada Tablo 3.11 de yeniden yapıldı. Onlar 2 farklı noktada su akışını ölçtüler.

Her j değişkeni için;

$$s_j = 1{,}4826 \underset{i}{med} \left| x_{ij} \right| \qquad (3.18)$$



Şekil 3.10. Libby ve Newgate'teki su debisinin saçılım grafiğinin EKK (kesikli doğru) ve LMS (düz doğru) modeli.

Bu yüzden 0 dan farklı sapmalar düşünülmesi mantıklıdır. Aynı neden için veri her x_{ij} yi standartlaştırarak

$$\frac{x_{ij}}{s_j} \tag{3.19}$$

biçiminde yapıyor.

3.6. Basit Regresyon İçin Diğer Robust Teknikleri

Son 50 yıla kadar birçok teknik basit regresyon modelinde tahmin edilen eğim ve keseni ileri sürmüştür. Bunların çoğu % 30 luk kırılma noktasına ulaşmadı.

Emerson ve Hooglin (1983) tarihi bir inceleme verdler. Bu inceleme Wald (1940), Nair ve Shrivastava (1942), Bartlett (1949) ve Brown ve Mood (1951) in teknikleri hakkında bazı açıklamalar içerir. Bu regresyon yöntemleri tamamen veri kümesinden ayrılma fikrine dayanıyor ve o zaman her grup için özet değer olarak tanımlanıyor. Onlar EKK regresyonunun robust olmayışının merakı ile ilgili kanıyı doğruladılar. Fakat sadece basit regresyona başvurabilirler.

Tukey (1970/1971) in sağlam doğru yöntemi robust tahminine ilk yaklaşımların bir değişimidir. Tek değişkenli veriye göre modellenmiş bir doğru için bu " kalem ve kağıt " tekniği hemen hemen eşit genişlikte 3 parça içinde veri kümesinin bölünmesinden başlar. Bu pay ayrımı en küçük, orta ve en geniş değerlerine göre uygulanır. X değerlerine bağlı olarak aynı grup seçilir. O zaman robust eğim tanımlanır. Öyle ki en dıştaki gruplarda denktir. Yani

$$\underset{i\in L}{med}(y_i - \hat{\theta}_1 x_i) = \underset{i\in R}{med}(y_i - \hat{\theta}_1 x_i) \qquad (3.20)$$

Burada L left (sol), R right (sağ)' ı simgeler. Burada kesen değer her iki grubun medyan rezidüsünü 0 yapmak için seçilir. Bu teknikle sapanlara karşı var olan korumanın en kötü durum sınırlaması 1/6 dır. Çünkü her iki grup için de kırılma noktası ½ olan medyan kullanılır. Velleman ve Hooglin (1981) sağlam doğruları bulmak için bir algoritma ve portatif program sağladılar. Bu tahmin edici üzerine bazı teorik ve Monte Carlo sonuçları Johnstone ve Velleman (1985b) tarafından sağlandı. Brown ve Mood (1951) tekniği aynı şekilde tanımlanır fakat 2 grup yerine 3 grupla verilmiştir. Sonuç olarak bu teknik için kırılma noktası ¼ 'e artar.

Andrews (1974) elde edilen düz bir doğru modeli için başka bir medyana dayalı yöntem geliştirdi. O, x değerlerini en küçükten en büyüğe doğru sıralayarak başladı, sonra c değerlerinden en küçük ve en büyüklerin belli bir sayısını eledi ve de bu değerlerin belli sayıda ki medyan (x değeri) komşuluklarındaki değerini x değerlerinin kalan iki kümesinde medyanları hesapladı. (med x_1 ve med x_2). y değerine göre de medyanlar hesaplandı. (med y_1 ve med y_2). Modelin eğimi o zaman şu şekilde tanımlanır:

$$\hat{\theta}_1 = \frac{medy_2 - medy_1}{medx_2 - medx_1}$$

Bu yöntemin kırılma noktası en fazla %25 tir. Çünkü her iki veri kümesinin yarısı modellenmiş doğruyu açıklayabiliyor. Andrews bu tekniği çoklu regresyona genelledi ve buna süpürme operasyonu dedi. Yani her itme de bir değişkenin diğerine

bağımlılığı değişkenin en erken aşamada tanımlanan bir katı tarafından belli bir değişkene uyarlanması ile kaldırılır. Aynı fikir robust doğruya genellemek için kullanılır.

Basit regresyon tahmin edicilerinin diğer bir grubu, veri kümesinden ayırmaksızın onların tanımında yapı taşları gibi eğim çiftleri kullanılır. Theil 1950, c_n^2 eğimlerinin $\hat{\theta}_1$ meydanını şu şekilde tanımladı.

$$\hat{\theta}_1 = \underset{1\le i<j\le n}{med} \frac{y_j - y_i}{x_j - x_i} \tag{3.21}$$

Bu tahmin edicisi yüksek asimtotik etkiye sahiptir. Sen(1968) x_i lerin arasındaki ilişkileri idare etmek için bu tahmin ediciyi geliştirdi. Bu tekniklerin kırılma noktasının yaklaşık %29,3 olduğu ortaya çıktı.Bu değer şu sebeple açıklanabilir:İyi bir eğimin oranı en azından ½ olmak zorunda olduğu için en az ½ olan $(1-\varepsilon)^2$ ifadesine ihtiyaç vardır.Burada ε verideki sapanların kırılmasıdır.Bundan dolayı şu ifade yazılır:

$$\varepsilon \le 1 - (1/2)^{1/2} \approx 0{,}293 . \tag{3.22}$$

Daha yakınlarda Siegel (1982) son gelişmelerle kırılma noktasını % 50 ye arttırdı. Bu da bizi tekrarlı medyan tahmin edicisine götürür. Bu yöntemle (3.21) 'deki tek medyan yerine eğim çiftlerinin medyanını iki aşamada hesaplamak zorunda kalındı. Eğim ve kesen aşağıdaki gibi tanımlandı:

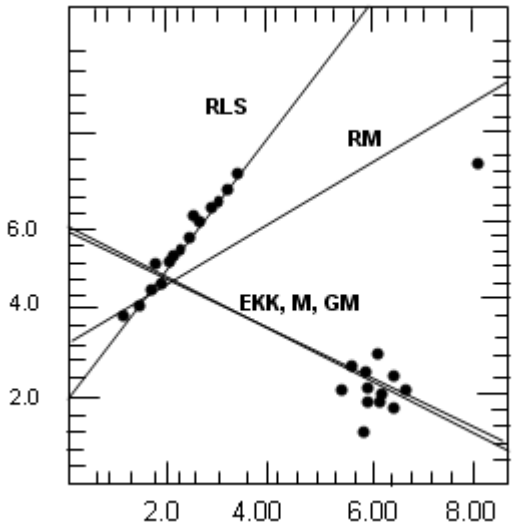
$$\left.\begin{aligned} \hat{\theta}_1 &= \underset{i}{med}\, \underset{j\ne i}{med} \frac{y_j - y_i}{x_j - x_i} \\ \hat{\theta}_2 &= \underset{i}{med}\, (y_i - \hat{\theta}_1 x_i) \end{aligned}\right\} \tag{3.23}$$

Basit regresyon tekniklerinin bir kısmı da bölüm 4 te verilecektir. Farklı regresyon teknikleriyle elde edilen modelleri kirlenmeye karşı kıyaslamak amacıyla

yapay bir örnek düşünelim: 30 tane iyi gözlem lineer ilişkiye göre meydana getirildi ve

$$y_i = 1{,}0x_i + 2{,}0 + e_i \tag{3.24}$$

şeklinde modellendi. Burada x_i ,[1,4] aralığında hep aynı şekilde dağılır ve e_i, 0 ortalamalı ve 0,2 standart sapmalı normal dağılımlıdır. O zaman 20 kötü gözlemin bir sınıfı eklendi. Bunlar 7,2 ortalamalı ve 0,5 standart sapmalı tek değişkenli normal dağılıma sahiptir. Bu ise birleştirilmiş örnekteki kirlenmenin %40 olduğunu verir. Bu da çok yüksektir. Aslında bu miktar eğer basit regresyon için tahmin edicilerinin çoğunun kırılma noktasını yukarı çekiyorsa ne olduğunu göstermek için seçilir. Şimdi de hangi tahmin edicinin verinin çoğunluğunun modelini açıklayacağının en iyisinin başaracağını görelim. Klasik EKK yöntemi $\hat{\theta}_1$=-0,47 ve $\hat{\theta}_2$ =5,62 yi verir. İyi ve kötü noktalara uymaya çalıştığı için bunun başarısız olduğu açıktır.



Şekil 3.11. Altı yöntem kullanılarak yapay veri için regresyon doğruları. (RLS, LMS 'e dayalı yeniden ağırlıklandırılmış en küçük kareler; EKK, en küçük kareler; M,Huber 'in M tahmin edicisi; GM, Mallows 'un ve Schweppe 'nin M tahmin edicileri; RM, tekrarlı medyan) .Kaynak: Rousseeuw (1984) .

3 robust tahmin edicisi uygulandı: Huber'in $\psi(x) = \min(1.5, \max(-1.5, x))$ M-tahmin edicisi, Hampel ağırlıklarıyla Mallows'un genelleştirilmiş M-tahmin edicisi Hampel –Krasker ağırlıklarıyla Schweppe'nin genelleştirilmiş M-tahmin edicisi (Mallows ve Schweppe'nin tahmin edicileri aynı Huber fonksiyonunu kullandılar). Yine de bu 3 yöntem EKK çözümünden nerdeyse ayırt edilemez sonuçlar verir.

Bazı M-tahmin edicileri şunlardır:

Huber minimax :

$$\psi(t) = \begin{cases} t & ;|t| < b \\ b\,\mathrm{sgn}(t) & ;|t| \geq b \end{cases}$$ b,sabittir.

Inen minimax :

$$\psi(t) = \begin{cases} t & ;|t| < a \\ b\,\mathrm{sgn}(t).\tanh\left[\frac{1}{2}b(c-|t|\right] & ;a \leq |t| < c \\ 0 & ;d.y \end{cases}$$

a,b ve c sabitler

Hampel :

$$\psi(t) = \begin{cases} t & ;|t| < a \\ a\,\mathrm{sgn}(t) & ;a \leq |t| < b \\ \left(c-|t|\right)/\left(c-b\right)a\,\mathrm{sgn}(t) & ;b \leq |t| < c \\ 0 & ;d.y. \end{cases}$$

a ,b ve c sabitler

Andrew :

$$\psi(t) = \begin{cases} \sin(t) & ;-\pi \le t < \pi \\ 0 & ;d.y. \end{cases}$$

Tukey :

$$\psi(t) = \begin{cases} t\left(1 - \left(t/c\right)^2\right)^2 & ;|t| < c \\ 0 & ;\ d.y \end{cases}$$

dır.

4. ÇOKLU REGRESYON

4.1.Giriş

Çoklu regresyonda y_i yanıt değişkeni modeldeki p tane $x_{i1},....,x_{ip}$ açıklayıcı değişkenlerine bağlıdır.

$$y_i = x_{i1}\theta_1 + ... + x_{ip}\theta_p + e_i \qquad (i = 1,\ldots, n) \tag{4.1}$$

Basit regresyondaki gibi bilinmeyen $\theta_1,...,\theta_p$ parametrelerinin tahmini için EKK tekniği sapanların varlığına oldukça hassastır. Böyle noktaların bulunması, saçılım grafiğinde etki noktalarının belirlenmesi artık mümkün olmadığı için daha zor olur. Bu nedenle böyle noktaları belirleyen bir araca sahip olmak önemlidir.

Son birkaç 10 yıldır, birçok istatistikçi robust regresyonuna önem verdi. Buna karşılık diğerleri dikkatlerini regresyon diagnostiklerine çevirdiler. Her iki yaklaşım da hemen hemen aynı iki önemli yaygın amaca dayanır. Yani sapanları ve modelin yetersizliğini belirlemek. Fakat onlar farklı bir yönde ilerlediler. Regresyon diagnostikleri bir regresyon modelini uygulamadan önce veri kümesinden silinmiş olan noktaları belirlemek için uğraştı. Robust regresyonu ise bu problemleri ters bir basamakta ele alarak diğer durumlarda oldukça yüksek etkili olan noktaların etkisini azaltmayı tasarlayarak yapılır. Bir robust yöntemi verinin çoğunluğunu bir modele uydurmaya çalışır. İyi noktalarla biçimlenmiş modelden uzakta yerleşen kötü noktalar sonuç olarak robust modelde geniş rezidülere sahip olacaktır. Bu nedenle sapanların hassas olmayışına ek olarak bir robust regresyon tahmin edicisi bu noktaların bulunmasını kolay bir iş olarak yapar. Elbette EKK daki rezidüler bu amaçla kullanılamazlar çünkü sapanlar çok küçük EKK rezidülerine sahipken EKK modeli bu sapan noktaları çok fazla çekebilir. EKK ya alternatif bir robust a ihtiyacı göstermek için bazı örneklere bakalım. İlk önce Brownlee (1965) tarafından bulunan ve en iyi bilinen stackloss (yığın kaybı) veri kümesidir. Biz bu örneği gerçek verilerin bir kümesi olduğu için ve birçok yöntemle çok sayıda istatistikçi tarafından incelendiği için seçtik.

Veri tablo 1 de listelenen 4 ölçekli 21 gözlemden oluşur. Veri amonyağın nitrik asite oksidasyonu için bir bitkideki işlemi tanımlıyor. Yığın Kaybı (y), x_1 oran işlemi, x_2 soğutulmuş suyun giriş sıcaklığı ve x_3 asit konsantrasyonu ile açıklandı. Özetle literatür atıf almış buluşların gösterdiği 1,3,4 ve 21. gözlemlerin sapan olduğu sonucunu birçok insanın çıkardığı söylenebilir. Bazılarına göre 2. gözlem de bir sapandır. EKK regresyonu şu denklemi verir:

$$\hat{y} = 0{,}716\, x_1 + 1{,}295\, x_2 - 0{,}152\, x_3 - 39{,}9,$$

Tablo 4.1 Yığın Kaybı Verisi (Stackloss Data)

İndeks (i)	Oran (x_1)	Sıcaklık (x_2)	Asit Konsantrasyonu (x_3)	Yığın Kaybı (y)
1	80	27	89	42
2	80	27	88	37
3	75	25	90	37
4	62	24	87	28
5	62	22	87	18
6	62	23	87	18
7	62	24	93	19
8	62	24	93	20
9	58	23	87	15
10	58	18	80	14
11	58	18	89	14
12	58	17	88	13
13	58	18	82	11
14	58	19	93	12
15	50	18	89	8
16	50	18	86	7
17	50	19	72	8
18	50	19	79	8
19	50	20	80	9
20	56	20	82	15
21	70	20	91	15

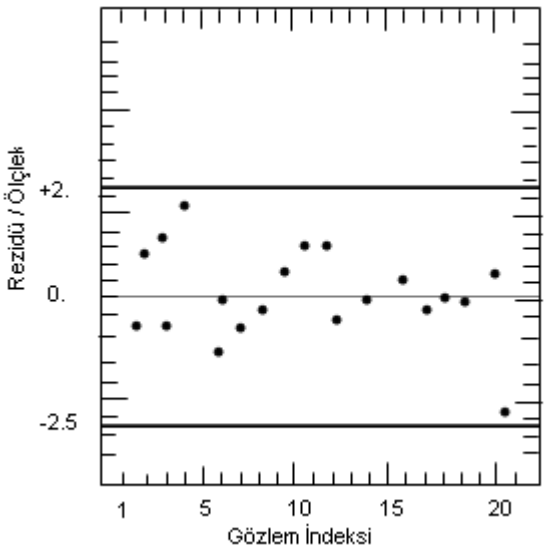
Kaynak: Brownlee (1965)

EKK gösterge grafiği Şekil 4.1 de gösteriliyor. Rezidülerin standartlaştırılması rezidülerin her bir kısmına modele uygun ölçek tahmini ile uygulanır. Yatay şerit – 2.5 ile 2.5 arasında standartlaştırılmış rezidülerin etrafını çeviriyor. Böyle Şekil 4.1 de olduğu gibi sapan göze çarpmıyor. EKK gösterge grafiğinden standartlaştırılmış EKK rezidüleri şeridin içine tamamen düştüğü için

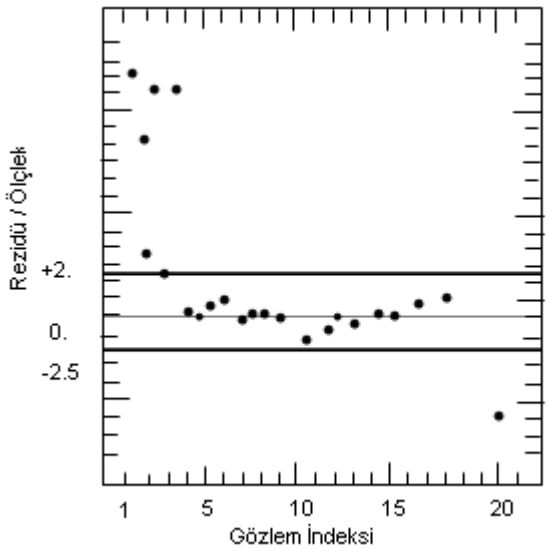
veri kümesinin sapan içermediği sonucuna varılır. Fakat Şekil 4.2 ye bakarsak en küçük medyan kareleri (LMS) ile ilişkili olan gösterge grafiği

$$\hat{y} = 0{,}714\, x_1 + 0{,}357\, x_2 - 0{,}000\, x_3 - 34{,}5,$$

Bu grafik robust modeline dayanır ve gerçekten zararlı noktaların varlığını ortaya çıkarır. Bu gösterge grafiğinden 1, 3, 4 ve 21. gözlemlerin en uzakta olduğu



Şekil 4.1. Yığın kaybı verisi: EKK ‘ye göre indeks grafiği



Şekil 4.2. Yığın kaybı verisi: LMS ‘e göre indeks grafiği

ve 2. gözlemin sapanların olduğu bölgenin sınırında olduğu için arada olduğu hemen açık bir biçimde görülür. Bu durum robust regresyon tekniğimizin bu veriyi tek bir hamlede nasıl analiz edebildiğini gösteriyor. Bunlar aynı veri kümesinin ilk analizlerinden zahmetli ve uzun olan bazılarıyla kıyaslanır.

Bu örnek sadece EKK rezidülerine bakılmasındaki tehlikeyi bir kere daha gösteriyor. Her regresyon analizinde EKK ve robust yönteminin her ikisinin standartlaştırılmış rezidülerini kıyaslamanın gerekli olduğunu söyleyebiliriz. Eğer iki yönteminde sonuçları hemen hemen aynı ise o zaman EKK güvenilir olabilir. Eğer farklı ise robust yöntemi sapanları ortaya çıkarmak için uygun bir araç olarak kullanılabilir. Burada o zaman sapanlar tamamen araştırılabilir ve belki düzeltilebilir (eğer orijinal ölçülere göre biri yakınsa) ya da silinebilir. Başka bir olasılık ise modeli değiştirmektir (örneğin kareler ekleyerek ya da ters çarpım terimleri ekleyerek ve / veya yanıt değişkeni değiştirerek). Bu şekilde Atkinson (1985, s.129-136) x_1, x_2, x_1x_2 *ve* x_1^2 aracılığıyla log (y) yi açıklayan modelleri kurarak yığın kaybı verisini analiz etti.

Sonraki örnek sosyal bilimden geliyor. Bu veri kümesi orta atlas okyanusu ve yeni İngiltere bölgesinden 20 tane okulun bilgisini içeriyor. Coleman tarafından incelenen bir populasyondan alınmıştır (1966). Mosteller ve Tukey (1977) 6 farklı değişken üzerindeki ölçümlerden oluşan bu örneği analiz ettiler bunlardan biri yanıt değişken olarak alındı ve aşağıdaki gibi tanımlandı.

x_1 = Kişi başı düşen personel maaşı

x_2 = Memur babaların yüzdesi

x_3 = Sosyo ekonomik statüden sapma durumları: Ailenin serveti, ailenin zarar görmüşlüğü, babanın eğitim durumu, annenin eğitim durumu ve ev eşyaları.

x_4 = Öğretmenlerin sözlü sınav puanlarının ortalamasın

x_5 = Annenin eğitim seviye ortalaması (her birim 2 okul yılıdır.)

y = Sözlü sınav ortalaması

Sıradan EKK regresyonu ile 20 okul için model

$$\hat{y} = -1{,}79\,x_1 + 0{,}044\,x_2 + 0{,}556\,x_3 + 1{,}11\,x_4 - 1{,}81\,x_5 + 19{,}9,$$

olarak verilir. En küçük medyan kareler regresyonunun modeli ise

$$\hat{y} = 0{,}580\,x_1 + 0{,}058\,x_2 + 0{,}637\,x_3 + 0{,}740\,x_4 - 2{,}32\,x_5 + 25{,}1$$

olarak verilir. Tablo 4.3 ün sol kısmı EKK tahminlerini ve ortak rezidülerini listeliyor. Bu sonuçlar EKK denkleminin 3 ve 11 okulları için yanıtın biraz altındaki tahmini ve
18 okulu için ise tahminin üstünde olduğunu ortaya çıkardı. Sadece EKK sonucunu

Tablo 4.2. Orta Atlas ve Yeni İngiltere Bölgesinden 20 Okulun Bilgisini İçeren Coleman Veri Kümesi

İndeks	x_1	x_2	x_3	x_4	x_5	y
1	3,83	28,87	7,20	26,60	6,19	37,01
2	2,89	20,10	-11,71	24,40	5,17	26,51
3	2,86	69,05	12,32	25,70	7,04	36,51
4	2,92	65,40	14,28	25,70	7,10	40,70
5	3,06	29,59	6,31	25,40	6,15	37,10
6	2,07	44,82	6,16	21,60	6,41	33,90
7	2,52	77,37	12,70	24,90	6,86	41,80
8	2,45	24,67	-0,17	25,01	5,78	33,40
9	3,13	65,01	9,85	26,60	6,51	41,01
10	2,44	9,99	-0,05	28,01	5,57	37,20
11	2,09	12,20	-12,86	23,51	5,62	23,30
12	2,52	22,55	0,92	23,60	5,34	35,20
13	2,22	14,30	4,77	24,51	5,80	34,90
14	2,67	31,79	-0,96	25,80	6,19	33,10
15	2,71	11,60	-16,04	25,20	5,62	22,70
16	3,14	68,47	10,62	25,01	6,94	39,70
17	3,54	42,64	2,66	25,01	6,33	31,80
18	2,52	16,70	-10,99	24,80	6,01	31,70
19	2,68	86,27	15,03	25,51	7,51	43,10
20	2,37	76,73	12,77	24,51	6,96	41,01

Kaynak: Mosteller ve Tukey (1977)

inceleyerek 3, 11 ve 18 okullarının lineer modelden en uzak kalan okullar olduğu sonucu elde edilir. Fakat Tablo 4.3 ün sağ kısmındaki 11 okulun robust modelinden tamamen sapmadığı görülür.

Robust regresyonu 3, 17 ve 18 okullarını geniş standartlaştırılmış rezidülerine ayırarak sapanlar olarak seçti. Sonradan bu standartlaştırılmış rezidüler mutlak standartlıştırılmış (2.5 tan daha geniş olanlar) rezidüye sahip bütün gözlemlere göre 0 ağırlığını veren ağırlıkları hesaplamak için kullanılır.

Coleman ın veri kümesine bunu yapmak tablo 3 te son sütunu verir. Bu ağırlıklar yeniden ağırlaklandırılmış EKK analizini (RLS) kullandırır. 1 ağırlıklı sadece 17 noktayı içeren veri kümesine EKK regresyon uygulayarak aynı şeye varılır. Modellenmiş denklemden ve ortak rezidülerden ayrı olarak sapanlarda t ye

Tablo 4.3. Coleman Verisi: Tahmin Edilmiş Sözlü Sınav Sonuçlarının Ortalaması ve EKK ve LMS Modeli için Ortak Rezidüler

	EKK Sonuçları		LMS Sonuçları		
İndeks	Tahmin edilmiş yanıt	Standartlaştırılmış rezidüler	Tahmin edilmiş yanıt	Standartlaştırılmış rezidüler	ağırlıklar
1	36,661	0,17	38,883	-1,59	1
2	26,860	-0,17	26,527	-0,01	1
3	40,460	-1,90	41,273	-4,04	0
4	41,174	-0,23	42,205	-1,28	1
5	36,319	0,38	37,117	-0,01	1
6	33,986	-0,04	33,917	-0,01	1
7	41,081	0,35	41,628	0,15	1
8	33,834	-0,21	32,922	0,41	1
9	40,386	0,30	41,519	-0,43	1
10	36,990	0,10	34,847	2,00	1
11	25,580	-1,06	23,169	0,11	1
12	33,454	0,84	33,514	1,43	1
13	35,949	-0,51	34,917	-0,01	1
14	33,446	-0,17	32,591	0,43	1
15	24,479	-0,86	22,717	-0,01	1
16	38,403	0,63	40,041	-0,29	1
17	33,240	-0,69	35,121	-2,82	0
18	26,698	2,41	24,918	5,76	0
19	41,977	0,54	42,662	0,37	1
20	40,747	0,13	41,027	-0,01	1

dayalı anlamlılık seviyelerinde etkilenmiştir. Bu durum güven aralıklarının oluşturulmasında ve regresyon katsayılarının hipotez testleri için önemlidir. Coleman ın testleri için regresyon katsayılarının anlamlılığının EKK modeli ve RLS modelindekinden oldukça farklı olduğu ortaya çıktı. Tablo 4.4 teki t değerli sıfır hipotezi $H_0 : \theta_j = 0$, alternatif hipoteze $H_1 = \theta_j \neq 0$ $x_3 ve x_4$ değişkenleri x = % 5

için önemli derecede 0 dan farklı EKK regresyon katsayılarına sahiptir. Çünkü t değerleri 14 (= n-p) serbestlik dereceli Student dağılımının 2,1448 değerinden büyüktür ve bu nedenle p değerleri 0,05' in altındadır.

Şimdi de bu verinin RLS' li analizini yapalım. Bu Tablo 4.4 ün sağ yanındaki katsayılara ve t değerlerine sebep olur. Temizlenmiş veri için şaşırtıcı olan durum şimdi bütün açıklayıcı değişkenler α = % 5 anlamlılık seviyesinde 0 dan

Tablo 4.4. Coleman Verisi: EKK ve RLS Modellerine Göre t Değerleri

	EKK Sonuçları			RLS Sonuçları		
Değişken	Katsayı	t – Değeri	p – Değeri	Katsayı	t – Değeri	p – Değeri
x_1	-1,793	-1,454	0,1680	-1,203	-2,539	0,0275
x_2	0,044	0,819	0,4267	0,082	4,471	0,0009
x_3	0,556	5,979	0,0000	0,659	19,422	0,0000
x_4	1,110	2,559	0,0227	1,098	7,289	0,0000
x_5	-1,810	-0,893	0,3868	-3,898	-5,177	0,0003
St Terim:	19,949	1,464	0,1652	29,750	6,95	0,0001

St: sabit

önemli derecede farklı regresyon katsayılarına sahip olmasıdır. (11 serbestlik dereceli t dağılımının %97.5i 2.2010 a eşittir. Bu nedenle bütün p değerleri 0,05 ten azdır).

Birçok örnekte bazı EKK regresyon katsayılarının önemliliğine sadece sapanlar neden olur ve o zaman karşılık gelen RLS katsayıları artık 0 dan önemli derecede farklı değildir.

Sıradan EKK yöntemi gizlenmiş etkiye karşı sağlam değildir. Yani bir veya daha fazla etkili noktanın silinmesinden sonra başka bir gözlem ilk başta görülmeyip sonra etkili olarak çıkabilir. Bu nedenle yüksek kırılmalı regresyon yönteminin (LMS gibi) kullanımı ağırlıkların tanımlanması için zorunludur. Bir örnek olarak Ruppert

ve Carroll (1980) ın tablo 5 te listenen tuzluluk verisini düşünelim. Sudaki tuzun ve kuzey Carolina 'nın Pamlico Boğazından akıntısı tutulmuş nehrin (örneğin tuz konsantrasyonu) ölçüm kümesidir. 2 haftaya kadar geride kalan tuzluluğa (x_1) karşılık geride kalmış tuzluluğun olduğu bir model oluşturacağız. Burada eğim yani bahar mevsiminin başlangıcından beri geçen 2' şer haftalık periyodların sayısıdır (x_2). x_3 ise nehrin hacminin boğazdaki boşalmasıdır. Carroll ve Ruppert (1985) verinin fiziksel altyapısını tanımladılar. Durum 5 ve 16 nın çok ağır boşaltma

Tablo 4.5. Tuzluluk Verisi

İndeks (i)	Kalan Tuz (x_1)	Meyil (x_2)	Boşalma (x_3)	Tuzluluk (y)
1	8,2	4	23,005	7,6
2	7,6	5	23,873	7,7
3	4,6	0	26,417	4,3
4	4,3	1	24,868	5,9
5	5,9	2	29,895	5,0
6	5,0	3	24,200	6,5
7	6,5	4	23,215	8,3
8	8,3	5	21,862	8,2
9	10,1	0	22,274	13,2
10	13,2	1	23,830	12,6
11	12,6	2	25,144	10,4
12	10,4	3	22,430	10,8
13	10,8	4	21,785	13,1
14	13,1	5	22,380	12,3
15	13,3	0	23,927	10,4
16	10,4	1	33,443	10,5
17	10,5	2	24,859	7,7
18	7,7	3	22,686	9,5
19	10,0	0	21,789	12,0
20	12,0	1	22,041	12,6
21	12,1	4	21,033	13,6
22	13,6	5	21,005	14,1
23	15,0	0	25,865	13,5
24	13,5	1	26,290	11,5
25	11,5	2	22,932	12,0
26	12,0	3	21,313	13,0
27	13,0	4	20,769	14,1
28	14,1	5	21,393	15,1

Kaynak: Ruppert ve Carroll(1980)

periyotlarına karşılık geldiğini gösterdiler. Yaptıkları analizler 5. gözlemin sadece 3. ve 16. durumların silinmesinden sonra etkili olduğu hatırlanabildi. Bu gizli etkinin bir örneğidir. Diğer taraftan LMS bu doğal olaydan etkilenmedi ve tek hamlede 5 ve 16 yı sapan olarak belirledi. EKK modeli şu şekilde verilir.

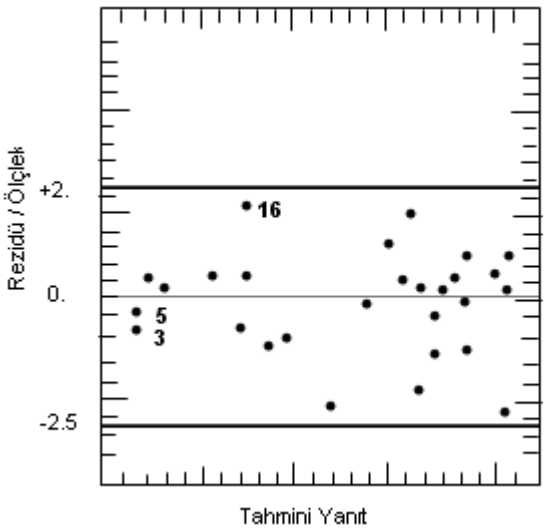
$$\hat{y} = 0{,}777\, x_1 - 0{,}026\, x_2 - 0{,}295\, x_3 + 9{,}59,$$

Buna karşılık LMS modeli şu şekilde verir:

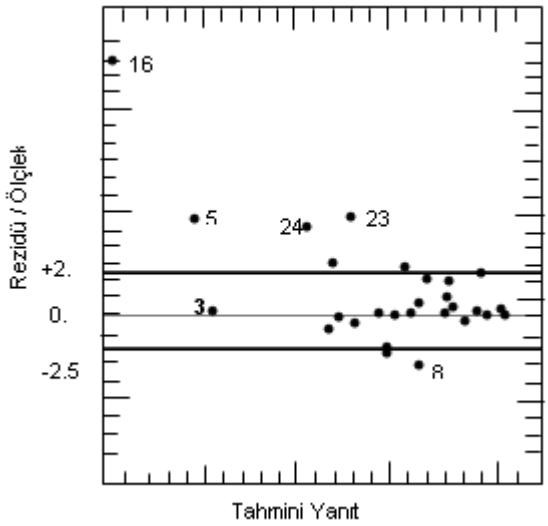
$$\hat{y} = 0{,}356\, x_1 - 0{,}073\, x_2 - 1{,}30\, x_3 + 36{,}7,$$

EKK a karşılık gelen rezidü grafiği Şekil 4.3 te vardır. Bu şekilde standartlaştırılmış rezidüler tahmin edilmiş yanıt değişkene karşılık grafiklenmiştir.Şekil 4.4 te böyle bir rezidü gafiği LMS için verilmiştir. Şekil 4.3 ten görülüyor ki modelde bir yanlışlık yoktur. Fakat sadece kaldıraç noktaları yüzünden küçük EKK rezidüleri üretmeye meyilli kaldıraç noktaları olduğu akılda tutulmalıdır.

Şekil 4.4 uzaktaki gözlemlerin varlığına tanıklık ediyor. Çünkü bazı noktalar şeritten uzağa düştü. Bu örnekte EKK rezidü grafiği robust modele göre olandan çok daha fazla farklı olduğu için güvenilir değildir.



Şekil 4.3. Tuzluluk verisi: EKK modeline göre rezidü grafiği



Şekil 4.4. Tuzluluk verisi: LMS modeline göre rezidü grafiği

Önceden söylediğimiz gibi rezidü grafikleri modellenmiş değerler yönünden modelin fonksiyonunda olası kusurları da gösterebilir.Bir rezidü grafiği,örneğin yanıtın monoton fonksiyonu olan varyans modelini,gösterir.Eğer modelin fonksiyonel kısmı belirli değilse o zaman grafikteki standartlaştırılmış LMS rezidüleri $\hat{y}_i$ ya yapısız diye bakılan noktaların düşey şeridinde bir artışı verir. Yine de modeldeki anormallikler modeldeki değişkenlerin dönüşümü ileri sürebilir. Örneğin bir rezidü grafiğinde eğri modeli y_i nin bazı fonksiyonları tarafından yerini alması işe sonuçlanır ($\log y_i$ veya y_i belli bir gücü arttırır).

4.2.Çoklu Regresyonda En Küçük Medyan Karelerinin Hesaplanması

Bu kısımda başarısız modelleri iyileştirmek için rezidü grafiklerinin kullanımını göstereceğiz. Çoklu regresyonla birlikte yapılan etkileşim dönemi tamamen benzerdir. Kayıp değerli veri kümesinin özel bir işlemi henüz yapılmadı. Bu durumda etki girişi biraz uzun olur. New York ulusal hava koruma ve ulusal hava durumu servisinden " hava kalitesi " verisinin bölümü için bu durumu göstereceğiz. Bu veriler tamamen Chambers te rapor edilmiştir (1983). Bütün veri kümesi 1 mayıs 1973 ten 30 eylül 1973 e kadarki hava kalitesi değerlerinin günlük okunmasından oluşuyor. Değerler Roosevelt adasında 1300 den 1500 saate kadar ozon

yoğunluğunun ortalaması (milyarda bir, Ozone ppb), merkez parktan 0800 den 1200 saate kadar 4000 - 7700 $\overset{0}{A}$ frekans bandında Longlays ten solar radyasyon (Solar

Tablo 4.6. Mayıs 1973 için Hava Kalitesi Verisi Kümesi

İndeks (i)	Solar Radyasyon (x_1)	Rüzgar Hızı (x_2)	Sıcaklık (x_3)	Ozon ppb (y)
1	190	7,4	67	41
2	118	8,0	72	36
3	149	12,6	74	12
4	313	11,5	62	18
5	9999	14,3	56	999
6	9999	14,9	66	28
7	299	8,6	65	23
8	99	13,8	59	19
9	19	20,1	61	8
10	194	8,6	69	999
11	9999	6,9	74	7
12	256	9,7	69	16
13	290	9,2	66	11
14	274	10,9	68	14
15	65	13,2	58	18
16	334	11,5	64	14
17	307	12,0	66	34
18	78	18,4	57	6
19	322	11,5	68	30
20	44	9,7	62	11
21	8	9,7	59	1
22	320	16,6	73	11
23	25	9,7	61	4
24	92	12,0	61	32
25	66	16,6	57	999
26	266	14,9	58	999
27	9999	8,0	57	999
28	13	12,0	67	23
29	252	14,9	81	45
30	223	5,7	79	115
31	279	7,4	76	37

x_1 in 9999 değeri kayıp ölçüleri gösterir. y nin 999 sayısı kayıp değere karşılık gelir.

Kaynak: Chamblers (1983)

Radi), ortalama rüzgar hızı (saatteki mili), La Guardia hava alanında 0700 ve 1000 saat arası ortalama rüzgar hızı (Windspeed) ve maksimum günlük sıcaklık (Fahrenayt derecesinde, Temperature) La Guardia hava alanında. Veri Tablo 4.6 da gösteriliyor.

Bu analizin amacı diğer değişkenler amacıyla ozon konsantrasyonunu açıklamaktır. Tablo 4.6 da ozone ppb ve / veya solar radyasyon ölçümleri (uzun bir gün için kaydedilen ölçümler). Araştırmacıların açtığı ilk yol kayıp olan en az bir değişkenin kayıp olduğu durumları elemektir. Bu 1 e eşit olan kayıp değer seçeneğini ayarlayarak başarılabilir. Hava kalitesi verisi için 9999 ile solar radyasyon ve ozone ppb yi 999 ile değişkenlerin kayıp değerlerini kodlayacağız. Bu kodlar değişkenler gözlenmediği için kabul edilebilir. Şimdi bu veri kümesi için etki çıktısının listesine bakalım:

Tablo 4.7. Hava Kalite Verisi: Tahmini Yanıt ve Standartlaştırılmış Sonuçlarla EKK Modeli

Değişken	Katsayı	Standart Hata	t Değeri
Solar Rad.	-0,01868	-0,03628	-0,51502
Rüzgar Hızı	-1,99577	1,14092	-1,74926
Sıcaklık	1,96332	0,66368	2,95823
Sabit Değer	-79,99270	46,81654	-1,70864

İndeks	Tahmini "Ozon ppb"	Standartlaştırılmış Rezidüler
1	33,231	0,43
2	43,195	-0,40
3	37,362	-1,41
4	12,933	0,28
7	24,873	-0,10
8	6,452	0,70
9	-0,700	0,48
12	31,334	-0,85
13	25,807	-0,82
14	26,639	-0,70
15	6,321	0,65
16	16,468	-0,14
17	19,901	0,78
18	-6,263	0,68
19	24,545	0,30
20	21,552	-0,59
21	16,335	-0,85
22	24,221	-0,73
23	19,944	-0,89
24	14,101	0,99
28	27,357	-0,24
29	44,591	0,02
30	59,567	3,08
31	49,238	-0,68

Azaltılmış veri kümesinin EKK analizi Tablo 4.7 de yazılmıştır. Modellenmiş ozone ppb değerlerinin bazıları (9 ve 18) negatiftir ki bu fiziksel olarak imkansızdır.

Bu modelin tuhaf davranışı Tablo 4.8 deki robust analizi ile kıyaslandığında kolayca anlaşılır. Bunlardan ilki her iki modelin denklemi esasen birbirinden farklıdır. Tablo 4.8 deki standartlaştırılmış rezidüleri içine alan sütunda 30 durumu bir sapan olarak çıkar. Bu tek sapan EKK hiperdüzlemini eğdiği için EKK modelinin kötü bir durumudur. Bu yüzde diğer noktalar EKK tarafından artık iyi modellenemiyor.

Elbette negatif tahmini değerler lineer denklem her ne zaman modellenirse beklenilebilir. Gerçekten de regresyon yüzeyi yatay olmadığı zaman tahmini

Tablo 4.8. Hava Kalitesi Verisi: LMS ye Dayalı RLS Modeli

Değişken	Katsayı	Standart Hata	t Değeri
Solar Rad.	0,00559	0,02213	0,25255
Rüzgar Hızı	-0,74884	0,71492	-1,04745
Sıcaklık	0,99352	0,42928	2,31438
Sabit Değer	-37,51613	28,95417	-1,29571

İndeks	Tahmini “Ozon ppb”	Standartlaştırılmış Rezidüler
1	24,571	1,52
2	28,686	0,68
3	27,402	-1,42
4	17,220	0,07
7	22,294	0,07
8	11,321	0,71
9	8,143	-0,01
12	25,204	-0,85
13	22,788	-1,09
14	23,413	-0,87
15	10,587	0,69
16	19,325	-0,49
17	20,786	1,22
18	5,772	0,02
19	23,232	0,63
20	17,065	-0,56
21	13,883	-1,19
22	24,369	-1,24
23	15,965	-1,11
24	14,617	1,61
28	20,137	0,26
29	33,210	1,09
30	37,950	7,13
31	34,010	0,28

$\hat{y}$ negatif olduğunda x vektörleri daima vardır. Bir x_1 kaldıraç noktasında negatif bir $\hat{y}_i$ ile kolayca karşılaşılabilecek robust regresyonu hala doğrudur. Fakat yukarıdaki örnekte EKK tahminlerinin iyi gözlemlerinde negatifdir.

4.3.Örnekler

Önceki bölümlerde yüksek kırılmalı modellerin daha fazla düşünülmesi gerektiğini gördük. Özellikle robust modele bağlı standartlaştırılmış rezidüler güçlü diagnostik araçlar verdi. Örneğin rezidü grafiklerinde görülebilir. Bu grafikler uzaktaki sapan değerlerin bulunmasını ve modelin başarısızlığına dikkat çekmeyi kolaylaştırır. Yine de rezidüler her nokta için bir ağırlık tanımlama da kullanılabilir. Böyle ağırlıklar RLS için sapanların etkisini sınırlamayı mümkün kılar. Yeniden ağırlıklandırılmayla gelen model verinin çoğunluğu tarafından takip edilen eğimi tanımlar. Modele bağlı istatistikler (F ve T değerleri gibi) EKK regresyonuyla hesaplananlardan daha güvenilirdir.

Örnek 4.3.1: Hawkins – Bradu – Kass verisi

Robust tekniğinin meziyetlerinin bazılarını göstermek için Hawkins, Bradu ve Kass (1984) tarafından üretilen veriyi kullandı. Böyle yapay bir veri en azından kötü noktaların durumunun tam olarak bilinmesi avantajını sunar. Burada gerçek veri analizlerinde doğal oluşlarından dolayı tartışmaların bazılarından kaçınılır. Bu şekilde bu yöntemin etkililiği ölçülebilir.

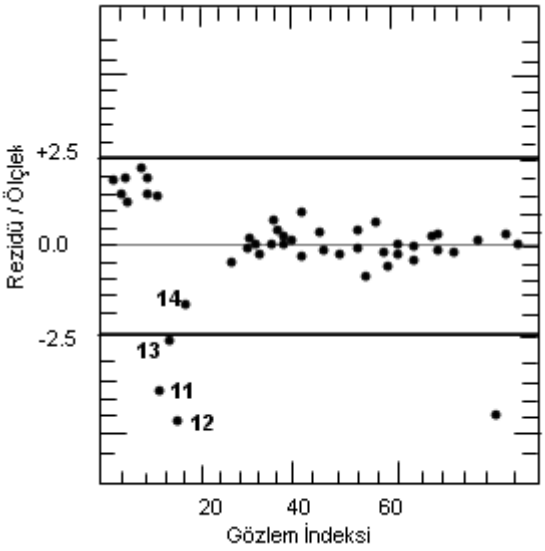
Veri kümesi tablo 9 da listeleniyor ve 4 ölçekli 75 gözlemden oluşuyor. (1 yanıt ve 3 açıklayıcı değişken) ilk 10 gözlem kötü kaldıraç noktalarıdır. (x_i leri sapandır fakat karşılık gelen y_i ler modele oldukça iyi uyar). Biz robust regresyonu ile EKK 'yı kıyaslayacağız. Hawkins (1984) de, M tahmin edicilerinin beklenen sonuçları üretemediğinden bahsediliyor. Çünkü sapanlar (kötü kaldıraç noktaları) gizlenildi ve 4 iyi kaldıraç noktası modele göre geniş rezidülere sahip olduğu için sapan olarak görüldü. Bu bizi şaşırtmamalı çünkü M tahmin edicileri kaldıraç noktaları olduğundan dolayı erken bozulur. Hawkins in temel kümelerin belli bir

Tablo 4.9. Hawkins, Bradu Ve Kass ın Yapay Veri Kümesi

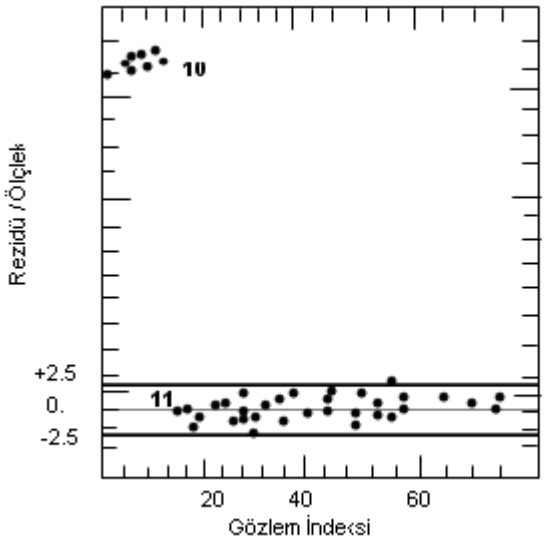
İndeks	x_1	x_2	x_3	y	İndeks	x_1	x_2	x_3	y
1	10,1	19,6	28,3	9,7	39	2,1	0,0	1,2	-0,7
2	9,5	20,5	28,9	10,1	40	0,5	2,0	1,2	-0,5
3	10,7	20,2	31,0	10,3	41	3,4	1,6	2,9	-0,1
4	9,9	21,5	31,7	9,5	42	0,3	1,0	2,7	-0,7
5	10,3	21,1	31,1	10,0	43	0,1	3,3	0,9	0,6
6	10,8	20,4	29,2	10,0	44	1,8	0,5	3,2	-0,7
7	10,5	20,9	29,1	10,8	45	1,9	0,1	0,6	-0,5
8	9,9	19,6	28,8	10,3	46	1,8	0,5	3,0	-0,4
9	9,7	20,7	31,0	9,6	47	3,0	0,1	0,8	-0,9
10	9,3	19,7	30,3	9,9	48	3,1	1,6	3,0	0,1
11	11,0	24,0	35,0	-0,2	49	3,1	2,5	1,9	0,9
12	12,0	23,0	37,0	-0,4	50	2,1	2,8	2,9	-0,4
13	12,0	26,0	34,0	0,7	51	2,3	1,5	0,4	0,7
14	11,0	34,0	34,0	0,1	52	3,3	0,6	1,2	-0,5
15	3,4	2,9	2,1	-0,4	53	0,3	0,4	3,3	0,7
16	3,1	2,2	0,3	0,6	54	1,1	3,0	0,3	0,7
17	0,0	1,6	0,2	-0,2	55	0,5	2,4	0,9	0,0
18	2,3	1,6	2,0	0,0	56	1,8	3,2	0,9	0,1
19	0,8	2,9	1,6	0,1	57	1,8	0,7	0,7	0,7
20	3,1	3,4	2,2	0,4	58	2,4	3,4	1,5	-0,1
21	2,6	2,2	1,9	0,9	59	1,6	2,1	3,0	-0,3
22	0,4	3,2	1,9	0,3	60	0,3	1,5	3,3	-0,9
23	2,0	2,3	0,8	-0,8	61	0,4	3,4	3,0	-0,3
24	1,3	2,3	0,5	0,7	62	0,9	0,1	0,3	0,6
25	1,0	0,0	0,4	-0,3	63	1,1	2,7	0,2	-0,3
26	0,9	3,3	2,5	-0,8	64	2,8	3,0	2,9	-0,5
27	3,3	2,5	2,9	-0,7	65	2,0	0,7	2,7	0,6
28	1,8	0,8	2,0	0,3	66	0,2	1,8	0,8	-0,9
29	1,2	0,9	0,8	0,3	67	1,6	2,0	1,2	-0,7
30	1,2	0,7	3,4	-0,3	68	0,1	0,0	1,1	0,6
31	3,1	1,4	1,0	0,0	69	2,0	0,6	0,3	0,2
32	0,5	2,4	0,3	-0,4	70	1,0	2,2	2,9	0,7
33	1,5	3,1	1,5	-0,6	71	2,2	2,5	2,3	0,2
34	0,4	0,0	0,7	-0,7	72	0,6	2,0	1,5	-0,2
35	3,1	2,4	3,0	0,3	73	0,3	1,7	2,2	0,4
36	1,1	2,2	2,7	-1,0	74	0,0	2,2	1,6	-0,9
37	0,1	3,0	2,6	-0,6	75	0,3	0,4	2,6	0,2
38	1,5	1,2	0,2	0,9					

değişimi diagnostikçi sapanların yerini tam olarak belirler fakat bu teknik kirlenmenin daha geniş bir kırılması ile başa çıkamaz.

Şimdi EKK ve LMS yi tartışmayla sınırlayalım EKK ya bağlı indeks grafiğinden (Şekil 4.5 e bakınız) 72.5 şeridinin dışına düştüğü için 11., 12. ve 13.



Şekil 4.5. Hawkins-Bradu-Kass verisi: EKK regresyonuna göre indeks grafiği



Şekil 4.6. Hawkins-Bradu-Kass verisi: LMS regresyonuna göre indeks grafiği

gözlemlerin sapan olduğu görülüyor. Maalesef önceki modelden bunların iyi gözlem olduğu biliniyor. Kötü kaldıraç noktaları EKK modelini yönünden tamamen çıkarıp eğerler. Bu nedenle ilk 10 nokta küçük standartlaştırılmış EKK rezidülerine sahiptir. (M tahmin edicilerine göre indeks grafikleri EKK nın kinden çok daha küçüktür.) Diğer taraftan LMS nin indeks grafiği (Şekil 4.6) ilk 10 noktayı etkili gözlemler

olarak açıklıyor. 4 iyi kaldıraç noktası 0 dolaylarında kesikli doğrunun komşuluğuna düşer. Yani bu noktalar LMS modeline iyi uymuşlardır. Açıkçası LMS indeks grafiğinden çizilen sonuç verinin yapısıyla uyuşuyor.

Diğer örnek model çeşitliliği için rezidü grafiklerinin kullanımını gösteriyor.

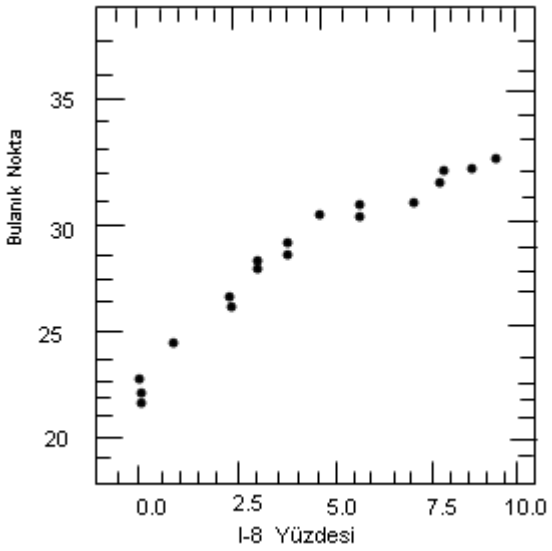
Örnek4.3.2: Bulanık nokta verisi

Tablo 4.10 bir sıvının bulanıklık noktasıyla alakalı ölçümlerin bir kümesini gösteriyor. Bu bulanık nokta bir yığındaki kristalleştirme seviyesinin bir ölçüsüdür ve kırıcı indeksle ölçülebilir.

Tablo 4.10. Bir sıvının bulanıklık verisi

İndeks (i)	I-8 oranı (x)	Bulanıklık noktası (y)
1	0	22,1
2	1	24,5
3	2	26,0
4	3	26,8
5	4	28,2
6	5	28,9
7	6	30,0
8	7	30,4
9	8	31,4
10	0	21,9
11	2	26,1
12	4	28,5
13	6	30,3
14	8	31,5
15	10	33,1
16	0	22,8
17	3	27,3
18	6	29,8
19	9	31,8

Kaynak: Draper ve Smith (1969,s.162)



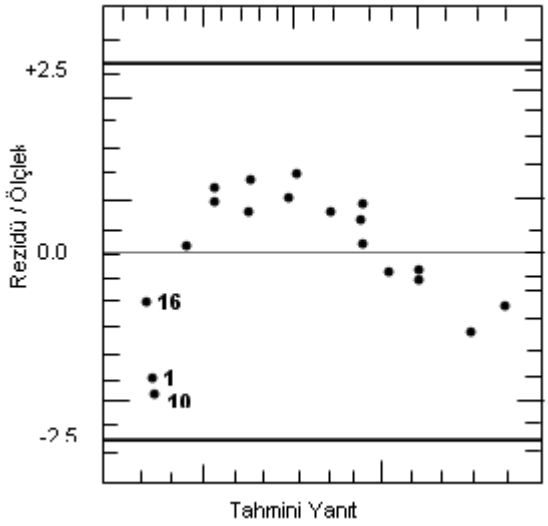
Şekil4.7. Bulanık nokta verisi: saçılım grafiği

Burada amaç bulanık nokta için bir tahmin olarak kullanılabilen ana yığında I-8 in yüzdesinin olduğu bir model oluşturmaktır. Veri sadece iki değişken içerdiği için saçılım grafiğinde bu değişkenler arasındaki ilişkiyi araştırması mümkündür. Tablo 4.10 daki saçılım grafiği Şekil 4.7 den bulunabilir.

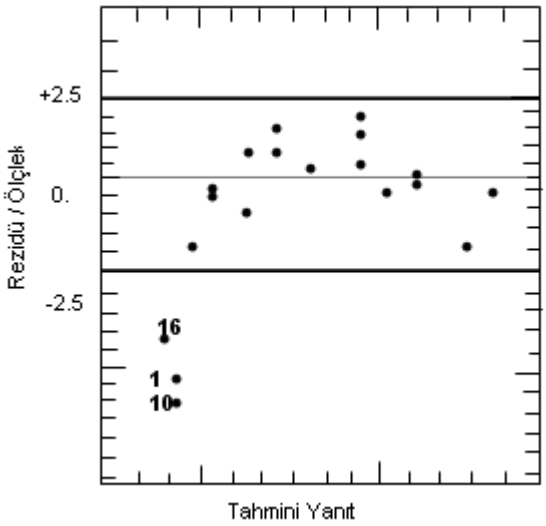
Şekil 4.7 den eğri/kavisli modelde basit lineer bir modelin yeterli olmadığı gösteriliyor. Şimdi lineer modeldeki rezidü grafiğinin bunu önerip önermediğini inceleyeceğiz. Şekil 4.8 deki EKK doğrusunun rezidü grafiğinden rezidülerin sıfır doğrusu civarında rasgele dağılmadığı görüldüğü için doğru modelinin kusurlu olduğu açıktır

Bu modelin sapanların varlığına sebep olmadığına emin olmak için RLS doğrusundaki (Şekil 4.9) rezidü grafiği şekil8 ile kıyaslayacağız. Lineerlikten ayrılış bu grafikte büyüdü ve $\mp 2{,}5$ şeridi içindeki rezidüler parabolik takip etmeye eğilimlidir. Durum 1,10 ve 16 ya bağlı olanlar şeridin dışına düşerler ki bu RLS doğrusundan dolayı geniş rezidülere sahip olduğunu gösterir. EKK ve RLS doğrularının her ikisinin eğimlerinin birbirinden önemli derecede farklı olması gerçeğine rağmen (Tablo 4.10 da ki t-değerlerine bakınız.) lineer model uygun değildir.

Bundan başka modelin yeterlilik ölçümü olan R^2 değeri yüksektir. EKK için R^2 =0.955 tir. RLS için ise R^2 =0.977 dir. Bu da gösteriyor ki yanıt değişkendeki toplam değişimin %97.7 sini açıklayıcı değişken açıklıyor. Elbette R^2 sadece tek modelin fonksiyonel biçimdeki bütün olası bütün kusurlarını gösteremez. Geniş bir R^2 verinin iyi modellenmesini garantilemez. Diğer yandan rezidü grafikleri modelin fonksiyonel yanını somutlaştırmaz. Bu nedenle bu grafiklerin gözden geçirilmesi R^2 geniş ya da t-değerlerinin önemliliği durumunda bile regresyon analizinin zorunlu bir parçasıdır. Bu grafiksel gösterim beklenmedik bir model ortaya çıkardığında modele uyarak hasarı onarmalıdır. Rezidülerin örneğinde aykırılıklara bağlı olduğunda açıklayıcı değişken ilave edilebilir veya modeldeki bazı değişkenler çıkarılabilir. Bütün düzeltilmiş modeller katsayıların lineerliklerini sınırlandırmak zorundadır ya da lineer regresyon kısmından çıkarırız. Fakat bazı lineer olmayan fonksiyonlar uygun bir dönüşüm kullanılarak lineerleştirilebilir.



Şekil 4.8. Bulanık nokta verisi. EKK ‘ye göre rezidü grafiği



Şekil.4.9. Bulanık nokta verisi: LMS 'e dayalı LMS 'e göre rezidü grafiği

Tablo 4.11. Bulanık Nokta Verisi: EKK ve RLS Regresyonuyla Tahmini Eğim ve Kesen ile t Değerleri

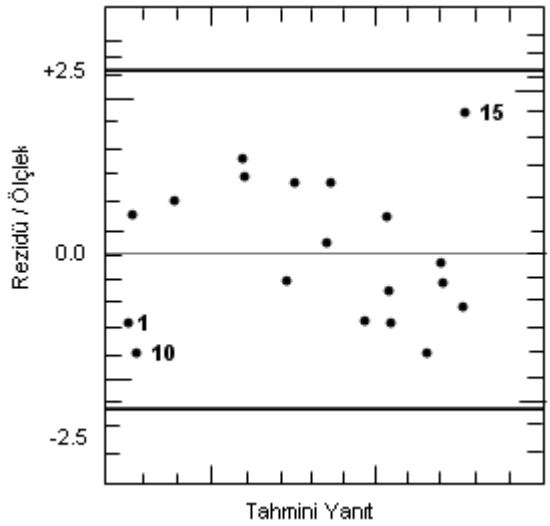
	EKK Sonuçları			RLS Sonuçları		
Değişken	$\hat{\theta}$	Standart Hata	t-Değerleri	$\hat{\theta}$	Standart Hata	t-Değerleri
I-8 Yüzdesi	1,05	0,055	18,9	0,89	0,036	24,7
Sabit	23,35	0,297	78,7	24,37	0,212	115,2
	$R^2 = 0{,}955$			$R^2 = 0{,}977$		

Şekil 4.9 gibi bir eğri grafiği nonlineerliği göstermek için tek yoldur. Bu görünürdeki eğriliği hesaplamak için alışılmış yaklaşım

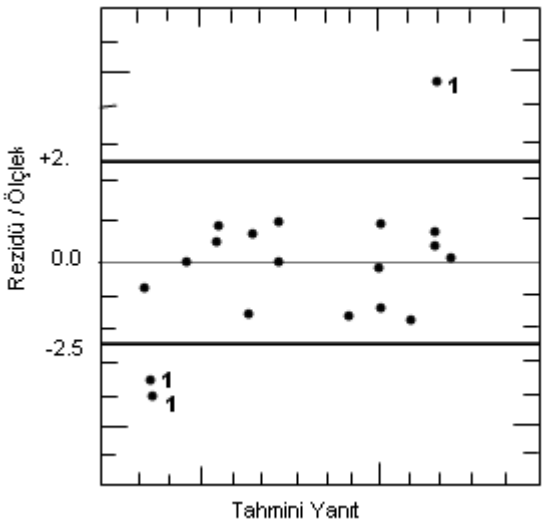
$$y = \theta_1 x + \theta_2 x^2 + \theta_3 + e \tag{4.2}$$

Şeklinde bir ikinci dereceden modelin kullanımıdır. Bu model için EKK ve RLS tahmin edicileri bazı özet istatistikleri boyunca Tablo 4.12 de veriliyor.

EKK ve RLS nin her ikisinin R^2 değeri bir parça artıyor. Buna karşılık regresyon katsayılarının t-değerleri hemen hemen değişmedi. Her iki model için katsayılar $\alpha = \%5$ seviyesinde sıfırdan önemli derecede farklıdır. EKK ve RLS ye göre rezidülerin dağılımına Şekil 4.10 ve Şekil 4.11 de bakabiliriz.



Şekil 4.10. Bulanık nokta verisi: Kuadratik model için EKK rezidü grafiği



Şekil 4.11. Bulanık nokta verisi: Kuadratik model için RLS rezidü grafiği

EKK rezidü grafiği incelendiğinde eklenmiş karesi alınmış terimin rezidü dağılımının onarılması tamamen başarılamadı. Model tamamen yansız değildir.

Modeli tekrar değiştirmeye karar vermeden önce RLS rezidü grafiğinden sağlanan bilgiyi analiz edelim. Bu grafikten bazı sapanlar göze çarpar.1,10 ve 15 gözlemleri yatay şeridin dışında kalıp LMS modelinden sıfır ağırlığını elde eder.Sıfır ağırlıklarına rağmen sapanlar grafikte hal görünür.Bu noktaların varlığı EKK tahmin edicilerini etkiler çünkü EKK bütün rezidüleri (sapanların rezidülerini bile) θ_1 in tahmininin artması pahasına küçültmeye çalışır.Diğer taraftan uzaktaki gözlemler geniş bir rezidü elde eden RLS tahmin edicilerine göre hareket etmez.Diğer gözlemlere göre Şekil 4.11 de şeridin içine düşen rezidüler değişimin sistematik bir örneğini göstermiyor. Şekil 4.10 ve Şekil 4.11 den varılan özetle ikinci dereceden denklemin basit bir doğrudan başka uygun bir model tanımladığı sonucuna varılır.Bundan başka RLS ikinci dereceden modelin rezidü grafiği 3 gözlem değeri gösterir.Bunlar EKK ikinci dereceden modelinin rezidü grafiğinde biçimi bozan değerlerdir.

Sapanların varlığı veya uygun bir modelin seçimi regresyon analizinin tek problemi değildir. İki veya daha fazla değişken arasındaki hemen hemen lineer bağımlılık tahmin edilmiş regresyon görüntüsünü ciddi bir biçimde de bozabilir ya da regresyon katsayılarını açıklayamayabilir. Bu olay çoklu lineerlik(multicollinearity)diye adlandırılır. Faktör uzayında değişkenler arasında ilişki olmaması ideal bir durum olacaktır. Bu durumda karşılık gelen diğer değişkenler sabitken açıklayıcı değişken tek bir genişlediğinde yanıttaki değişimin miktarı için bir regresyon katsayısını yorumlamak kolaydır.Çoklu lineerlik veride mevcut olduğunda regresyon denklemine tek bir açıklayıcı değişkenin katkısını tahmin etmek zordur.

Çoklu lineerliği saptamak çok karışık olabilir. Sadece 2 değişken olduğunda o zaman çoklu lineerlik Pearson korelasyonunun yüksek mutlak değerine ya da alternatif(daha robust) Spearman rank korelasyon katsayısına götürür. Daha yüksek ölçekli faktör uzaylarında çoklu lineerlik birkaç değişkeni içerebildiği için ikişer ikişer korelasyon katsayıları çoklu lineerliği keşfetmek için daima yeterli değildir. Bu yüzden kalan x değerleri üzerinde x_j nin regresyonunun karesi alınmış çoklu regresyon katsayıları R_j^2,diğer açıklayıcı değişkenlere bağlı herhangi x_j nin derecesini ölçmek için diagnostikler mümkündür.

Tablo 4.12. Bulanık Nokta Verisi: Özet Değerlerle 2. Dereceden Model İçin EKK ve RLS Tahminleri

	EKK Sonuçları			RLS Sonuçları		
Değişken Değerleri	$\hat{\theta}$	Standart Hata	t-Değerleri	$\hat{\theta}$	Standart Hata	t-Değerleri
I-8 Yüzdesi	1,67	0,099	16,9	1,57	0,084	18,8
(I-8 Yüzdesi)^2	-0,07	0,010	-6,6	-0,07	0,009	-7,5
Sabit	22,56	0,198	113,7	22,99	0,172	133,7
	$R^2 = 0{,}955$			$R^2 = 0{,}977$		

Örnek4.3.3: Kalp sondası verisi

Tablo 4.13 teki kalp sondası verisi çoklu lineerliğin sebep olduğu etkisinin belli bir kısmını gösterir. Sonda bütün kalça kemiği bölgesinde toplardamarın veya atardamarın içinden geçiriliyor ve kalbin içine hareket ediyor. Sonda kalp fonksiyonuyla ilgili bilgi sağlamak için özel bir bölge içinde hareket edebiliyor. Bu teknik bazen doğuştan kalbi kusurlu çocuklara uygulanır. Söylenilen bu sondanın uygun uzunluğu doktorlar tarafından belirlenmelidir. 12 çocuk için uygun sonda uzunluğu(y) fluoroskop(röntgen perdesi) ile kontrol edilerek tanımlanır. Bu alet sonda ucuna uygun durum için ulaştırılır.

Sonda uzunluğu ve hastanın boyu(x_1) ve hastanın kilosu (x_2) arasındaki ilişkiyi tanımlamak amacımızdır. RLS deki hesaplar gibi EKK nın ki için de model

$$y = \theta_1 x_1 + \theta_2 x_2 \ + \theta_3 + e$$

Tablo 4.14 'te verilmiştir.

Tablo 4.13. Kalp sondası verisi

İndeks (i)	Boy (x_1)	Ağırlık (x_2)	Sonda uzunluğu (y)
1	42,8	40,0	37
2	63,5	93,5	50
3	37,5	35,5	34
4	39,5	30,0	36
5	45,5	52,0	43
6	38,5	17,0	28
7	43,0	38,5	37
8	22,5	8,5	20
9	37,0	33,0	34
10	23,5	9,5	30
11	33,0	21,0	38
12	58,0	79,0	47

Hastanın boyu inç, ağırlığı libre(453 gr)ve sonda uzunluğu santimetre olarak verilmiştir.
Kaynak: Weisberg(1980,s.218)

Her iki regresyon için F-değerinin genişliği yeterince büyüktür. Bu da $\hat{\theta}_1$ ve $\hat{\theta}_2$ nın birlikte yanıt değişkeninin tahminine katkısı vardır sonucuna götürür. Her regresyon katsayısı için t-testine bakılırsa EKK nın $\hat{\theta}_1$ $\alpha = \%5$ anlamlılık seviyesinde sıfırdan önemli derecede farklı değildir. Aynı şey $\hat{\theta}_2$ için de söylenilebilir. Fakat F-istatistiğinde tümüyle görülen iki x değişkeninin önemli olduğu görülür. RLS modeli için $\hat{\theta}_1$ sıfırdan önemli derecede henüz farklı değildir. Yani karşılık gelen açıklayıcı

Tablo 4.14. Kalp Sondası Verisi: EKK ve RLS Sonuçları

	EKK Sonuçları			RLS Sonuçları		
Değişken	$\hat{\theta}$	t - değeri	p -Değeri	$\hat{\theta}$	t - değeri	p - Değerleri
Yükseklik	0,211	0,6099	0,5570	-0,723	-2,0194	0,0900
Ağırlık	0,191	1,2074	0,2581	0,514	3,7150	0,0099
Sabit	20,38	2,4298	0,0380	48,02	4,7929	0,0030
	$R^2 = 21{,}267$ (p=0,0004)			$R^2 = 32{,}086$ (p=0,0006)		

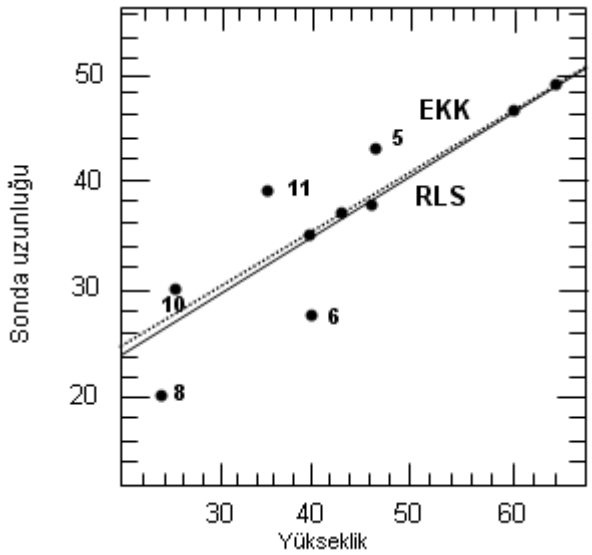
değişkenin modele çok az katkısı vardır. Üstelik $\hat{\theta}_1$ işaretinin metinde bir anlamı yoktur. Böyle olaylar genelde çoklu lineerliğin olduğu durumlarda meydana gelir. Gerçekten Tablo 4.15 değişkenler arasındaki yüksek ilişkiyi gösteriyor.

Boy ve kilo arasındaki korelasyon çok yüksektir. Burada her iki değişken neredeyse tamamen diğerini açıklar. Bundan başka regresyon katsayıları için düşük t-değeri doğruluyor ki açıklayıcı değişkenlerden herhangi biri modelde ihmal edilebilir. Kalp sondası verisi için biz ağırlık değişkenini atacağız ve tepedeki sonda uzunluğunun basit regresyonuna bakacağız. Bu iki ölçekli veri kümesinin saçılım grafiği EKK ve RLS modelleriyle Şekil 4.12 de veriliyor.

EKK için $\hat{y} = 0{,}612x + 11{,}48$ (şekil 12 de kesikli çizgi ile veriliyor.)

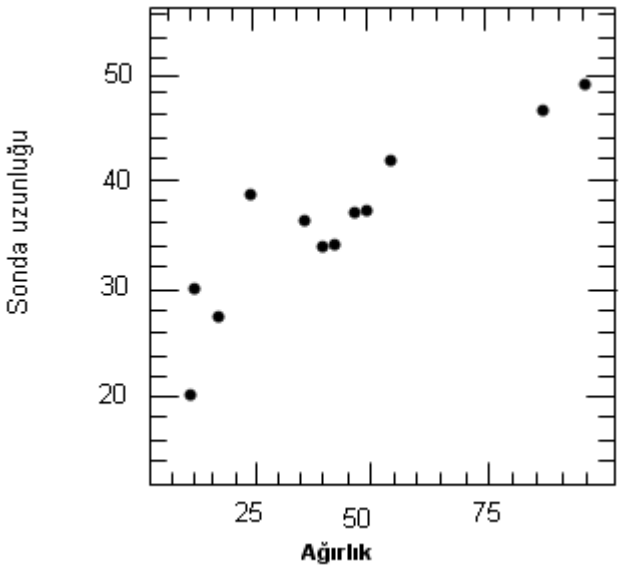
RLS için $\hat{y} = 0{,}614x + 11{,}11$ ($\hat{\theta}_1$ şimdi pozitiftir.) EKK a yakın uzanıyor.

RLS modeline nazaran uzakta kalan noktalar 5, 6, 8,10 ve 11 dir. Çünkü RLS doğrusunun yukarısında ve aşağısında tamamen dengelenmişler. EKK ve RLS doğruları bu örnekte aşikar bir biçimde farklı değillerdir.



Şekil 4.12. Kalp sondası verisi. 12 çocuğun boylarına karşı uygun sondanın uzunluğu

Alternatif seçim, boy yerine ağırlığı kullanmak olacaktır.Sonda uzunluklarının ölçümlerinin hastanın ağırlığına karşı olan grafik Şekil 4.13 te gösteriliyor. Bu saçılım grafiği bir lineer model yaklaşımını akla getirmiyor.Bir dönüşümün gerekliliği açıktır. Fiziksel zeminde lineerliği elde etmek amacıyla ağırlığın kendisi yerine ağırlığın küp kökünü kullanmaya çalışabiliriz. Hatta çoklu lineerlik için kanıt olmadığı zaman açıklayıcı değişkenlerin tüm kümesini modelde tamamen kullanarak çok geniş yapılabilir.İlk önce tanımlanan işleyiş biçimini anlatmak için çok sayıda değişken kullanmak bunu zorlaştırır.Bununla birlikte bu durum mümkün olduğu kadar yanıttaki değişimi açıklamadaki objektif ile birleşmesi gerekir.



Şekil 4.13. Kalp sonda verisi: 12 çocuk için ağırlığa karşı uygun sonda uzunluğunun saçılım grafiği

Bu ise daha fazla açıklayıcı değişkenin önemine götürür.Fakat istatistiksel yönden modelin tamlığını iyileştirmek için bazen değişkenlerin azaltılmasını söyleyebiliriz. Gerçekten sıfırdan önemli derecede farklı olmayan regresyon katsayılarına karşılık gelen açıklayıcı değişkenler tahminlerin varyansını arttırabilir. Diğer taraftan denklemlerden en uygununa karar vermek için öznel olay bilgisi yeterlidir. Diğer taraftan değişkenlerin uygun bir altkümesini bulmak için istatistiksel bir yöntem uygulamak zorundadır.Bu problem değişken seçimi diye adlandırılır. Bu

ise model çeşitliliği problemine sıkı sıkıya bağlıdır. Gerçekten de değişken seçimi problemi denkleme hangi biçimde açıklayıcı değişken girmesi gerekir(karesi alınmış terim,logaritmik terim, … , vb.) sorusunu içerir.

Tablo 4.15. Kalp Sonda Verisi:Değişkenler Arasındaki Korelasyon

	Pearson Korelasyon Katsayısıları		
Yükseklik	1,00		
Ağırlık	0,96	1,00	
Sonda uzunluğu	0,89	0,90	1,00
	Yükseklik	Ağırlık	Sonda uzunluğu
	Spearman Rank Korelasyon Katsayıları		
Yükseklik	1,00		
Ağırlık	0,92	1,00	
Sonda uzunluğu	0,80	0,86	1,00
	Yükseklik	Ağırlık	Sonda uzunluğu

Basit olarak her iki problemin eş zamanlı olarak işleyişinden ibaret fakat çözülmesi zor olan "ideal" yaklaşımından kaçınmak amacıyla ayrı olarak bu problemlere genelde bakılır. İster y de ister x te olan sapanların varlığı çok olsa bile güçleştiriyor. Bu yüzden gözlemlerin ağırlıklarını açıklamak amacıyla değişkenlerin tüm kümesinde yüksek kırılmalı regresyon analizinin başlamasını tavsiye edelim. O zaman ilk yaklaşım olarak temizlenmiş örnekte değişken seçimi için klasik bir yöntem kullanılabilir. Böyle değişken seçim yöntemleri literatürde geniş bir biçimde araştırılmış. Sonda değişkenleri kümesi çok geniş olduğu zaman bu kümeden başlayan bütün olası modeller düşünülebilir. Bu ise tüm altkümeler regresyonu yöntemidir.Bu bizi

$$2^{mevcut \quad aciklayici \ \text{degisken} \quad sayisi}$$

kadar farklı modele götürür. Bu sayı çok hızlı bir biçimde artar. İlgi EKK tahmin edisi üzerine odaklansa bile(Ya da ilk kez yapılacak çok robust bir yöntemi varsayarak RLS) bu sayı tamamen uygulanamaz. Bu durum birçok insanı daha etkili algoritmalar geliştirmeye itti ve böylece başarılı altkümeler için EKK tahminlerini hesaplayarak sayısal yöntemlere başvurdular. Genelde bu yöntemler Gauss – Jordan

indirgemesini ya da süpürme (sweep) yöntemine dayanır. Her ne zaman denklemlerin modellenmesindeki çeşit birinin yok edilmesi ise en iyi modeli veren değişkenlerin alt kümesine karar vermek için bir kritere ihtiyaç vardır. Hocking (1976) ortalama karelenmiş hata, açıklayıcılık katsayısı ve C_p istatistiğini içeren bu amaçtaki farklı ölçümleri tarif etmiştir.

Bütün alt kümelerin araştırılmasının en iyi kümeyi üreteceği sezilmesine rağmen hesaplama maliyeti yüzünden kullanılan en yaygın yöntem değildir. Açıklayıcı değişkeni birer birer ekleyerek ya da silerek oluşan sözde stepwise (merdivensi) yöntemlerinin her hususta kullanılması tercih edilir. Ön seçim ve arka eleme merdivensi yöntemleri ve her ikisinin kombinasyonu birbirinden ayrılır. Bu çeşitlerdeki değişimler BMDP, SAS ve SPSS gibi birçok istatistiksel paketlerde uygulanır.

Ön seçimde açıklayıcı değişkenin yanıtla en iyi ilişki içinde olan değişken olduğu basit bir regresyonla başlanır. O zaman sonraki gelen her aşama için modele bir değişken eklenir. Herhangi bir aşamada seçilen x_j değişkeni sondalar arasında en geniş F_j oranını etek üretendir.

$$F_j = \frac{SSE_k - SSE_{k+(j)}}{\hat{\sigma}^2{}_{k+(j)}}$$

Şeklinde tanımlanır. Burada SSE_k k-terimli modele karşılık gelen rezidü hatalarının kareleri toplamıdır ve $SSE_{k+(j)}$ ise x_j nin eklendiği modele karşılık gelir. Eğer F_j önceden belirtilmiş değerden daha geniş ise x_j değişkenini denklem içerir. Bu önceden belirlenmiş değer genelde durdurucu kural olarak ifade edilir. Bu değer tam bir akışla çalışacak bir yöntemdir. Her genişliğin bir alt kümesinin elde edileceği bir yoldur. Bundan başka bu alt kümelerin en iyisini seçmek için bir kriter kullanılmalıdır.

Arkadaki seçim metotlarında (ters yönde çalışılan) bütün değişkenleri içeren bir denklemden başlar. Her aşamada en kötü değişken elenir Tabi en küçük F_j oranı üretildiğinde

örneğin x_j değişkeni şimdiki modelden elenecek bir sonda olacaktır (bu model k tane terimden oluşuyor).

$$F_j = \frac{SSE_{k-(j)} - SSE_k}{\hat{\sigma}_k^2}$$

Tekrar ileri kısım için olanlara benzer birkaç durdurucu kural önerildi. Efroymson (1960) her iki fikri birleştirdi: Metodu basit olarak ileri kısım tipiydi. Fakat her aşamada bir değişkenin elenmesi de mümkündü.

Bir değişkenin seçimi problemi ile karşı karşıya kalınmasından dolayı mevcut tekniklerle zayıf noktaların farkına varmak zorundayız. Örneğin merdivensi yöntemler verilen genişlikte mutlaka en iyi alt kümeyi vermez. Bundan başka bu teknikler pratikte sıklıkla yanlış yerde kullanılan açıklayıcı değişkenlerin sıralanışını içerir. Silinme veya dahil edilme basamağı çok aldatıcıdır. Çünkü arka elemedeki silinen ilk değişken (aynı şekilde ön seçimde ilk elenenler) mutlak bir biçimde en kötü (ya da en iyi) olmak zorunda değildir. Örneğin ön seçimde giren ilk değişken diğer değişkenlerin varlığında gerekli olabilir. Yine de ön ve arka merdivensi (stepwise) teknikleri değişkenlerin en iyi alt kümelerinden tamamen farklı sonuçlanabilir. Berg (1978) bütün alt kümeler regresyonunda merdivensi yöntemleri kıyasladı. Eğer ön eleme her alt küme genişliğinde tüm alt kümeler regresyonuyla aynı fikirdeyse o zaman arka elemede her alt küme genişliği için bütün alt kümeler regresyonuyla da aynı fikirde olacaktır bunun tersi de doğrudur. Tüm alt kümeler regresyonu yine de merdivensi tekniklerinin dezavantajlarından kaçmak için ideal bir yol değildir. Gerçekten bütün mümkün alt kümelerin değerlendirilmesi geniş bir biçimde kullanılan kritere bağlıdır. Bundan başka tüm alt kümeler yöntemi her alt küme genişliği için en iyi kümeyi üretmesine rağmen bütün populasyonda bu durum mutlaka olmayacaktır. O sadece bu örnekte en iyisi Gorman ve Toman (1966) tarafından yapılan gözlem değişken seçimi konusunu bitirmek için belki de uygundur: “ Tek bir en iyi alt küme olduğu farklıdır fakat az çok birkaç tane iyisi var.

Örnek 4.3.4: Eğitim Giderleri Verisi

Bu örnek bu örnek heteroscedasticity nin ilginç bir örneğini verir. Bu veri Amerika' daki 50 tane eyaletin eğitim harcamalarından oluşur. Veri Tablo 4.16 da yeniden oluşturuldu.

Tablo 4.16 daki y değişkeni bir eyaletteki halk eğitiminde kişi başına düşen giderdir (1975 için tasarlanmıştır). Burada amaç y yi x_1 açıklayıcı değişkeni (1970 te kente oturan her 1000 kişiye düşen ev sayısı), x_2 (1973 te kişi başına düşen gelir) ve x_3 (1974 te 18 yaşının altında her 1000 kişiye karşılık gelen ev sayısı).

Genelde bir gözlemin i indeksi zamana bağlıdır. Bu şekilde bir indeks grafiği örneği zamana karşı rezidülerin yayılmasının değişimini gösterir. Fakat rezidülerin büyüklüğü $\hat{y}_i$ ile veya bir açıklayıcı değişkenle veya zaman sırasından başka durumların diğer sıralamasıyla sistematik bir biçimde değiştirilebileceği görülür. Şimdiki veri kümesinin objektifliği durumların uzaydaki sıralaması ile ilişkilerin değişmezliğini analiz edecek. Tablo 4.16 daki veri coğrafi bölgelerle gruplandırılmıştır. 4 gruba ayrılır: Kuzeydoğu eyaletleri (1-9 ile gösterilir), kuzey merkez eyaletleri (10-21 ile gösterilir), güney eyaletleri (22-37 ile gösterilir) ve batı eyaletleri (38-50 ile gösterilir).

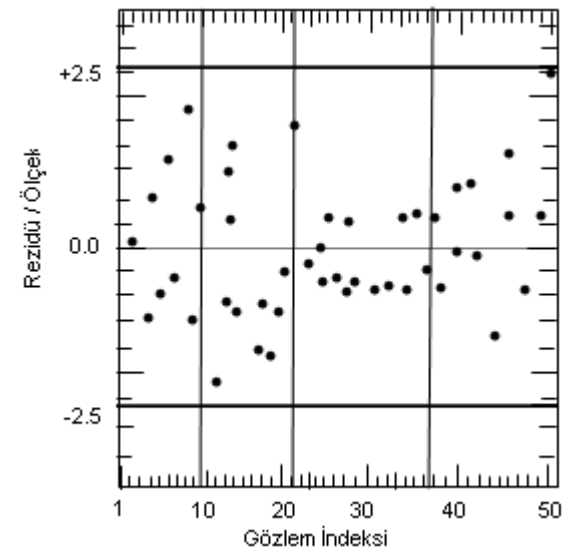
Bu veri kümesi için klasik olarak EKK başvurma Tablo 4.17 deki katsayıları verir. Burada bu katsayılar LMS ye dayalı yeniden ağırlıklandırılmış EKK nın sonuçlarını da içerir. Şimdi EKK indeks grafiği (Şekil 4.14a) ile RLS indeks grafiğini (Şekil 4.14b) kıyaslayalım.

Temel farkları durum 50 dir. Durum 50 (Alaska) RLS grafiğinde bir sapan olarak tanımlanırken aksine EKK grafiğinde açıkça göze çarpmıyor (Alaska' daki eğitim harcamalarının x_1, x_2 ve x_3 ün tek başlarına populasyon karakteristikleri esasında beklenenden çok daha yüksektir.). Diğer taraftan her iki grafikte farklı coğrafi bölgelerdeki değişikliklerin rezidülerin dağılımında görüldüğünü gösteriyor. Bu olay heteroscedasticity için bir tip örnektir. Buradaki eksiklik 4 ayrı sınıfı ele alarak iyileştirilebilir. Fakat o zaman durum sayısı bu örnek için çok sınırlı olur.

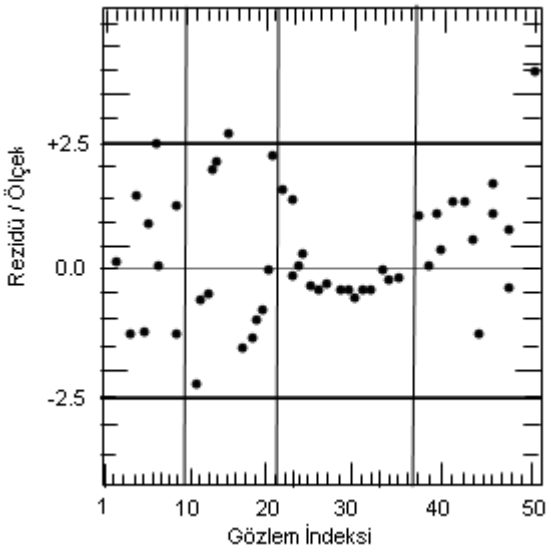
Tablo 4.16. Eğitim Giderleri Verisi KB: Kuzeybatı, KM: Kuzey Merkez, G: Güney, B:Batı

İndeks	Yer		x_1	x_2	x_3	y
1	ME		508	3944	325	235
2	NH		564	4578	323	231
3	VT		322	4011	328	270
4	MA	KB	846	5233	305	261
5	RI		871	4780	303	300
6	CT		774	5889	307	317
7	NY		856	5663	301	387
8	NJ		889	5759	310	285
9	PA		715	4894	300	300
10	OH		753	5012	324	221
11	IN		649	4908	329	264
12	H.		830	5753	320	308
13	MI		738	5439	337	379
14	WI		659	4634	328	342
15	MN	KM	664	4921	330	378
16	IA		572	4869	318	232
17	MO		701	4672	309	231
18	ND		443	4782	333	246
19	SD		446	4296	330	230
20	NB		615	4827	318	268
21	KS		661	5057	304	337
22	DE		722	5540	328	344
23	MD		766	5331	323	330
24	VA		631	4715	317	261
25	WV		390	3828	310	214
26	NC		450	4120	321	245
27	SC		476	3817	342	233
28	GA		603	4243	339	250
29	FL	GÜNEY	805	4647	287	243
30	DY		523	4967	325	216
31	TN		588	3946	315	212
32	AL		584	3724	332	208
33	MS		445	3448	358	215
34	AR		500	3680	320	221
35	LA		661	3825	355	244
36	OK		680	4189	306	234
37	TX		797	4336	335	269
38	MT		534	4418	335	302
39	HD		541	4323	344	268
40	WY		605	4813	331	323
41	CO		785	5046	324	304
42	NM		698	3764	366	317
43	AZ		796	4504	340	332
44	UT	BATI	804	4005	378	315
45	NV		809	5560	330	291
46	WA		726	4989	313	312
47	OR		671	4697	305	316
48	CA		909	5438	307	332
49	AK		831	5309	333	311
50	HI		484	5613	386	546

Kaynak: Chatterjee ve Price (1977)



(a)



(b)

Şekil 4.14. Eğitim harcamaları verisi: (a) EKK 'e göre indeks grafiği. (b) LMS 'e dayalı RLS 'e göre indeks grafiği

Chatterjee ve Price (1977) ağırlıklandırılmış EKK regresyonunun diğer bir çeşidini kullanarak bu veriyi analiz ettiler. Onlar karelenmiş rezidülerin ağırlıkları toplamını hesaplamak amacıyla 4 bölgenin her biri için ağırlıkları ayırdı. Bu ağırlıklar klasik EKK dan sonuçlanmış rezidülerin karelerinin ortalamasını (belli bir

bölge içindeki) kullanarak ilk eyalette tahmin edilir.Chatterjee ve Price de bir sapan olarak Alaska'yı düşündü ve ihmal etmeye karar verdi.

Tablo4.17. Eğitim Harcamaları Verisinin EKK ve RLS Sonuçları: Katsayılar, Standart Hatalar,ve t-Değerleri

	EKK Sonuçları			RLS Sonuçları		
Değişken	$\hat{\theta}$	Standart Hata	t-Değeri	$\hat{\theta}$	Standart Hata	t-Değeri
x_1	-0,004	0,051	-0,08	0,075	0,042	1,76
x_2	0,072	0,012	6,24	0,038	0,011	3,49
x_3	1,552	0,315	4,93	0,756	0,292	2,59
Sabit	-556,6	123,2	- 4,52	-197,3	117,5	-1,68

4.4. LMS, LTS ve S-Tahmin Edicilerinin Özellikleri

Bu bölüm bası teorik sonuçları içerir ve sadece robust regresyonu uygulanmasıyla ile ilgilenen kişiler tarafından atlanabilir ve matematiksel yönüyle ilgilenmez. LMS tahmin edicisiyle ilgili ilk kısım

$$\min_{\hat{\theta}} med_i^2 \tag{4.3}$$

ile verilir.

(p+1)-ölçekli satır vektörlerinin lineer uzayını ait n gözlem

$$(x_i\, y_i) = (x_{i1},..., x_{ip}, y_i)$$

olsun. Bilinmeyen θ parametresi p boyutlu $\theta_1,...,\theta_p)^t$ sütun vektörüdür. (istifini bozmayan) lineer model $N(0,\sigma^2)$ ye göre dağılımlı e_i için $y_i = x_i\theta + e_i$ yi gösterir. Her durumda bu bölümde x_i = 0 olduğu bütün gözlemlerin silindiği varsayılır. Çünkü θ için hiçbir bilgi verilmiyor. Eğer model kesere sahipse bu durumu otomatikman / kendiliğinden sağlar. Çünkü o zaman her x_i nin en son koordinatı 1 e

eşittir. Bundan başka $(x_i y_i)$ nin (p+1) boyutlu uzayında n / 2 gözlemden daha fazlasını içeren 0 dan dolayı düşey hiper düzlem yoktur. (Böyle bir hiperdüzlem (0,…,0) ve (0,…,0,1) i içeren p ölçekli bir alt uzaydır. Bu alt uzaya hiper düzlem diyeceğiz. Bu tüm uzayın boyutundan daha azdır. [q] notasyonu q ya eşit ya da daha az olan en büyük tam sayıya karşılık gelir.)

4.5. İzdüşüm Aramayla İlişki

İzdüşüm arama (PP) yöntemlerinin amacı daha düşük ölçekli uzayda bu verinin iz düşümünü alarak çok değişkenli veri kümesinde yapıyı keşfetmek. Böyle teknikler orijinal olarak Roy (1953), Kruskal (1969) ve Switzer (1970) tarafından bulundu. " İzdüşüm arama ismi Friedman ve Tukey (1974) tarafından bulundu. Onlar algoritmayı başarılı bir biçimde geliştirdiler. Esas problem iyi izdüşümler bulmaktır. Çünkü keyfi izdüşümler belirgin bir biçimde çok öğretici değildir. Friedman ve Stuetzle (1982) güçlü yapılarla oluşan noktaların bazı örneklerini verdiler. Burada yapının olmadığı yerde izdüşüme sahip olduğu görüldü. Bu nedenle PP, en ilginç olanını araştırma da belli bir objektif fonksiyonunu sayısal olarak en elverişli hale getirerek yüksek boyutlu nokta kümesinden düşük boyutlu bir çok izdüşüm dener (objektif fonksiyonu, izdüşüm indeksi olarak ta adlandırılır. Bazı önemli uygulamaları PP sınıflaması (Friedman ve Stuetzle, 1980), PP regresyonu (Friedman ve Stuetzle, 1981), robust temel bileşenleri (Ruymgaart 1981) ve PP yoğunluk tahmini (Friedman 1984).

Şimdi robust regresyonu ve PP arasındaki ilişkinin varlığını gösterelim. Buna bakmak için $(x_i y_i)$ nin (p + 1) boyutlu uzayını düşünelim. Burada x_i nin son bileşeni sabit terimli regresyon durumunda 1 e eşittir. Bu uzayda lineer modeller şu şekilde tanımlanır:

$$\blacklozenge, y\begin{pmatrix} 0 \\ -1 \end{pmatrix} = 0 \qquad (4.4)$$

belli bir p – ölçekli θ sütun vektörü için iyi bir $\hat{\theta}$ bulmak amacıyla herhangi bir θ için $(\theta,-1)^t$ e ortogonal şekilde y ekseninde nokta kümesini izdüşümleyerek başlarız. Yani her $(x_i\, y_i)$, $(\theta,\ r_i(\theta))$ üzerine izdüşümlenir. Burada $r_i(\theta) = y_i - x_i\theta$ izlenildiğinde

$$s(r_1(\theta),..., r_n(\theta)) \tag{4.5}$$

dağılımında böyle herhangi bir izdüşümün farklılığını ölçeriz. Burada s objektifi bir ölçek equivariant (eşit değişken)tır.

$$s(r_1,..., \tau r_n) = |\tau|\, s(r_1,..., r_n) \qquad \text{her } \tau \text{ için} \tag{4.6}$$

Fakat dönüşüm değişmez. PP tahmini $\hat{\theta}$, o zaman (4.5) indeksinin izdüşümünü minimumlaştırarak elde edilir. Eğer $s(r_1,...,r_n) = \left(\sum_{i=1}^{n} r_i^2 / n\right)^{1/2}$ ise bu en farklı $\hat{\theta}$ sade bir biçimde EKK katsayılarının vektörüdür.

Benzer bir biçimde $s(r_1,...,r_n) = \sum_{i=1}^{n} |r_i| / n$ ifadesi L_1 tahmin edicisini verir ve

$\sum_{i=1}^{n} |r_i|^q / n$ ifadesini minimumlaştırmak L_q tahmin edicisini (Gentleman 1965, Sposito 1977) verir. Çok robust olan S yi kullanmak yüksek kırılmalı regresyon tahmin edicilerini geri getirir: $s = (med_i r_i^2)^{1/2}$ LMS yi verir. $s = \left(\sum_{i=1}^{h} r^2_{in} / n\right)^{1/2}$ LTS yi verir ve ölçümün robust M tahmin edicisine eşit olan s yi koyarak S tahmin edicilerini elde ederiz. Dikkat edilmelidir ki (4.6) ü sağlayan herhangi bir s nin minimumlaştırılası regresyon, ölçek ve benzer eşit değişken (affine equivariant) olan bir regresyon tahmin edicisini verir. Bu nedenle PP ailesine ait regresyon tahmin edicilerinin tam bir sınıfının elde edildiği s yi değiştirerek iyi bir s regresyon tahmin edicisinin bir çeşidini tanımlar. Bu nedenle PP unsuru lineer regresyon modelini kapsamayı ve her iki klasik yöntem ve yüksek kırılmalı yöntemleri sormaya devam eder.

Şu ana kadar bilinen (LMS, LTS ve S – tahmin edicileri) sadece benzer eşit değişkenli yüksek kırılmalı regresyon tahmin edicilerinin PP ye bağlı olduğu görülür. (GM – tahmin edicileri yüksek kırılmalı değildir ve tekrarlı medyan benzer eşit değişkenli değildir.) Yüksek kırılma ve bu PP arasındaki görünen ilişkinin bir sebebi vardır. Gerçekten de Donoho, Rousseeuw ve Stahel tam model durumlarının yakın davranışlarıyla tanımlanan kırılma oranlarını buldular: Bunlar verinin çoğunluğunun regresyon hiperdüzleminde tam olarak uzanması durumlarıdır. Böyle genel durumlar kesin olarak verinin çoğunluğunun başarısızlığa uğradığı noktadaki izdüşüme sahiptirler. Diğer bir biçimde yüksek kırılma izdüşümlerin bazı özel çeşitlerinin olduğu durumlardaki bir tahmin edicinin davranışına bağlı olduğu görülür. Esas olarak PP böyle izdüşümleri araştırmada kullanılabildiği için yüksek kırılmalı yöntemlerin sentezlenmesinde PP nin kullanılışlığı şaşırtıcı değildir.

Fakat PP yüksek kırılmalı eşit değişkeni elde etmek için mecburi tek yol değildir. En azından Rousseeuw (1983) ün verinin en az yarısını içeren elipsoid minimal ses örneğinde çoklu durumda vardı. Yine de benzer eşit değişken tahmin edicisine dayanan her PP yüksek kırılmaya sahip olmayacaktır.

4.6.Robust Çoklu Regresyonuna Diğer Yaklaşımlar

Gerçekteki durumları nadiren karşılayan optimal EKK ölçüsünün altındaki durumları göreceğiz. Daha robust regresyon alternatiflerini tanımlamak amacıyla birçok istatistikçi çok büyük değerlerin basit meydanının dayanıklılığını çürüttü. Örneğin;L_1 ölçüsü tek değişkenli medyanın genellemesi olarak görülür.Çünkü $\sum_{i=1}^{n}\left|y_i-\hat{\theta}\right|$ nın minimumlaştırılması y_i nin n gözleminin meydanı olarak tanımlanır.Regresyon probleminde L_1 tahmin edicisi

$$\min_{\hat{\theta}} \sum_{i=1}^{n}|r_i| \tag{4.7}$$

olarak verilir.

Karesini alma yerine mutlak değeri koyma robustlıkta önemli bir kazanca götürür. Fakat kırılma noktasının terimlerinde L_1 gerçekten EKK dan daha iyi değildir. Çünkü L_1 ölçüsü y_i yönünde sapanlara karşı robusttır fakat hala kaldıraç noktaları yaralanabilir. Üstelik Wilson(1978) n artarken L_1 tahmin edicisinin etkisinin azaldığını gösterdi.Görüntünün sayısal bir noktasından (4.7) nin minimumlaştırılması lineer programın çözümüne eşittir:

$$\min_{\hat{\theta}} \sum_{i=1}^{n} (u_i + v_i)$$

sınırı altında

$$y_i = \sum_{k=1}^{p} \theta_k x_{ik} + u_i - v_i \quad , \quad u_i \geq 0 \text{ ve } v_i \geq 0$$

Rezidülerin $1 \leq q \leq 2$ için L_q normunun minimumlaştırılması Gentleman (1965), Forsythe (1972) ve Sposito (1977) tarafından düşünüldü. Bu kişiler doğrunun L_q modeli için bir algoritma verdiler. Dodge(1984) $L_1 - L_2$ normlarının konveks kombinasyonuna dayanan şu şekilde bir tahmin edici ileri sürdü:

$$\min_{\hat{\theta}} \sum_{i=1}^{n} \left((1-\delta)\frac{r_i^2}{2} + \delta |r_i| \right) \quad , \quad 0 \leq \delta \leq 1$$

Maalesef ki bütün bu önerilen tahmin ediciler sıfır kırılma noktasına sahiptir. Bölüm3 ün 6.kısmında medyana da dayalı basit regresyon için bazı tahmin ediciler sıralandı. Onların çoğu süpürme yöntemi aracılığıyla çoklu regresyonda üretiliyor

Theil in (1950) tahmin edicisinin arkasındaki fikir yayılmalar ve değişikliklerin kaynağıdır.Bu tahmin edici bütün çiftli eğimlerin meydanına bakmaktan oluşur.Son önerilen fikir Oja ve Niinimaa (1984) dan gelir.p indekslerini içeren {1,2,…,n} in J={ $i_1, i_2, ..., i_p$ } her altkümesi için

$$\theta_j = \theta(i_1, i_2, ..., i_p) \tag{4.8}$$

İfadesini tanımladılar. $(x_{i1}, y_{i1}), \ldots, (x_{ip}, y_{ip})$ p tane nokta ile tam olarak gelecek hiperdüzleme karşılık gelen parametre vektörü olarak. Bu θ_j lere yapay gözlemler olarak isimlendirdiler ve C_n^p leri vardır. Fikirleri şimdi bütün bu θ_j lerin(p boyutlu uzayda) çok değişkenli yer tahmin edicisini hesaplamaktır. θ_j nin belli ağırlıklandırılmış ortalaması EKK yı verir fakat elbette robust çok değişkenli bir tahmin edicinin içine koymak istiyor. Eğer

$$\hat{\theta}_j = \underset{j}{med}(\theta_j) \qquad \text{her j=1,…,p} \tag{4.9}$$

hesaplanırsa (Bütün J altkümeleri üzerine koordinatlanmış medyan) o zaman benzer eşit değişkenli olamayan bir regresyon tahmin edicisi elde edilir. Basit regresyon için (4.9) gerçekten Theil-Sen tahmin edicisini verir. Benzer eşit değişkenli regresyon tahmin edicisini elde etmek amacıyla çok değişkenli yer tahmin edici T ye başvurulur. Burada T,kendi kendine benzer eşit değişkenlidir. Yani,

$$T(z_1 A + b, \ldots, z_k A + b) = T(z_1, \ldots, z_k) A + b \tag{4.10}$$

p boyutlu satır vektörünün $\{z_1, \ldots, z_k\}$ herhangi bir örneği için singüler olmayan karesel A matrisi için ve herhangi p-boyutlu b vektörü için tanımlanır.

Bu amaçla genelleştirilmiş medyanı uygulamayı amaçladılar. Benzer eşit değişkenli olan Oja nın marifetli bir yapısıdır. Maalesef genelleştirilmiş medyanın hesaplanmasının karmaşıklığı çok büyüktür.(ve θ_j vektörlerinin çok geniş bir kümesini uygulamak zorundadır.)

(6.3) koordinatsal medyanı uygulandığında bile Oja-Niinimaa regresyon tahmin edicileri C_n^p yapay gözlemleri sayesinde hesaplama zamanı düşünülmelidir. Her iki durumda tüm θ_j leri düşünmek imkansızdır. Bu nedenle yapay gözlemleri rasgele bir alt popülasyonunu düşünmek kullanışlı olabilir.

Şimdi bu tekniğin kırılma noktasını düşünelim. Sadece $(x_{i1}, y_{i1}), ..., (x_{ip}, y_{ip})$ p noktanın hepsi iyi olduğunda $\theta_j = \theta(i_1, i_2, ..., i_p)$ yapay gözlemin iyi olduğuna emin olacağız. Orijinal veride sapanların bir ε kırılması varsa o zaman "iyi" yapay gözlemlerin $(1-\varepsilon)$ oranından sadece emin olacağız. Bu nedenle çok değişkenli yer tahmin edicisinin mümkün en iyi kırılma noktası %50 olduğu için

$$(1-\varepsilon) \geq \frac{1}{2} \tag{4.11}$$

e sahip olmalıyız. Yani "iyi" yapay gözlemlerin hala %50 sine ihtiyaç vardır. (4.11) formülü orijinal veri kümesinde izin verilen yani;

$$\varepsilon^* = 1 - \left(\frac{1}{2}\right)^{1/p} \tag{4.12}$$

Bozulmanın üst sınırını verir.

p=2 için Theil in tahmin edicisinin kırılma noktası tekrar bulunacaktır.(4.12) daki ε^* değeri Tablo 4.21 de gösterildiği gibi p ile çok hızlı azaldı.

M-tahmin edicisi(Huber 1973) robust tahmininde ilerlemede önemli bir aşamaya dikkat çeker. Birçok araştırmacı diğer taraftan mümkün olduğu kadar robust olan M-tahmin edicilerine uygun böyle p ve ψ yapısal fonksiyonlara odaklandılar. Fakat diğer taraftan açıkça etkili değildir(normal hata durumunda).EKK da ψ(t)=t ile bir M-tahmin edicidir ki L_1 regresyonu ψ(t)=sgn(t) ye bağlıdır. Huber(1964) aşağıdaki ψ fonksiyonunu ileri sürdü:

$$\psi(t) = \begin{cases} t & ,|t| < b \\ b\,\mathrm{sgn}(t) & ,|t| \geq b \end{cases} \tag{4.13}$$

Burada b sabittir. Bu tahmin edicinin asimtotik oranları Huber (1973) de tartışıldı. Hampel (1974) uzaktaki gözlemlere karşı daha da güçlü bir şekilde modeli koruyor.

$$\psi(t)=\begin{cases} t & ,|t|<a \\ a\,\mathrm{sgn}(t) & ,a\le|t|\le b \\ a\,\mathrm{sgn}(t)\,(c-|t|)/(c-b) & ,b\le|t|\le c \\ 0 & ,d.y. \end{cases} \tag{4.14}$$

Bu fonksiyona 3 parçalı indirgenmiş M tahmin edicisi denilir. Şekil 4.19 her iki tipin ψ fonksiyonlarını gösteriyor.

Tablo 4.18. p 'nin bazı değişkenleri için $\varepsilon^* = 1-\left(\frac{1}{2}\right)^{1/p}$ değeri

p	ε^*	p	ε^*
1	%50	6	%11
2	%29	7	% 9
3	%21	8	% 8
4	%16	9	% 7
5	%13	10	% 7

Elbette ki yeni tahmin ediciler tanımlamak ve asimtotik oranlarına çalışmak yeterli değildir. Bunun yanında tahmin edicileri hesaplamak için bir yöntem geliştirmek gerekir. M – tahmin edicilerinin çözümü genelde itme yöntemi ile uygulanır. Her aşamada katsayıların tahminini ve eşzamanlı ölçümünü yapmak zorundayız. Fakat " iyi " bir başlangıç değeri ile itmeye başlamak çok önemlidir. Yani hala robust olan yeterli bir tahmin edicidir. Bu tedbir olmaksızın beklenen robust çözümüne tamamen uymayan yerel minimumda kolayca bitirilebilir. Dutter (1977), M tahmin edicilerine göre sayısal problemlerin çözümü için bazı algoritmalar ileri sürdü. GM – tahmin edicilerinin veya sınırlandırılmış etki tahmin edicilerinin hesaplanması benzer problemleri gösterir.

M – ve GM – tahmin edicilerinin kırılma noktasının oldukça farklı olduğuna dikkat edelim. Kaldıraç noktalarının yaralanabilirliği yüzünden M – tahmin edicileri için ε^* tekrar % 0 olduğunda GM tahmin edicileri 0 olmaz. P sonsuza giderken GM -

tahmin edicilerinin kırılma noktası % o a düşer. Yohai (1985) iyi kaldıraç noktalarının varlığında çok düşük etkiye sahip bazı GM – tahmin edicilerini inceledi. GM – tahmin edicilerinin tam kırılma noktaları Martin (1987) tarafından hesaplandı.

Robust regresyonuna diğer bir yaklaşım rezidülerin rankına dayanıyor. Tek değişkenli yerin çatısında bu R tahmin edicileri Hodges ve Lehmann (1963) sayesindedir. Rank istatistiklerini kullanma fikri Adichie (1967), Jureckova (1971) ve Jaeekel (1972) tarafından çoklu regresyon alanına genişletildi. Jaeekel in önerisi aşağıdaki tanıma götürdü: Eğer R_i, r $r_i = y_i - x_i\theta$ nın rankı ise o zaman objektifi

$$\min_{\hat{\theta}} \sum_{i=1}^{n} a_n(R_i) r_i \ , \tag{4.15}$$

Burada $a_n(i)$ skor fonksiyonu monotondur ve $\sum_{i=1}^{n} a_n(i) = 0$ eşitliğini sağlar. $a_n(i)$ skoru için bazı olasılıklar

Wilcox' un skorları: $a_n(i) = i - (n+1)/2$

Van der Waerden skorları: $a_n(i) = \phi^{-1}(i/(n+1))$

Medyan skorları : $a_{n(i)} = \mathrm{sgn}(i - (n+1)/2$

Sınırlandırılmış Normal Skorlar: $a_n(i) = \min(c_1 \max \phi^{-1}(i/(n+1)), c)$

Kesenli regresyon durumunda ayrı olarak sabit terimi tahmin etmek gerekir. Çünkü objektif fonksiyonu kesenlere karşı sabittir. Bu, rezidülerin robust yer tahminini kullanarak yapılabilir. R tahmin edicilerinin M tahmin edicilerine kıyasla önemli bir avantajı otomatik olarak ölçümünün eşit değişimli olmasıdır. Bu nedenle eş zamanlı ölçüm tahmin edicisine bağlı değiller. Buna rağmen Jureckova (1977) R tahmin edicilerinin asimtotik olarak M tahmin edicilerine denk olduğunu gösterdi. Heiler ve Willers (1979), Jureckova tarafından yaptırılanlardan daha zayıf koşullar altında R tahmin edicilerinin asimtotik normalliğini ispatladı. Lecher (1980) R tahmin edicileri için bir algoritma geliştirdi ve tamamladı. Bu algoritmada (4.15) in minimumlaştırılması Rosenbrock un bulduğu algoritmayla gerçekleştirildi. Cheng ve

Hettmansperger (1983) (4.15) un çözümü için tekrarlı yeniden ağırlıklandırılmış en küçük karaler algoritmasını ileri sürdü.

L_1 - tahmin edicilerinin sınıfı robust tek değişkenli yerde belirgin bir rol de oynar. Onlar sıralı istatistiklerin lineer bir kombinasyonuna dayanıyor ve popülariteleri basit hesaplamaları esasına dayanır. Bickel (1973) regresyon için tek aşamalı L – tahmin edicilerinin bir sınıfını ileri sürdü. Bu θ nın ilk tahminine dayanır. Koenker ve Bassett (1978) L – tahmin edicilerinin diğer bir önerisini formüle etti. Bu öneri lineer regresyon için örnek niceliklerin benzerliğini kullanarak yapılır. $\hat{\theta}_n$ çözümü olarak α - regresyon niceliği (quantite) özelliğini tanımlıyorlar. Burada

$$P_n(r_i) = \begin{cases} \alpha r & r_i \geq 0 \\ (\alpha - 1)_{ri} & r_i \leq 0 \end{cases}$$

(x = 0.5 için L_1- tahmin edicisi elde edilir.) Koenker ve Bassett o zaman bu $\hat{\theta}_\alpha$ nın lineer kombinasyonlarını hesaplamayı ileri sürdü. Portnoy (1983) bu tahmin edicilerin bazı asimtotik oranlarını ispatladı. Fakat kırılma noktalarının hala sıfır olduğuna dikkat edilmelidir.

Ruppert ve Carroll 'ın (1980) budanmış en küçük kareler tahmin edicileride L-tahmin edicileridir. Onlar iyi bilinen yer L-tahmin edicilerinin budanmış ortalamalarına karşılık gelir. Onlar budanmış olan gözlemleri seçmek için iki yol ileri sürdüler. Bunlardan biri regresyon özelliklerinin düşüncesini kullanır. Buna karşılık diğeri başlangıç tahmin edicisinden rezidüleri kullanır.

Lineer regresyon için M-,L- ve R-tahmin edicileri hakkında benzer çalışmaların sonuçlarını tanımlanır. Farklı tasarımlar ve hata dağılımları için tahmin edicilerin sınıflarının davranışı kıyaslandı. Genelleştirilmiş örneklerin oldukça küçük, $n \leq 40$ ve $p \leq 3$ ve kaldıraç noktalarının yapılandırılmadığına dikkat etmek önemlidir. Bu çalışmadaki sonuç aşağıdaki gibi özetlenir: Normallikten sapmalar çok hafif olsa bile EKK çok yetersizdir. Redescending ψ fonksiyonlu M-tahmin edicilerinin oldukça iyi iş olduğu meydana çıktı. Wilcoxon skorlarıyla R-tahmin edicileri,ki bunlar ölçüm eşit değişkeni olma avantajına sahiptirler,(ve önceden sabit

olmak zorunda olduğu için kullanımının kolaylığı)iyi bir alternatif L-tahmin edicileri daha az yeterli sonuçları başardı.

Maksimum kırılma noktalı ilk regresyon tahmin edicisi Siegel(1982) sayesinde tekrarlı medyandır. Yukarıda söylenen Oja ve Niinimaa (1984) nın önerisi gibi p noktalarının bütün alt kümelerine dayanıyor. Herhangi p gözlem $(x_{i1}, y_{i1}), \ldots, (x_{ip}, y_{ip})$ tek parametre vektörünü tanımlıyor. Bu parametre vektörü $\theta_j = \theta(i_1, i_2, \ldots, i_p)$ li ifadenin j-inci koordinatıdır. O zaman tekrarlı medyan aşağıdaki gibi koordinatsal ifade edilir:

$$\hat{\theta}_j = \underset{i_1}{med}\left(\ldots \left(\underset{i_{p-1}}{med}\left(\underset{i_p}{med}\, \theta_j(i_1, \ldots, i_p) \right) \right) \ldots \right) \tag{4.16}$$

Bu tahmin edici kolay bir biçimde hesaplanabilir. Gerçekte %50 kırılma noktalı tahmin ediciyi veren medyanlar sıralı olarak hesaplanır. Burada Oja-Niinimaa nın önerisinden geniş bir biçimde daha niyidir. Maalesef iki dezavantajı vardır: Eşit değişkenli olmasının yokluğu ve altkümelerin geniş sayısı.Fakat ikinci problemden onların tüm C_n^p lerini kullanmak yerine rasgele bazı altkümeleri seçerek kaçınılabilir.

Son zamanlarda Yohai (1985) LMS ve LTS gibi yüksek kırılmalı tahmin ediciler için daha yüksek etkiye karşı yeni bir gelişmeyi getirdi. Bu yeni sınıfı MM – tahmin edicileri olarak isimlendirdi. Yohai nin tahmin edicileri 3 aşamada tanımlanır. İlk açama da yüksek kırılmalı tahmin edici θ^* hesaplanır. Bu amaçla robust tahmin edicisinin etkisinin olmasına gerek yoktur. O zaman & 50 kırılmalı s_n ölçümünün M – tahmini robust modelden $r_i(\theta^*)$ rezidüleri üzerinden hesaplanır. Son olarak $\hat{\theta}$ MM – tahmin edicisi;

$$\sum_{i=1}^{n} \psi(r_i(\theta)/s_n)x_i = 0$$

İfadesinin herhangi bir çözümü olarak ifade ediliyor. Bu ifade

$$s(\theta) \le s(\theta^*)$$

sağlar. Burada

$$s(\theta) = \sum_{i=1}^{n} P(r_i(\theta)/s_n) \quad \text{dir.}$$

Yohai ilk aşamanın %50 kırılma noktasının MM – tahmin edicilerinden geçtiğini ve tam model özelliğine de sahip olduğunu gösterdi. Buna rağmen hatalar normal dağılımlı olduğunda MM – tahmin edicilerinin yüksek etkiye sahip olduğunu ispatladı. Küçük sayısal bir çalışmada GM – tahmin edicileriyle elverişli olarak kıyaslandıklarını gösterdi. Yüksek etkiyle yüksek kırılmanın birleşmesi durumunda Yohai ve Zamar (1986)

$$\min_{\hat{\theta}} s_n^2 \frac{1}{n}\sum_{i=1}^{n} p\left(\frac{r_i}{s_n}\right),$$

tarafından tanımlanan " T – tahmin " edicilerini düşündüler. Burada tekrar p, s_n ölçek tahmininden farklı olabilir. Bu yöntem $r_i(\theta)$ rezidülerine uygulanır. Asimtotik olarak, bu yapıda kullanılan T – tahmin edicisi 2 tane p fonksiyonunun ağırlıklandırılmış ortalamasının M tahmin edicisi gibi davranır. Henüz araştırmadığımız diğer olasılık

$$\left[\frac{1}{n}\sum_{i=1}^{n} r_i^2(\theta)\right] \wedge \left[^2 S^2(\theta)\right] \qquad (4.17)$$

İfadesinin objektif fonksiyonunu minimum yapmaktır. Burada $\wedge$ ikisinin minimumunu ifade eder. k>1 ve $S(\theta)$ $r_1(\theta),..., r_n(\theta)$ rezidüleri üzerindeki ölçümün yüksek kırılmalı tahmin edicisidir. $S(\theta)$, $med_i r_i^2(\theta)$ veya $/n\sum_{i=1}^{n}(^2(\theta)$ nın bir katı olabilir veya $S(\theta)$ ölçümün uygun bir M – tahmin edicisidir. (4.17) in minimumlaştırılması yüksek kırılmalı bir noktayla yüksek asimtotik etkinin birleşeceği görülür. Çünkü en sık ilk parça (EKK) iyi bir biçimde kullanılır buna karşılık 2. parça kötü şekillerden korur. Fakat bu tahmin edicinin esas davranışının basitçe iyi tamamlamak için yeterli olup olmayacağını doğrulamak geriye kalıyor.

5. SONUÇLAR VE ÖNERİLER

Bu çalışma 4 ana başlık altında toplanmaktadır. Birinci bölümde sapanlarla beraber bazı kavramlar açıklanmıştır. İkinci bölümde tahmin edicilerin tarihsel süreçlerinden bahsedilip tahmin edicilerin sapanlardan nasıl etkilendiği örneklerle açıklandı. Üçüncü bölümde basit regresyon başlığı altında LMS ve RLS tahmin edicileri EKK tahmin edicisiyle karşılaştırıldı. Dördüncü bölümde çoklu regresyon başlığı altında LMS, LTS, L_1 ,S, M ve GM tahmin edicileri EKK ve birbirleriyle kıyaslandı ve en iyi kırılma noktasına S tahmin edicisi ulaştı. LMS tahmin edicisinin diğerlerine kıyasla daha kolay hesaplanabilmesinden dolayı daha kullanışlı olduğu ortaya çıktı.

Bu tez çalışmasından hareketle bundan sonra verilerle bir işleme başlamadan önce sapan değerleri bulmak amacıyla robust tahmin edicilerinden herhangi birinin kullanılması önerilebilir.

KAYNAKLAR

ADICHIE, J.N., 1967. Estimation of regression coefficients based on rank tests, Ann. Math. Stat., 38, 894-904.

AFIFI, A.A., ve AZEN, S.P., 1979. Statistical Analysis, A Computer Oriented Approach, 2nd ed., Academic Pres, New York.

ANDREWS, D.F., 1974. A robust method for multiple linear regression, Technometrics, 16, 523-531.

ATKINSON,A.C., 1985. Plots, Transformations, and Regression, Clarendon Press, Oxford.

BARTLETT, M.S.,1949. Fitting a straigh line when both variables are subject to error. Biometrics, 5, 207-212.

BERK, K.N., 1978. Comparing subset regression procedures, Technometrics, 20, 1-6.

BICKEL, P.J., 1973. On some analogues to linear combination of order statistics in the linear model, Ann. Stat., 1, 597-616.

BLOOMFIELD, P., ve STEIGER, W.L.,1980. Least Absolute Deviations Curve-Fitting. SIAM J. Sci. Stat. Comput.,1, 290-301.

BLOOMFIELD, P., ve STEIGER, W.L.,1983. Least Absolute Deviations: Theory,Applications and Algorithms, Birkhauser Verlag, Boston.

BOX, G.E., 1953. Non-normality and tests on variances, Biometrica, 40, 318-355.

BROWN, G.W., and MOOD, A.M., 1951. On median tests for linear hypotheses, in Proceedings of the 2nd Berkeley Symposium on Mathematical Statistics and Probability, edited by J. Neyman, University of California Press Berkeley and Los Angeles, pp. 159-166.

BROWNLEE, K.A., 1965. Statistical Theory and Methodology in Science and Engineering, 2nd ed., John Wiley & Sons, New York.

CARROLL, R.J., ve RUPPERT, D., 1985. Transformations in regression: A robust analysis, Technometrics, 27, 1-12.

CHAMBERS, J.M., CLEVELAND, W.S., KLEINER, B., ve TUKEY, P.A., 1983. Graphical Methods for Data Analysis, Wadsworth International Group, Belmont, CA.

CHATTERJEE, S., ve PRICE, B., 1977. Regression Analysis by Example, John Wiley & Sons, New York.

CHENG, K.S., and HETMANSPERGER, T.P., 1983.Weighted least-squares rank estimates, Commun.Stat. (Theory and Methods), 12, 1069-1086.

COLEMAN, J., et al., 1966. Equality of Educational Opportunity, two volumes, Office of Education, U.S. Department of Health, Washington, D.C.

DANIEL, C., ve WOOD, F.S., 1971. Fitting Equations to Data, John Wiley & Sons, New York.

DEVROYE , L., ve GYORFI, L., 1984. Nonparametric Density Estimation: The L_1 View, Wiley-Interscience, New York.

DODGE, Y., 1984. Robust estimation of regression coefficients by minimizing a convex combination of least squares and least absolute deviations, Computational Statistics Quarterly, 1, 139-153.

DONOHO, D.L., 1984. High Breakdown Made Simple, subtalk presented at the Oberwolfach Conference on Robust Statistics, West Germany, 9-15 September.

DONOHO, D.L., JOHNSTONE, L., ROUSSEEUW, P.J., ve STAHEL, W., 1985. Discussion on "Projection Pursuit" of P. Huber, Ann. Stat., 13, 496-500.

DRAPER, N.R., and SMITH, H., 1969. Applied Regression Analysis, John Wiley & Sons, New York.

DUTTER, R., 1977. Numerical solution of robust regression problems: Computational aspects, a comparison, J. Stat. Comput. Simulation, 5, 207-238.

EDGEWORTH, F.Y.,1887. On observations relating to several quantities, Hermathena, 6, 278-285.

EFROYMSON, M.A., 1960. Multiple regression analysis, in Mathematical Methods for Digital Computers, edited by A. RALSTON ve H.S. WILF, John Wiley & Sons, New York, pp. 191-203.

EMERSON, J.D., ve HOAGLIN, D.C., 1983. Resistant lines for y versusu x, in Understanding Robust and Explaratory Data Analysis, edited by D. Hoaglin,

F. Mosteller, ve J. Tukey, John Wiley & Sons, New York, pp. 129-165.

EZEKIEL, M., ve FOX, K.A., 1959. Methods of Correlation and Regression Analysis, John wiley & Sons, New York.

FORSYTHE, A.B., 1972. Robust estimation of straight line regression coefficients by minimizing p-th power deviations, Techonometrics, 14, 159-166.

FRIEDMAN, J.H., ve STUETZLE, W., 1980. Projection pursuit classification, unpublished manuscript.

FRIEDMAN, J.H., ve STUETZLE, W., 1981. Projection pursuit regression, Am. Stat. Assoc., 76, 817-823.

FRIEDMAN, J.H., ve STUETZLE, W., 1982. Projection pursuit methods for data analysis, in Modern Data Analysis, edited by R. LAUNER and A.F. SIEGEL, Academic Pres, New York.

FRIEDMAN, J.H., STUETZLE, W. ve SCHROEDER, A., 1984. Projection pursuit density estimation, J. Am. Stat. Assoc., 79, 599-608.

FRIEDMAN, J.H., and TUKEY, J.W., 1974. A projection pursuit algorithm for exploratory data analysis, HEE Trans. Comput., C-23, 881-889.

GENTLEMAN, W.M., 1965. Robust Estimation of Multivariate Location by Minimizing p-th Power Transformations, Ph.D. dissertation, Princeton University, and Bell Labaratories memorandum M65-1215-16.

GORMAN, J.W., ve TOMAN, R.J., 1966. Selection of variables for fitting equations to data, Technometrics, 8, 27-51.

HEİKKİLA, J.H.,2005. Robust regression, Graduate course on Advanced statistical signal processing, Information processing laboratory,Department of Electrical Engeneering, pages: 3-38, P.O. Box 4500, 90014 University of Oulu,jth@ee.oulu.fi, Finland.

HAMPEL, F.R., 1974. The influence curve and its role in robus estimation, J. Am. Stat. Assoc., 69, 383-393.

HAMPEL, F.R., 1975. "Beyond location parameters: Robust concepts and methods, Bull. Int. Stat. Inst., 46, 375-382.

HAMPEL, F.R., 1978. Optimally bounding the gross errors sensitivity and the influence of position in factor space, in Proceedings of the Statistical

Computing Section of the American Statistical Association, ASA, Washington, D.C.,59-64.

HAMPEL, F.R., Ronchetti, E.M., Rousseeuw, P.J., ve Stahel, W.A., 1986. Robust Statistics: The Approach Based on Influence Functions, John Wiley & Sons, New York.

HAWKINS, D. M., BRADU, D., KASS, G.V., 1984. Location of several outliers in multiple regression data using elemantal sets, Technometrics, 26, 197-208.

HEILER, S., ve WILLERS, R., 1979. Asymptotic Normality of R-Estimates in the Linear Model, Forschungsbericht No. 79/6, Universitat Dortmund, Germany.

HILL, R.W., 1977. Robust Regression When There Are Outliers in the Carriers, unpublished Ph.D. dissertation, Harvard University, Boston, MA.

HOCKING, R.R., 1976. The analysis and selection of variables in linear regression, Biometrics, 32, 1-49.

HODGES, J. L., Jr., ve LEHMANN, E.L.,1963. Estimates of location based on rank tests, Ann. Math. Stat., 34, 598-611.

HOLLAND, P.W., ve WELSCH, R.E., 1977. Robust Regression Using Iteratively Reweighted Least Squares, Commun. Stat. (Theory and Methods) , 6, 813-828.

HUBER, P.J., 1964. Robust estimation of a location parameter, Ann. Math. Stat., 35, 73-101.

HUBER, P.J., 1973. Robust Regression: Asymptotics, Conjectures And Monte Carlo, Ann. Stat.,1, 799-821.

HUBER, P.J., 1981. Robust Statistics, John Wiley & Sons, New York.

HUBER, P.J., ve DUTTER, R., 1974. Numerical Solutions Of Robust Regression Problems, in: COMPSTAT 1974, Proceedings in Computational Statistics, edited by G. Bruckmann, Physika Verlag, Vienna.

HUMPHREYS, R.M., 1978. Studies of luminious stars in nearby galaxies. I. Supergiants and O stars in the milky way, Astrophys. J. Suppl. Ser., 38, 309-350.

JAECKEL, L.A., 1972. Estimating regression coefficients by minimizing the dispersion of residuals, Ann. Math. Stat., 5, 1449-1458.

JERISON, H.J., 1973. Evolution of the Brain and Intelligence, Academic Pres, New York.

JOHNSTONE, I.M., ve VELLEMAN, P.F., 1985b. The resistant line and related regression methods, J. Am. Stat. Assoc., 80, 1041-1059.

JURECKOVA, J., 1971. Nonparametric estimate of regression coefficients, Ann. Math. Stat., 42, 1328-1338.

JURECKOVA, J., 1977. Asymptotic relations of M-estimates and R-estimates in linear regression model, Ann. Stat., 5, 364-372.

KOENKER, R., ve BASSETT, G.J., 1978. Regression Quantiles, Econometrica, 46, 33-50.

KRASKER, W.S., 1980. Estimation in linear regression models with disparate data points, Econometrica, 48, 1333-1346.

KRASKER, W.S., ve Welsch, R.E., 1982. Efficient bounded-influence regression estimation, J. Am. Stat. Assoc., 77, 595-604.

KRUSKAL, J.B., 1969. Toward a practical method which helps uncover the structure of a set of multivariate observations by finding the linear transformation which optimizes a new index of condensation, in Statistical Computation, edited by R.C. Milton and J.A. Nelder,Academic Press, New York.

LECHER, K., 1980. Unthersuchung von R-Schatzern im Linearen Modell mit Simulationen, Diplomarbeit, Abt. Statistik, University of Dortmund, West Germany.

MALLOWS, C.L., 1975. On Some Topics in Robustness, Unpublished Memorandum, Bell Telephone Laboratories, Murray Hill, NJ.

MARAZZI, A., 1980. ROBERTH, A Subroutine Library For Robust Statistical Procedures, in COMPSTAT 1980, Proceedings in Computational Statistics, Physica Verlag, Vienna.

MARTIN, R.D., YOHAI, V., ve ZAMAR, R., 1987. Minimax bias robust regression estimation, Technical Report, in preparation.

MOSTELLER, F., and TUKEY, J.W., 1977. Data Analysis and Regression, Addison-Wesley, Reading. MA.

NAIR, K.R., ve SHRIVASTAVA, M.P., 1942. On a simple method of curve fitting. Sankhya, 6, 121-132.

NARULA, S.C., ve WELLINGTON, J.F., 1982. The Minimum Sum Of Absolute Errors Regression: A State Of The Survey, Int. Stat. Rev.,50, 317-326.

OJA, H., ve NIINIMAA, A., 1985. Asymptotic properties of the generalized median in the case of multivariate normality, J. R. Stat. Soc. Ser. B, 47, 372-377.

PEARSON, E.S., 1931, The analysis of variance in cases of non-normal variation, Biometrica, 23, 114-133.

PLACKETT, R.L.,1972. Studies in the history of probability and statistics XXIX: The discovery of the method of least squares. Biometrika, 59, 239-251.

PORTNOY, S., 1983. Tightness of the sequence of empiric c.d.f. processes defined from regression fractiles, Research Report. Division of Statistics. University of Illinois.

RONCHETTI, E., ve ROUSSEEUW, P.J., 1985. Change-of-variance sensitivities in regression analysis, Z. Wahrsch. Verw. Geb., 68, 503-519.

ROSENBROCK, H.H., 1960. An automatic method for finding the greatests or lowest value of a function, Comput. J., 3, 175-184.

ROUSSEEUW, P.J., 1983. Multivarate Estimation With High Breakdown Point, paper presented at Fourth Pannonian Symposium on Mathematical Statistics and Probability, Bad Tatzmanndorf, Austria, September 4-9, 1983. Abstract in IMS Bull., 1983, 12, p.234. Appeared in(1985), Mathematical Statistics and Applications, Vol. B., edited by W. Grossmann, G. Pflug, I. Vineze, ve W. Wertz, Reidel, Dordrecht, The netherlands, pp. 283-297.

ROUSSEEUW, P.J., 1984. Least Median of Squares Regression, J. Am. Stat. Assoc., 78, 871-880.

ROY, S.N., 1953. On a heuristic method of test construction and its use in a multivariate analysis, Ann. Math. Stat., 24, 220-238.

RUPPERT, D.,ve CARROLL, R.J., 1980. Trimmed least squares estimation in the linear model,J. Am. Stat. Assoc., 75, 828-838.

RUYMGAART, F.H., 1981. A robust principal component analysis, J. Multivar. Anal., 11, 485-497.

SAMAROV, A.M., 1985. Bounded-influence regression via local minimax mean square error, J. Am. Stat. Assoc., 80, 1032-1040.

SEN, P.K., 1968. Estimates of the regression coefficient based on Kendall's tau, J. Am. Stat. Assoc., 63, 1379-1389.

SIEGEL, A.F., 1982. Robust regression using repeated medians, Biometrica, 69, 242-244.

SIMKIN, C.G.F., 1978. Hyperinflation and Nationalist China, Stability and Inflation, A Volume of Essays to Honour the Memory of A.W.H. Philips, John Wiley & Sons, New York.

SPOSITO, V.A., KENNEDY, W.J., ve GENTLE, J.E., 1977. "L_p norm fit of a straight line, Appl. Stat., 26, 114-116.

STEELE, J.M. ve STEIGER, W.I., 1986. Algorithms and complexity for least median of squares regression, Discrete Appl. Math., 14, 93-100.

STİGLER, S.M., 1981. Gauss and the inverntion of least squares, Ann. Stat.,9, 465-474.

STUDENT, 1927. Errors of routine analysis, Biometrica, 19, 151-164.

SWITZER, P., 1970. Numerical classification, in Geostatistics, Plenum, New York.

THEIL, H., 1950. A rank-invariant method of linear and polynomial regression analysis (Parts 1-3), Ned. Akad. Wetensch. Proc. Ser. A. 53, 386-392, 521-525, 1397-1412.

TUKEY, J.W.,1960. A survey of sampling from contaminated distributions. Contributions to Probability and Statistics, edited by I. Olkin, Stanford University Pres, Stanford. CA.

TUKEY, J.W., 1970/1971. Exploratory Data Analysis (Limited Preliminary Edition), Addison-Wesley, Reading, MA.

VANSINA, F., ve DE GREVE, J.P., 1982. Close binary systems before and after mass transfer, Astrophys. Space Sci., 87, 377-401.

VELLEMAN, P.F., ve WELSH, R.E., 1981. Efficient computing of regression diagnostics, Am. Stat., 35, 234-242.

WALD, A., 1940. The fitting of straight lines if both variables are subject to error: Ann. Math. Stat. , 11, 284-300.

WEISBERG, S., 1980. Applied Linear Regression, John Wiley & Sons, New York.

WILSON, H.C., 1978. Least squares versus minimum absolute deviations estimation in linear models, Decision Sci., 322-335.

YALE, C., ve FORSYTHE, A.B., 1976. Winsorized Regression, Technometrics, 18, 291-300.

YOHAI, V.J., 1985. High breakdown-point and high efficiency robust estimates for regression, to appear in Ann. Stat.

ÖZGEÇMİŞ

1982 yılında Adana'da doğdum. İlk öğrenimimi Ömer Haluk Özuçak İlkokulu'nda tamamladım. Ortaokul öğrenimimi Ramazanoğlu İlköğretim Okulu'nda, Lise öğrenimimi Şehit Temel Cingöz Lisesi'nde tamamladım. Şehit Temel Cingöz Lisesi'nden 1999 yılında mezun olup aynı yıl Gazi Üniversitesi Fen Edebiyat Fakültesi Matematik Bölümünü kazandım. Dört yıllık lisans eğitiminden sonra 2003 yılında mezun oldum. Aynı yıl Çukurova Üniversitesi Fen Bilimleri Enstitüsü İstatistik Anabilim Dalı'nda yüksek lisans yapmaya başladım.