ÇOK DEĞİŞKENLİ LİNEER REGRESYON MODELİNDE ETKİLİ GÖZLEMLERİN SAPTANMASINA İLİŞKİN ÖLÇÜLER

GÜLSEN KIRAL

Ç.Ü.

FEN BİLİMLERİ ENSTİTÜSÜ

MATEMATİK ANABİLİM DALI

YÜKSEK LİSANS TEZİ

ADANA

EYLÜL-1996

Ç.Ü. Fen Bilimleri Enstitüsü Müdürlüğüne,

Bu Çalışma Jürimiz tarafından Matematik Anabilim Dalında YÜKSEK LİSANS tezi olarak kabul edilmiştir.

Başkan : Yrd.Doç.Dr. Nedret Billor

Üye : Prof.Dr. Fikri Akdeniz

Üye : Yrd.Doç.Dr. Sadullah Sakallıoğlu

Kod No: 1159

Yukarıdaki imzaların adı geçen öğretim üyelerine ait olduğunu onaylarım.

T.C. ÇUKUROVA ÜNİVERSİTESİ FEN BİLİMLERİ ENSTİTÜSÜ MÜDÜRLÜĞÜ 1973

Prof. Dr. Ural Dinç
Enstitü Müdürü

İÇİNDEKİLER ***SAYFA NO***

TABLO LİSTESİ ***SAYFA NO***

ÖZ

Bu çalışmada basit, çoklu ve çok değişkenli doğrusal regresyon modelleri tanıtıldıktan sonra gözlemlenmiş veri kümesi içinde etkili gözlemleri saptamamızı sağlayan tanılama yöntemleri anlatılmıştır. Ayrıca *Billor ve Loynes (1992)*'un çoklu doğrusal regresyonda tanımladıkları LD^* ölçüsü çok değişkenli doğrusal regresyon modellerine uyarlanmıştır.

ABSTRACT

In this study, after giving some introductory information about simple, multiple, and multivariate linear regression models, diagnostic methods are given for assessing influential (outlier) observations in the observed data set. In addition to this, LD^* method proposed by *Billor* and *Loynes (1992)* is adapted to multivariate linear regression models.

1. BÖLÜM

DOĞRUSAL REGRESYON MODELLERİ

1.1 GİRİŞ

Fizik, ekonomi, biyoloji, mühendislik, sosyal bilimler gibi pekçok alanda kullanılan regresyon analizi değişkenler arasında var olan karışık ilişkilerin ortaya çıkarılmasını sağlayan son derece kullanışlı yinelemeli bir yöntemdir. Bu yöntem yardımıyla eldeki verilere en uygun model kurulmaya çalışılır. Bu konu üzerine yapılan çalışmalar son yirmi yıldır yoğunluk kazanmıştır.

En küçük kareler (e.k.k.) ile kestirilmiş doğrusal regresyon modelleri istatistiksel işlemlerde geniş bir kullanıma sahiptir. Bu nedenle çalışmamızda e.k.k. ile kestirilmiş doğrusal regresyon modellerini kullanıp, kestirimi belirleyen çarpanları ve bu çarpanların kestirim üzerindeki etkilerini araştıracağız.

Kestirilmiş regresyon eşitliğini belirleyen elemanlar,

- değişkenler,
- gözlemler ve
- model varsayımlarıdır.

Modeli tam olarak temsil edemeyen değişkenlerin modelde kalması regresyon katsayılarının değişmesine neden olabilir. Kestirimde meydana gelen bu değişim ve model üzerindeki etkisinin değerlendirilmesi önemlidir. Bu konu ile ilgili çok sayıda yöntem bulunmaktadır. Bu yöntemlerin ortaya çıkışı rezidü analizi çalışmaları ile başlar. Bu çalışmaların çoğu *1960*'lı yıllardan önce *F. Anscombe, J.W. Tukey, G.E.P. Box, D.R. Cox, C. Daniel ve K.S. Srikantan* tarafından yapılmıştır.

1970'li yıllarda bireysel gözlemlerin etkisinin değerlendirilmesi ve ilgili yöntemlerin ortaya çıkışı ve gelişimi ile rezidü analizine duyulan ilgi büyük ölçüde artırmıştır.

1970 ve *1975*'li yıllar arasında yapılan çalışmalar yoğunluk kazanmış ve bugüne kadar yoğunluğunu sürdürmüştür.

Yapılan çalışmalardan birkaçı;

Nelder ve Wedderburn (1972): Genelleştirilmiş doğrusal modeller için kestirilmiş modelin duyarlılığını ölçme,

Chatterjee ve Price (1977) ve Belsley ve ark. (1980): Çoklu iç ilişkiye neden olan gözlemlerin saptanması,

Chatterjee ve Price (1977), Judge ve ark. (1985) : Sabit varyanslı olmama problemi ve oto korelasyon,

Huber (1981) ve Rausseeuw ve Leroy (1987): Robust regresyon,

Cook ve Weisberg (1982): Rezidülerin regresyon içinde etkinin saptanması,

Moolgavkor, Lustbader ve Venson (1984): Üssel aileler üzerinde satır silmenin etkisi,

Storer ve Crowley (1985): Genel koşullu likelihooddaki parametre tahminlerindeki değişiklik üzerinde satır silmenin etkisi,

Atkinson (1985) ve Carroll ve Ruppert (1988): Açıklayıcı ve cevap değişkenlerinin dönüşümleri,

Chatterjee ve Hadi (1988): Doğrusal regresyonda etkili satırların saptanması, konuları üzerinde çalışmışlardır.

Eldeki verilere uygun doğrusal modelin e.k.k. ile uydurulması işlemi veri kümesinin uzağında bulunan bir veya birkaç gözlemin modele dahil edilmesi (veya çıkarılması) işleminden büyük ölçüde etkilenebilir. Modelin verileri sağlıklı bir şekilde temsil edebilmesi için bu gözlemlerin ortaya çıkarılması ve etkilerinin saptanması gereklidir. Bu gözlemlerin belirlenmesini sağlayan çok sayıda yöntem bulunmaktadır. Bu yöntemlerden kullanımı en yaygın olanları:

- *Cook (1977)* 'un D_i uzaklığı,
- *Hoaglin ve Welsch (1978)* 'in H şapka matrisi,
- *Andrews ve Pregibon (1978)* 'un AP_i si,
- *Welsch (1982)* 'in W_i sı,
- *Belsley ve ark. (1980)* 'ın $DFBETAS_i$, $COVRATIO_i$, $FVARATIO_i$, $DFFITS_i$ ları ve
- *Gentleman ve Wilk (1975)* 'in Q_i sıdır.

Bu yöntemlerin çoğu *SAS, SPSS, MINITAB* gibi bir çok istastiksel paket programlarda uygulanabilmektedir.

Tanılama yöntemleri adı altında inceleyeceğimiz bu ölçülerin çoğu satır silinmesi ya da eklenmesi yöntemi kullanılarak yapılmaktadır. *Belsley ve ark. (1980) ve Cook (1977, 1979)* tarafından yapılan açıklamalarda satır silmesi yerine satır ya da satırların alt kümesini bozulma terimi etkilettirerek etkinin araştırılmasıyla daha sağlıklı sonuçlar elde edileceği söylenmiştir. Bu mantıkla *Cook* tarafından $D_i(w)$ *ve Cook ve Weisberg (1982)* tarafından da $LD(w)$ ölçüleri tanımlanmıştır.

$D_i(w)$ ve $LD(w)$ ölçüleri her ne kadar diğer ölçülere göre daha iyi sonuç verse de model bozulmasında bir takım eksik tanımlamalara neden olabilir. Bu eksik tanımlamalar *Billor ve Loynes (1992)*'un tanımladığı LD^* ölçüsünün kullanılmasıyla giderilebilir.

Tek bir satırın etkisinin araştırılması işlemi yukarıda tanımladığımız ölçülerin kullanılmasıyla çok kolay bir şekilde yapılırken, birden fazla satır için bu işlemlerin yapılması o kadar kolay olmamaktadır. Bu problem *Jones ve Ling (1988)* tarafından, *Gray (1983,1985,1986) ve Hocking (1984)* çalışmalarından yararlanılarak çözülmeye çalışılmıştır.

Gray (1983,1985,1986) tek satır üzerindeki etkiyi ölçmemizi sağlayan ölçüleri rezidü, etki (leverage) elemanları ve basit regresyon parametrelerinin bir fonksiyonu olarak yazılabileceğini göstermiş, *Hocking (1983)* tarafından da buna benzer çalışmalar yapılmıştır. Bu sayede birbirleri ile ilişkisiz görülen birçok ölçü birbiri cinsinden ifade edilebilmiştir. *Jones ve Ling (1988)* bu çalışmalardan yola çıkarak çoklu regresyon içinde tanımlı ölçüleri, (leverage) etki, rezidü ve skalerler cinsinden ifade etmiş ve bu ölçüleri J_I sınıfı adını verdiği bir sınıfa dahil etmiştir. Bu konunun çok değişkenli doğrusal regresyona uyarlanması *Barrett ve Ling (1992)* tarafından yapılmıştır. Bu sayede doğrusal regresyon problemlerinde satırların alt kümeleri üzerinde etki hesaplamada fazla işlem yapmadan kurtuluruz.

Bu çalışma başlıca beş bölümden oluşmaktadır. Birinci bölümde basit ve çoklu doğrusal regresyon modelleri tanıtılıp standart e.k.k. regresyon sonuçlarının bir özeti, bu sonuçların bağlı olduğu varsayımlar ve çalışma içinde kullanılacak gösterimler verilecektir. İkinci bölümde bir regresyon eşitliği içinde satırların rolünü incelememizi sağlayan tanılama yöntemleri verilip kestirim üzerinde bireysel ya da grup halinde gözlemlerin etkisi incelenecektir. Üçüncü bölümde çok değişkenli doğrusal regresyon

modeli tanıtılıp bu model içinde etkili gözlemleri saptamaya ve değerlendirmeye yarayan tanılama yöntemleri anlatılacaktır. Dördüncü bölümde *Cook (1986)* tarafından tanımlanan yerel etki, *Billor ve Loynes (1992)* tarafından tanımlanan yerel etkiye yeni yaklaşım yöntemi tanıtılarak çok değişkenli regresyon modellerine uyarlamaları incelenecektir. Son olarak beşinci bölümde ise anlatılan tüm ölçüler bir veri grubu üzerinde uygulanacaktır.

1.2 Gösterimler

Birim matris, sıfır matrisi ve birlerin vektörü sırasıyla I, 0 ve 1 olarak tanımlanırken, herhangi bir M matrisinin transpozu, tersi, rankı, izi (trace), determinantı ve spektral normu; M', M^{-1}, $rank(M)$, $trace(M)$, $det(M)$ ve $\|M\|$ ifadeleri ile gösterilir.

Y nin $i.$ satırı y_i, X'in $i.$ satırı x_i' ve $j.$ kolonu x_j $(i = 1,.....n)$, $(j = 1,...k)$ ile gösterilir.

İndise yazılan (i) gösterimi, ilgili matris ya da vektörün $i.$ satırının iptal edilmiş olduğunu gösterir. Birden fazla satırın silinmesi durumu ise; I silinmek istenen satırların indis kümesini göstermek üzere, i indisi yerine I nın yazılması ile ifade edilir.

$\hat{\beta}_{(i)}$ $i.$ satır silindikten sonra elde edilen parametrelerin e.k.k. kestirimi ve $s^2_{(i)}$ i. satır silindikten sonra elde edilen varyansın kestirimini ifade etmek üzere

$$\hat{\beta}_{(i)} = \hat{\beta} - (X'X)^{-1} x_i \frac{e_i}{1 - h_{ii}} \qquad (1.1)$$

$$s^2_{(i)} = \left(\frac{e'_{(i)} e_{(i)}}{n - p - 1} \right) \qquad (1.2)$$

eşitlikleri ile tanımlanır.

$E(.)$, $V(.)$ ve $cov(.,.)$ gösterimleri sırasıyla rastgele değişkenlerin beklenen değeri, varyansı ve kovaryansını temsil eder.

Regresyon analizinde önemli yeri olan diğer bir gösterim H şapka matrisi olup

$$H = X(X'X)^{-1} X' \qquad (1.3)$$

ile gösterilir. H nin $i.$ köşegen elemanı etki (leverage) elemanı olarak adlandırılıp

$$h_{ii} = x_i'(X'X)^{-1}x_i, \qquad (i = 1,.....n) \tag{1.4}$$

eşitliği ile gösterilir.

Rezidülerin vektörü

$$e = Y - X\hat{\beta} \tag{1.5}$$

ve σ^2 nin yansız kestiricisi

$$s^2 = \frac{(Y-\hat{Y})'(Y-\hat{Y})}{n-p} = \frac{e'e}{n-p} \tag{1.6}$$

eşitlikleri ile tanımlanır.

$\hat{y}_{i(i)}$, $i.$ satır silindikten sonra elde edilen uydurulmuş değerler vektörünün $i.$ satırını ifade eder ve

$$\hat{y}_{i(i)} = x_i'\hat{\beta}_{(i)} \tag{1.7}$$

ile tanımlanır.

Çok değişkenli doğrusal regresyonda karşılaşacağımız gösterimler:

H_I, I ile indisli H nın temel minörüne karşılık gelen matris olup

$$H_I = X_I(X'X)^{-1}X_I' \tag{1.8}$$

ile temsil edilir.

E adi (ordinary) rezidülerin matrisini ifade etmek üzere, Q nun I satırını içeren alt matrisi

$$Q_I = E_I(E'E)^{-1}E_I' \tag{1.9}$$

ile gösterilir. Burada I, silinmek istenen satırların indis kümesi ve

$$E_{(I)} = Y_{(I)} - X_{(I)}\hat{B}_{(I)} \qquad (1.10)$$

olarak tanımlıdır.

$\hat{B}_{(I)}$; veri kümesini temsil eden matrisin I satırı silindikten sonra elde edilen parametre kestirim matrisidir.

q_{ii}; i. standart rezidü olup

$$q_{ii} = e_i'(e'e)^{-1}e_i \qquad (1.11)$$

eşitliği ile tanımlanır.

A; nxp tipinde bir matris olmak üzere A nın kolonlarının ilk kolondan başlayarak alt alta yazılması ile elde edilen $npx1$ tipindeki vektör $Vec(A)$ ile ifade edilir. Kısaca

$$Vec(A) = A^* = (a_1', a_2', \ldots, a_n')' \qquad (1.12)$$

ile gösterilir.

X in herhangi I satırını içeren alt matris X_I ile gösterilirken, $X_{(I)}$ ile; I satırı silinmiş X in alt matrisi ifade edilir. $Y_{(I)}$ ve Y_I içinde benzer tanımlamalar yapılabilir.

$\otimes$; sembolü ise direkt çarpımı ifade etmektedir *(Graybill (1983))*.

1.3 Tek Değişkenli Doğrusal Regresyon Modelleri

Bu bölümde basit doğrusal regresyon ve çoklu doğrusal regresyon modeli tanıtılıp, modeller ile ilgili varsayımlar ile kestirim yöntemleri kısaca anlatılacaktır.

1.3.1 Basit Doğrusal Regresyon Modeli

Üzerinde çalıştığımız değişken sayısı iki ve değişkenler arasında doğrusal bir ilişki varsa bu ilişkiyi gösteren regresyon eşitliği X bağımsız ve Y bağımlı değişken olmak üzere;

$$Y = f(X) = \beta_0 + \beta_1 X \qquad (1.13)$$

şeklindedir.

(1.13) eşitliği ile gösterilen model, yalnızca iki değişken içermesi ve değişkenler arasındaki ilişkinin doğrusal olması nedeniyle ***basit doğrusal regresyon modeli*** olarak tanımlanır.

Örneklem verilerinden yararlanarak oluşturduğumuz model her zaman kitleyi tam olarak temsil etmeyebilir. Bu nedenle ε hata terimi modele eklenerek önceden hata yapılabileceği kabul edilir. Bu durumda model

$$Y = \beta_0 + \beta_1 X + \varepsilon \qquad (1.14)$$

formuna sahip olup, *kitle regresyon modeli* olarak adlandırılır.

Burada;

Y: cevap değişkeni,

β_0 ve β_1 : regresyon parametreleri,

X: gözlem değerlerini veren açıklayıcı değişken,

ε : sıfır ortalamalı, σ^2 varyanslı, rastgele ilişkisiz hata değişkenlerdir.

Y cevap değişkenleri rastgele değişkenler olmak üzere X açıklayıcı değişkenlerini önemsiz hata ile ölçülen ve veri analizi ile kontrol edilebilen değişkenler olarak düşünmek gerekir. Yani; X in mümkün olan her değerinde Y için bir olasılık dağılımı vardır. Bu dağılımın ortalaması

$$E(Y \mid X = x) = \beta_0 + \beta_1 x$$

ve varyansı

$$V(Y \mid X = x) = V(\beta_0 + \beta_1 x + \varepsilon) = \sigma^2$$

dir.

Görüldüğü gibi Y nin ortalaması, X in doğrusal bir fonksiyonu olmasına rağmen Y nin varyansı X in değerlerinden bağımsızdır. Hata terimleri ilişkisiz olduğundan cevap değişkenleri de ilişkisizdir.

(1.14) nolu regresyon modelindeki regresyon parametreleri, regresyon katsayıları olarak adlandırılırlar. β_1, regresyon doğrusunun eğimi olup X içindeki bir

birimlik değişmeye karşılık Y nin olasılık dağılımının ortalamasında meydana gelen değişimi gösterir. β_0 parametresi ise regresyon doğrusunun Y kesişim noktasıdır.

(1.14) modeli içindeki β_0 ve β_1 regresyon parametrelerinin değerleri bilinmez. Bu regresyon parametrelerinin kestirimleri (x_i,y_i) $(i=1,2,...,n)$ örneklem verisi kullanılarak en küçük kareler (e.k.k.) ya da maksimum likelihood yöntemlerinden elde edilir.

1.3.1.1 Basit Doğrusal Regresyon Modellerinde Parametrelerin En Küçük Kareler Kestiricileri

$(x_1,y_1),(x_2,y_2),...,(x_n,y_n)$ gibi, n veri çiftine sahip olduğumuzu düşünelim. Bu veriler için örneklem regresyon modeli

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad (i = 1,2,...,n) \qquad (1.15)$$

olarak yazılmak üzere bu verilere uygun olarak regresyon doğrusu, β_0 ve β_1 in kestirimi ile elde edilir. Kestirimler en küçük kareler (e.k.k.) yöntemi yardımıyla; y_i gözlemleri ile uydurulmuş doğru arasındaki uzaklıkların kareleri toplamının minimum yapılmasıyla elde edilir.

β_0 ve β_1 parametrelerinin kestirimi;

$$S(\beta_0,\beta_1) = \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$$

fonksiyonunun minimize eden değerlere karşılık gelir. Bu nedenle bu minimizasyon problemi yardımıyla

$$\left.\frac{\partial S}{\partial \beta_0}\right|_{\hat{\beta}_0,\hat{\beta}_1} = -2\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

ve

$$\left.\frac{\partial S}{\partial \beta_0}\right|_{\hat{\beta}_0,\hat{\beta}_1} = -2\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)x_i = 0$$

denklemlerinin β_0 ve β_1 e göre çözülmesi bize istenen kestirimleri verir. Bu iki eşitliğin düzenlenmesi ile e.k.k. normal eşitlikleri

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

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

formuna sahip olur.

Normal eşitliklerin çözümü ile;

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X} \qquad (1.16a)$$

ve

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} y_i (x_i - \overline{X})}{\sum_{i=1}^{n} (x_i - \overline{X})^2} \qquad (1.16b)$$

olarak elde edilir. $\left(\overline{Y} = \frac{\sum_{i=1}^{n} y_i}{n}, \overline{X} = \frac{\sum_{i=1}^{n} x_i}{n} \right)$

(1.16a,b) deki $\hat{\beta}_0$ ve $\hat{\beta}_1$ eşitlikleri sırasıyla β_0 ve β_1 in e.k.k. kestiricileridir. Bu durumda uydurulmuş basit doğrusal regresyon modeli

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

formuna sahip olacaktır.

β_0 ve β_1 kestirimlerine ek olarak hata varyansı σ^2 nin kestirimi rezidü kareler toplamından *(SSE)* elde edilir ve

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

olarak tanımlanır.

(1.14) nolu denklemi Y, X, ε ve β nın

$$Y=\begin{bmatrix} y_1 \\ .. \\ .. \\ y_n \end{bmatrix} \qquad X=\begin{bmatrix} 1 & x_1 \\ .. & .. \\ .. & .. \\ 1 & x_n \end{bmatrix} \qquad \varepsilon=\begin{bmatrix} \varepsilon_1 \\ .. \\ .. \\ \varepsilon_n \end{bmatrix} \text{ ve} \qquad \beta=\begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$$

matrisel formlarının kullanılmasıyla

$$Y = X\beta+\varepsilon \qquad (1.17)$$

olarak ta ifade edilebilir.

En küçük kareler sonuçları ve bunlara dayalı istatistiksel analiz aşağıdaki varsayımları gerektirir.

- ***Doğrusallık Varsayımı:*** Tüm gözlemlenmiş Y cevap değişkenlerinin vektörü X' in lineer fonksiyonu olarak yazılabilir.

$$Y = X\beta+\varepsilon \qquad (1.18a)$$

- ***Hesaplamadaki Varsayımlar:*** β 'nın bir tek kestiricisini bulmak için *rank(X)* $= 2$ olmalıdır. *(1.18b)*

- ***Dağılımla ilgili varsayımlar:*** E.k.k.'e dayalı istatistiksel analizler (örneğin; t-testleri, F-testleri) yapılırken bir takım varsayımlara gereksinim duyarız. Bunlar;

  a) X ler ölçüm hatasız değişkenlerdir.

  b) ε_i ler *(i=1,2,..n)* x_i den bağımsızdır. *(1.18c)*

  c) $\varepsilon_i \sim N(0,\sigma^2)$ birbiriyle ilişkisiz rastgele hata terimleridir.

(1.18a,b,c) de verilen varsayımlar altında e.k.k kestiricileri aşağıdaki özelliklere sahiptir *(Searle (1971) ve Graybill (1976))*,

a) $\hat{\beta}$, β nın yansız kestiricisidir.

$E(\hat{\beta}) = \beta$

b) $N(\mu,\sigma^2)$; μ ortalamalı σ^2 varyanslı normal dağılımı ifade etmek üzere

$$\hat{\beta} \sim N(\beta, \sigma^2 (X'X)^{-1})$$

dağılımına sahiptir.

c) *E.k.k.* kestiricileri $\hat{\beta}$ ile ilgili diğer bir özellik *Gauss-Markoff* teoremi olarak bilinir. Bu teoreme göre $E(\varepsilon)=0$, $V(\varepsilon)=\sigma^2$ ve ilişkisiz hatalar varsayımı altında kurulan *(1.17)* nolu regresyon modelinin e.k.k. kestiricileri $(\hat{\beta})$; yansız ve y_i nin doğrusal kombinasyonu olarak yazılabilen tüm diğer yansız kestiricilerle karşılaştırıldığında minimum varyansa sahiptir.

$\hat{Y}=[\hat{y}_i]$ ve $e=[e_i]$ i=1,2,...,n ile ilgili özellikler:

- $\hat{Y}$ *nx1* tipinde uydurulmuş değerlerin vektörü ve $H= X(X'X)^{-1}X'$ olarak tanımlanmak üzere

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

olarak ifade edilip aşağıdaki özelliklere sahiptir.

a) $E(\hat{Y})=X\beta$

b) $Var(\hat{Y})=\sigma^2 H$

c) $\hat{Y} \sim N(X\beta, \sigma^2 H)$

- *e, nx1* tipinde adi rezidü vektörünü ifade etmek üzere

$$e = Y - \hat{Y}$$
$$= (I-H)Y$$

aşağıdaki özelliklere sahiptir.

a) $E(e)=0$

b) $Var(e)=\sigma^2(I-H)$

c) $e \sim N(0,\sigma^2(I-H))$

d) $\frac{e'e}{\sigma^2} \sim \chi_{(n-p)}^2$

Burada $\chi_{(n-p)}^2$; n-p serbestlik dereceli χ^2 dağılımını ifade eder.

- Herhangi bir regresyon modeli içinde rezidüler toplamı sıfırdır.

$$\sum_{i=1}^{n}(y_i - \hat{y}_i) = \sum_{i=1}^{n} e_i = 0 .$$

- y_i gözlemlenmiş değerleri toplamı; $\hat{y}_i$ uydurulmuş değerleri toplamına eşittir.

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{y}_i .$$

- Açıklayıcı değişkenler ile onlara karşılık gelen rezidülerin çarpımlarının toplamı her zaman sıfırdır.

$$\sum_{i=1}^{n} x_i e_i = 0 .$$

- Rezidülerin karşılık gelen uydurulmuş değerlerle çarpımlarının toplamı her zaman sıfırdır.

$$\sum_{i=1}^{n} \hat{y}_i e_i = 0 .$$

1.3.1.2 Maksimum Likelihood Yöntemi ile Kestirim

Hata terimleri olan ε_i $(i=1,2,..,n)$ lerin 0 ortalamalı, σ^2 varyanslı bir dağılımdan geldiği ve bu dağılımın şekli biliniyorsa parametre kestirimleri yapılırken e.k.k. yerine maksimum likelihood yöntemi kullanılır. Biz özel olarak hataların normal dağılımlı olması durumunu inceleyeceğiz.

ε_i ler normal dağılımdan $(NID(0,\sigma^2))$ geldiğinden y_i lerde normal dağılıma sahiptir. y_i nin olasılık yoğunluk fonksiyonu

$$f(y_i,\theta)=\frac{1}{\sqrt{2\Pi\sigma^2}}e^{-\frac{1}{2}\left(\frac{y_i-(\beta_0+\beta_1x_i)}{\sigma}\right)^2} \quad -\infty<y_i<\infty$$

olup, parametre kestirimleri y_i lerin ortak olasılık yoğunluk fonksiyonunun kullanılması ile elde edilir.

Bu yöntemde

$$L=L(y_1,y_2,..,y_n|\beta_0,\beta_1,\sigma^2)=\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2}\left(\frac{y_i-(\beta_0+\beta_1x_i)}{\sigma}\right)^2} \qquad (1.19)$$

eşitliğiyle tanımlı likelihood fonksiyonu ya da buna denk olarak tanımlı

$$l(\beta,\sigma^2)=\ln L=-\frac{n}{2}\ln 2\pi-\frac{n}{2}\ln\sigma^2-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i-\beta_0-\beta_1x_i\right)^2 \qquad (1.20)$$

log-likelihood fonksiyonunu maksimum yapan değerler $(\tilde{\beta}_0,\tilde{\beta}_1,\tilde{s}^2)$ araştırılır. Bu değerler;

$$\left.\frac{\partial \ln L}{\partial\beta_0}\right|_{\tilde{\beta}_0,\tilde{\beta}_1,\tilde{s}^2}=0$$

$$\left.\frac{\partial \ln L}{\partial\beta_1}\right|_{\tilde{\beta}_0,\tilde{\beta}_1,\tilde{s}^2}=0$$

$$\left.\frac{\partial \ln L}{\partial\sigma^2}\right|_{\tilde{\beta}_0,\tilde{\beta}_1,\tilde{s}^2}=0$$

eşitliklerinin çözümü ile bulunur.

Eşitliklerin çözümü sonunda, $\tilde{\beta}_0,\tilde{\beta}_1,\tilde{s}^2$ maksimum likelihood kestiricileri (m.l.k.) sırasıyla

$$\tilde{\beta}_0=\bar{Y}-\tilde{\beta}_1\bar{X} \qquad (1.21a)$$

$$\tilde{\beta}_1 = \frac{\sum_{i=1}^{n} y_i (x_i - \bar{X})}{\sum_{i=1}^{n} (x_i - \bar{X})^2} \qquad (1.21b)$$

$$\tilde{s}^2 = \frac{\sum_{i=1}^{n} (y_i - \tilde{\beta}_0 - \tilde{\beta}_1 x_i)^2}{n} \qquad (1.21c)$$

olarak elde edilir.

$\tilde{\beta}_0$, $\tilde{\beta}_1$ ve $\tilde{s}^2$ maksimum likelihood kestiricileri için aşağıdaki eşitlikler yazılabilir.

$$\tilde{\beta}_0 = \hat{\beta}_0$$

$$\tilde{\beta}_1 = \hat{\beta}_1 \qquad (1.22)$$

$$\tilde{s}^2 = s^2\left(\frac{n-2}{n}\right)$$

1.3.2 Çoklu Doğrusal Regresyon Modeli

Basit doğrusal regresyon modelinde X açıklayıcı değişkenlerin vektörü nxp tipinde bir matris ve β bilinmeyen parametreleri px1 tipinde bir vektör (p>2) ise bu model **"çoklu doğrusal regresyon modeli"** olarak adlandırılır.

k tane açıklayıcı değişken içeren çoklu doğrusal regresyon modeli kitle için

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon \qquad (1.23)$$

olarak tanımlanırken örneklem için

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + \varepsilon_i$$

$$= \beta_0 + \sum_{j=1}^{k} \beta_j X_{ij} + \varepsilon_i \quad (i=1,2,\ldots n\,,\ j=1,2,..,k)$$

formunda ifade edilebilir *(p=k+1)*. *(1.23)* denklemi *Y*, *X*, ε ve β nın

$$Y=\begin{bmatrix} y_1 \\ .. \\ .. \\ y_n \end{bmatrix}_{nx1} \quad X=\begin{bmatrix} 1 & x_{11} & .. & x_{1k} \\ 1 & x_{21} & .. & x_{2k} \\ 1 & .. & .. & .. \\ 1 & x_{n1} & .. & x_{nk} \end{bmatrix}_{nx(k+1)} \quad \varepsilon=\begin{bmatrix} \varepsilon_1 \\ .. \\ .. \\ \varepsilon_n \end{bmatrix}_{nx1} \quad ve \quad \beta=\begin{bmatrix} \beta_0 \\ .. \\ .. \\ \beta_k \end{bmatrix}_{(k+1)x1}$$

matrisel formlarının kullanılmasıyla

$$Y = X\beta+\varepsilon \qquad (1.24)$$

olarak ifade edilebilir.

Burada ;

Y : gözlemlerin vektörü,

X : tam ranklı sabitlerin (veya açıklayıcı değişkenlerin) matrisi,

β : parametrelerin vektörü,

ε : sıfır ortalama ve $\sigma^2 I$ kovaryans matrisli bağımsız rastgele hata terimlerinin vektörüdür.

Kestirilmiş regresyon katsayılar vektörü $\hat{\beta} = \left[\hat{\beta}_0,...\hat{\beta}_k\right]'$

$$S(\beta) = \sum_{i=1}^{n} \varepsilon_i^{\,2} = \varepsilon'\varepsilon$$

$$= (Y - X\beta)'(Y - X\beta)$$

fonksiyonunun sırasıyla β_0, β_1, ...,β_k ya göre minimize edilmesiyle

$$\hat{\beta} = (X'X)^{-1} X'Y \qquad (1.25)$$

olarak elde edilir.

Bu kestiriciler basit doğrusal regresyon modelinin kestiricileri için tanımladığımız tüm özelliklere sahiptir.

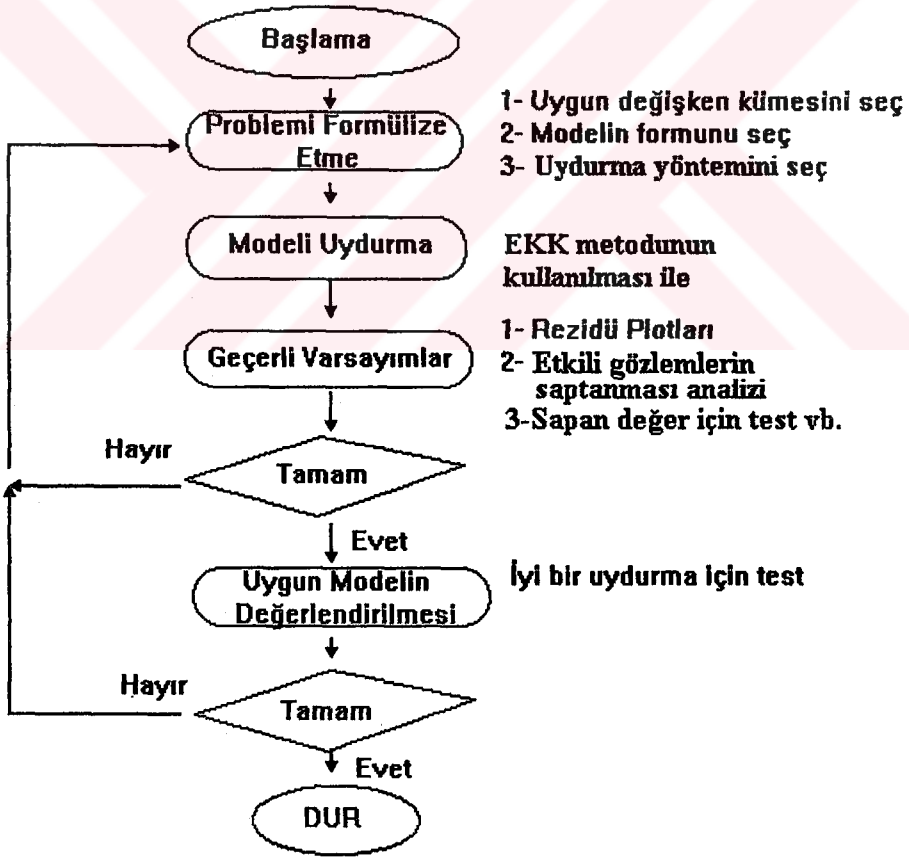
2. BÖLÜM

TEK DEĞİŞKENLİ DOĞRUSAL REGRESYON MODELLERİNDE TANILAMA YÖNTEMLERİ

2.1 GİRİŞ

Eldeki verilere uygun regresyon modelinin kurulması ve modelin uygunluğunun araştırılması regresyon analizinin temel konularından biridir. Model uygunluğunun kontrolü *(1.18a,b,c)* varsayımlarının geçerliliğinin araştırılması ile yapılır.

Regresyon analizi bir dizi analitik tekniklerin kullanımı olup, bu analiz yinelemeli bir süreç olarak değerlendirilebilir. Bu yinelemeli süreç bir şema ile verilebilir *(Chatterjee ve Hadi (1988))*.



Şekil 2.1 Yinelemeli Regresyon Yönteminin Akış Diagramı ile Gösterimi

Model kurulduktan sonra uygunluğunun test edilmesi için çok sayıda yöntem bulunmaktadır. Bunların çoğu, sapan değerlerin model üzerindeki etkilerini saptamada kullanılır.

Tanılama yöntemleri adı altında inceleyeceğimiz bu ölçülerin her biri farklı etkiyi ölçtüğünden herbiri ayrı ayrı ele alınıp incelenecektir. Ama öncelikle sapan değer, etkili nokta, yüksek etki kavramları verilecektir.

• Sapan Değer (Outlier):

Doğrusal regresyonda uydurulmuş regresyon doğrusunun uzağında olan yani uydurulmuş regresyon modeli ile uyuşmayan rezidüsü büyük olan noktalar sapan değer olarak adlandırılır.

•Yüksek Etkili Gözlemler (High Leverage Observations):

Eldeki veri grubunun yoğunlaştığı bölgenin uzağında bulunan gözlemler yüksek etkiye sahiptir. Bu gözlemler genellikle X uzayının uzağında bulunup sapan değer özelliği gösterirler.

•Etkili Gözlemler (Influential Observations) :

Veri kümesi içinde diğer veriler ile karşılaştırıldığında bireysel olarak ya da grup halinde uydurulmuş regresyon eşitliğini büyük ölçüde etkileyen noktalardır.

Bir gözlem her regresyon sonucu üzerinde aynı etkiyi göstermeyebilir. Öncelikle etkinin ne üzerine olduğu önemlidir. Örneğin; $\hat{\beta}$ üzerinde etki, $\hat{\beta}$ nın varyansı üzerinde etki, tahmin edilen değer üzerindeki etki ya da kestirilmiş değer üzerine etkiyi araştırabiliriz. Bu analizde asıl amacımız hangi etkinin tercih edileceği sorusuna cevap bulmaktır. Örneğin; β üzerindeki kestirim asıl hedefimiz ise o zaman $\hat{\beta}$ üzerinde gözlemlerin etkilerinin ölçüsü, ön kestirim (prediction) asıl amacımız ise o zaman ön kestirilen değerler $(\hat{y}_i)$ üzerinde etkinin ölçüsü ile uğraşırız.

Bu tanımladığımız kavramlar üzerinde bazı önemli noktalar aşağıdaki gibi ifade edilebilir.

a) Sapan değerler etkili gözlem olmak zorunda değillerdir.

b) Etkili gözlemler sapan değer olmak zorunda değillerdir.

c) E.k.k. kestiricileri belirlenirken veri kümesi içindeki noktaların birbirleri ile uyuşabilme olasılıklarının artırılması için büyük rezidüye sahip gözlemler veri kümesine dahil edilmek istenmez.

d) Sapan değerde olduğu gibi yüksek etkili noktalarda etkili gözlem olmak zorunda değildir. Aynı şekilde etkili gözlemlerde yüksek etkiye sahip olmak zorunda değillerdir. Buna karşın yüksek etkiye sahip gözlemler etkili gözlemler olma olasılığı vardır.

(Chatterjee ve Hadi, 1988)

2.2 Tek Değişkenli Dogrusal Regresyon Modellerinde Tanılama Yöntemleri

En küçük kareler (e.k.k.) yöntemi ile model uydurma işleminde ve bundan kaynaklanan istatistiksel modellerde etkinin belirlenmesi işlemleri üzerine yapılan çalışmalar son 25 yıldır büyük artış göstermiştir.

Bir regresyon denklemini belirleyen elemanlar,

- gözlemler,
- değişkenler,
- model varsayımlarıdır.

Veriler ile varsayılan model arasındaki çeşitli uyumsuzluk tiplerinin belirlenmesi istatistiksel modellerde esastır.

Verilen bir veri kümesine genel bir doğrusal modelin e.k.k. ile uydurulmasının doğurduğu sonuçlar bir veya birkaç gözlemin çıkarılması (veya eklenmesi) ile önemli olarak değişebilir (veya etkilenebilir). Bu nedenle bu tip gözlemlerin belirlenmesi ve analizin çeşitli açılardan etkilerinin saptanması bir veri analisti için önemlidir.

Analiz sonuçlarını etkileyen bu gözlemlerin belirlenmesi için birbirleriyle ilişkili çok sayıda yöntem vardır. Bu yöntemler aşağıda verilen niceliklere dayalıdır.

- rezidüler,
- x-y uzayında noktaların uzaklığı,

- etki eğrisi,
- güven elipsoidlerinin hacmi,
- likelihood fonksiyonu

olarak verilebilir.

2.2.1 Tek Değişkenli Doğrusal Regresyon Modellerinde Tek Bir Satırın Etkisini Ölçmemizi Sağlayan Tanılama Yöntemleri

- **Rezidüler**

Rezidüler regresyon tanılama yöntemleri içinde önemli bir rol oynar. Hiç bir analiz rezidülerin tam bir incelemesi olmaksızın tamamlanamaz. Regresyon sonuçlarının standart analizi *(1.6)* varsayımlarına dayalıdır. Doğru analizin yapılabilmesi için bu varsayımların geçerliliğinin kontrol edilmesi gereklidir.

Rezidü vektörleri *(e)*, hata vektörleri *(ε)* cinsinden

$$e = (I-H)\varepsilon \qquad (2.1)$$

ifade edilebilirler. Bu eşitlikte e nin ε için uygun bir nicelik olabilmesi için H nin köşegen üzerinde olmayan noktalarının küçük olması gereklidir. Hata terimleri ε_i ler birbirleri ile ilişkisiz ve aynı varyansa sahip olmalarına rağmen rezidü terimleri e_i lerin bağımsızlık (H köşegenel olmadıkça), aynı varyansa sahip olma (H nin köşegenel elemanları eşit olmadığı sürece) özellikleri yoktur.

Sonuç olarak rezidülerin ε_i lerin yerini alabilmesi için X in satırlarının homojen bu nedenle H nin köşegen elemanları yaklaşık olarak eşit ve köşegen haricinde bulunanlar (off-diagonal) da yeteri kadar küçük olmalıdır.

Bu gibi durumlarda varyansı sabitleştirmek için varyansın dönüşüm yapılmış formları kullanılır.

Bunlar, σ_i i. rezidünün standart sapması olmak üzere e_i yerine,

$$f(e_i,\sigma_i) = \frac{e_i}{\sigma_i} \qquad (2.2a)$$

kullanabilir. *(2.2a)* nın dört özel durumu vardır.Bunlar;

- normalleştirilmiş rezidüler,

olarak ifade edilir.

Küçük veri kümelerinde Student rezidüler daha kullanışlıdır. Çünkü rezidü varyanslarındaki farklılıklar daha önemlidir. n büyükse gözlemlerin uydurulmuş regresyon doğrusuna olan uzaklıkları olarak tanımlanan rezidülerin büyük değerli olması etkinin olabileceği uyarısını verir. Bu durumda, ölçeklenmiş ve ölçeklenmemiş rezidü arasında pek fark yoktur.

- **Şapka Matrisi (H)**

Hoaglin ve Welsch (1978) tarafından tanımlanan bu ölçü ile; $H = X(X'X)^{-1}X'$ matrisinin (şapka matrisi), köşegenel elemanları olan $h_{ii} = x_i'(X'X)^{-1}x_i$ leri kullanılarak X uzayının uzağındaki noktalar saptanır.

Bu ölçü için eşik değer;

$$\sum_{i=1}^{n} h_{ii} = RankH = RankX = p$$

eşitliği yardımıyla 2p/n olarak elde edilmiştir *(Chatterjee ve Hadi (1988) sh:100).*

Daha kaba bir inceleme yapılıyorsa; $h_{ii} > 1$ eşitliğini sağlayan noktaların X uzayının uzağında olduğu söylenebilir.

- **DFBETAS**

Belsley ve ark. (1980) tarafından tanımlanan bu ölçü *i.* gözlemin parametre kestirimi üzerindeki etkisini ölçer.

$\hat{\beta}_{(i)}$; *i.* gözlem çıkarıldıktan sonra elde edilen parametre kestiricisi olmak üzere bu ölçü

$$DFBETAS_i = \hat{\beta} - \hat{\beta}_{(i)} \qquad (2.3)$$

olarak tanımlanır. Bu eşitlik

$$X'_{(i)}X_{(i)} = X'X - x_i'x_i \qquad (2.4a)$$

$$Y'_{(i)}Y_{(i)} = Y'Y - y'_i y_i \qquad (2.4b)$$

eşitlikleri ve *Sherman-Morrison ve Woodbury* teoreminin *(Rao (1973) sh:33)* bir sonucu olarak yazılabilen

$$(X'_{(i)}X_{(i)})^{-1} = (X'X)^{-1} + \frac{(X'X)^{-1}x'_i x_i (X'X)^{-1}}{1-h_{ii}} \qquad (2.4c)$$

formu yardımıyla,

$$DFBETAS_i = \frac{(X'X)^{-1}x'_i e_i}{1-h_{ii}} \qquad (2.5)$$

olarak ifade edilebilir *(Miller 1964)*.

(2.5) eşitliğinden elde edilen değer çok büyük ise *i.* gözlemin, parametre kestirimi üzerine etkisinin büyük olduğu söylenir.

Bu ölçü, *i.* gözlemin *j.* parametre kestirimi $(\hat{\beta}_j)$ üzerindeki etkisini de ölçer. Bunun için;

$$DFBETAS_{ij} = \frac{(\hat{\beta}_j - \hat{\beta}_{j(i)})}{\sqrt{s^2_{(i)}(X'X)^{-1}_{jj}}} = \frac{(\hat{\beta}_j - \hat{\beta}_{j(i)})}{\sqrt{s^2_{(i)}C_{jj}}} \qquad (2.6)$$

formundan yararlanılır. *(2.6)* eşitliğinde, C_{jj}; $(X'X)^{-1}$ matrisinin *j.* köşegenel elemanıdır. Bu ölçü için eşik değer $2/\sqrt{n}$ olup, *(2.6)* dan elde edilen değer bu eşik değerden büyükse *i.* gözlemin *j.* katsayı üzerinde etkisinin olduğu söylenir.

- **DFFITS**

Belsley ve ark. (1980) (sh:15) tarafından tanımlanan bu ölçü, *i.* gözlemin uydurulmuş değer üzerindeki etkisini ölçer.

$\hat{y}_i$, *i.* gözlemin uydurulmuş değeri ve $\hat{y}_{i(i)}$, *i.* gözlem çıkarıldıktan sonra elde edilen *i.* gözlemin uydurulmuş değeri olmak üzere bu ölçü

$$DFFIT_i \equiv \hat{y}_i - \hat{y}_{i(i)} = x_i\left[\hat{\beta} - \hat{\beta}_{(i)}\right] = \frac{h_{ii}e_i}{1-h_{ii}} \qquad (2.7)$$

şeklinde ifade edilebilir.

(2.7) eşitliğinin, $\hat{y}_i$ nin standart hatası $\sigma(\sqrt{h_{ii}})$ ye bölünmesi ve σ yerine *i.* gözlemin silinmesi durumunda elde edilen σ nın kestirici olan $s_{(i)}$ nin konulmasıyla bu ölçü tam olarak ifade edilebilir.

$$DFFITS_i = \left[\frac{h_{ii}}{1-h_{ii}}\right]^{1/2} \frac{e_i}{s_{(i)}\sqrt{1-h_{ii}}} \qquad (2.8)$$

Bu ölçü için eşik değer $2\sqrt{p/n}$ olarak bulunmuştur *(Belsley ve ark. (1980) sh:28).*

(2.8) eşitliğinin karesi alınıp gerekli düzenlemeler yapılırsa

$$DFFITS_i^2 = \frac{(\hat{\beta} - \hat{\beta}_{(i)})' X'X(\hat{\beta} - \hat{\beta}_{(i)})}{s_{(i)}^2} \qquad (2.9)$$

ifadesi elde edilir. Bu formun kullanımı ileride J_1 sınıfı anlatılırken detaylı olarak incelenecektir.

- **Cook Uzaklığı *(D_i)***

Bir çok ölçü, verinin *X* veya *Y* uzayındaki yerini inceleyerek potansiyel olarak etkili olan noktaları ortaya çıkarırken; esas olarak, hem noktanın bulunduğu yerin hem de cevap değişkenlerinin ölçü içindeki etkisinin birlikte değerlendirilmesi istenir.

Cook (1977, 1979) bu amaçla D_i *Cook* istatistiğini tanımladı. $\hat{\beta}_{(i)}$, *i.* gözlem çıkarıldıktan sonra elde edilen parametre kestirimi olmak üzere bu ölçü en genel formda;

$$D_i(M,c) = \frac{(\hat{\beta} - \hat{\beta}_{(i)})' M (\hat{\beta} - \hat{\beta}_{(i)})}{c}, \quad i = 1....n \qquad (2.10)$$

olarak tanımlanabilir.

(2.10) eşitliğindeki M bir norm, c de standartlaştırma için kullanılan bir skalerdir. Amaca uygun olarak M ve c için farklı seçimler yapılabilir. Bunlar *Tablo 2.1* de görülmektedir.

M	*c*	*İstatistik*
X′X	ps^2	$D(X'X, ps^2)$
X′X	$ps^2_{(i)}$	$D(X'X, ps^2_{(i)})$
$X'_{(i)}X_{(i)}$	ps^2	$D(X'_{(i)}X_{(i)}, ps^2)$
$X'_{(i)}X_{(i)}$	$ps^2_{(i)}$	$D(X'_{(i)}X_{(i)}, ps^2_{(i)})$

Tablo 2.1 Normlu Etki Ölçüleri

Bunların yorumları sonraki bölümlerde yapılacaktır. Ama en çok kullanılan norm $M = X'X$ ve $c = ps^2$ olması durumudur. Bu durumda *(2.10)* eşitliği,

$$D_i(M,c) \equiv D_i = \frac{(\hat{\beta} - \hat{\beta}_{(i)})' X'X (\hat{\beta} - \hat{\beta}_{(i)})}{ps^2}, \quad i = 1.....n \qquad (2.11)$$

formuna dönüşür.

D_i nin büyük değerli olmasına neden olan noktalar, e.k.k. kestiricisi $\hat{\beta}$ üzerinde etkili olarak ele alınır.

D_i nin büyüklüğü $F_{\alpha,p,n-p}$ tablo değeri ile karşılaştırılarak da değerlendirilebilir. $D_i \cong F_{0.5,p,n-p}$ ise *i.* gözlemin silinmesi ile $\hat{\beta}$ tüm verilere bağımlı olarak elde edilen %50 lik güven bölgesi sınırlarına taşınır. $F_{0.5,p,n-p} \cong 1$ olduğu için, $D_i > 1$ olan noktalar e.k.k. kestiricisi $\hat{\beta}$ üzerinde etkili olarak varsayılırlar.

D_i istatistiği *(1.1)* ve *(1.6)* eşitlikleri yardımıyla

$$D_i = \frac{t_i^2}{p}\frac{h_{ii}}{1-h_{ii}}, \quad i = 1,.....n \qquad (2.12)$$

olarak ifade edilebilir.

Görüldüğü gibi D_i istatistiği, Student rezidüler (t_i) ve etki elemanları (leverageler) (h_{ii}) nin çarpımı formunda ifade edilir. Bu oran sayesinde X yada Y yönünde bir problem olup olmadığı kolaylıkla değerlendirilebilir. Bu değerlendirme D_i nin büyük ya da küçük değerli oluşuna göre yapılır.

- **$COVRATIO_i$**

Şimdiye kadar incelediğimiz tanılama yöntemleri tek bir satırın model üzerindeki etkisini ortaya çıkartmakta ama kestirimin kesinliği hakkında tam olarak bir bilgi vermemektedir.

Bu amaçla *Belsley ve ark. (1980)*, kesinliği gösteren uygun bir skaler olarak genelleştirilmiş varyans kavramından yararlanmış ve determinantal oran fikri altında *COVRATIO* ölçüsünü tanımlamıştır

i. gözlem çıkartıldığında bu gözlemin güven elipsoidinin hacminde ne kadarlık değişim ortaya çıkardığını gösteren bu ölçü, kovaryans matrisler oranı olarak bilinir ve

$$COVRATIO_i = CVR_i = \left|\frac{v(\hat{\beta}_{(i)})}{v(\hat{\beta})}\right|, \quad i = 1,.....n \qquad (2.13)$$

olarak tanımlanır.

(2.13) eşitliği $v(\hat{\beta}_{(i)}) = s^2_{(i)}(X'_{(i)}X_{(i)})^{-1}$ ve $v(\hat{\beta}) = s^2(X'X)^{-1}$ eşitlikleri yardımıyla

$$COVRATIO_i = \left|\frac{(X'_{(i)}X_{(i)})^{-1}s^2_{(i)}}{(X'X)^{-1}s^2}\right|, \quad i = 1,....n \qquad (2.14)$$

olarak düzenlenebilir.

Bu ölçü sonuçları kabaca aşağıdaki gibi değerlendirilebilir.

a) $COVRATIO_i > 1$ ise *i.* gözlem modelin kesinliğini gereksiz yere vurgular.

b) $COVRATIO_i < 1$ ise *i.* gözlemin modele katılması kestirimin kesinliğini azaltır.

c) $COVRATIO_i = 1$ ise tüm gözlemler kovaryans matrisi üzerinde eşit etkiye sahiptir.

Hesaplamada, *(2.14)* eşitliği *(2.4c)* eşitliğinin yerine konulması ve bazı matris özelliklerinin (*Graybill (1983)*) kullanılması ile

$$COVRATIO_i = \left(\frac{s^2_{(i)}}{s^2}\right)^p \frac{1}{1-h_{ii}} = \frac{1}{\left(\frac{n-p-1}{n-p}\right)+\frac{t_i^2}{n-p}} \frac{1}{1-h_{ii}} \qquad (2.15)$$

formunda ifade edilebilir.

Bu ölçü için eşik değer, *Belsley ve ark. (1980)* tarafından p parametre sayısı, n gözlem sayısı olmak üzere;

$$COVRATIO_i > 1+\frac{3p}{n} \qquad (2.16a)$$

ya da

$$COVRATIO_i < 1+\frac{3p}{n} \qquad (2.16b)$$

olarak tanımlanmıştır.

- **$FVARATIO_i$**

COVRATIO ölçüsünün tanımlanmasındaki mantıkla, *Belsley ve ark. (1980)* tarafından sunulan bu ölçü ile , *i.* gözlemin $\hat{y}_i$ nin varyansında meydana getirdiği değişim saptanır.

Genel olarak;

$$FVARATIO_i = \frac{V(\hat{y}_{i(i)})}{V(\hat{y}_i)} \qquad (2.17)$$

olarak tanımlanır.

(2.17) de $V(\hat{y}_{i(i)}) = s^2_{(i)} \frac{h_{ii}}{1-h_{ii}}$ ve $V(\hat{y}_i) = s^2 h_{ii}$ değerlerinin yerine konulmasıyla

$$FVARATIO_i \equiv \frac{s^2_{(i)}}{s^2(1-h_{ii})} \qquad (2.18)$$

olarak daha açık formda yazılabilir.

Bu ifade $\frac{s^2_{(i)}}{s^2}$ nin *p.* kuvveti dışında, *COVRATIO* tanımlaması ile aynıdır. Bu ölçü *COVRATIO* da olduğu gibi şapka matrisinin köşegenel elemanları h_{ii} ve Student rezidülerin aldığı değerlere bağlı olarak yorumlanırlar.

- **W_i**

Welsch (1980) veri kümesinden silme ile indirgenmiş veri yerine alt küme üzerinde satır eklemenin etkisini araştırmak istemiş ve bu nedenle W_i ölçüsünü tanımlamıştır. Bu ölçü karesel formda

$$W_i^2 = \frac{B_i'(X_{(i)}'X_{(i)})B_i}{(n-1)s^2_{(i)}} \qquad (2.19)$$

eşitliği ile gösterilir. Burada

$$B_i = (n-1)(X_{(i)}'X_{(i)})^{-1}x_i'(y_i - x_i\hat{\beta}_{(i)}') \qquad (2.20)$$

olarak tanımlanır.

(2.20),(2.4c) eşitliklerinin yerine konması ve $h_{ii} = x_i'(X'X)^{-1}x_i$, $q_{ii} = e_i'(e'e)^{-1}e_i$ eşitlikleri yardımı ile

$$W_i^2 = (n-1-p)(n-1)h_{ii}q_{ii}(1-h_{ii}-q_{ii})^{-1}(1-h_{ii})^{-2} \qquad (2.21)$$

şeklinde ifade edilebilir.

Bu ölçü hakkındaki yorumlar sonraki bölümde J_1 sınıfı içinde anlatılacaktır.

- **AP_i**

Andrews ve Pregibon (1978), β nin güven elipsoidinin hacminin $det(X'X)$ e bağımlı oluşu ve büyük rezidüye sahip gözlemlerin veri kümesinden atılmasıyla rezidü kareler toplamı *(SSE)* içinde büyük bir düşüş olacağı düşüncesinden hareketle *i.* gözlemin model üzerindeki etkisi $det(X'X)$ ile *SSE* değerlerinde meydana gelen değişimi ölçerek elde etmek istemiş ve bu nedenle

$$AP_i = \frac{SSE_{(i)}\, det(X'_{(i)}X_{(i)})}{SSE\, det(X'X)}, \qquad i = 1,2,.....n \qquad (2.22)$$

oranını tanımlamışlardır.

X matrisinin kolonuna y vektörünü ekleyerek elde edilen matris Z olmak üzere $(Z = (X;y))$

$$Z'Z = \begin{bmatrix} X'X & X'y \\ y'X & y'y \end{bmatrix}$$

matrisel formu yazılabilir.

Matris özellikleri yardımıyla *(Graybill, 1983)*

$$det(Z'Z) = det(X'X)SSE$$

ve benzer olarak

$$det(Z'_{(i)}Z_{(i)}) = det(X'_{(i)}X_{(i)})SSE_{(i)}$$

eşitlikleri elde edilebilir.

Böylece *(2.22)* eşitliği

$$\frac{SSE_{(i)}\, det(X'_{(i)}X_{(i)})}{SSE\, det(X'X)} = \frac{det(Z'_{(i)}Z_{(i)})}{det(Z'Z)} \qquad (2.23a)$$

formuna dönüşür. Bu oran *i.* gözlem silindiğinde $det(Z'Z)$ içinde meydana gelen değişimi ölçer. Verilerin yoğunlaştığı bölgenin dışında kalan noktaların veri kümesinden

atılması determinant oranında büyük bir düşüşe, dolayısıyla da hacimdeki büyük artışa neden olacaktır. Bu durumda *(2.23a)* nın küçük değerleri özel bir inceleme gerektirmektedir. Bu nedenle *(2.23a)* oranı daha uygun bir tanımla

$$AP_i = 1 - \frac{det(Z'_{(i)}Z_{(i)})}{det(Z'Z)} \qquad (2.23b)$$

olarak verilmiştir *(Andrews ve Pregibon (1978))*.

(2.23b) eşitliğinin büyük değerleri özel bir incelemenin yapılması gerekliliğini ortaya çıkarır.

Chatterje ve Hadi (1988) Lemma 2.4 den yararlanırsak

$$\begin{aligned} det(Z'_{(i)}Z_{(i)}) &= det(Z'Z - z_i z_i') \\ &= det(Z'Z)(1 - z_i'(Z'Z)^{-1} z_i) \\ &= det(Z'Z)(1 - h_{z_{ii}}) \end{aligned}$$

eşitliği elde edilir.

Dolayısıyla

$$\frac{det(Z'_{(i)}Z_{(i)})}{det(Z'Z)} = 1 - h_{z_{ii}}$$

dir. Gerekli düzenleme ile

$$AP_i = h_{z_{ii}} \qquad (2.24)$$

olacaktır. Burada $h_{z_{ii}}$; Z nin şapka matrisinin i. köşegen elemanıdır.

(2.24) den görüldüğü gibi AP_i nin, $h_{z_{ii}}$ ye denk olduğu görülür. $0 \leq h_{z_{ii}} \leq 1$ olduğundan $0 \leq AP_i \leq 1$ eşitsizliği yazılabilir. Dolayısıyla bu ölçü için eşik değer 1 olacaktır.

- **Likelihood Fonksiyonuna Bağlı Ölçüler**

n gözlem üzerine kurulan, paremetreleri β ve σ^2 olan log-likelihood fonksiyonu ve maksimum likelihood kestiricileri $\tilde{\beta}$ ve $\tilde{s}^2$ *(1.21)* eşitlikleri ile tanımlanmıştı. i. gözlemin silinmesi durumunda bu kestiriciler $\tilde{\beta}_{(i)}$ ve $\tilde{s}_{(i)}^2$ ile ifade edilirler.

$$\tilde{\beta}_{(i)} = \hat{\beta}_{(i)} \qquad (2.25a)$$

$$\tilde{s}_{(i)}^{\,2} = s_{(i)}^{\,2}\frac{n-p-1}{n-1} \qquad (2.25b)$$

(1.21) ve (2.25a,b) eşitlikleri ile tanımlanan kestiriciler için log-likelihood fonksiyonları sırasıyla

$$l(\tilde{\beta},\tilde{s}^2) = -\frac{n}{2}ln2\pi - \frac{n}{2}ln\tilde{s}^2 - \frac{(Y-X\tilde{\beta})'(Y-X\tilde{\beta})}{2\tilde{s}^2} \qquad (2.26)$$

ve

$$l(\tilde{\beta}_{(i)},\tilde{s}_{(i)}^{\,2}) = -\frac{n}{2}ln2\pi - \frac{n}{2}ln\tilde{s}_{(i)}^{\,2} - \frac{1}{2\tilde{s}_{(i)}^{\,2}}\sum_{r=1}^{n}(y_r - x_r'\tilde{\beta}_{(i)})^2 \qquad (2.27)$$

olarak yazılabilir. x_r', X in r-inci satırı, y_r, y nin r-inci satırını ifade eder.

i. gözlemin likelihood fonksiyonu üzerindeki etkisi *(2.26)* ile *(2.27)* eşitlikleri arasındaki farkın alınması ile ölçülür. Bu fark *Cook ve Weisberg (1982)* tarafından tanımlanan

$$LD_i(\beta,\sigma^2) = 2\left[l(\tilde{\beta},\tilde{s}^2) - l(\tilde{\beta}_{(i)},\tilde{s}^2{}_{(i)})\right], \qquad i=1,2,\ldots n \qquad (2.28)$$

ölçüsünün kullanılması ile elde edilir.

(2.28) ile verilen eşitlik β ve σ^2 için $100(1-\alpha)\%$ lik asimtotik güven aralığına

$$\left\{(\beta,\sigma^2): 2\left[l(\tilde{\beta},\tilde{\sigma}^2) - l(\beta,\sigma^2)\right] \le \chi^2{}_{\alpha,p+1}\right\}$$

benzediğinden *(Cox-Hinkley (1974), Lehman (1982))*, etki değerlendirmesi, $LD_i(\beta,\sigma^2)$ değerinin $\chi^2{}_{\alpha,p+1}$ dağılımı ile karşılaştırılması ile yapılabilir.

Likelihood displacement yöntemi olarak tanımlanan $LD_i(\beta,\sigma^2)$ formu β ve σ^2 nin ortak kestirimi yapılmak istenildiğinde kullanılan kullanışlı bir yöntemdir.

Likelihood fonksiyonunu kullanarak sadece β ya da σ^2 üzerindeki etki de incelenebilir. Bu durumda kullanılması gereken likelihood displacement formları sırasıyla

$$LD_i(\beta \mid \sigma^2) = 2\left[l(\tilde{\beta},\tilde{s}^2) - \max_{\sigma} l(\tilde{\beta}_{(i)},\tilde{s}^2)\right], \quad i=1,2,\ldots,n$$

$$LD_i(\sigma^2 \mid \beta) = 2\left[l(\tilde{\beta}, \tilde{s}^2) - \max_{\beta} l(\tilde{\beta}, \tilde{s}^2_{(i)}) \right] \quad i=1,2,..,n$$

olacaktır. Burada

$$\max_{\sigma} l(\tilde{\beta}_{(i)}, \tilde{s}^2) = -\frac{n}{2} ln\, 2\pi - \frac{n}{2} ln \tilde{s}^2(\tilde{\beta}_{(i)}) - \frac{n}{2}$$

ve

$$\max_{\beta} l(\tilde{\beta}, \tilde{s}^2_{(i)}) = -\frac{n}{2} ln\, 2\pi - \frac{n}{2} ln \tilde{s}^2_{(i)} - \frac{n}{2} \frac{\tilde{s}^2}{\tilde{s}^2_{(i)}}$$

olarak tanımlanmıştır.

Burada $LD_i(\beta, \sigma^2)$ ve $LD_i(\beta \mid \sigma^2)$ kullanılan olasılık modeline dayalıdır. Halbuki, daha önce verilen etki ölçüleri tamamen sayısal değerlerdir. Likelihood uzaklığının bir avantajı, normal doğrusal modelin dışındaki modellere de uygulanabilir olmasıdır.

2.2.2 Tek Değişkenli Doğrusal Regresyon Modellerinde Birden Fazla Satırın Ortak Etkisini Araştırmamızı Sağlayan Tanılama Ölçüleri ve J_I Sınıfı

Gray (1983-1986) tek bir satır üzerine yapılan etki ölçülerinin etki elemanları (leverage values), rezidüler ve temel regresyon parametrelerinin bir fonksiyonu olduğunu göstermişti. *Hocking (1983)*'de etki elemanları ve rezidülere dayalı olarak benzer tek satırlı tanılamaları ve etki ölçülerini verdi. Fakat çoklu satır için bu işlemler yapılmamıştır. Bu nedenle *Jones ve Ling (1988)* yapılan çalışmalardan yola çıkarak çoklu satırın etkisinin araştırılmasında kullanılan ölçüleri bir sınıf altında toplayıp, bunların etki, rezidü elemanları ve skalerlerin bir fonksiyonu olarak ifade etmiştir.

J_I sınıfı adı altında incelenen bu ölçüler I silinen satırların indis kümesini göstermek üzere genel olarak

$$J_I(f;u,v,c) = \left[e_I' (I - H_I)^{-u} H_I^{v} e_I \right] f(m,n,p) / c \qquad (2.29)$$

formunda tanımlanır.

f : problemin büyüklüğünün yani m, n ve p nin, bir fonksiyonu (m : silinen alt kümelerin sayısı, p: parametre sayısı, n : gözlem sayısı),

I : mxm tipinde birim matris,

H_I : H nin alt matrisi $(H_I = X_I(X'X)^{-1}X_I')$,

u ve v : sıfırdan farklı tam sayılar,

c : standartlaştırma için kullanılan bir skalerdir.

J_I sınıfının tanımlı olabilmesi için, $(I-H_I)$ nın tersinin olduğu varsayılır.

Tüm veri kümesinden I ile indisli bir grup gözlemin silinmesi ile ortaya çıkacak etkinin belirlenmesi ile ilgili pek çok farklı ölçü literatürde bulunabilir.

Burada mevcut olan ölçülerin gerçekte çoğunun J_I sınıf ölçülerinin özel durumları olduğunu göstereceğiz. Bu işlemler J_I sınıfındaki argümanların uygun seçimi vasıtasıyla kesin karşılıkları verilerek yapılabilir. Bununla beraber, bu yaklaşımda bazı etki ölçüleri vardır ki J_I sınıfının üyeleri olarak ifade edilemezler.(*Cook (1986)' un LD* si, bazı satır silme ile elde edilen etki ölçüleri (*Johnson ve Geisser (1983) in ölçüsü*) gibi.

Bu bölümde hangi ölçülerin J_I sınıfına ait oldukları işlemler yapılarak gösterilecektir.

•Cook Uzaklığı (D_I)

Tek bir satırın etkisinin incelenmesi için, *Cook (1977)* tarafından tanımlanan *Cook uzaklığı* (D_I), birden fazla satırın etkisinin değerlendirilmesi amacıyla

$$D_I(M,c) = \frac{(\hat{\beta} - \hat{\beta}_{(I)})' M (\hat{\beta} - \hat{\beta}_{(I)})}{c} \qquad (2.30)$$

olarak tanımlanabilir *(Cook ve Weisberg, (1982))*.

Burada M bir norm ve c bir ölçekleme faktörü olup en fazla $M = X'X$ ve $c = ps^2$ formu kullanılmaktadır.

$$D_I(X'X, ps^2) = \frac{(\hat{\beta} - \hat{\beta}_{(I)})' X'X (\hat{\beta} - \hat{\beta}_{(I)})}{ps^2} \qquad (2.31)$$

formunda yazılabilir $(s^2 = e'e / n - p)$.

Bir takım düzenlemelerden sonra;

$$D_I(X'X, ps^2) = e_I'(I - H_I)^{-1} H_I (I - H_I)^{-1} e_I / ps^2$$

elde edilir *(Cook ve Weisberg (1982))*.

Bu eşitlik, H_I ve $(I-H_I)^{-1}$ in değişebilirlik özelliğinden yararlanarak,

$$D_I(X'X, ps^2) = \frac{e_I'(I - H_I)^{-2} H_I e_I}{ps^2} \qquad (2.32)$$

formuna dönüşür.

Bu form *(2.29)* eşitliğinde $f(.) = \frac{1}{p}$, $c = s^2$, $u = 2$ ve $v = 1$ seçilmesi ile elde edilebilir.

•D_I Cook Uzaklığının Diğer Formları

Cook (1979) ve Cook ve Weisberg (1982); (2.30) eşitliği ile verilen $D_I(M,c)$ içindeki ölçekleme faktörleri ve normların farklı kullanımlarını ele almışlardır. Diğer araştırmacılarda, $D_I(M,c)$ nin özel durumları olarak gösterilebilen ölçüleri diğerlerinden bağımsız olarak ileri sürmüşlerdir. *Belsley ve ark. (1980)* in $MDFFIT_I$ ve *Little (1985)* tarafından sunulan L_I ölçüsü bunlara örnek olarak verilebilir. Her iki ölçümde M için $X'_{(I)}X_{(I)}$ ve c için 1 değerinin kullanılması durumuna karşılık gelir. Bu nedenle

$$MDFFIT_I = L_I = D_I(X_{(I)}' X_{(I)}, I)$$

$$= (\hat{\beta} - \hat{\beta}_{(I)})' X_{(I)}' X_{(I)} (\hat{\beta} - \hat{\beta}_{(I)}) \qquad (2.33)$$

Cook ve Weisberg (1982) (sh:132) (3.6.4.) eşitliğinden yararlanırsak *(2.33)*

$$MDFFIT_I = e_I'(I - H_I)^{-1} H_I e_I \qquad (2.34)$$

formuna dönüşür.

Bu istatistik $f(.) = 1$, $c = 1$, $u = 1$, $v = 1$ değerleri ile J_I sınıfının bir elemanıdır. *(2.33)* ve *(2.34)* den

$$D_I(X_{(I)}'X_{(I)}, ps^2) = e_I'(I - H_I)^{-1} H_I e_I / ps^2 \qquad (2.35)$$

$$D_I(X_{(I)}'X_{(I)}, ps_{(I)}^2) = e_I'.(I - H_I)^{-1} H_I e_I / ps_{(I)}^2 \qquad (2.36)$$

eşitlikleri yazılabilir.

(2.35) ve *(2.36)* da $f(.) = 1/p$, $u = v = 1$ ve sırasıyla $c = s^2$ ve $c = s_{(I)}^2$ olarak seçilirse bu ölçüler J_I sınıfının özel durumları olurlar.

Son olarak *(2.30)* daki M için $X_{(I)}'X_{(I)}$ yı kullanırsak

$$D_I(X_{(I)}'X_{(I)}, c) = e_I'(I - H_I)^{-2} H_I^2 e_I / c \qquad (2.37)$$

eşitliği elde edilir. Burada da $u = v = 2$ olarak seçilirse *(2.37)* eşitliği J_I sınıfının özel formudur deriz.

- **$DFFITS_I^2$**

Belsley ve ark. (1980 sh:15) tek bir satır silindiğinde *(i)*, y_i nin uydurulmuş değerinde nasıl bir değişim olduğunu ölçmek için $DFFIT_i$ *(2.7)* ölçüsünü ve bunun standartlaştırılmış formunda ölçekleme faktörü olarak s^2 yerine $s_{(i)}^2$ kullanarak

$DFFITS_i$ *(2.9)* ölçümünü tanıtmışlardı. Birden fazla satırın silinmesi durumunda aynı işlemi yapmak istediğimizde

$$DFFITS_I^{\,2} = \frac{(\hat{\beta} - \hat{\beta}_{(I)})' X'X (\hat{\beta} - \hat{\beta}_{(I)})}{s^2_{(I)}} \qquad (2.38)$$

şekline dönüşür. Bu nedenle $DFFITS_I^2 = D_I(X'X, s^2_{(I)})$ dır. *(2.32)* den yararlanarak

$$DFFITS_I^{\,2} = e_I'(I - H_I)^{-2} H_I e_I \,/\, s_{(I)}^{\,2} \qquad (2.39)$$

eşitliği yazılabilir.

Bu eşitlik $f(.) = 1$, $c = s_{(I)}^2$, $u = 2$, $v = 1$ olarak alınırsa *(2.29)* in özel durumuna dönüşür.

•Welsch'in W_I^2 ölçümü

Welsch (1982) sonsuz büyüklükteki örnekleme satırların bir alt kümesinin eklenmesi ile ortaya çıkacak etkinin uygun şekilde normlu ölçüsüne sonlu bir örneklem yaklaşımı verdi. Bu ölçü

$$W_I^{\,2} = B_I'(X'_{(I)}X_{(I)})B_I \,/\, ((n-m)s_{(I)}^{\,2}) \qquad (2.40)$$

eşitliği ile tanımlıdır. Burada

$$B_I = \frac{1}{m}\sum_{i \in I}(n-m)(X'_{(I)}X_{(I)})^{-1} X_i'(Y_i - X_i\hat{\beta}_{(I)}) \qquad (2.41)$$

dir.

Bu form gerekli düzenlemelerin yapılmasıyla

$$W_I^{\,2} = \frac{(n-m)e_I'(I - H_I)^{-3} H_I e_I}{m^2 s_{(I)}^{\,2}} \qquad (2.42)$$

şeklinde ifade edilebilir.

$f(.) = (n-m)/m^2$, $c = s_{(I)}^2$, $u = 3$ ve $v = 1$ olarak seçilirse, *(2.42)* J_I sınıfının bir üyesi olur.

•Gentleman ve Wilk'in Q_I ölçümü

Gentleman ve Wilk (1975), I ile indislenen satırların bir alt kümesinin silinmesi durumunda, indirgenmiş rezidü kareler toplamına bağımlı bir ölçü tanımladı. Bu ölçü en büyük Q_I değerine sahip m büyüklükteki alt kümenin m büyüklüğündeki "en çok olası sapan değer küme" olacağı düşüncesinden hareketle

$$Q_I = \left\| Y - X\hat{\beta} \right\|^2 - \left\| Y_{(I)} - X_{(I)}\hat{\beta}_{(I)} \right\|^2 \qquad (2.43)$$

olarak yazılabilir. Bu ölçü gerekli düzenlemeler yapıldıktan sonra

$$Q_I = e_I'(I - H_I)^{-1} e_I \qquad (2.44)$$

olarak tanımlanabilir.

$f(.) = c = u = 1$ ve $v = 0$ olarak seçildiğinde bu ölçünün *(2.29)* un özel bir durumu olduğu kolaylıkla görülebilir.

•Basit ve Ön Kestirilmiş Rezidü Kareler Toplamı

Silinen kümenin, basit rezidülerinin kareleri toplamı olan

$$SSE_I = e_I' e_I$$

eşitliği $f(.) = c = 1$ ve $u = v = 0$ ile *(2.29)* un özel bir durumudur. Diğer taraftan ön kestirilmiş rezidülerin kareleri toplamı olan;

$$SSPE_I = \left\| Y_I - X_I \hat{\beta}_{(I)} \right\|^2$$

da kolaylıkla

$$e_I'(I - H_I)^{-2} e_I$$

formunda ifade edilebileceği açıktır ve bu nedenle f(.) = c = 1, u = 2 ve v = 0 seçilmesiyle bu ölçünün *(2.29)* in özel bir durumu olduğu kolaylıkla görülebilir.

Şimdiye kadar gördüğümüz J_I sınıfına ait bütün ölçüler *Tablo 2.2* de özetlenmiştir.

	v		
u	**0**	**1**	**2**
0	SS (Silinen Rezidüler) $= e_I' e_I (SSE_I)$	----------	----------
1	Q_I (Gentleman ve Wilk 1975) $= \left\| Y - X\hat{\beta} \right\|^2 - \left\| Y_{(I)} - X_{(I)}\hat{\beta}_{(I)} \right\|^2$	$MDFFIT_I$ (Belsley ve ark. (1980)) = L_I (Little (1985)) $D_I\left(X_{(I)}'X_{(I)}, ps^2\right)$ $D_I\left(X_{(I)}'X_{(I)}, ps_{(I)}^2\right)$ (Cook ve Weisberg (1982)	----------
2	SS (ön kestirilmiş rezidüler) $= \left\| Y_I - X_I \hat{\beta}_{(I)} \right\|^2$	$DFFITS_I^2$ (Belsley ve ark. (1980) $D_I\left(X'X, s_{(I)}^2\right)$ $D_I\left(X'X, ps^2\right)$ (Cook 1979)	$D_I\left(X_I'X_I, ps^2\right)$ $D_I\left(X_I'X_I, ps_{(I)}^2\right)$
3	----------	W_I^2(Welsch, 1982)	----------

Tablo 2.2. (2.29) daki J_I sınıfının özel durumlarını gösteren tanılama ölçüleri

2.2.3 J_I Sınıfının Özellikleri

J_I sınıfına ait tüm ölçüler, bu sınıfın çekirdeği olarak tanımlanan

$$K_I(u,v) = e_I'\left(I - H_I\right)^{-u} H_I^v e_I \qquad (2.45)$$

nin birer fonksiyonu olarak yazılabilirler. Burada $H_I = X_I(X'X)^{-1}X_I'$ ve $e_I = Y_I - X_I\hat{\beta}$ olarak tanımlıdır.

Tablo 2.3 de; J_I sınıfına ait ölçülerin, çekirdek $K_I(u,v)$ fonksiyon $f(m,n,p)$ ve skalerler $(1, ps^2, ps_{(I)}^2)$ ya bağlı olarak nasıl ifade edilebilecekleri özetlenmiştir.

Ölçüler	u	v	Ölçülerin $K_I(u,v)$ cinsinden ifadeleri
Cook's D_I	2	1	$K_I(2,1)/ps^2$
$DFFITS_I^2$	2	1	$K_I(2,1)/s_{(I)}^2$
$MDFFIT_I = L_I$	1	1	$K_I(1,1)$
$D_I(X_{(I)}'X_{(I)}, ps^2)$	1	1	$K_I(1,1)/ps^2$
$D_I(X_{(I)}'X_{(I)}, ps_{(I)}^2)$	1	1	$K_I(1,1)/ps_{(I)}^2$
$D_I(X_I'X_I, ps^2)$	2	2	$K_I(2,2)/ps^2$
$D_I(X_I'X_I, ps_{(I)}^2)$	2	2	$K_I(2,2)/ps_{(I)}^2$
W_I^2	3	1	$(n-m)/m^2\left[K_I(3,1) / s^2{}_{(I)}\right]$
$SSPE_I$	2	0	$K_I(2,0)$
Q_I	1	0	$K_I(1,0)$
SSE_I	0	0	$K_I(0,0)$

Tablo 2.3. Çekirdek $K_I(u,v)$ ye bağlı olarak yazılabilen etki ve tanılama ölçüleri ve skaler çarpanları

Bu tablo kullanımı ve J_I sınıfının $K_I(u,v)$ fonksiyonu ile ilgili özellikleri yardımıyla ölçülerin değerlerinin hesaplanması daha kolaylıkla yapılabilir.

Özellik 1: Seçilen bir I alt kümesi üzerinde u, v ve k nın verilen değerleri için

$$K_I\left[(u+1),v\right] \geq K_I\left[u,(v+k)\right] \qquad (2.46)$$

$$K_I[(u+1),v] = \sum_{k=0}^{\infty} K_I[u,(v+k)]$$

olması

$$K_I[u,v] + K_I[(u+1),(v+1)] = K_I[(u+1),v] \qquad (2.47)$$

ifadesini gerektirdiği gösterilebilir.

Bu eşitlikler sayesinde seçilen I alt kümesi üzerinde, J_I sınıfına ait ölçüler birbiri cinsinden rahatlıkla ifade edilebilir ya da karşılaştırılabilir.

Örnek 1: *(2.46)* nolu eşitsizliği u=2, v=1 ve k=0 için

$$K_I[3,I] \geq K_I[2,I]$$

ve u=1, v=1 ve k=0 için

$$K_I[2,I] \geq K_I[1,I]$$

olarak yazılabilir. Bu eşitsizlikleri birleştirirsek

$$K_I[3,I] \geq K_I[2,I] \geq K_I[1,I]$$

elde edilir. *Tablo 2.3* ün kullanımıyla

$$\frac{m^2}{n-m} W_I^2 \geq DFFITS_I^2 \geq D_I\left(X'_{(I)}X_{(I)}, s^2_{(I)}\right)$$

yazılabilir. Böylece verilen I alt kümesi üzerinde W_I^2 , $DFFITS_I^2$, $D_I\left(X'_{(I)}X_{(I)}, s^2_{(I)}\right)$ ölçülerin birbirleri ile ilişkisi görülebilir.

Örnek 2: *(2.47)* nolu eşitlik $u=0$, $v=0$ için

$$K_I[0,0] + K_I[1,1] = K_I[1,0]$$

olarak yazılabilir. Bu eşitlik *Tablo 2.3* yardımıyla

$$SSE_I + MDFFIT_I = Q_I$$

olarak yazılabilir .

Buradan SSE_I ve $MDFFIT_I$ değerleri bilindiğinde Q_I ölçüsünü hesaplamak için sadece bir toplama işlemi yapmamız gerektiği kolaylıkla görülebilir.

Özellik 2: Herhangi iki I ve J alt kümeleri üzerinde, u ve v nin verilen değerlerine göre

$$K_I[u,v] \leq K_J[u,v]$$

$$K_I[(u+1),v] \geq K_J[(u+1),v]$$

eşitsizlikleri sağlanırsa

$$K_I[(u+1),(v+1)] \geq K_J[(u+1),(v+1)] \qquad (2.48)$$

eşitsizliği yazılabilir.

Örnek 3 : Verilen I ve J alt kümeleri için $SSE_I \leq SSE_J$ ve $Q_I \geq Q_J$ ise seçilen bir ölçü üzerinde, I alt kümesinin elemanlarının outlier (sapandeğer) olması olasılığı J alt kümesinden daha yüksektir. *(2.48)* özdeşliğinden I alt kümesi $MDFFIT_I$ ve $D_I\left(X'_{(I)}X_{(I)}, ps^2_{(I)}\right)$ ölçülerine göre J alt kümesinden daha etkili olduğunu söyleyebiliriz.

Buna ek olarak, şayet her iki alt küme aynı büyüklükte ve $Q_I \geq Q_J$ eşitliği sağlanıyorsa $s^2_{(I)} \leq s^2_{(J)}$ olur. O zaman da I alt kümesi $D_I\left(X'_{(I)}X_{(I)}, ps^2_{(I)}\right)$ ölçüsüne göre J alt kümesinden daha etkili olur. Aksine $Q_I \leq Q_J$ fakat I alt kümesi J alt kümesi silindikten sonra kötü bir kestirim veriyorsa *(ör: $SSPE_I \geq SSPE_J$)* o zaman Cook uzaklığına göre I alt kümesinin J alt kümesinden daha etkili olduğu sonucuna varılır.

2.2.4 J_I Sınıfının Kullanımındaki Yararlar

J_I sınıfı özellikle birkaç ölçü üzerinden satırların etkili alt kümeleri araştırılmak istenildiğinde hesaplamada büyük kolaylıklar sağlamaktadır.

Varsayılan birçok ölçünün ilk tanımlamalarında, tanımlamalar direkt olarak kullanılıyorsa ve özellikle tek bir ölçü kullanılarak bu işlem yapılacaksa satırların tekil alt kümelerinin etkili değerlerini elde etmede oldukça çok miktarda hesaplama yapılması gerekmektedir. Örneğin; I ile indislenen bir alt küme için D_I Cook uzaklığını hesaplamada kişi $\hat{\beta}_{(I)}$ yeni kestirimini hesaplayıp $(\hat{\beta} - \hat{\beta}_{(I)})$ vektörünü pxp tipindeki $X'X$ matrisi ile sağdan ve soldan çarpması gerekmektedir. p büyük olduğunda $\hat{\beta}_{(I)}$ yı elde etmek için mantıklı bir algoritma kullanılsa bile aritmetik işlemlerin sayısı artabilmektedir. Bütün bu işlemleri yapma yerine *(2.32)* eşitliği ile tanımladığımız Cook uzaklığı formunu kullanarak hesaplamaların büyük bir bölümünden kurtulabiliriz. $X'X$ in sonuçlarının önceden verildiği varsayımı altında $D_I(X'X, ps^2)$ Cook uzaklığının değerleri *(2.31)* in kullanımı ile p^3 üncü dereceden hesaplama yapılarak elde edilir. Buna karşın *(2.32)* eşitliğinin kullanılması sadece m^3 üncü dereceden bir hesaplama gerektirmektedir. Dolayısıyla $m << p$ olduğunda *(2.32)* hesaplamada büyük kolaylıklar sağlayacaktır.

J_t sınıfı farklı etki ölçülerin kümesi için, I ile indislenen bir alt kümenin etkisi araştırılmak istenildiğinde de büyük hesaplama kolaylıkları sağlamaktadır. *(2.35)* formu ve *Tablo 2.3* yardımıyla hesaplamanın büyük bir bölümü birbirinden farklı $\{u, v\}$ tamsayıları için $K_I(u,v)$ nin değerlendirilmesi ile yapılır. *Tablo 2.2* içinde ölçülerin birbirlerinin kombinasyonu olarak sırayla hesaplanması ile aritmetik işlemlerin sayısının mümkün olduğunca azalmaktadır. Örneğin; $(I - H_I)^{-1}$, $(I - H_I)^{-1} H_I$, $(I - H_I)^{-2} H_I$ ve $(I - H_I)^{-2} H_I^2$ formları sağdan ve soldan e_I' ve e_I ile çarpılır ve sırasıyla 1, 1, ps^2 ve ps^2 ye bölünürse Q_I, $MDFFIT_I$, D_I, $D_I(X_I' X_I, ps^2)$ ölçü değerleri elde edilebilir.

Bu sınıfı kullanmadaki diğer bir avantaj, seçilen her alt küme üzerinde nümerik bir üst sınır bulabileceğimizdir. Bu konu hakkındaki detaylı bilgi sonraki bölümde verilecektir.

3. BÖLÜM

ÇOK DEĞİŞKENLİ DOĞRUSAL REGRESYON MODELLERİNDE TANILAMA YÖNTEMLERİ

3.1 ÇOK DEĞİŞKENLİ DOĞRUSAL REGRESYON MODELİ

Doğrusal regresyon modeli uydurulurken *j.* denemede elde edilen açıklayıcı değişkenler (x_j), cevap değişkenleri (y_j) ve hatalar (ε_j) sırasıyla

$$x_j = \left[x_{j1},.....,x_{jp}\right]$$

$$y_j = \left[y_{j1},y_{j2},.....y_{jr}\right]$$

$$\varepsilon_j = \left[\varepsilon_{j1},\varepsilon_{j2},.....\varepsilon_{jr}\right]$$

eşitlikleri ile gösterilirse $(j=1,2,..,n)$; Y, X, B ve ε matrisleri

$$Y = \begin{bmatrix} y_{11} & \cdot & \cdot & y_{1r} \\ y_{21} & \cdot & \cdot & y_{2r} \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ y_{n1} & \cdot & \cdot & y_{nr} \end{bmatrix}_{nxr} = \begin{bmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ y_n \end{bmatrix}$$

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdot & \cdot & \cdot & x_{1p} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ x_{n1} & x_{n2} & \cdot & \cdot & \cdot & x_{np} \end{bmatrix}_{nxp} = \left[x_1\ x_2\ x_3\ \ldots\ x_n\right]'$$

$$B=\begin{bmatrix}\beta_{11} & \beta_{12} & . & . & \beta_{1r}\\ . & . & . & . & .\\ . & . & . & . & .\\ \beta_{p1} & \beta_{p2} & . & . & \beta_{pr}\end{bmatrix}_{pxr}=\begin{bmatrix}\beta_1\\ .\\ .\\ \beta_p\end{bmatrix}$$

$$\varepsilon=\begin{bmatrix}\varepsilon_{11} & . & . & \varepsilon_{1r}\\ \varepsilon_{21} & . & . & \varepsilon_{2r}\\ . & . & . & .\\ . & . & . & .\\ \varepsilon_{n1} & . & . & \varepsilon_{nr}\end{bmatrix}_{nxr}=\begin{bmatrix}\varepsilon_1\\ \varepsilon_2\\ .\\ .\\ \varepsilon_n\end{bmatrix}$$

matris formlarına sahip olacaktır.

Bu tanımlamalar ve $E(\varepsilon_i) = 0,\ cov(\varepsilon_i, \varepsilon_k) = \sigma_{ik}I \qquad i, k = 1,\ldots,r$ varsayımları altında kurulan

$$Y = XB + \varepsilon \tag{3.1}$$

modeli **çok değişkenli doğrusal regresyon modeli** olarak adlandırılır.

Burada B ve σ_{ik} bilinmeyen parametreler olup, j. deneme üzerinde p gözlemin varyans kovaryans matrisi $\Sigma = \{\sigma_{ik}\}$ ile ifade edilir. Farklı denemelerden elde edilen gözlemler ilişkisizdir.

Bu modelde tek değişkenli regresyonda olduğu gibi e.k.k. yöntemi kullanılarak B nin elemanları tahmin edilmek istenir. Bunun içinde

$$Tr\left[(Y - XB)'(Y - XB)\right]$$

eşitliğini minimize eden B yi bulmak gerekir. Bu eşitliğin çözümü ile elde edilen normal eşitlikler

$$(X'X)B = X'Y$$

ve e.k.k. çözümü

$$\hat{B} = (X'X)^{-1}X'Y \tag{3.2a}$$

veya u_j, Y veri matrisinin j. kolonunu ifade etmek üzere

$$\hat{B} = (X'X)^{-1} X'[u_1 \quad u_2 \quad . \quad . \quad u_r] = [\hat{\beta}_1 \quad \hat{\beta}_2 \quad . \quad . \quad \hat{\beta}_r] \qquad (3.2b)$$

olarak ifade edilebilir.

Y nin herbir kolonu bir değişken olarak düşünülürse β_j nin e.k.k. kestiricisi

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

olarak verilir.

Bu durumda uydurulmuş değerlerin matrisi

$$\hat{Y} = X\hat{B} \qquad (3.3)$$

ve rezidü terimlerinin matrisi

$$E = Y - X\hat{B} \qquad (3.4)$$

ile ifade edilir.

Σ'nın bir yansız kestiricisi

$$S = \frac{(Y - X\hat{B})'(Y - X\hat{B})}{n - p} \qquad (3.5a)$$

$$= \frac{Y'(I_n - X(X'X)^{-1}X')Y}{n - p} \qquad (3.5b)$$

eşitliğiyle tanımlanır.

(3.1) eşitliği ile tanımlı model çeşitli tanılama yöntemlerinin elde edilmesinde işlem kolaylığı sağlaması açısından

$$VecY = Y^*_{nrx1} = (I_r \otimes X)_{nrxpr} B^*_{prx1} + \varepsilon^*_{nrx1}$$

olarak da ifade edilebilir. *VecY; Y* nin kolonlarının alt alta yazılmasıyla elde edilen vektörü ifade ediyor olup

$$VecY = Y^* = \left(y_1' \quad y_2' \quad . \quad . \quad y_n'\right)'$$

matrisel eşitliği ile tanımlıdır. Aynı tanımlamalar *VecB* ve *VecE* için de geçerlidir.

Vec'li formun kullanılması durumun da E^* ve Y^* ın beklenen değer ve varyans-kovaryans matrisleri sırasıyla;

$$E(\varepsilon^*) = 0$$

$$Cov(\varepsilon^*) = \Sigma \otimes I_n$$

ve

$$E(Y^*) = (I_r \otimes X)VecB$$

$$Cov\,(Y^*) = (\Sigma \otimes I_n)$$

olarak verilebilir.

Çok değişkenli doğrusal regresyon modelleri için vermemiz gereken varsayımlar ve kestirim teorileri birinci bölümde tek değişkenli regresyon modelleri için tanımladıklarımızla aynı olup burada tekrar verilmeyecektir.

3.2 Çok Değişkenli Doğrusal Regresyon Modellerinde Tanılama Yöntemleri

Tek değişkenli doğrusal regresyon modelleri için sıradan e.k.k. regresyon analizi içinde tanımladığımız tanılama ölçülerinin çok değişkenli regresyon modelerine uyarlaması yapılabilir. Bu ölçüler dışında etki saptamaya yarayan farklı yöntemler de geliştirilmiştir. *Naik (1987)*'in çok değişkenli doğrusal regresyon modellerinde dönüştürülmüş rezidülerin kullanımıyla çok değişkenli yüksekliğine (kurtosisine) dayalı olarak sapan değerlerin araştırılması için tanımladığı yöntemi bunlara örnek olarak

verebiliriz. Bu bölümde sadece üçüncü bölümde verilen tanılama yöntemlerinin pek çoğunun çok değişkenli doğrusal regresyon modellerine uyarlaması ele alınacaktır.

3.2.1 Çok Değişkenli Doğrusal Regresyon Modellerinde Bir Satırın Etkisinin Araştırılmasını Sağlayan Tanılama Yöntemleri

- **Rezidüler, h_{ii}, $DFBETAS_i$**

Tek değişkenli doğrusal regresyon modelleri için tanımladığımız bu yöntemler çok değişkenli doğrusal regresyon modellerinde de benzer formlara sahip olup burada tekrar anlatılmayacaktır.

- **D_i Cook İstatistiği**

Cook ve Weisberg (1982) tarafından *(2.10)* eşitliği ile tanımlanan bu ölçü, çok değişkenli doğrusal regresyonda, M ve V pozitif tanımlı matrisler olmak üzere

$$D_i = \left[vec(\hat{B} - \hat{B}_{(i)})\right]' (V^{-1} \otimes M)\left[vec(\hat{B} - \hat{B}_{(i)})\right] \qquad i = 1,2,.....n \quad (3.6)$$

şeklinde ifade edilir. Burada $\hat{B}_{(i)}$; *i.* gözlem çıkarıldıktan sonra elde edilen parametre kestirimidir.

Bu ölçü $M = X'X$ ve $V = pS$ için

$$D_i = \left[vec(\hat{B} - \hat{B}_{(i)})\right]' ((pS)^{-1} \otimes X'X)\left[vec(\hat{B} - \hat{B}_{(i)})\right] \qquad i = 1,2,.....n \quad (3.7)$$

formuna dönüşür.

Bu eşitlik,

$$(Vec\ C)'(A \otimes B)(Vec\ D) = tr(C'BDA') \qquad (3.7a)$$

ve

$$S = \frac{E'E}{n-p}$$

ve *(1.1)* nolu eşitliğin $\hat{B}$ ya uyarlanması ile

$$D_i = \frac{n-p}{p} h_{ii}(1-h_{ii})^{-2} q_{ii} \qquad i = 1,2,.....n \qquad (3.8)$$

olarak h_{ii} ve q_{ii} ler cinsinden ifade edilebilir. Burada h_{ii} *ve* q_{ii} sırasıyla *(1.4)* ve *(1.11)* eşitlikleri ile tanımlıdır.

- **DFFITS$_i$**

Tek değişkenli doğrusal regresyon modellerinde *Belsley ve ark. (1980)* tarafından karesel formda *(2.9)* eşitliği ile tanımlanan bu ölçü çok değişkenli doğrusal regresyonda

$$DFFITS_i^2 = \left[vec(\hat{B} - \hat{B}_{(i)})\right]' ((pS_{(i)})^{-1} \otimes X'X)\left[vec(\hat{B} - \hat{B}_{(i)})\right] \qquad (3.9)$$

olarak tanımlanır. Burada $S_{(i)}$; *i.* gözlem çıkarıldıktan sonra elde edilen varyans-kovaryans matrisinin kestirimidir.

Bu form *(1.1)*, *(1.2)* ve *(3.7a)* eşitlikleri yardımıyla

$$DFFITS_i^2 = \frac{(n-p-1)}{p} h_{ii} q_{ii} (1-h_{ii}-q_{ii})^{-1} (1-h_{ii})^{-1} \qquad (3.10)$$

olarak yazılabilir.

- **COVRATIO$_i$**

Tek değişkenli regresyonda *(2.13)* eşitliği ile tanımlanan bu ölçü çok değişkenli doğrusal regresyonda

$$COVRATIO_i = \frac{det\left\{cov\left[vec\left(\hat{B}_{(i)}\right)\right]\right\}}{det\left\{cov\left[vec\left(\hat{B}\right)\right]\right\}} = \frac{det\left\{S_{(i)} \otimes \left(X'_{(i)}X_{(i)}\right)^{-1}\right\}}{det\left\{S \otimes (X'X)^{-1}\right\}}$$

$$= \frac{(det\, S_{(i)})^p (det(X'_{(i)}X_{(i)})^{-1})^r}{(det\, S)^p (det(X'X)^{-1})^r}$$

$$= \frac{\left[det\left[\left(E'_{(i)}E_{(i)}\right)(n-p-1)^{-1}\right]\right]^p \left[det(X'X)\right]^r}{\left[det\left[(E'E)(n-p)^{-1}\right]\right]^p \left[det(X'_{(i)}X_{(i)})\right]^r}$$

$$= \left(\frac{n-p}{n-p-1}\right)^{rp} \frac{\left[det\left(E'_{(i)}E_{(i)}\right)\right]^p}{\left[det(E'E)\right]^p} \frac{\left[det(X'X)\right]^r}{\left[det\left(X'_{(i)}X_{(i)}\right)\right]^r}$$

$$= \left(\frac{n-p}{n-p-1}\right)^{rp} \left(1-h_{ii}-q_{ii}\right)^p \left(1-h_{ii}\right)^{-r-p} \qquad (3.11)$$

olarak ifade edilir.

- **FVARATIO$_i$**

Tek değişkenli doğrusal regresyonda *(2.17)* eşitliği ile tanımlanan bu ölçü çok değişkenli regresyonda

$$FVARATIO_i = \frac{det\left\{cov\left[vec\left(x_i\hat{B}_{(i)}\right)\right]\right\}}{det\left\{cov\left[vec\left(x_i\hat{B}\right)\right]\right\}} = \frac{det\left[S_{(i)} \otimes (1-H_i)^{-1}H_i\right]}{det\left[S \otimes H_i\right]}$$

$$= \left(\frac{n-p}{n-p-1}\right)^{rp} \left(1-h_{ii}-q_{ii}\right)\left(1-h_{ii}\right)^{-r-1} \qquad (3.12)$$

olarak tanımlanır. Burada $H_i=h_{ii}$ *dir.* Dikkat edilecek olursa $p = 1$ için $FVARATIO_i \equiv COVRATIO_i$ dır.

- **Welsch'in W_i^2 Ölçüsü**

Tek değişkenli doğrusal regresyon modellerinde *Welsch (1980)* tarafından karesel formda *(2.19)* eşitliği ile tanımlanan bu ölçü çok değişkenli doğrusal regresyonda

$$W_i^2 = \frac{1}{n-m}\left(vecG_i\right)'\left[S_{(i)}^{-1} \otimes \left(\left(X'_{(i)}X_{(i)}\right)\right](vec(G_i)) \qquad (3.13)$$

eşitliği ile gösterilebilir. Burada

$$G_i = (n-1)(X'_{(i)}X_{(i)})^{-1}x_i'(y_i - x_i\hat{B}'_{(i)})$$

olarak tanımlıdır. Gerekli düzenlemelerin yapılması ile

$$W_i^2 = (n-p-1)(n-1)h_{ii}q_{ii}(1-h_{ii}-q_{ii})^{-1}(1-h_{ii})^{-2} \qquad (3.14)$$

formuna dönüşür.

- **AP_i**

Andrews ve Pregibon (1978) tarafından *(2.23a)* eşitliği ile verilen bu ölçüde Y gözlemlerin vektörünün, çoklu gözlem matrisi Y ile yer değişmesiyle çok değişkenli doğrusal regresyona uyarlanabilir. Bu durumda

$Z = [X \mid Y]$ ve $Z'Z = \begin{bmatrix} X'X & X'Y \\ Y'X & Y'Y \end{bmatrix}$ formlarına sahip olup *Jones ve Ling (1992) Lemma A.2* yardımıyla

$$\begin{aligned} det(Z'Z) &= det(X'X)det(Y'Y - Y'X(X'X)^{-1}X'Y) \\ &= det(X'X)det(E'E) \end{aligned} \qquad (3.15)$$

eşitliği elde edilir.

Benzer olarak

$$det(Z'_{(i)}Z_{(i)}) = det(X'_{(i)}X_{(i)})det(E'_{(i)}E_{(i)}) \qquad (3.16)$$

yazılabilir.

(3.15) ve *(3.16)* nin *(2.23)* de yerine konulması ve bazı matris özellikleri yardımıyla *(Barrett ve Ling (1992), sh 190)*

$$\begin{aligned} AP_i &= (1-h_{ii})(1-h_{ii}-q_{ii})(1-h_{ii})^{-1} \\ &= 1-h_{ii}-q_{ii} \end{aligned} \qquad (3.17)$$

elde edilir.

3.2.2 Çok Değişkenli Doğrusal Regresyonda Birden Fazla Satırın Etkisini Araştırmamızı Sağlayan Tanılama Ölçüleri ve J_I Sınıfı

Jones ve Ling (1988) tarafından tek değişkenli doğrusal regresyon içinde tanımlanan etki ölçülerinin birleşiminden meydana gelen J_I sınıfının çok değişkenliye uyarlanması *Barrett ve Ling (1992)* tarafından yapılmıştır. Bu sınıf, ölçülerin özelliklerine göre iki farklı formda tanımlanmıştır (J_I^{det}, J_I^{tr}).

Bu sınıflarda ölçüler I silinen satırların kümesini göstermek üzere H_I etki (leverage) ve Q_I rezidü matrislerinin birer fonksiyonu olarak ifade edilirler.

$$J_I^{tr}(f;a,b) = f(.)tr\left[H_I Q_I (I - H_I - Q_I)^a (I - H_I)^b\right] \qquad (3.18)$$

$$J_I^{det}(f;a,b) = f(.)det\left[(I - H_I - Q_I)^a (I - H_I)^b\right] \qquad (3.19)$$

Burada;

a *ve* b : ölçü özelliklerine bağlı olarak değişen tamsayılar,

f: matrislerin boyutlarının skaler bir fonksiyonu,

H_I: I indisli şapka matrisi $(H_I = X_I(X'X)^{-1}X_I')$,

Q_I: I indisli rezidülerin matrisidir $(Q_I = E_I(E'E)^{-1}E_I')$.

3.2.2.1 J_I^{tr} Sınıfına Ait Ölçüler

- **D_I Cook Uzaklığı ve Diğer Formları**

Cook (1978) tarafından tanımlanan bu istatistiğin çok değişkenli regresyonda birden fazla satırın etkisini araştırmak için kullanılan formu, M ve V pozitif tanımlı matrisler olmak üzere

$$D_I = vec(\hat{B} - \hat{B}_{(I)})(V^{-1} \otimes M)vec(\hat{B} - \hat{B}_{(I)}) \qquad (3.20)$$

şeklindedir.

Bu form

$$(vecC)'(A \otimes B)(vecD) = tr(C'BDA')$$

$$\text{tr(AB)} = (\text{vecA}')'(\text{vecB})$$

$$vec(ABC) = (C' \otimes A)vecB$$

matris özelliklerinden yararlanarak yazılabilen

$$tr(C'(BDA')) = (vecC)'vec(BDA') = (vecC)'(A \otimes B)(vecD) \qquad (3.21)$$

eşitliği yardımıyla

$$D_I(M,V) = tr\left[(I - H_I)^{-1} X_I (X'X)^{-1} M (X'X)^{-1} X_I' (I - H_I)^{-1} E_I V^{-1} E_I'\right] \qquad (3.22)$$

olarak ifade edilebilir.

M ve V nin seçimlerine göre bu istatistik farklı formlara sahip olur.

- $M = X'X$ ve $V = pS$ için $D_I(X'X, pS)$ istatistiğini

$$D_I(X'X, pS) = \left[Vec(\hat{B} - \hat{B}_{(I)})\right]' ((pS)^{-1} \otimes X'X)\left[Vec(\hat{B} - \hat{B}_{(I)})\right]$$

$$= tr\left[(I - H_I)^{-1} X_I (X'X)^{-1} X'X (X'X)^{-1} X_I' (I - H_I)^{-1} E_I (pS)^{-1} E_I'\right]$$

$$= \frac{n-p}{p} tr\left[(I - H_I)^{-1} X_I (X'X)^{-} X'X (X'X)^{-} X_I' (I - H_I)^{-1} E_I (E'E)^{-1} E_I'\right]$$

$$= \frac{n-p}{p} tr\left[(I - H_I)^{-2} H_I Q_I\right] \qquad (3.23)$$

olarak yazabiliriz.

- $M = X'X$ ve $V = pS_{(I)}$ için $D_I(X'X, pS_{(I)})$ istatistiğini

$$D_I\left(X'X, pS_{(I)}\right) = \left[vec\left(\hat{B} - \hat{B}_{(I)}\right)\right]' \left(\left(pS_{(I)}\right)^{-1} \otimes X'X\right)\left[vec\left(\hat{B} - \hat{B}_{(I)}\right)\right]$$

$$= tr\left[\left(I - H_I\right)^{-1} X_I (X'X)^{-1}(X'X)(X'X)^{-1} X_I' \left(I - H_I\right)^{-1} E_I \left(pS_{(I)}\right)^{-1} E_I'\right]$$

$$= tr\left[\left(I - H_I\right)^{-1} X_I (X'X)^{-1} X_I' \left(I - H_I\right)^{-1} E_I \left(p\frac{E_{(I)}'E_{(I)}}{n-p-m}\right)^{-1} E_I'\right]$$

$$= \frac{n-p-m}{p} tr\left[\left(I - H_I\right)^{-1} H_I \left(I - H_I\right)^{-1} Q_I \left(I - H_I - Q_I\right)^{-1}\left(I - H_I\right)\right]$$

$$= \frac{n-p-m}{p} tr\left[H_I Q_I \left(\left(I - H_I - Q_I\right)^{-1}\left(I - H_I\right)^{-1}\right]\right. \qquad (3.24)$$

olarak elde ederiz.

- $M = X_{(I)}'X_{(I)}$ ve $V = pS_{(I)}$ için $D_I(X_{(I)}', X_{(I)}, pS_{(I)})$ istatistiğini

$$D_I\left(X_{(I)}'X_{(I)}, pS_{(I)}\right) =$$

$$tr\left[\left(I - H_I\right)^{-1} X_I (X'X)^{-1}(X'X)^{-1} X_{(I)}'X_{(I)}(X'X)^{-} X_I' \left(I - H_I\right)^{-1} E_I \left(pS_{(I)}\right)^{-1} E_I'\right]$$

$$= tr\left[\left(I - H_I\right)^{-1} X_I (X'X)^{-1}\left(X'X - X_I'X_I\right)(X'X)^{-} X_I' \left(I - H_I\right)^{-1} E_I \left(\frac{E_{(I)}'E_{(I)}}{p\left(n-p-m\right)^{-1}}\right)^{-1} E_I'\right]$$

$$= \frac{n-p-m}{p} tr\left[\left[\left(I - H_I\right)^{-1} X_I (X'X)^{-1} X'X (X'X)^{-} X_I' \left(I - H_I\right)^{-1} E_I \left(E_{(I)}'E_{(I)}\right)^{-1} E_I'\right]\right.$$

$$\left.-\left[\left(I - H_I\right)^{-1} X_I (X'X)^{-1} X_I'X_I (X'X)^{-} X_I' \left(I - H_I\right)^{-1} E_I \left(E_{(I)}'E_{(I)}\right)^{-1} E'\right]\right]$$

$$= \left(\frac{n-p-m}{p}\right) tr\left[\left(I - H_I\right)^{-1} H_I \left(I - H_I\right)^{-1} Q_I \left(I - H_I - Q_I\right)^{-}\left(I - H_I\right)\right.$$

$$\left.-\left(I - H_I\right)^{-1} H_I H_I \left(I - H_I\right)^{-1} Q_I \left(I - H_I - Q_I\right)^{-1}\left(I - H_I\right)\right]$$

$$= \left(\frac{n-p-m}{p}\right) tr\left[\left(I - H_I\right)^{-2} H_I Q_I \left(I - H_I - Q_I\right)^{-1}\left(I - H_I\right)\right.$$

$$\left.-\left(I - H_I\right)^{-2} H_I^2 Q_I \left(I - H_I - Q_I\right)^{-1}\left(I - H_I\right)\right]$$

$$= \left(\frac{n-p-m}{p}\right) tr\left[\left(I - H_I\right)^{-1} H_I Q_I \left(I - H_I - Q_I\right)^{-1} - \left(I - H_I\right)^{-1} H_I Q_I \left(I - H_I - Q_I\right) H_I\right]$$

$$=\left(\frac{n-p-m}{p}\right)tr\left[\left((I-H_1)^{-1}H_1Q_1(I-H_1-Q_1)^{-1}\right)(I-H_1)\right]$$

$$=\left(\frac{n-p-m}{p}\right)tr\left[H_1Q_1(I-H_1-Q_1)^{-1}\right] \qquad (3.25)$$

olarak elde ederiz.

- $M = X'_{(1)}X_{(1)}$ ve $V = pS$ için $D_1(X'_{(1)}X_{(1)}, pS)$ istatistiği

$$D_1\left(X'_{(1)}X_{(1)}, pS\right)=\left[vec\left(\hat{B}-\hat{B}_{(1)}\right)\right]'\left((pS)^{-1}\otimes X'_{(1)}X_{(1)}\right)\left[vec\left(\hat{B}-\hat{B}_{(1)}\right)\right]$$

$$=tr\left[(I-H_1)^{-1}X_1(X'X)^{-1}\left(X'_{(1)}X_{(1)}\right)(X'X)^{-1}X'_1(I-H_1)^{-1}E_1(pS)^{-1}E'_1\right]$$

$$=tr\left[(I-H_1)^{-2}X_1(X'X)^{-1}\left(X'X-X'_1X_1\right)(X'X)^{-1}X'_1E_1\left(\frac{E'E}{n-p}\right)^{-1}E'_1\right]$$

$$=\frac{n-p}{p}\left[tr\left[(I-H_1)^{-2}X_1(X'X)^{-1}X'X(X'X)^{-1}X'_1E_1(E'E)^{-1}E'_1\right]\right.$$

$$\left.-tr\left[(I-H_1)^{-2}X_1(X'X)^{-1}X'_1X_1(X'X)^{-1}X'_1E_1(E'E)^{-1}E'_1\right]\right]$$

$$=\frac{n-p}{p}tr\left[(I-H_1)^{-2}X_1(X'X)^{-1}X'_1Q_1\right]-tr\left[(I-H_1)^{-2}H_1H_1Q_1\right]$$

$$=\frac{n-p}{p}tr\left[(I-H_1)^{-2}H_1Q_1-tr(I-H_1)^{-2}H_1^2Q_1\right]$$

$$=\frac{n-p}{p}tr\left[(I-H_1)^{-2}H_1Q_1(I-H_1)\right]$$

$$=\frac{n-p}{p}tr\left[H_1Q_1(I-H_1)^{-1}\right] \qquad (3.26)$$

olarak gösterilebilir.

Bütün bu işlemler *Tablo 3.1* de özetlenmiştir.

(3.29), (3.14), (3.25), (3.26) eşitliklerinin herbiri *(3.18)* in farklı tanımlamaları olup J_1^{tr} sınıfının üyeleridir. *Tablo 3.2* de bu istatistiklerin hangi tanımlamalar altında J_1^{tr} nin elemanı olduğu ifade edilmiştir.

M	V	İstatistik	Formu
$X'X$	pS	$D_I(X'X,pS)$	$\frac{n-p}{p} tr\left[(I-H_I)^{-2} H_I Q_I\right]$
$X'_{(I)}X_{(I)}$	pS	$D_I(X'_{(I)}X_{(I)},pS)$	$\frac{n-p}{p} tr\left[H_I Q_I (I-H_I)^{-1}\right]$
$X'X$	$pS_{(I)}$	$D_I(X'X,pS_{(I)})$	$\frac{n-p-m}{p} tr\left[H_I Q_I (I-H_I-Q_I)^{-1}(I-H_I)^{-1}\right]$
$X'_{(I)}X_{(I)}$	$pS_{(I)}$	$D_I(X'_{(I)}X_{(I)},pS_{(I)})$	$\frac{n-p-m}{p} tr\left[H_I Q_I (I-H_I-Q_I)^{-1}\right]$

Tablo 3.1 M ve V nin seçimlerine göre Cook istatistiğinin formları

Ölçü	f(n,p,r,m)	a	b	Sınıf
$D_I(X'X,pS)$	$\frac{n-p}{p}$	0	-2	J_I^{tr}
$D_I(X'_{(I)}X_{(I)},pS)$	$\frac{n-p}{p}$	0	-1	J_I^{tr}
$D_I(X'X,pS_{(I)})$	$\frac{n-p-m}{p}$	-1	-1	J_I^{tr}
$D_I(X'_{(I)}X_{(I)},pS_{(I)})$	$\frac{n-p-m}{p}$	-1	0	J_I^{tr}

Tablo 3.2 Cook istatistiğinin formlarının J_I^{tr} sınıfının elemanları olduğunu gösteren tablo.

- **$DFFITS_I$**

$D_I(X'X,pS_{(I)})$ formu, *Belsley ve ark. (1980)* tarafından tanımlanan $DFFITS_I$ in karesinin çok değişkenliye uyarlanmış formu olup burada tekrar incelenmeyecektir.

$D_I\left(X'X, pS_{(I)}\right)$ ölçüsü J_I^{tr} sınıfının bir üyesi olduğundan, $DFFITS_I^2$ de bu sınıfın bir üyesidir.

- **W_I**

(3.42) eşitliği ile tanımlı bu istatistiğin çok değişkenli regresyonda birden fazla satırın etkisinin incelenmesi için kullanılan formu

$$W_I = \left(vecG_I\right)'\left[\left(S_{(I)}^{-1} \otimes X'_{(I)}X_{(I)}\right)\right]\left(vecG_I\right)/n-m \qquad (3.27)$$

dır. Burada

$$G_I = \frac{n-m}{m}\left(X'_{(I)}X_{(I)}\right)^{-1}X'_I\left(Y_I - X_I\hat{B}_{(I)}\right)$$

olarak tanımlıdır.

(3.27) formunun *(3.21)* eşitliğinin kullanımı ve gerekli düzenlemelerin yapılmasıyla

$$W_I = \frac{1}{n-m}tr\left(G'_I\left(X'_{(I)}X_{(I)}\right)G_I\left(S_{(I)}^{-1}\right)'\right)$$

$$= \frac{(n-m)^2}{(n-m)m^2}tr\left[\left(Y_I - X_I\hat{B}_{(I)}\right)'X_I\left[\left(X'_{(I)}X_{(I)}\right)^{-1}\right]'\left(X'_{(I)}X_{(I)}\right)\left(X'_{(I)}X_{(I)}\right)^{-1}\right.$$
$$\left.X'_I\left(Y_I - X_I\beta_{(I)}\right)\left(S_{(I)}^{-1}\right)'\right]$$

$$= \frac{n-m}{m^2}tr\left[\left(\frac{E_I}{\left(I-H_I\right)}\right)'H_I\left(I-H_I\right)^{-1}\left(\frac{E_I}{\left(I-H_I\right)}\right)\left[\left(\frac{E'_{(I)}E_{(I)}}{n-p-m}\right)^{-1}\right]'\right]$$

$$= \frac{(n-p-m)(n-m)}{m^2}tr\left[E'_I\left(E'_{(I)}E_{(I)}\right)^{-1}E_I\left(I-H_I\right)^{-3}H_I\right]$$

$$= \frac{(n-p-m)(n-m)}{m^2} tr\left[Q_I\left(I-H_I-Q_I\right)^{-1}\left(I-H_I\right)\left(I-H_I\right)^{-3}H_I\right]$$

$$= \frac{(n-p-m)(n-m)}{m^2} tr\left[H_I Q_I\left(I-H_I-Q_I\right)^{-1}\left(I-H_I\right)^{-2}\right] \qquad (3.28)$$

formuna dönüşür. Bu form

$f = \frac{(n-m-p)(n-m)}{m^2}$, $a = 1$ ve $b = -2$ için J_I^{tr} sınıfının bir üyesi olduğu

görülebilir.

3.2.2.2 J_I^{det} Sınıfına Ait Ölçüler

- **AP_I**

Andrews ve Pregibon (1978)'un, (2.23a) eşitliği ile tanımladıkları bu ölçünün çok değişkenli regresyonda birden fazla satırın etkisinin araştırılmasını sağlayan formu:

$$AP_I = \frac{det\left(Z'_{(I)}Z_{(I)}\right)}{det(Z'Z)}$$

şeklindedir. Bu form

$$det\left(Z'_{(I)}Z_{(I)}\right) = det\left(X'_{(I)}X_{(I)}\right)det\left(E'_{(I)}E_{(I)}\right)$$

ve

$$det(Z'Z) = det(X'X)\,det(E'E)$$

eşitlikleri yardımıyla aşağıdaki şekilde düzenlenebilir.

$$AP_I = \frac{det\left(X'_{(I)}X_{(I)}\right)det\left(E'_{(I)}E_{(I)}\right)}{det(X'X)\,det(E'E)}$$

$$= \frac{det(X'X)\,det\left(I - H_I\right)det\left(E'_{(I)}E_{(I)}\right)}{det(X'X)\,det(E'E)}$$

$$= det\left(I - H_I\right)\frac{det\left(E'_{(I)}E_{(I)}\right)}{det(E'E)}$$

$$= det\left(I - H_I\right)\frac{det(E'E)\,det\left[\left(I - H_I - Q_I\right)\left(I - H_I\right)^{-1}\right]}{det(E'E)}$$

$$= det\left(I - H_I\right)det\left[\left(I - H_I - Q_I\right)\left(I - H_I\right)^{-1}\right]$$

$$= det\left(I - H_I - Q_I\right) \qquad (3.29)$$

(3.29) eşitliğinde

$f = 1$, $a = 1$ ve $b = 0$ olarak seçilirse bu ölçünün J_I^{det} sınıfının bir üyesidir.

- **$COVRATIO_I$**

Belsley ve ark. (1980) tarafından *(2.13)* eşitliği ile tanımlanan bu ölçünün çok değişkenli regresyonda birden fazla satırın etkisinin araştırılmasını sağlayan formu

$$COVRATIO_I = \frac{det\left(cov\left(vec\left(\hat{B}_{(I)}\right)\right)\right)}{det\left(cov\left(vec\hat{B}\right)\right)} = \frac{det\left(S_{(I)} \otimes \left(X'_{(I)}X_{(I)}\right)^{-1}\right)}{det\left(S \otimes (X'X)^{-1}\right)} \qquad (3.30)$$

olarak verilebilir.

Bu form $det\left(A_{nxn} \otimes B_{pxp}\right) = [det(A)]^p [det(B)]^n$ matris özelliği yardımıyla

$$COVRATIO_I = \frac{\left[det(n-p-m)^{-1} E'_{(I)} E_{(I)}\right]^p [det(X'X)]^n}{\left[det(n-p)^{-1} E'E\right]^p \left[det\left(X'_{(I)} X_{(I)}\right)\right]^n}$$

olarak yazılabilir. Burada

$$det\left(E'_{(I)} E_{(I)}\right) = det(E'E)\, det\left[\left(I - H_I - Q_I\right)\left(I - H_I\right)^{-1}\right]$$

ve

$$det\left(X'_{(I)} X_{(I)}\right) = det(X'X)\, det\left(I - H_I\right)$$

eşitlikleri yerine konulur ve gerekli düzenleme yapılırsa

$$COVRATIO_I = \left(\frac{n-p}{n-p-m}\right)^m det\left[\left(I - H_I - Q_I\right)^p \left(I - H_I\right)^{-(r+p)}\right] \quad (3.31)$$

formuna dönüşecektir.

(3.31) eşitliği

$f = \left(\frac{n-p}{n-p-m}\right)^m$, $a = p$ ve $b = -(r+p)$ için J_I^{det} sınıfının bir üyesidir.

- **FVARATIO$_I$**

Belsley ve ark (1980) tarafından tanımlanan bu ölçünün çok değişkenli regresyonda birden fazla satırın etkisini araştırmamızı sağlayan formu;

$$FVARATIO_1 = \frac{det\left(cov\left(vec\left(X_1\hat{B}_{(1)}\right)\right)\right)}{det\left(cov\left(vec\left(X_1\hat{B}\right)\right)\right)} = \frac{det\left[S_{(1)} \otimes \left(I - H_1\right)^{-1} H_1\right]}{det\left(S \otimes H_1\right)} \quad (3.32)$$

olup bu form *(1.2), (1.8), (1.6)* eşitliklerinin çok değişkenliye uyarlanmış formu ve gerekli matris özellikleri yardımıyla

$$FVARATIO_1 = \frac{\left[det\left[\left(I - H_1\right)^{-1} H_1\right]\right]^r \left[det\left[S_{(1)}\right]\right]^m}{\left[det(S)\right]^m \left[det\left(H_1\right)\right]^r}$$

$$= \frac{\left[det\left(I - H_1\right)\right]^{-r} \left[det\left(\frac{E'_{(1)}E_{(1)}}{n-p-m}\right)\right]^m}{\left[det\left(\frac{E'E}{n-p}\right)\right]^m}$$

$$= \left(\frac{n-p}{n-p-m}\right)^{rm} \left[det\left(I - H_1\right)\right]^{-r} \left[det\left(E'_{(1)}E_{(1)}\right)\right]^m \left[det(E'E)\right]^{-m}$$

$$= \left(\frac{n-p}{n-p-m}\right)^{rm} \left[det\left(I - H_1\right)\right]^{-r} \left[det\left(I - H_1 - Q_1\right)\right]^m \left[det\left(I - H_1\right)\right]^{-m}$$

$$= \left(\frac{n-p}{n-p-m}\right)^{rm} \left[det\left(I - H_1\right)\right]^{-(r+m)} \left[det\left(I - H_1 - Q_1\right)\right]^m \quad (3.33)$$

$$= \left(\frac{n-p}{n-p-m}\right)^{rm} det\left[\left(I - H_1\right)^{-(r+m)} \left(I - H_1 - Q_1\right)^m\right]$$

olarak düzenlenebilir.

(3.33) nolu eşitlik

$f = \left(\frac{n-p}{n-p-m}\right)^{rm}$, $a = m$ ve $b = -(r+m)$ için J_I^{det} sınıfının bir üyesidir.

ÖLÇÜ	SINIF	f(n,p,r,m)	a	b
$D_I(X'X,pS)$	J_I^{tr}	$\frac{n-p}{p}$	0	-2
$D_I(X'_{(I)}X_{(I)},pS)$	J_I^{tr}	$\frac{n-p}{p}$	0	-1
$D_I(X'X,pS_{(I)})$	J_I^{tr}	$\frac{n-p-m}{p}$	-1	-1
$D_I(X'_{(I)}X_{(I)},pS_{(I)})$	J_I^{tr}	$\frac{n-p-m}{p}$	-1	0
W_I	J_I^{det}	$\frac{(n-m-p)(n-m)}{m^2}$	-1	-2
AP_I	J_I^{det}	1	1	0
$COVRATIO_I$	J_I^{det}	$\left(\frac{n-p}{n-p-m}\right)^{rp}$	p	-(r+p)
$FVARATIO_I$	J_I^{det}	$\left(\frac{n-p}{n-p-m}\right)^{rm}$	m	-(r+m)

Tablo 3.3 J_I^{tr} ve J_I^{det} sınıfları içinde tanımlı tanılama ölçüleri

3.3 Etki ve Rezidü Elemanlarının Belirlenmesi

Bir ve birden fazla satırın etkisini ölçen tanılama yöntemleri, rezidü ve (veya) etki elemanlarına bağlı olarak yazılabilirler *(Gray (1985))*. Bu yöntemlerin sonuçlarının değerlendirilmesinde grafiksel gösterimler büyük kolaylıklar sağlar. *Atkinson (1986)* ve

Fung (1993) veri analizi yapılırken robust kestiriminin kullanılması ve etkili gözlemlerin ve sapan değerlerin saptanmasında grafiksel yöntemlerin kullanılmasını önermişlerdir. *McCulloch ve Meeter (1983), Gray (1983,1985,1986) ve Hadi (1992) de* bir satırın etkisinin araştırılmasında leverage(etki)-rezidü plotlarını kullanmışlardır.

Etki ve rezidü bileşenlerine dayalı olarak ortak çoklu satırın etkisinin saptanması, hem tek değişkenli hem de çok değişkenli durumlarda hala geniş bir şekilde araştırılması gereken bir konudur. Bu konu üzerine yapılan çalışmalar, etki (leverage) ve rezidü elemanlarının rollerinin, ayrı ayrı etkinin birinden mi yoksa onların birleşimlerinden mi olduğunun incelenmesiyle yapılır.

Tanılama ölçülerinin formlarında rezidülere bağlı olmayan ifade genel etki (leverage) matrisi *(L_I)*, geri kalan matris çarpımı ise genel rezidü matrisi *(R_I)* olarak tanımlanır. R_I ve L_I'nın belirlenmesi rezidü formunun araştırmacı tarafından seçimine bağlıdır. Tek değişkenli durumda farklı rezidü tanımlamaları çok değişkenliye ve çoklu satır etkisinin araştırılmasına genelleştirilmişi ele alınarak yapılır.

Örneğin:

i. dahili standart rezidü $\left(e_i(e'e)^{-1/2}\right)$ nün karesel formu;

$$Q_I = E_I(E'E)^{-1}E_I',$$

i. harici standart rezidü $\left(e_i\left(e'_{(i)}e_{(i)}\right)^{-1/2}\right)$ nün karesel formu;

$$E_I\left(E'_{(I)}E_{(I)}\right)^{-1}E_I' = Q_I\left(I - H_I - Q_I\right)^{-1},$$

i. dahili Student rezidü $\left(e_i/s(1-h_{ii})^{1/2}\right)$ nün karesel formu;

$$\left(I - H_I\right)^{-1/2}Q_I\left(I - H_I\right)^{-1/2},$$

i. harici Student rezidü $\left(e_i/s_{(i)}(1-h_{ii})^{1/2}\right)$ nün karesel formu;

$$\left(I - H_I\right)^{-1/2}Q_I\left(I - H_I - Q_I\right)^{-1}\left(I - H_I\right)^{-1/2}$$

olarak çoklu satır için çok değişkenliye genelleştirilebilir.

Student rezidülerin genelleştirilmiş formu kullanımı ile verilen J_I^{tr} içindeki ölçüler için R_I genel rezidü ve L_I genel etki matrisi *Tablo 3.4* de verildi. Bu tabloya göre J_I^{tr} sınıfına ait ölçüler $(f(.)tr(L_IR_I))$ formuna sahiptir ve J_I^{det} sınıfına ait ölçüler, $f(.)det(L_IR_I)$ olarak ifade edilebilir.

Tek satır etkisinin belirlenmesi için etki ve rezidü bileşenlerinin benzer bir parçalanışı diğer ölçüler için *Tablo 3.4* deki matrislerin skaler karşılıklarıyla yer değiştirilmesi sayesinde elde edilebilir.

Çok değişkenli doğrusal regresyon modellerinde birden fazla satır etkisinin incelenmesinde etki ve rezidü elemanlarının ayrı ayrı belirlenmesi oldukça zordur. Çünkü *mxm* tipindeki etki veya rezidü matrisi içindeki bilgiler bir skalere indirgenip iç çarpıma dayalı olarak J_I^{tr} içindeki ölçüler için etkinin karakterleri incelenmelidir.

$\mathrm{tr}(AB) = (\mathrm{vec}(A'))'(\mathrm{vec}(B))$ matris özelliğinden yararlanarak,

$$tr(L_IR_I) = [vec(L_I)]'[vec(R_I)]$$

eşitliği yazılabilir. Bu form skaler çarpma özelliği yardımıyla,

$$tr(L_IR_I) = \|vec(L_I)\|_2 \|vec(R_I)\|_2 Cos\theta_I \qquad (3.34)$$

olarak da ifade edilebilir. Burada θ_I; $vec(L_I)$ ile $vec(R_I)$ arasındaki açı, $\|.\|_2$ de; Euclid normu ifade eder. İşlemleri biraz daha basite indirgemek için $\|vecA\|$ yerine Frobenius matris formu olan

$$\|vecA\| = \|A\| = \left(\sum_{i,j} a_{ij}^2\right)^{1/2}$$

eşitliğini yazalım. L_I *ve* R_I pozitif semi definite matrisler *(Graybill (1983))* olduklarından bunların normları, özdeğerleri $\lambda_k(.)$ lara bağlı olarak

$$\|vecL_I\| = \|L_I\| = \left(\sum_{k=1}^{m} \lambda_k^2(L_I)\right)^{1/2}$$

$$\|vecR_I\| = \|R_I\| = \left(\sum_{k=1}^{m} \lambda_k^2(R_I)\right)^{1/2}$$

ifade edilebilirler.

Kısaca $\|L_I\|$ için Γ_I ve $\|R_I\|$ için $\Re_I$ formlarını kullanalım. Bu niceliklerin her biri etki (leverage) ve rezidü bileşenlerinin toplam etkiye göreceli katkısını temsil etmektedir. Her bir bileşenin göreceli katkısı açı ile değişmemesine rağmen, gerçek etki $Cos\theta_I$ ya bağımlıdır. Alt kümeler arasında anlamlı karşılaştırmalar yapmak için her bir alt küme için Γ_I ve $\Re_I$ 'nın ölçeklendirilmesi gerekir ve böylece göreceli katkı gerçek etkiyi yansıtır.

Bu ölçeklendirilmiş nicelikleri,

$$\Gamma_I^* = \Gamma_I(Cos\theta_I)^{1/2}$$

ve

$$\Re_I^* = \Re_I(Cos\theta_I)^{1/2}$$

olarak tanımlarsak,

$$tr(L_I R_I) = \Gamma_I^* R_I^*$$

eşitliği elde edilir.

J_I^{det} içindeki ölçülerin, skaler leverage ve rezidü bileşenlerinin ayrıştırılması, *Tablo 3.4* içinde verilen rezidü ve leverage matrislerin determinantlarının basit bir şekilde kullanımı ile elde edilebilir. Örneğin, $1/AP_I$ ölçüsünün etki bileşeni

$$det(X'X)/det\left(X'_{(I)}X_{(I)}\right) = det\left[\left(I - H_I\right)^{-1}\right]$$

ve rezidü bileşeni,

$$det(E'E)/det\left(E'_{(I)}E_{(I)}\right) = det\left[\left(I - H_I - Q_I\right)^{-1}\left(I - H_I\right)\right]$$

ile belirlenebilir.

ÖLÇÜLER	ETKİLER(L_I)	REZİDÜLER(R_I)
$D_I(X'_{(I)}X_{(I)},pS)$	H_I	$(I-H_I)^{-1/2}Q_I(I-H_I)^{-1/2}$
$D_I(X'X,pS)$	$H_I(I-H_I)^{-1}$	$(I-H_I)^{-1/2}Q_I(I-H_I)^{-1/2}$
$D_I(X'_{(I)}X_{(I)},pS_{(I)})$	H_I	$(I-H_I)^{-1/2}Q_I(I-H_I-Q_I)^{-1}(I-H_I)^{1/2}$
$D_I(X'X,pS_{(I)})$	$H_I(I-H_I)^{-1}$	$(I-H_I)^{-1/2}Q_I(I-H_I-Q_I)^{-1}(I-H_I)^{1/2}$
W_I	$H_I(I-H_I)^{-2}$	$(I-H_I)^{-1/2}Q_I(I-H_I-Q_I)^{-1}(I-H_I)^{1/2}$
$1/AP_I$	$(I-H_I)^{-1}$	$(I-H_I-Q_I)^{-1}(I-H_I)$
$1/COVRATIO_I$	$(I-H_I)^r$	$\left[(I-H_I-Q_I)^{-1}(I-H_I)\right]^p$
$1/FVARATIO_I$	$(I-H_I)^r$	$\left[(I-H_I-Q_I)^{-1}(I-H_I)\right]^m$

Tablo 3.4. J_I^{tr} ve J_I^{det} sınıfları içinde tanımlı tanılama ölçüleri için rezidü ve etki matrislerinin genel formları

Etki ve rezidü bileşenlerinin ayrılabilir olması birden fazla satırın etkisinin incelenmesi için, grafiksel bir gösterimin yapılabileceğini ortaya çıkarır. J_I^{det} sınıfı içinde $det(L_I)$'in $det(R_I)$'e karşı J_I^{tr} sınıfı içinde ise Γ_I^*'ın $\Re_I^*$'a karşı serpilme diyagramlarının oluşturulması ile etki ve rezidü niceliklerinin rolleri kolaylıkla incelenebilir.

Sabit etkinin eğrileri $y=\frac{1}{x}$ formuna sahip olup, noktalar grafiğin sol alt köşesine doğru kümeleştiklerinden J_I^{tr} içinde; Γ_I^* yerine $log\Gamma_I^*$, $\Re_I^*$ yerine $log\Re_I^*$ formu ve J_I^{det} içinde; $det(L_I)$ yerine $logdet(L_I)$, $det(R_I)$ yerine $logdet(R_I)$ formu kullanılarak kümeleşme ortadan kaldırılabilir. Bu durumda eğriler eğimi -1 olan doğrulardan meydana gelir ve koordinatların toplamı, etkinin logaritmasını vermektedir. *McCulloch ve Meeter (1983);* tek satırın Cook uzaklığı yardımıyla etkisinin incelenmesi için; Student rezidülerin geri kalan leverage fonksiyonlarına karşın

rezidü elemanlarının göreceli katkılarını temsil eder). Γ_I ölçülerinin maksimum ortak etkisi ve $\Re_I$ ölçülerinin maksimum ortak rezidüsü (sapan değerliliği) olmak üzere $log\Gamma_I$ nın $log\Re_I$ ya karşı benzer grafiğini incelemeye ilgi duyabiliriz. Önceki grafik gerçek etkiyi gözler önüne sererken sonraki etki için bir üst sınır vermemektedir. $m=1$ olduğunda bunlar birbirleriyle uyuşmaktadır.

$Cos\theta_I$ nın rolü, hangi alt kümenin en büyük olası etkiye $(\|L_I\|\|R_I\|)$ sahip olduğunu yansıtmak içindir. Etki ve rezidü elemanları ile $Cos\theta_I$ tam olarak ölçülebilir. Fakat $\sum_{i,j\in I} h_{ij}q_{ij}$ terimi ile bu elemanların biri diğerinden üstün olabilir ve bu nedenle bunların herbiri h_{ij} ve q_{ij} nin benzer ya da farklı işaretlere sahip olduğunun derecesini ölçer. *(Chatterjee ve Hadi (1988 sh: 97-99))* basit regresyonda terimlerin ortak etkileri içinde satır ekleme veya çıkarma olaylarının nasıl olacağı hakkında güzel açıklamalar sunmuşlardır.

Bu grafiklerin karşılaştırılması ortak etkili gözlemler olduğunda birinin diğerini iptal eden alt kümelerinin belirlenmesine yardımcı olur. Yani etki ya da rezidü elemanları içinde uç olan herhangi bir alt küme incelenebilir ve bu alt küme için gerçek etki ile bu değerler karşılaştırılabilir. Hatta üç boyutlu plotların hareketlendirilmesini sağlayan programlar bulunursa, Γ_I karşın $\Re_I$ karşın $Cos\theta_I$ grafiği hareket ettirilip incelenebilir ve belki de renk katarak etkinin şiddeti saptanabilir.

3.4 J_I Sınıfının Kullanımındaki Yararlar

I indisli verilen bir alt küme üzerinde bir tek ölçünün hesaplanması pek çok tanılama ölçülerinin, çok değişkenliye uzanımları ve tek değişkenliler için verilen ifadelerin orijinal formları sayısal hesaplamalar için uygun değildir. Örneğin; $X'X$ in sonucunun önceden verildiği varsayımı ile Cook uzaklığı $(D_I(X'X, pS))$ nın hesaplaması için p^3. dereceden hesaplama yapılması gerekirken, Cook uzaklığının çok değişkenliye uyarlanmış formu için benzer olarak p^3xr^3 üncü dereceden hesaplama yapılması gerekir. Diğer taraftan *(3.20)* e karşın *(3.23)* ün kullanımı sadece m satırı silinen bir matrisin tersinin alınması ve çarpma işlemlerini içermektedir. Bu sadece m^3 üncü dereceden bir hesaplama gerektirmektedir. Silinen alt kümenin büyüklüğü m genellikle p den çok daha küçük olacağından hesaplamada önemli kolaylıklar olacaktır.

Satırların her bir alt kümesi üzerinde çeşitli etki ölçüleri düşünüldüğü zaman çarpmalar akıllıca yapılırsa *(3.18)* ve *(3.19)* un kullanımı bize oldukça önemli ek hesaplama kolaylığı verir.

Örnek; *Tablo3.3* den görülebilir ki Cook uzaklığı $(D_I(X'X, pS))$'nın hesaplanmasından sonra $(I-H_I)^{-2}H_IQ_I$ nun değeri mevcuttur. Öyle ki tek bir $(I-H_I-Q_I)^{-1}$ ek çarpımı ile $(I-H_I)^{-2}H_IQ_I(I-H_I-Q_I)^{-1}$ formu kolaylıkla elde edilebilir.

Genel olarak, eğer alt kümelerin bazı kolleksiyonunun her alt kümesi üzerinde elde edilen belirli bir ölçü kümesi, önceden bilinirse bir alt küme için J_I^{tr} ve J_I^{det} sınıfının çeşitli üyelerinin elde edilmesindeki marjinal maliyet önemsenmez.

Bu sınıfın kullanılmasındaki diğer bir avantaj incelenen herbir alt küme için istenen ölçü üzerinde bir üst sınırın hesaplanabilmesidir. Önemli sayıda alt küme üst sınır temel alınarak elendiği durumda, etkili alt kümelerin belirlenmesi için yapılması gereken işlem sayısı azalacaktır.

Örneğin; çok değişkenli doğrusal regresyonda tanımlanan $(D_I(X'X, pS))$ Cook uzaklığı

$$D_I(X'X, pS) = \frac{n-p}{p} tr\left[\left(I-H_I\right)^{-2} H_IQ_I\right]$$

için bir üst sınır elde edelim. Bu form $tr(ABC) \le tr(A)tr(B)tr(C)$ özelliğinden yardımıyla,

$$\frac{p}{n-p} D_I\left(X'X, pS\right) \le \frac{tr(H_I)tr(Q_I)}{\left[1-tr(H_I)\right]^2} \qquad (tr(H_I) < 1)$$

olarak yazılabilir. Dolayısıyla Cook uzaklığı için üst sınır $\frac{tr(H_I)tr(Q_I)}{\left[1-tr(H_I)\right]^2}$ olarak elde edilir. Bu ifade *Cook ve Weisberg (1982)* in tek değişkenli doğrusal regresyon modelleri için verdiği sınırın genelleştirilmesi olarak düşünülebilir.

Benzer bir yaklaşım kullanılarak J_I^{tr} sınıfındaki ölçüler için üst sınırlar elde edilebilir *(Barrett ve Ling (1992))*.

4. BÖLÜM

YEREL ETKİ YAKLAŞIMI

4.1 GİRİŞ

Tanılama yöntemlerinin çoğu satır silme işlemine dayalı olarak yapılırken, son yıllarda satır ya da satırların alt kümesine bozulma (perturbation) uygulayarak etkinin değerlendirilmesi işlemi yoğunluk kazanmıştır.

Bu yöntem ilk olarak *Cook (1986)* tarafından sunulmuş daha sonra birçok araştırmacının çalışmalarına ışık tutmuştur. Bu yöntemden yararlanılarak yapılan çalışmalardan birkaçını şöyle sıralayabiliriz.

***Tsai (1986)*:** İstatistiksel model üzerinde küçük bozulmaların yerel etkisi ve score (ölçekleme) test istatistiği arasındaki ilişkiyi incelemiştir.

***Beckman ve ark. (1987)*:** Varyans analiz modeli içindeki genel varsayımlardan hareketle bozulmanın etkisini değerlendirmede yerel etki yöntemini kullanmıştır.

***Lawrence (1988)*:** Sabit varyans varsayımı üzerinde küçük bozulmaların neden olduğu parametre kestirim dönüşümlerinin yerel değişikliklerini incelemek için tanımladığı tanılama yönteminde *Cook*'un yaklaşımından yararlanmıştır. *Tsai ve Wu (1992)* de *Box-Cox* regresyon modeli içinde kuvvet kestirim dönüşümü üzerinde satırlardaki bozulmaların etkisini değerlendirmede *Lawrence*'nin yerel etki analizini kullanmış ve bu yöntemin *Cook (1986)* tarafından önerilen yerel etki analizi ile aynı olduğunu göstermişlerdir.

***Thomas ve Cook (1989)*:** Genelleştirilmiş doğrusal modellerde verilerin cevap değişkenlerinin vektöründe, satırların ağırlıklarında ve açıklayıcı değişkenlerde meydana gelen küçük bozulmaların etkisini değerlendirmede *Cook* yaklaşımını kullanmışlardır.

***Thomas (1990)*:** Genelleştirilmiş doğrusal modellerde regresyon katsayılarının güven elipsoidinin hacmi üzerindeki etkisi için bir ölçü düşünmüş ve yerel etki yönteminin kullanımı ile veri içindeki küçük değişikliklerin güven elipsiodi üzerindeki etkilerini incelemiştir.

***Thomas ve Cook (1990)*:** Genelleştirilmiş doğrusal modelden özelleştirilmiş nokta kestirimi üzerinde verideki küçük bozulmaların etkisini değerlendirmede bu yöntemi kullanmıştır.

Schwarzmann (1991): Lineer regresyon modellerinde Y cevap değişkeninin satırlarında meydana gelen bozulma için e rezidülerin vektörü ve maksimum eğriliğin doğrultusu arasındaki ilişkiyi incelemiştir. $w_0=0$ noktasındaki LD fonksiyonunun maksimum eğriliğinin doğrultusu olan l_{max} ın e rezidü vektörü ile orantılı olduğunu ve varyansın maksimun likelihood kestiricisi $\tilde{s}^2$ olmak üzere, $l_{max} = e/\sqrt{n\tilde{s}}$ eşitliğinin yazılabileceğini göstermiştir.

Schall ve Dunne (1992): Varyans şişirme çarpanı ile yerel etki arasındaki ilişkinin varlığını ortaya çıkarmışlardır.

Paula, G.A. (1993, 1995): Kısıtlanmış regresyon modellerinde yerel etkinin değerlendirilmesi *(1993)*, daha sonra da; kısıtlanmış genelleştirilmiş lineer modellerde etki ve rezidüler *(1995)* konusu üzerine çalışmıştır.

Bozulma içeren tanılama yöntemleri daha çok ***Cook ve Weisberg (1982)*** ve ***Welsch (1982)*** tarafından incelenmiştir. İşlemler doğru dağılım fonksiyonu altında, bir fonksiyonel yaklaşım ile eğrinin etkinliği incelenerek yapılmaktadır.

Bu bölümde istatiksel bir modelin varsayımlarında meydana gelen küçük bozulmalarının meydana getirdiği etkiyi saptamak için ***Cook (1986)*** tarafından tanımlanan **'yerel etki'** yöntemi ve bu yöntemdeki eksiklikleri ortadan kaldırmak amacıyla ***Billor ve Loynes (1992)*** tarafından tanımlanan **'yerel etkiye yeni yaklaşım'** yöntemi verilecektir. Ama öncelikle bu konulara temel oluşturacak bozulma ve etki grafiği kavramlarını verelim.

4.2 Bozulma

Veride ya da model varsayımlarında yapılan değişiklikler bozulma (perturbation) olarak adlandırılır

İki farklı bozulma tipi vardır:

1) Model Bozulması

2) Veri Bozulması

1) Model Bozulması: Model varsayımlarında yapılan değişiklikler model bozulması olarak adlandırılır. Örneğin; normal dağılıma sahip hata terimlerinin σ^2 sabit varyansına sahip olması özelliği, sabit varyanslı olmama özelliği ile yer değiştirirse bir model bozulması yapılmış olur. Daha açık olarak, $\varepsilon \sim N(0,\sigma^2 I)$ varsayımı; w_i bir bozulma terimi olmak üzere $(i=1,2,...,n)$; $\varepsilon \sim N(0,\sigma^2 diag^{-1}(w_i))$ ile yer değişirse model bozulmuştur.

2) Veri Bozulması: Mümkün olan bozulmanın tipi veri kümesine bağımlı olursa veri bozulması yapılmıştır. Örneğin; normal doğrusal regresyonda açıklayıcı değişkenlerin matrisi ya da cevap değişkenlerinin vektöründe değişiklik yaparsak veri bozulması yapmış oluruz.

Veri bozulmasının yapılmasını gerektiren iki ana neden vardır.

- Ölçüm hatalarının varlığı,
- Sapan değerlerin varlığı.

Sapan değerler ölçüm hatalarından kaynaklanabildiği gibi fonksiyonel formun uydurulması ile de ortaya çıkabilirler.

Model yada veri bozulduğunda, bozulma tiplerinin etkisinin nasıl incelenmesi gerektiği *Cook (1986)* 'un tanımladığı ölçüde yeterince açık değildir. Bu konu *Billor ve Loynes (1992)* 'un tanımlamış olduğu yöntem anlatılırken detaylı olarak verilecektir.

4.3 Cook'un $D_i(w)$ Ölçüsü

(1.12) eşitliği ile tanımlı $Y=X\beta+\varepsilon$ çoklu doğrusal regresyon modelini ele alalım. *nx1* tipindeki ε vektörünün ε_i elemanları sıfır ortalamalı σ^2 bilinen varyanslı bağımsız normal rastgele değişken, *X; nxp* tipinde açıklayıcı değişkenlerin matrisi ve *Y; nx1* tipinde cevap değişkenlerinin bir vektörü olarak tanımlansın. Bu model üzerinde *i.* satırın uydurulmuş değer üzerindeki etkisi D_i Cook istatistiği kullanılarak hesaplanabiliyordu. Bu istatistik

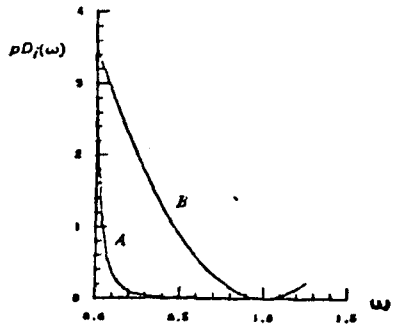
$$D_i = \frac{\left\|\hat{Y} - \hat{Y}_{(i)}\right\|^2}{p\sigma^2} \qquad (4.1)$$

olarak ifade edilebilir. Bu yöntemde D_i nin büyük olması ile *i.* satırın analiz sonuçlarını etkilemektedir sonucuna varılır ve satır veri kümesinden atılarak model yeniden düzenlenir. Fakat bu şekilde hareket etmek her zaman için sağlıklı sonuç vermeyebilir. Çünkü satırın etkili çıkmasının nedeni veriden değil de model varsayımlarındaki bazı eksikliklerden kaynaklanmış olabilir. Cook bu olayları düşünerek D_i istatistiğini tanımladıktan sonra varyans eşitsizliği olayını da beraber inceleyebilmek için ağırlığın (bozulmanın) önemli olduğu $D_i(w)$ istatistiğini tanımladı. Bu istatistik, varyansın sabit olması konusunda şüphe edildiğinde model varsayımlarından ε nun σ^2 sabit varyanslı olması özelliği yerine *i.* satıra *w* bozulma terimi etkilettirilip sabit varyanslılık özelliği bozularak uygulanmaktadır. Bu durumda *(4.1)* ile ifade edilen Cook istatistiği

$$D_i(w) = \frac{\left\|\hat{Y} - \hat{Y}_w\right\|^2}{p\sigma^2} \qquad (4.2)$$

olarak tanımlanır. Burada $\hat{Y}_w$; *i.* satıra *w* bozulma terimi etkilettirildikten sonra elde edilen uydurulmuş değerler vektörüdür. *w* nun bazı değerleri için D_i, $D_i(w)$ ya eşit sonuçlar verir. Yapılan işlem sonucu $D_i(w)$ değeri büyük bulunursa *i.* gözlemin varyansının σ^2 değilde σ^2/w olduğu söylenir. Bozulma tek bir satıra etkilettirilebildiği gibi birden fazla satıra da etkilettirilebilir.

(1.24) eşitliği ile tanımladığımız modelde *A* ve *B* gibi iki satıra bozulma terimi etkilettirelim. Yapılan bozulmalar ve bozulma sonucu istatistiğin aldığı değerler *Şekil 4.1* de görülmektedir.



Şekil 4.1. (1.24) nolu modelde A ve B satırları için w karşın $pD_i(w)$ grafiği

Bu şekil incelenerek her iki satır için *w* ya 0 ve 1 değerleri verilerek yapılan bozulmaların, eşit etkiye sahip olduğu, ama bozulma *0<w<1* arasında değişiyorken, *B* satırının sabit varyanslılık özelliğini yitirmiş olması olasılığının, *A* satırınınkinden daha çok olduğu söylenebilir. Bozulma tüm satırlar için ayrı ayrı yapılırsa karşımıza yukarıdakine benzer n tane grafik çıkar. Bu grafiklerin büyük bir bölümü üst üste yer alırken bazıları o gruptan farklı yerlerde bulunabilirler. Farklı yerlerde bulunan bu grafikler hangi satırı temsil ediyorsa o satır için modelin sabit varyanslılık özelliğini büyük ölçüde etkileyen satırdır yorumu yapılır.

Görüldüğü gibi satır silme ölçüleri tek başına büyük hataları ortaya çıkarma dışında ki problemleri çözmek için yeterli olamazlar.

(4.2) eşitliği ile tanımlı ifade gerekli düzenlemelerin yapılması ile

$$D_i(w) = \left[\frac{\left\| Y - \hat{Y}_w \right\|^2 - \left\| Y - \hat{Y} \right\|^2}{p\sigma^2} \right]$$

olarak ta ifade edilebilir. Bu eşitlik aşağıdaki düzenlemelerin yapılması ile

$$pD_i(w) = \left[\frac{\left\| Y - \hat{Y}_w \right\|^2 - \left\| Y - \hat{Y} \right\|^2}{\sigma^2} \right]$$

$$= \left[\frac{\sum_{i=1}^{n} (y_i - \hat{Y}_w)^2 - \sum_{i=1}^{n} (y_i - \hat{Y})^2}{\sigma^2} \right]$$

$$= 2\left[\left(-\frac{1}{2}\log 2\pi - \frac{1}{2}\log\sigma^2 - \frac{\sum_{i=1}^{n} (y_i - \hat{Y})^2}{2\sigma^2} \right) - \left(-\frac{1}{2}\log 2\pi - \frac{1}{2}\log\sigma^2 - \frac{\sum_{i=1}^{n} (y_i - \hat{Y}_w)^2}{2\sigma^2} \right) \right]$$

$$= 2\left[L(\hat{\beta}) - L(\hat{\beta}_w) \right] \qquad (4.3)$$

formuna dönüşür. Burada $\hat{\beta}$; β nin m.l.k.si ve $\hat{\beta}_w$; β nın *i.* satırı *w* bozulma terimini içeriyorken elde edilen m.l.k.si dir.

4.4 Cook Yaklaşımında Model Tanıtımı ve Etki Grafiği

Cook (1986) 'un tanımlamış olduğu yöntemde,

θ : *px1* tipinde, m.l.k.si $\hat{\theta}$ olan bilinmeyen parametrelerin vektörü,

$L(\theta)$: gözlemlenmiş veri ve bozulmamış model için log-likehood fonksiyonu,

Ω : küçük bozulmaları içeren açık bir küme,

w : Ω içinde tanımlı model içindeki bozulmaları ifade eden *mx1* tipinde bir vektör,

$L(\theta/w)$: modele bozulma uygulandıktan sonra elde edilen log-likelihood fonksiyonu,

$\hat{\theta}_w$: modele bozulma uygulandıktan sonra elde edilen θ nın m.l.k.si olarak tanımlı olsun.

Ω içinde öyle bir w_0 noktası vardır ki bu nokta hiçbir bozulmayı temsil etmez. Yani $L(\theta/w_0) = L(\theta).\ L(\theta/w)$, $(\hat{\theta}', w_0')$ nün bir komşuluğu içinde iki kez türevlenebilir sürekli bir fonksiyon olsun. w bozulmasının etkisi, $\hat{\theta}$ ile $\hat{\theta}_w$ karşılaştırılarak elde edilmek isteniyorsa *Cook ve Weisberg (1982)* tarafından tanımlanan likelihood displacement

$$LD(w) = 2\left[L(\hat{\theta}) - L(\hat{\theta}_w)\right]$$

kullanılır.

$LD(w)$ nun w'a karşın grafiği, bozulmanın meydana getirdiği etkiyi ortaya çıkarır. Bu grafik w; Ω içinde değişiyorken, $\alpha(w)$ vektörünün değerleri ile oluşturulan

$$\alpha(w) = \begin{pmatrix} w \\ LD(w) \end{pmatrix}_{(m+1)x1}$$

geometrik yüzeyidir.

Diferansiyel geometride bu formun yüzeyi *MONGE PATCH (Milliman ve Parker (1977))* ile adlandırılırken, w nun boyutu olan m, 1 e eşit olduğunda $\alpha(w)$ **"etki grafiği"** olarak ifade edilir.

Parametrelerin kümesi $\theta = (\theta_1, \theta_2)$ olarak parçalanır. θ nın bir alt kümesi üzerinde etkinin saptanması işlemi de yerel etki yönteminin kullanılması ile incelenmiştir *(Cook, 1986)*.

Bozulma uygulanan satır sayısı m, *2* den küçük olduğunda gerekli bilgi direkt olarak etki grafiğinden elde edilir. Ama $m \geq 2$ ise bozulmanın etkisi birtakım yöntemlerden yararlanılarak bulunur. Bu yöntemlerden biri *Cook (1986)* tarafından sunulan **"yerel etki"** yöntemidir.

4.5 Yerel Etki

Herhangi bir w_0 noktası etrafındaki etki grafiğinin davranışlarını karakterize etmek için yerel etki yönteminden yararlanılır. Bu yöntem etki grafiğinin w_0 etrafındaki yerel hareketlerine bağlı olarak yürütülmektedir. Bu hareketler incelenirken geometrik

normal eğri kullanılır. Likelihood displacement yüzeyinin normal eğriliği (C_l), w_0 daki eğriye en iyi uyan çemberin yarıçapının tersi ya da eğri boyunca istenen yay uzunluğundaki teğet vektörün yatay eksenle yaptığı açıdaki değişim oranı olarak düşünülebilir.

4.5.1 Eğriliğin (C_l) Formu

w_0 noktasındaki düz eğrilerin eğrilikleri ;

$$C = \frac{|\dot{X}x\ddot{X}|}{|\dot{X}|^3}$$

eşitliği yardımıyla hesaplanır. Burada $\dot{X}$ ve $\ddot{X}$ sırasıyla X vektörünün w_0 noktasındaki birinci ve ikinci türevleri, "x" ise vektörel çarpımı ifade etmektedir.

$X = \alpha(w) = (w, LD(w))'$ olarak seçilirse w_0 noktasındaki etki grafiğinin eğriliği

$$C_l = \frac{|\dot{w}L\ddot{D}(w) - L\dot{D}(w)\ddot{w}|}{(\dot{w}^2 + L\dot{D}(w))^{3/2}}$$

olarak elde edilir. $\dot{w} = 1$, $\ddot{w} = 0$, $L\dot{D}(w) = 0$ eşitliklerinden yararlanırsak eğrilik

$$C_l = |L\ddot{D}(w)|$$

formuna indirgenir. $L\ddot{D}(w)$ türevi zincir kuralı yardımıyla hesaplanırsa

$$C = 2|l'\ddot{F}l| \qquad (4.4)$$

eşitliği elde edilir. Burada $\ddot{F}$; mxm tipinde elemanları

$$\frac{\partial^2 L(\hat{\theta}_w)}{\partial w_k \partial w_j}, \ (j,k = 1,...,m)$$

olan bir matristir.

$\ddot{F}$ nın formu yine zincir kuralı yardımıyla

$$\ddot{F} = J'\ddot{L}J$$

olarak yazılabilir. Burada J; pxm tipinde elemanları $\frac{\partial \hat{\theta}_{iw}}{\partial w_j}$, $i = 1,2,...,p$, $j = 1,2,...,m$ ve $\ddot{L}$; pxp tipinde elemanları $\frac{\partial^2 L(\hat{\theta})}{\partial \hat{\theta}_i^2}$ $i = 1,2,..,p$ olan bir matristir. ($-\ddot{L}$; $w=w_0$ noktasında bozulmuş model için gözlemlenmiş bilgi matrisidir (observed information matrix).)

İşlem kolaylığı açısından J

$$J = -(\ddot{L})^{-1}\Delta$$

olarak ifade edilirse $\ddot{F}$,

$$\ddot{F} = \Delta'(\ddot{L})^{-1}\Delta \qquad (4.5)$$

formuna dönüşür. Burada Δ; pxm tipinde elemanları $\left.\frac{\partial^2 L(\theta \mid w)}{\partial \theta_i \partial w_j}\right|_{\theta=\hat{\theta}, w=w_0}$ olan bir matristir.

Bütün bu eşitliklerin *(4.4)* de yerine konulması ile eğrilik;

$$C_l = 2\left|l'\Delta'(\ddot{L})^{-1}\Delta l\right| \qquad (4.6)$$

olarak karesel formda ifade edilir.

Yerel etki yöntemini geometrik olarak şöyle açıklayabiliriz. Bozulma hedef modele uygulanır. Ω içinde, w_0 noktasındaki tanjant düzlemine normal olan l doğrultusu seçilir. (Öyle ki bu doğrultu $\|l\| = 1$ koşulunu sağlar.) Seçilen doğrultunun belirlediği düzlemle yüzey kesiştirilip kesişim kümesindeki doğrular belirlenir. Bu

doğrular "ağırlık doğrusu" (lifted line) olarak adlandırılır. l nin herbir doğrultusu için farklı bir ağırlık doğrusu elde edilir. l doğrultusundaki ağırlık doğrularının normal eğrilikleri *(4.4)* eşitliği ile hesaplanır $(C_l = 2|l'\ddot{F}l|)$. Bu hesaplama $\ddot{F}$ nın özdeğerlerinin bulunması ile yapılır. C_l nin büyük değerleri, l doğrultusundaki bozulmanın model üzerinde etkili olduğunu ifade eder. Hesaplanan özdeğerlerden en büyüğü maksimum eğriliği ifade eder ve

$$C_{max} = \max_{\|l\|=1} 2|l'\ddot{F}l| \qquad (4.7)$$

ile formülize edilir.

Maksimum özdeğere karşılık gelen özvektör ise maksimum doğrultu vektörü olup l_{max} ile gösterilir. l_{max} vektörünün elemanları incelenerek en büyük yerel etkiye neden olan satır belirlenir.

Maksimum eğrilik (C_{max}), maksimum doğrultu l_{max} ve l_{max} doğrultusundaki ağırlık doğrularının grafiği *(ör: $LD(w_0+al_{max})$ karşın a, $a \in R^l$) Cook*'un yerel etki yönteminde etkinin değerlendirilmesini sağlayan en önemli niceliklerdir.

İncelemeler sonucu elde edilen C_{max} ve l_{max} için aşağıdaki yorumlar yapılabilir.

- C_{max} ın büyük değerleri l doğrultusundaki bozulmanın duyarlı oluşunun bir göstergesidir.
- Maksimum doğrultu l_{max}, likelihood displacement içindeki en büyük yerel değişikliği elde etmek için modeli nasıl bozabiliriz sorusunu yanıtlar.
- C_{max} maksimum eğriliği gibi l_{max} doğrultusundaki ağırlık doğrusunun grafiği yerel etkinin değerlendirilmesini sağlayan diğer önemli bileşendir.

4.6 Normal Doğrusal Regresyonda Satırların Bozulması

(1.12) eşitliği ile tanımlı çoklu doğrusal regresyon modelini ele alalım. *w; nx1* tipinde bozulma terimlerini içeren bir vektör olmak üzere bozulmuş model için log-likelihood fonksiyonu

$$L(\beta \mid w) = Sabit - \frac{1}{2\sigma^2}\sum_{i=1}^{n} w_i (y_i - x_i'\beta)^2 \qquad (4.8)$$

olarak ifade edilir. Burada w_i ve y_i, sırasıyla w ve y nin *i.* elemanları ve x_i'; X'in *i.* satırını ifade etmektedir. Bu model için kullanılması gereken **eğriliğin formu,**

a) σ^2 ***biliniyorken;***

elemanları $\frac{\partial^2 L(\beta \mid w)}{\partial \beta_i \partial w_j}$ $i = 1,2,...,p$ $j = 1,2,...,m$ olan pxm tipindeki

$$\Delta = \frac{X'D(e)}{\sigma^2}$$

matrisi ve elemanları $\frac{\partial^2 L(\beta)}{\partial \beta_i \partial \beta_j}$, $i, j = 1,2,...,p$ olan pxp tipindeki;

$$\ddot{L}(\hat{\beta}) = -\frac{X'X}{\sigma^2}$$

matrislerinin *(4.6)* da yerine konulması ile

$$C_i = \frac{2\left|l'D(e)HD(e)l\right|}{\sigma^2} \qquad (4.9)$$

olarak elde edilir. Burada $D(e) = diag(e_1, e_2, ..., e_n)$ ve $H = X(X'X)^{-1}X'$ olarak tanımlıdır.

b)σ^2 ***bilinmediğinde;***

$$\Delta = \begin{bmatrix} \frac{\partial^2 L(\beta \mid w_1)}{\partial\beta\partial w_1} & .. & .. & \frac{\partial^2 L(\beta \mid w_n)}{\partial\beta\partial w_n} \\ \frac{\partial^2 L(\sigma^2 \mid w_1)}{\partial\beta\partial w_1} & .. & .. & \frac{\partial^2 L(\sigma^2 \mid w_n)}{\partial\beta\partial w_n} \end{bmatrix} = \begin{bmatrix} \frac{X'D(e)}{\tilde{s}^2} \\ \frac{e'_{sq}}{2\tilde{s}^4} \end{bmatrix}$$

ve

$$\ddot{L}(\hat{\theta}) = \begin{bmatrix} \frac{\partial^2 L(\theta)}{\partial\beta^2} & \frac{\partial^2 L(\theta)}{\partial\sigma^2\partial\beta} \\ \frac{\partial^2 L(\theta)}{\partial\beta\partial\sigma^2} & \frac{\partial^2 L(\theta)}{\partial(\sigma^2)^2} \end{bmatrix} = \begin{bmatrix} -\frac{X'X}{\tilde{s}^2} & 0 \\ 0 & -\frac{n}{2\tilde{s}^4} \end{bmatrix}$$

eşitliklerinin *(4.6)* da yerine konulması ile;

$$C_l = \frac{2l'\left[D(e)HD(e) + e_{sq}e'_{sq} / 2n\tilde{s}^2\right]l}{\tilde{s}^2} \qquad (4.10)$$

olarak elde edilir. Burada $\tilde{s}^2;\sigma^2$ nin m.l.k.'sı, e_{sq} ise elemanları e_i^2 olan $nx1$ tipinde bir vektörü ifade etmektedir.

(4.9) ve *(4.10)* ile tanımlanan eğriliklerden elde edilecek olan l_{max} için analitik bir form bulunamamaktatır.

Cook (1986) tarafından, C_l için eşik değer; basit rastgele örneklem $(Y=\beta_0+\varepsilon)$ üzerine yapılan inceleme sonucu 2 olarak elde edilmiştir.

4.7 Yerel Etki Yaklaşımındaki Eksiklikler

Billor ve Loynes (1992) tarafından yapılan çalışmada; *Cook (1986)*'un yerel etki yöntemindeki eksiklikler ortaya çıkarılmış ve bu eksiklikleri ortadan kaldırıp daha güvenilir sonuç elde edilmesini sağlayan yeni bir yöntem sunulmuştur.

Cook (1986)'un yaklaşımında dört pratik ve teorik zorluk bulunmaktadır.

- Karşılaştırma yapmamızı sağlayan eşik değerin seçimi,
- Maksimum eğriliğin (C_{max}) değerlendirilmesi,
- Bozulmanın yeniden paremetrize edilmesi altında eğriliğin değişmezliğinin eksikliği (Bu problem *Schall veDunne* tarafından çözüldü.),
- Parametrelerin tanımlamasındaki eksiklik (en önemli problem).

1. Eşik Değerin Seçimi: *Cook (1986)* tarafından hesaplanan eşik değer basit rastgele örneklem modeli kullanılarak hesaplandığından veri kümesinden tamamıyla bağımsızdır. Bu nedenle verinin önemli olduğu konumlarda pek sağlıklı sonuç vermemektedir.

2. C_{max} ***ın Değerlendirilmesi:*** Etkili satırların belirlenmesinde sadece C_{max} a bakarak yorum yapılmamalıdır. Çünkü C_{max} çok küçükken l_{max} a baktığımızda çok etkili satırlarla karşılaşabiliriz. Yani C_{max} tek başına bir kriter değildir.

3. Eğriliğin Değişmezliğinin Eksikliği: Cook yerel etkiyi ölçerken model bozulmasını kullandığında değişen her parametre için farklı bir eğrilik elde etmektedir.

Parametrelerin değişmesi sadece eğrinin belli oranda büyüyüp küçülmesine neden olup eğrinin şeklinde bir değişiklik meydana getirmemektedir. Dolayısıyla bozulma altında eğriliklerin birbirine eşit çıkması beklenmektedir. Ama *Cook (1986)* un yaklaşımında değişen her parametre için farklı bir eğrilik elde edilmiştir. Bu da diğer bir çelişkidir *(Billor ve Loynes (1992))*.

4. Parametrelerin Tanımlanmasındaki Eksiklik: Model değişiminde bazı parametreler özelliklerini korurken bazıları koruyamaz. Bu problem veri bozulmasından değil model bozulmasından kaynaklanır.

$Y=X\beta+\varepsilon$ modelini düşünelim. Modeli β yerine $\beta_1=(1+w)\beta$ koyarak, datayı ise X yerine $X_1=(1+w)X$ koyarak bozalım.

$$Y=X\beta_1+\varepsilon \qquad (4.11)$$

$$Y=X_1\beta+\varepsilon \qquad (4.12)$$

Her iki durum için aynı kestirimler elde edilmesine rağmen, yorumlar farklı yapılmalıdır. *(4.12)* nolu modelden $\hat{\beta}$ parametre kestirimleri üzerinde veri bozulmasının etkileri, bozulmuş açıklayıcı değişkenlerin matrisine bakarak görülebilir. Ama *(4.11)* nolu modelde $\hat{\beta}$ parametresi üzerinde yaptığımız model bozulmasının etkisinin yorumlaması o kadar açık değildir. Bu nedenle problemlerle karşılaşmamak için model bozulduğunda parametrelerin değişebileceği fikrinden hareketle model bozulmasının etkisi *(4.11)* nolu model içindeki parametre tahminlerindense "uydurulmuş model" üzerinde hesaplama daha doğru sonuç verecektir.

Cook (1986) model bozulmasını yaparken, sadece model varsayımlarını değiştirmiş bu değişiklikten verinin etkilenebileceğini düşünmemiştir. Aynı şekilde veri bozulması yaparken de sadece veride değişiklik yapmış bundan model parametrelerinin etkilenebileceğini dikkate almamıştır. Bu da hesaplamalarda yanlış sonuçların çıkmasına neden vermektedir.

(4.11) ve (4.12) eşitlikleri ile tanımlanmış bozulmuş modeller açılırsa

Veri bozulması için; $Y=X(1+w)\beta+\varepsilon$

Model bozulması için; $Y=X(1+w)\beta+\varepsilon$

formları elde edilir.

Görüldüğü gibi iki farklı bozulma yapılmasına rağmen modeller birbirinin aynı kalıyor. Bu parametre tanımlanmasındaki eksiklikten kaynaklanmaktadır. Yapılan işlemde model bozulmasından verinin etkilenebileceği yada veri bozulmasında β nın etkileneceği düşünülmeden yapıldığından yanlış sonuçlarla karşılaşabiliriz.

Etki hakkında sağlıklı sonuç elde edilebilmesi için en doğru yol; etkinin uydurulmuş model üzerinde araştırılmasıdır.

Parametre tanımlamasındaki eksikliği aşağıdaki örneği inceleyerekte anlıyabiliriz.

Örnek : Bozulmamış model

$$Y=X\beta+\sigma\varepsilon \qquad (4.13)$$

bozulmuş model

$$Y=X\beta+(1+w)\sigma\varepsilon \qquad (4.14)$$

olarak tanımlansın. σ bilinmiyor ve w bir skaler olsun. *Cook (1986)* un tanımlamış olduğu yöntemle hareket edilirse $LD=nw-n\log(1+w)$ ve $C_{max} = L\ddot{D} = n$ olarak elde edilir.

(4.13) ve (4.14) modelleri σ bilinmediğinde aynı modelleri ifade etmektedir. Sadece σ^2 parametresi *(4.14)* de yeniden tanımlanmış formdadır yani bir dönüşüm söz konusudur. Bu durumda bozulmuş ve bozulmamış model arasında bir fark olmadığından LD nin sıfır çıkması beklenir ancak $L\ddot{D} = n = C_{max}$ elde edilmiştir. Bu hata parametrelerin tanımlanmasındaki eksiklikten kaynaklanmaktadır. Çünkü *Cook* un yönteminde model bozulduktan sonra parametrelerin bozulmuş olacağı düşünülmeden işlem yapılmaktadır.

4.8 Yerel Etkiye Yeni Yaklaşım

Bölüm 4.7. de ifade edilen problemleri ortadan kaldırmak için *Cook (1986)*'un tanımladığı LD'ye benzer daha kullanışlı bir form, *Billor ve Loynes (1992)* tarafından tanımlanmıştır. LD^* ile gösterilen bu form

$$LD^* = -2\left[L(\hat{\theta}) - L(\hat{\theta}_w \mid w)\right]$$

olarak ifade edilmektedir. Burada $L(\hat{\theta}_w \mid w)$ parametrelendirmenin seçiminden bağımsız olmak üzere bozulmuş model altında likelihood fonksiyonunu maksimize eden değerdir. Bir örnekle bu tanımlamanın parametre tanımlamasındaki eksikliği nasıl ortadan kaldırdığı görülebilir.

(4.13) ve (4.14) ile verilen normal doğrusal regresyon modelini düşünelim. *(4.13)* ün bozulmuş formu *(4.14)* e benzer biçimde $W=diag(w+1,w+1,...,w+1)$ şeklinde tanımlı iken $Y=X\beta+\varepsilon^*$ modeli için $Var(\varepsilon^*)=\sigma^2W^{-1}$ olarak yazılabilir. Bu durumda doğrusal regresyon için yeni ölçü

$$LD^* = -2\left[L\left(Y,\hat{\beta},\hat{\sigma}^2\right) - L\left(Y,\hat{\beta}_w,\hat{\sigma}_w^2 \mid w\right)\right]$$

olarak yazılabilir. Burada $\hat{\beta}$ *ve* $\tilde{s}^2$; β *ve* σ^2 nin bozulmamış model altındaki m.l.k.si, $\hat{\beta}_w$ *ve* $\hat{\sigma}_w^2$ ise β *ve* σ^2 nin bozulmuş model altındaki m.l.k. lerini ifade etmektedir. Kestirimlerden sonra bozulmamış model için log-likelihood fonksiyonu

$$L\left(Y,\beta,\hat{\sigma}^2\right) = -\frac{n}{2}log\,2\Pi - \frac{n}{2}log\,\hat{\sigma}^2 - \frac{n}{2}$$

ve bozulmuş model için log-likelihood

$$L\left(Y,\beta,\sigma^2 \mid w\right) = -\frac{n}{2}log\,2\Pi - \frac{n}{2}log(1+w)\sigma^2 - \frac{\left(Y-X\beta\right)'\left(Y-X\beta\right)}{2(1+w)\sigma^2}$$

olarak tanımlıdır. Likelihood kestiricileri

$$\hat{\beta}_w = (X'WX)^{-1}X'WY$$

ve

$$\hat{\sigma}_w^2 = (1+w)^{-1}\hat{\sigma}^2$$

nin kullanılmasıyla log-likelihood fonksiyonunun formu

$$L\left(Y,\hat{\beta}_w,\hat{\sigma}_w^2 \mid w\right) = -\frac{n}{2}log\,2\Pi - \frac{n}{2}log\,\hat{\sigma}^2 - \frac{n}{2}$$

olarak elde edilir. Burada $W=diag(1+w,1+w,\ldots,\ 1+w)$ olarak tanımlıdır. Dikkat edilecek olursa bu form w dan bağımsızdır. Varyans için parametrik formun değişmesi gerçeği altında model gerçekte değişmediğinden $LD^*=0$ bulunur. Böylece parametrelerin tanımlama eksikliği ortadan kalkar.

4.8.1 Yerel Etkiye Yeni Yaklaşım Yönteminde Satır Silme

Cook un *LD* sinde olduğu gibi satır silme işlemini LD^* için düşünebiliriz. Fakat eğer model bozulması ele alınırsa verilerin değişmemesi gerekir. Halbuki satır silme verinin boyutunu değiştirdiğinden bunu bozulma olarak ele almak uygun değildir. Bu nedenle Ortalaması Değişen Sapan Değer Modelin tüm veri kümesine uygulanmasıyla ortaya çıkacak kestiriciler i inci satırın silinmesi ile elde edilecek olan $\hat{\beta}_{(i)}$ ve $\hat{\sigma}^2$ ye denk olacağından LD^* için benzer motivasyon Ortalaması Değişen Sapan Değer Modelinden yararlanılarak elde edilebilir.

$Y=X\beta+\varepsilon$ standart doğrusal regresyon modelini düşünelim. Hata terimlerinin varyansının σ^2 olması haricinde tüm varsayımlarımız geçerli olsun. *Cook (1986)* 'un yerel etkisi, *Cook (1977)* tarafından tanımlanan D_i istatistiğinden esinleniyordu. *Billor ve Loynes (1992)'in* LD_i^* üzerine yaptıkları çalışmada, D_i istatistiğine benzer olarak tanımladıkları D_i^* istatistiğinden yararlanılmıştır. D_i^*; $\hat{\beta}$ *ile* $\hat{\beta}_{(i)}$ arasındaki uzaklığı ölçen D_i ye benzer bir ölçüdür. Burada $\hat{\beta}$; β nın m.l.k.si ve $\hat{\beta}_{(i)}$; *i.* gözlem silindikten sonra elde edilen β nin m.l.k dir. Böylece $L(\beta)$ orijinal model için log-likelihood ve $L_{ms}(\beta)$ ortalaması değişen sapan değer model

$$Y = X\beta + l_i\phi + \varepsilon$$

için log-likelihood olmak üzere LD_i^* ölçüsü

$$LD_i^* = -2\left[L\left(\hat{\beta}\right) - L_{ms}\left(\hat{\beta}_{(i)},\hat{\phi}\right)\right]$$

olarak tanımlanır. Burada l_i; $nx1$ tipinde i. elemanı 1 diğer tüm elemanları sıfır olan bir vektör, ϕ; $w_i'\beta$ formuna sahip i. satırdaki bozulmayı temsil eden elemanı içinde bulunduran bir değer ve $\hat{\phi}$, mean shift outlier model içinde ϕ nin m.l.k. dir.

$$LD_i^* = \frac{e_i^2}{(1-h_{ii})\sigma^2}$$

olarak ifade edilebilir. Görüldüğü gibi bu ölçü LD de olduğu gibi iki ana niceliğe (e_i, h_{ii}) bağımlıdır. Ayrıca LD_i^*; i-inci Student rezidünün karesi formundadır. D_i Cook uzaklığına benzer olarak

$$D_i^* = \frac{LD_i^*}{p}$$

şeklinde tanımlanabilir. *Cook (1986)*'un tanımladığı yöntem $D_i(w)$ ve likelihood uzaklığı LD arasındaki ilişkiyle motive edilmişken, *Billor ve Loynes (1992)*'un yöntemi D_i^* ve LD^* likelihood uzaklığı arasındaki ilişkiden yararlanarak tanımlanmıştır. Yani

$$pD_i^*(w) = -2\left[L(\hat{\beta}) - L(\hat{\beta}_w \mid w)\right]$$

dır. Burada $\hat{\beta}_w$; bozulmuş model altında β nin m.l.k. dir. *Billor ve Loynes (1992)*'un ölçüsünde tanımlanan D_i^* ölçüsü, D_i ölçüsüne göre büyük h_{ii} lerden daha az etkilenir.

Genel olarak; *Cook*'un LD için yaptığı işlemler LD^* içinde yapılabilir. Bunun için $(w, LD^*(w))$ yüzeyi ve bu yüzey üzerinde değişik doğrultularda ağırlık doğruları elde edilir ve sabit bir l için $w=w_0+al$ olmak üzere $LD^*(w_0+al)$ nin davranışları incelenir. LD^* ölçüsünde birinci türev sıfırdan farklı ve l nin bir fonksiyonu formundadır. Bu sayede LD^* ın yerel davranışları hakkındaki bilgiler birinci türevden elde edilebilir.

w nun boyutu (m) birden farklı olduğunda etki hakkındaki yorumlar maksimum eğim S_{max} ve bunun doğrultusuna bakılarak yapılır.

Maksimum eğim

$$S_{max} = |\nabla LD^*(w)|$$

ile gösterilir. Burada $\nabla LD^*(w_0)$; w_0 noktasındaki gradyanı ifade etmektedir.

$$\frac{\partial LD^*(w)}{\partial a} = \frac{\partial LD^*(w)}{\partial w}\frac{\partial w}{\partial a}$$

$$= \frac{\partial LD^*}{\partial w} l$$

$$= 2l\left[\frac{\partial L(\hat{\theta}_w \mid w)}{\partial w} + \frac{\partial L(\theta \mid w)}{\partial \theta}\frac{\partial \hat{\theta}_w}{\partial w}\right]$$

$$= 2l\frac{\partial L(\theta \mid w)}{\partial w}\Big|_{\theta=\hat{\theta}, w=w_0}$$

olup maksimum eğim S_{max}

$$S_{max} = 2|\nabla L(\hat{\theta} \mid w)|$$

olarak ta ifade edilebilir. Bu teorinin normal doğrusal regresyona uyarlanması *Billor ve Loynes (1992)* tarafından yapılmıştır. Bunun için σ^2 biliniyor ve W genel bozulma matrisi

$$W_g = diag(1+w_1, 1+w_2, \ldots, 1+w_n)$$

olarak tanımlanmak üzere $Y=X\beta+\varepsilon$ doğrusal regresyon modelinin ele alındığını varsayalım ($Var(\varepsilon)=\sigma^2 W^{-1}$). İlk olarak W_g nin tek bir satıra bozulma uygulanmış özel durumu olan

$$W=diag(1+w_1,1,\ldots,1) \qquad (4.15)$$

formunu ele alıp $LD^*(w)$ fonksiyonundaki değişim oranı ölçülür. $m=1$ olduğundan fonksiyonun değişim oranı tanjant doğrusunun incelenmesi ile elde edilebilir. LD^* ın w_1 noktasındaki tanjant doğrusu;

$$\frac{\partial LD^*}{\partial w_1} = -2\frac{\partial L(\beta \mid w)}{\partial w_1}$$

formuna sahip olup bu formun $\hat{\beta}$ m.l.k.si ve $w_1=0$ da ki değeri

$$1-\frac{e_1^2}{\sigma^2}$$

olarak elde edilir *(Billor ve Loynes (1992)).*

Daha sonra aynı problem için 1. ve 2. satırlara eşanlı olarak $W=(1+w_1,1+w_2,1,\ldots,1)$ bozulmasının etkilettirilmesi ile benzer sonuçlar elde edilmiştir. Son olarak da elde edilen sonuçlardan yola çıkarak tüm satırlara birden eşanlı olarak bozulma uygulanıp sonuç genelleştirilmiştir. Bu genelleştirme aşağıdaki işlemlerin yapılması ile görülebilir.

Tüm satırlara eşanlı olarak bozulma uygulanırsa W_g bozulma matrisi;

$$W_g=diag(1+w_1,1+w_2,\ldots,1+w_n)$$

formuna sahiptir. Bu bozulma altında LD^* fonksiyonunun değişim oranı

$$\nabla LD^*(w_0)=\left(\frac{\partial LD^*}{\partial w_1},\frac{\partial LD^*}{\partial w_2},\ldots,\frac{\partial LD^*}{\partial w_n}\right)$$

gradyanı ile ölçülür. $w_i=0$ da

$$\frac{\partial LD^*}{\partial w_i} = 1 - \frac{e_i^2}{\sigma^2}$$

olarak tanımlıdır *(Billor ve Loynes (1992))*. LD^* ın artışındaki maksimum oran $\nabla LD^*(w_0)$ ın doğrultusunda gerçekleşir. Bu değer

$$\left|\nabla LD^*(w_0)\right| = \sqrt{\left(\frac{\partial LD^*}{\partial w_1}\right)^2 + \left(\frac{\partial LD^*}{\partial w_2}\right)^2 + \ldots + \left(\frac{\partial LD^*}{\partial w_n}\right)^2}$$

$$= \sqrt{\sum_{i=1}^{n}\left(1 - \frac{e_i^2}{\sigma^2}\right)^2}$$

olarak verilir. LD^* ın maksimum oranda azalışı ise $-\nabla LD^*(w_0)$ ın doğrultusunda gerçekleşip $\left|-\nabla LD^*(w_0)\right|$ ile hesaplanır. S_{max}'ın kosinüs doğrultusu

$$Cos\theta_i = \frac{\frac{\partial LD^*}{\partial w_i}}{\left|\nabla LD^*\right|}$$

$$= \frac{1 - \frac{e_i^2}{\sigma^2}}{\sqrt{\sum_{j=1}^{n}\left(1 - \frac{e_j^2}{\sigma^2}\right)^2}}, \quad i = 1,2,\ldots,n$$

dır.

Bu yöntem herhangi bir w^* noktası etrafindaki grafiğin yerel davranışları incelenmek istenildiğinde kullanılır. Bu yöntemin hesaplamasında sadece birinci türev kullanıldığından, w_0 dan farklı bir nokta için işlem yapılmak istenildiğinde verilen noktadaki tanjant düzleminin elde edilip w^* daki maksimum artış oranının elde edilmesi yeterlidir.

Yerel etkinin değerlendirilmesinde bu yeni yaklaşım karışık işlemleri içermediğinden *Cook (1986)* 'un tanımladığı yöntemden daha kullanışlıdır.

4.9 S_{max} ın Öneminin Değerlendirilmesi

Yerel etki çalışmasının sonuçlarının yorumlanmasında, S_{max} ın belirlenmesi çok önemlidir.

Tek bir satır üzerinde (*i.* satır) bozulma uygulanıp yerel etki araştırılmak istenildiğinde $\frac{\partial LD^*}{\partial w_i}$ $i=1,2,...,n$ elemanlarını temsil eden vektör belirlenir. Vektör içindeki elemanlardan hangisi büyükse o elemanın temsil ettiği satırın etkili olduğu söylenir. S_{max} için bir eşik değerin kullanılması gerekir.

Eşik değerin belirlenmesinde birçok yol vardır. Bunlardan biri tüm satırlara bozulma uygulanıp normal doğrusal regresyon modeli kullanılarak elde edilmiştir.

Bunun için LD^* ın artışındaki maksimum oranı temsil edebilecek bir form düşünülmüş ve

$$A=\sum_{i=1}^{n}\left(1-\frac{e_i^2}{\sigma^2}\right)^2$$

formunun momenti araştırılmıştır. İşlem yapılırken karışık tanımlamaları ortadan kaldırmak için A içindeki e yerine ε hata terimi konulmuştur $(\varepsilon \sim N(0,\sigma^2))$. İşlem sonunda A nın beklenen değeri $2n$ olarak elde edilmiştir. ε_i ler A nın varyansından bağımsız olduğundan $Var(A)=56n$ olarak elde edilir. En uygun eşik değer $E(A)+2\sigma(A)$ değeridir. Dolayısıyla $\mu+2\sigma=2n+4\sqrt{14}\,n$ bulunur.

Daha iyi başka bir yaklaşım dönüşümler ile elde edilebilir. Bu durumda A nın dağılımı çarpıklaşabilir. Örneğin: A nın kesirsel kuvveti alınıp bu yaklaşımın en iyi kuvvetinin ne olacağı belirlenebilir. Fakat şimdi bulduğumuz $2n+4\sqrt{14}\,n$ değeri yerel etkinin ölçüsü içinde kabaca bir eşik değer olarak değerlendirilebilir.

4.10 Çok Değişkenli Doğrusal Regresyonda Yerel Etki

Cook (1986) tarafından tanımlanan yerel etki yöntemi, çok değişkenli doğrusal regresyona *Kim (1995)* tarafından uyarlanmıştır.

Çok değişkenli doğrusal regresyonda yerel etki, herbir hata terimi için kovaryans matrisine uygulanan bozulma terimlerinin etkilerinin değerlendirilmesi ile araştırılır.

4.10.1 Model Tanıtımı ve Bozulma

$y_i = B'x_i + \varepsilon_i, \quad i = 1,2,...,n$ çok değişkenli doğrusal regresyon modelinde;

y_i: *rx1* tipinde cevap değişkenlerinin vektörü,

x_i : *px1* tipinde açıklayıcı değişkenlerin vektörü,

ε_i : *rx1* tipinde sıfır ortalamalı Σ kovaryans matrisli bağımsız *r* değişkenli normal dağılıma sahip hata terimlerinin vektörü,

B : *pxr* tipinde bilinmeyen parametrelerin matrisi

olarak tanımlansın.

Modele $W=(w_1,w_2, \ldots,w_n)'$ bozulması uygulandığında hata terimlerinin birbirinden bağımsız, sıfır ortalama ve Σ/w_i kovaryans matrisli *r*-değişkenli normal dağılımlı olduğu varsayılır. Burada w_i lerin hepsi birden 1 olursa bozulmuş model bozulmamış modele indirgenir.

İşlem kolaylığı açısından Σ kovaryans matrisi AA' olarak tanımlanır. $A=(a_{ij})$ köşegenel elemanları pozitif olan alt üçgensel matris, $A^{-1} = C'$ eşitliğindeki $C=(c_{ij})$ ise üst üçgensel matristir. Σ kovaryans matrisinin tersi;

$$\Sigma^{-1} = CC'$$

ve bozulmamış model için B parametre vektörünün m.l.k. si

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

olarak tanımlanır. H şapka matrisi

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

olmak üzere, E rezidü matrisi

$$E = (I_n - H)Y$$

olarak ifade edilir. Bu tanımlamalar altında Σ'nın m.l.k. si

$$\hat{\Sigma} = \frac{E'E}{n}$$

ve m.l.k. nin değişmemesi temeli altında A nın m.l.k.si $E'E/n$ nin Cholesky köküdür. (Bir matrisin Cholesky kökü; alt ve üst üçgensel iki matrisin çarpımının karekökünü ifade eder.) Yani $E'E = n\hat{A}\hat{A}'$ dür. Bu durumda $\hat{C}' = \hat{A}^{-1}$ olarak tanımlanabilir.

B nin kolonları ile C matrisinin kolonlarındaki sıfırdan farklı elemanların alt alta sıralanması ile elde edilen vektör $\theta_{r(r+2p+1)/2x1}$ ile gösterilsin.

Bozulmuş ve bozulmamış modeller için log-likelihood fonksiyonları $L(\theta \mid w)$, $L(\theta)$ ve likelihood displacement $LD(w) = 2\left[L(\hat{\theta}) - L(\hat{\theta}_w)\right]$ ile gösterilir. $\hat{\theta}_w$ *ve* $\hat{\theta}$; θ'nın bozulmuş ve bozulmamış model altındaki m.l.k.leridir.

4.10.2 Eğrilik

w nun $LD(w)$ ya karşı grafiği, w değiştikce bir yüzey meydana getirirken bu yüzeyin maksimum eğriliği (C_{max}) ile ilişkili özdeğerde aşağıdaki işlemlerin yapılması ile elde edilir.

l_n; $nx1$ tipinde tüm elemanları 1 e eşit olan bir vektör olmak üzere $\Delta_{\left(\frac{r(r+2p+1)}{2} xn\right)}$ ve $\ddot{L}_{\left(\frac{r(r+2p+1)}{2} x \frac{r(r+2p+1)}{2}\right)}$ matrisleri, elemanları $\left.\frac{\partial^2 L(\theta \mid w)}{\partial\theta\partial w'}\right|_{\theta=\hat{\theta}, w=l_n}$ ve $\left.\frac{\partial^2 L(\theta)}{\partial\theta\partial\theta'}\right|_{\theta=\hat{\theta}}$ olacak şekilde tanımlanarak belirlenir. Belirlenen matrisler *(4.5)* eşitliğinde yerine konularak $\ddot{F}$ oluşturulur. $2\ddot{F}$ nın özdeğerleri elde edilir. Elde edilen mutlak değerce en büyük özdeğer, eğrinin maksimum eğriliği olup C_{max} ile, bu eğriliğe karşılık gelen özvektör ise l_{max} ile gösterilir. Bu eğri $l_{(n+1)}$ ve $(l'_{max},0)'$ *(Cook 1986, sh.138-139)* vektörlerinin gerdiği düzlem ile yüzeyin kesişiminden elde edilir. Burada $l_{(i)}$; *i.* satırı 1 diğer satırları sıfır olan $ix1$ tipinde bir vektördür. l_{max} ın büyük elemanları özel inceleme gerektirmektedir.

4.10.3 Formların Açık Bir Şekilde Gösterimi

Bozulmamış model için log-likelihood fonksiyonu sabit terimler hariç

$$L(\theta) = n\sum_{i=1}^{r} log(c_{ii}) - \frac{1}{2}\sum_{j=1}^{n}\left\{\sum_{i=1}^{r}\left(c_i'\left(y_j - B'x_j\right)\right)^2\right\}$$

olarak yazılır. $L(\theta)$ nın B_k ya göre 1. parçal türevi;

$$\left.\frac{\partial L(\theta)}{\partial B_k}\right|_{\theta=\hat{\theta}} = \sum_{j=1}^{n}\left\{\sum_{i=k}^{r}\left(c_i'\left(y_j - B'x_j\right)\right)c_{ki}x_j\right\}$$

ile hesaplanırken, 2. parçal türevi;

$$\left.\frac{\partial^2 L(\theta)}{\partial B_k \partial B_s'}\right|_{\theta=\hat{\theta}} = -\left(\hat{c}_{(k)}\hat{c}_{(s)}\right)X'X$$

eşitliği ile belirlenir. $c'_{(s)}$, C nin $s.$ satırını ifade eder.

Bu eşitlikler yardımıyla $L(\theta)$ nın $Vec(B)$ ye göre 2. parçal türevi $\frac{\partial^2 L(\theta)}{\partial Vec(B)\partial Vec(B)'}$ matrisel formda;

$$-\left(\hat{C}\hat{C}'\right)\otimes(X'X)$$

olarak yazılabilir.

E nin $j.$ satırı e'_j olmak üzere

$$X'\left(I_n - H\right)Y = \sum_{j=1}^{n} x_j e'_j = 0$$

dır. Dolayısıyla $L(\theta)$ nın B_k ve c_{ij} ye göre ikinci parçal türevleri $\theta = \hat{\theta}$ da tüm k, i ve j ler için sıfırdır.

$$\sum_{j=1}^{n} e_j e'_j = n\hat{A}\hat{A}'$$

eşitliği yardımıyla $L(\theta)$ nın $\theta = \hat{\theta}$ da c_{kl} ve c_{si} ye göre ikinci parçal türevleri;

$$\frac{\partial^2 L(\theta)}{\partial c_{il}^2} = -n\hat{a}_{il}^2 - n\hat{a}'_{(i)}\hat{a}_{(i)}$$

ve

$$\frac{\partial^2 L(\theta)}{\partial c_{kl}\partial c_{si}} = -n\hat{a}'_{(k)}\hat{a}_{(s)}$$

olarak elde edilirken, diğer yerlerde sıfırdır. Burada $\hat{a}'_{(k)}$; $\hat{A}$ nın $k.$ satırını ifade eder. *vech(C);* C matrisinin köşegen üzerinde ve üstündeki elemanlarının kolon kolon alt alta yazılması ile elde edilen vektörü ifade etmek üzere, $L(\theta)$ nın $Vech(C)$ ye göre 2. parçal türevi;

$$\frac{\partial^2 L(\theta)}{\partial Vech(C)\partial Vech(C)'}$$

elemanları

$$-ndiag\left(\hat{a}_{ii}^2 I_{(i)} I_{(i)}' + \hat{A}_i \hat{A}_i'\right)$$

olan blok-köşegenel matris formundadır. $\hat{A}_i$; A nın ilk i satır ve sütununu içeren bir alt matristir.

$\ddot{L}$ nın köşegen haricinde bulunan matrisleri sıfır matrisi olarak elde edilir.

Tüm bulunan sonuçların $\ddot{L}$ da yerine konulması ile $\ddot{L}$; blok-köşegenel matrise dönüşür.

$$\ddot{L} = \begin{bmatrix} -\hat{C}\hat{C}' \otimes X'X & 0 \\ 0 & -ndiag(\hat{a}_{ii}^2 I_{(i)} I_{(i)}' + \hat{A}_i \hat{A}_i' \end{bmatrix}$$

$\hat{a}_{ii}^2 I_{(i)} I_{(i)}' + \hat{A}_i \hat{A}_i'$ nün tersi $\hat{C}_i \hat{C}_i' - \hat{c}_{i(i+)} \hat{c}_{i(i+)}' / 2$ olup, $\ddot{L}$ nın tersi de

$$\ddot{L}^{-1} = -\frac{1}{n} diag\left(n\left(\hat{A}\hat{A}'\right) \otimes (X'X)^{-1}; \hat{C}_i \hat{C}_{(i)}' - \hat{c}_{i(i+)} \hat{c}_{i(i+)}' / 2\right)$$

formuna sahip blok-köşegenel matristir. Burada $c_{i(i+)}$; $ix1$ tipinde c_i nin ilk i elemanını içeren bir vektör olarak tanımlıdır.

Bozulmuş modelden Δ yı hesaplayabiliriz. Bozulmuş model altında log-likelihood fonksiyonu sabit terim hariç

$$L(\theta \mid w) = \frac{r}{2} \sum_{j=1}^{n} log\left(w_j\right) + n \sum_{i=1}^{r} log\left(c_{ii}\right) - \frac{1}{2} \sum_{j=1}^{n} w_j \left\{ \sum_{i=1}^{r} \left(c_i' \left(y_j - B'x_j\right)\right)^2 \right\}$$

olarak ifade edilir. θ nın elemanları $\left(B, c_{i(i+)}\right)'$ formunda olduğundan Δ nın elemanları belirlenirken türevler B ve $c_{i(i+)}$ ye göre alınmalıdır. Bu mantıkla $\Delta = \left(\Delta_0', \Delta_1', \ldots, \Delta_r'\right)'$ matrisi

$$\Delta_0 = \left.\frac{\partial^2 L(\theta \mid w)}{\partial Vec(B)\partial w'}\right|_{\theta=\hat{\theta}, w=l_n} \quad ve \quad \Delta_i = \left.\frac{\partial^2 L(\theta \mid w)}{\partial c_{i(i+)}\partial w'}\right|_{\theta=\hat{\theta}, w=l_n}$$

formlarının hesaplanması ile elde edilir.

Benzer olarak hareket edilirse

$$\left.\frac{\partial^2 L(\theta \mid w)}{\partial B_k \partial w_j}\right|_{\theta=\hat{\theta}, w=l_n} = \left(\hat{c}_{(k)}'\hat{C}' e_j\right) x_j$$

eşitliği elde edilir. Dolayısıyla Δ_0 ın j. kolonu

$$\left.\frac{\partial^2 L(\theta \mid w)}{\partial Vec(B)\partial w_j}\right|_{\theta=\hat{\theta}, w=l_n} = \left(\hat{C}\hat{C}' e_j\right) \otimes x_j \quad j=1,2,..,n$$

ile ifade edilirken, $i \neq 0$ için Δ_i nin elemanları

$$\Delta_i = -I_{(i)} E' D_i$$

olarak yazılır. $I_{(i)}$; ixr tipinde, ilk kolonları ixi tipinde birim matris geri kalan kolonları sıfır matrisi olan bir matris ve D_i; nxn tipinde köşegenel bir matris olup, köşegen üzerindeki elemanları $c_i' E'$ $(i = 1,2,\ldots,n)$ formundadır.

Δ *ve* $\ddot{L}$ için bulunan formlar *(4.5)* de yerine konulursa

$$\ddot{F} = -\left(E\hat{C}\hat{C}'E'\right) * H - \frac{1}{n}\sum_{i=1}^{r} D_i E Q_i E' D_i$$

eşitliği elde edilir. H *ve* $E\hat{C}\hat{C}'E'$ matrislerinin *ij.* elemanlarının çarpımı $\left(E\hat{C}\hat{C}'E'\right) * H$ matrisinin *ij.* elemanını verir. *Q; rxr* tipinde elemanları

$$Q_i = \begin{pmatrix} \hat{C}_i\hat{C}_i' & 0 \\ 0 & 0 \end{pmatrix} - \frac{1}{2}\begin{pmatrix} \hat{c}_{i(i+)}\hat{c}_{i(i+)}' & 0 \\ 0 & 0 \end{pmatrix}$$

olarak tanımlanmış bir matristir. Burada $\hat{C}_i$; $\hat{C}$ nın ilk *i* satır ve *i* sütununa sahip alt matristir.

r; 1 e eşit olduğunda; $\ddot{F}$ tek değişkenli doğrusal regresyondaki formüle dönüşür *(Cook (1986).*

l_{max} için analitik bir ifade yukarıda tanımlı $\ddot{F}$ için elde edilememiştir.

4.11 YEREL ETKİYE YENİ YAKLAŞIMIN ÇOK DEĞİŞKENLİ DOĞRUSAL REGRESYONA UYARLANMASI

(4.10) da verildiği gibi *Cook (1986)* tarafından tanımlanan yerel etki yöntemi *Kim (1995)* tarafından çok değişkenli doğrusal regresyona uyarlanmıştı. Bu bölümde ise *Billor ve Loynes (1992)* tarafından tanımlanan LD^*

$$LD^*(w) = -2\left[L(\hat{\theta}) - L(\hat{\theta}_w \mid w)\right]$$

ölçüsü çok değişkenli doğrusal regresyona uyarlanacaktır. Bunun için $\hat{B}_w$, $L(\hat{\theta})$ ve $L(\hat{\theta}_w \mid w)$ nun çok değişkenli doğrusal regresyona uyarlanması gerekir. $L(\theta)$ nın

$$L(\theta) = n\sum_{i=1}^{r} log(c_{ii}) - \frac{1}{2}\sum_{j=1}^{n}\left\{\sum_{i=1}^{r}\left(c_i'(y_j - B'x_j)\right)^2\right\}$$

olduğunu biliyoruz. Bu nedenle sadece $\hat{B}_w$ ve $L(\hat{\theta}_w \mid w)$ formları belirlenmelidir.

Bölüm 4.10.1 de tanımlanan çok değişkenli doğrusal regresyon modelini ele alalım. Tek bir satıra bozulma uygulayarak $I.(\hat{0}_w \mid w)$ nun formunun belirlenmesi için öncelikle $\hat{B}_w$ belirlenmelidir.

$\hat{B}_w$ nun belirlenmesi:

Bilindiği gibi,

$$\hat{B}_w = (X'WX)^{-1} X'WY$$

olarak ifade ediliyordu. Bu formun çok değişkenli doğrusal regresyonda açık olarak yazmak için $(X'WX)^{-1}$ ve $(X'WY)$ formlarının açık halini belirlemeliyiz.

$$X'WX = X'X + w_1 x_1 x_1'$$

ve $X'WY$:

$$X'WY = X'Y + w_1 x_1 y_1'$$

olarak ifade edilebilir. $(X'WX)^{-1}$, *Sherman, Morrison* ve *Woodbury* teoremi *(Rao 1973)* nin sonucu *(2.4c)* eşitliğine benzer olarak;

$$(X'X)^{-1} - \frac{(X'X)^{-1} w_1 x_1 x_1' (X'X)^{-1}}{1 + w_1 x_1' (X'X)^{-1} x_1}$$

formunda yazılabilir. O halde

$$\begin{aligned}
\hat{B}_w &= (X'WX)^{-1} X'WY \\
&= \left((X'X)^{-1} - \frac{w_1 (X'X)^{-1} x_1 x_1' (X'X)^{-1}}{1 + w_1 x_1' (X'X)^{-1} x_1} \right) \left(X'Y + w_1 x_1 y_1' \right) \\
&= (X'X)^{-1} X'Y + (X'X)^{-1} w_1 x_1 y_1' - \frac{w_1 (X'X)^{-1} x_1 x_1' (X'X)^{-1}}{1 + w_1 x_1' (X'X)^{-1} x_1} X'Y \\
&\quad - \frac{w_1 (X'X)^{-1} x_1' x_1 (X'X)^{-1}}{1 + w_1 x_1' (X'X)^{-1} x_1} w_1 x_1 y_1' \\
&= \hat{B} + w_1 \left[(X'X)^{-1} x_1 y_1' - \frac{(X'X)^{-1} x_1 x_1' \hat{B}}{1 + w_1 h_{11}} \right] - \frac{w_1^2 (X'X)^{-1} x_1' x_1 (X'X)^{-1} x_1 y_1'}{1 + w_1 x_1' (X'X)^{-1} x_1} \\
&= \hat{B} + \frac{w_1 (X'X)^{-1} x_1' (y_1' - x_1' \hat{B}) - w_1^2 (X'X)^{-1} x_1 y_1' h_{11}}{1 + w_1 h_{11}} \\
&\quad - w_1^2 \frac{(X'X)^{-1} x_1 x_1' (X'X)^{-1} x_1 y_1'}{1 + w_1 x_1' (X'X)^{-1} x_1} \\
\hat{B}_w &= \hat{B} + \frac{w_1 (X'X)^{-1} x_1 e_1'}{1 + w_1 h_{11}} - 2 \frac{w_1^2 (X'X)^{-1} x_1 y_1' h_{11}}{1 + w_1 h_{11}}
\end{aligned}$$

(w_1^2 'li terimleri $o(w_1)$ ile gösterirsek)

$$= \hat{B} + \frac{w_1 (X'X)^{-1} x_1 e_1'}{1 + w_1 h_{11}} + o(w_1)$$

olur. Burada h_{ii}; H şapka matrisinin $i.$ köşegen elemanı, $0 \le h_{ii} \le 1$ ve $0 \le w_1 \le 1$ olup $h_{11} w_1$ çok küçük bir değer olacağından önemsenmez ve payda 1 olur. Bu durumda

$$\hat{B}_w = \hat{B} + w_1 (X'X)^{-1} x_1 e_1' + o(w_1) \qquad (4.16)$$

olarak elde edilir. e_1 adi (ordinary) rezidü matrisinin ilk satır vektörüdür.

$L(\hat{\theta}_w | w)$ yu belirlemeden önce form içinde karşılaşacağımız bazı eşitlikleri belirleyelim.(θ ; *(4.10)* uncu bölümde tanıtıldığı gibi alınıyor.)

$(y_i - x_i' \hat{B}_w)^2$ nin belirlenmesi:

(4.6) eşitliği yardımıyla,

$$\begin{aligned} y_1 - \hat{B}_w' x_1 &\approx y_1 - (\hat{B} + w_1 (X'X)^{-1} x_1 e_1')' x_1 \\ &\approx y_1 - \hat{B}' x_1 - e_1 x_1' (X'X)^{-1} x_1 w_1 \\ &\approx y_1 - \hat{B}' x_1 - e_1 h_{11} w_1 \\ &\approx e_1 - e_1 h_{11} w_1 \\ &= e_1 - e_1 h_{11} w_1 + o(w_1) \end{aligned} \qquad (4.17)$$

dir. Bu eşitlik

$$(y_i - \hat{B}_w' x_i) = (e_i - e_1 h_{11} w_1 + o(w_1)) \qquad (4.18)$$

olarak genelleştirilirken, *(4.18)* in karesi

$$\begin{aligned} (y_i - \hat{B}_w' x_i)'(y_i - \hat{B}_w' x_i) &\approx (e_i - e_1 h_{11} w_1)'(e_i - e_1 h_{11} w_1) \\ &\approx e_i' e_i - e_i' e_1 h_{11} w_1 - w_1' h_{11}' e_1' e_i - w_1' h_{11}' e_1' e_1 h_{11} w_1 \\ &\approx e_i' e_i - e_i' e_1 h_{11} w_1 - (e_i' e_1 h_{11} w_1)' \\ &\approx e_i' e_i - 2 e_i' e_1 h_{11} w_1 \\ &= e_i' e_i - 2 e_i' e_1 h_{11} w_1 + o(w_1) \end{aligned}$$

formuna sahiptir.

$L(\hat{\theta}_w | w)$ ***nun belirlenmesi:***

Y matrisinin j inci kolonu için olasılık yoğunluk fonksiyonu

$$f(y_j;\theta_w) = \frac{1}{(2\pi)^{r/2}|\Sigma|^{1/2}|w_j|^{r/2}} e^{-\frac{1}{2}(y_j - B_w'x_j)\Sigma^{-1}w_j(y_j - B_w'x_j)}$$

formuna sahip olup Y matrisi için olasılık yoğunluk fonksiyonu

$$F(Y;\theta_w) = \prod_{i=1}^{n} f(y_j;\theta_w) = \frac{1}{(2\pi)^{nr/2}|diagw_j^{-r/2}||\Sigma|^{n/2}} e^{-\frac{1}{2}\sum_{j=1}^{n} w_j(y_j - B_w'x_j)\Sigma^{-1}(y_j - B_w'x_j)}$$

olarak yazılabilir. Üslü terimlerden kurtulabilmek için bu eşitliğin logaritmasını alalım.

$$\log F = \log \prod_{i=1}^{n} f(y_i;\theta_w) = -\frac{nr}{2}\log 2\pi - \log|diagw_j^{-r/2}| + \frac{n}{2}\log|\Sigma|^{-1} - \frac{1}{2}\sum_{j=1}^{n} w_j(y_j - B_w'x_j)\Sigma^{-1}(y_j - B_w'x_j)$$

Bu eşitlik,

$$|diagw_j^{-r/2}| = w_1^{-r/2}w_2^{-r/2}\ldots w_n^{-r/2} = \prod_{j=1}^{n} w_j^{-r/2}$$

ve

$$|\Sigma^{-1}| = |CC'| = c_{11}c_{22}\ldots c_{rr}c_{11}c_{22}\ldots c_{rr} = \left(\prod_{i=1}^{r} c_{ii}\right)^2$$

eşitlikleri yardımıyla

$$\log F = -\frac{nr}{2}\log 2\pi + \frac{r}{2}\sum_{j=1}^{n}\log(w_j) + n\sum_{i=1}^{r}\log c_{ii} - \frac{1}{2}\sum_{j=1}^{n} w_j \sum_{i=1}^{r}(c_i'(y_j - B_w'x_j))^2$$

olarak yazılabilir. Gerekli düzenlemeleri yapıp sabit terimi ihmal edersek,

$$\log F = \frac{r}{2}\sum_{j=1}^{n}\log w_j + n\sum_{i=1}^{r}\log(c_{ii}) - \frac{1}{2}\sum_{j=1}^{n} w_j \sum_{i=1}^{r}(c_i'(y_j - B_w'x_j))^2$$

eşitliğini elde ederiz. Bu eşitlik üzerinde w_j

$$w_j = \begin{cases} 1 + w_1 & j = 1 \text{ ise,} \\ 1 & j > 1 \text{ ise,} \end{cases}$$

olarak seçilsin (yani sadece 1. satıra bozulma uygulansın). Bu durumda

$$\log F = L(\theta_w|w) = \frac{r}{2}(\log(1+w_1) + \log 1 + \ldots + \log 1) + n\sum_{i=1}^{r}\log c_{ii}$$
$$- \frac{1}{2}(1+w_1)\sum_{i=1}^{r}(c_i'(y_1 - B_w'x_1))^2 - \frac{1}{2}\sum_{j=2}^{n}\sum_{i=1}^{r}(c_i'(y_j - B_w'x_j))^2$$
$$= \frac{r}{2}\log(1+w_1) + n\sum_{i=1}^{r}\log c_{ii} - \frac{1}{2}\sum_{i=1}^{r}(c_i'(y_1 - B_w'x_1))^2 - \frac{1}{2}\sum_{j=2}^{n}\sum_{i=1}^{r}(c_i'(y_j - B_w'x_j))^2$$
$$- \frac{w_1}{2}\sum_{i=1}^{r}(c_i'(y_1 - B_w'x_1))^2$$

olur.

$$LD^* = -2\left[L(\hat{\theta}) - L(\hat{\theta}_w|w)\right]$$
$$= -2\left[n\sum_{i=1}^{r}\log \hat{c}_{ii} - \frac{1}{2}\sum_{j=1}^{n}\sum_{i=1}^{r}(\hat{c}_i'(y_j - \hat{B}_w'x_j))^2 - \frac{r}{2}\log(1+w_1) - n\sum_{i=1}^{r}\log \hat{c}_{ii}\right.$$
$$\left. + \frac{1}{2}\sum_{i=1}^{r}(\hat{c}_i'(y_1 - \hat{B}_w'x_1))^2 + \frac{1}{2}\sum_{j=2}^{n}\sum_{i=1}^{r}(\hat{c}_i'(y_j - \hat{B}_w'x_j))^2 + \frac{w_1}{2}\sum_{i=1}^{r}(c_i'(y_1 - \hat{B}_w'x_1))^2\right]$$
$$= -2\left[-\frac{1}{2}\sum_{j=1}^{n}\sum_{i=1}^{r}(\hat{c}_i'(y_j - \hat{B}_w'x_j))^2 - \frac{r}{2}\log(1+w_1) + \frac{1}{2}\sum_{j=1}^{n}\sum_{i=1}^{r}(\hat{c}_i'(y_j - \hat{B}_w'x_j))^2\right.$$
$$\left. + \frac{w_1}{2}\sum_{i=1}^{r}(\hat{c}_i'^2(e_1'e_1 - 2e_1'e_1h_{11}w_1))\right]$$
$$r\log(1+w_1) - w_1\sum_{i=1}^{r}(\hat{c}_i'^2(e_1'e_1 - 2e_1'e_1h_{11}w_1))$$
$$= r\log(1+w_1) - w_1e_1'e_1\sum_{i=1}^{r}\hat{c}_i'^2 + 2w_1^2\sum_{i=1}^{r}\hat{c}_i'^2e_1'e_1h_{11}$$
$$= r\log(1+w_1) - w_1e_1'e_1\sum_{i=1}^{r}\hat{c}_i'^2 + o(w_1)$$

4.11.1 Çok Değişkenli Doğrusal Regresyonda LD^* Ölçüsünün Genel Formları

$LD^*(w)$, R^2 içinde basit bir eğri tanımlar. LD^* içindeki büyük yerel değişiklikleri ortaya çıkarmak için farklı doğrultularda eğrinin değişim oranının belirlenmesi gerekir. w_0 noktasındaki eğrinin eğimi (tanjant vektörü), $LD^*(w)$ nun w_1 'e göre 1.türevi ile elde edilir.

$$\left.\frac{\partial LD^*}{\partial w_1}\right|_{w_1=0} = \frac{r}{1+w_1} - e_1'e_1\sum_{i=1}^{r}(\hat{c}_i')^2\Bigg|_{w_1=0}$$

$$\approx r - e_1^2\sum_{i=1}^{r}(\hat{c}_i')^2$$

1. ve 2. satır birlikte bozulduğunda yani;

$$W = diag(1+w_1,\ 1+w_2,\ 1,\ \ldots,\ 1)$$

olarak tanımlandığında yapılan işlemler tekrarlanırsa, $\hat{B}_w$

$$\hat{B}_w = \hat{B} + w_1(X'X)^{-1}x_1'e_1 + w_2(X'X)^{-1}x_2'e_2 + o(w_1) + o(w_2)$$

olarak elde edilir.

n satıra birden bozulma uygularsak, bozulma matrisinin formu;

$$W = diag(1+w_1,\ 1+w_2,\ \ldots,\ 1+w_n)$$

olur. Yine aynı işlemlerin yapılmasıyla;

$$\hat{B}_w = \hat{B} + \sum_{j=1}^{n}(w_j(X'X)^{-1}x_j'e_j + o(w_j))$$

eşitliği elde edilir. Bu durumda likelihood displacementın formu;

$$LD^*(w) = \prod_{j=1}^{n} r\,log(1+w_j) - \sum_{j=1}^{n}(w_j e_j'e_j\sum_{i=1}^{r}(\hat{c}_i')^2 + o(w_j))$$

olur.

$m=1$ için tanımlanan tanjant doğrusu, $m=2$ için tanjant düzlemine dönüşür. Tanjant düzlemindeki değişim oranı gradyan yardımıyla,

$$\nabla LD^*(w_0)=\left(\frac{\partial LD^*}{\partial w_1},\ \frac{\partial LD^*}{\partial w_2},\ \ldots,\frac{\partial LD^*}{\partial w_n}\right)$$
$$=\left(r-e_1'e_1\sum_{i=1}^{r}(\hat{c}_i')^2,\ r-e_2'e_2\sum_{i=1}^{r}(\hat{c}_i')^2,\ \ldots,r-e_n'e_n\sum_{i=1}^{r}(\hat{c}_i')^2\right)$$

olarak yazılır. LD^* ın artışındaki maximum oran $\nabla LD^*(w_0)$ ın doğrultusunda gerçekleşir ve bu maksimum oran

$$S_{max}=\left|\nabla LD^*(w_0)\right|=\sqrt{\left(\frac{\partial LD^*}{\partial w_1}\right)^2+\left(\frac{\partial LD^*}{\partial w_2}\right)^2+\ \ldots.+\left(\frac{\partial LD^*}{\partial w_n}\right)^2}$$
$$=\sqrt{\sum_{j=1}^{n}\left(r-e_j'e_j\sum_{i=1}^{r}(\hat{c}_i')^2\right)^2}$$

olarak verilir. Bu durumda kosinüs doğrultusu,

$$Cos\theta_i=\frac{\dfrac{\partial LD^*}{\partial w_j}}{\left|\nabla LD^*\right|}=\frac{r-e_i'e_i\sum_{i=1}^{r}(\hat{c}_i')^2}{\sqrt{\sum_{i=1}^{n}\left(r-e_i'e_i\sum_{i=1}^{r}(\hat{c}_i')^2\right)}},\qquad (j=1,2,\ldots,n)$$

olarak ifade edilir.

5. BÖLÜM

UYGULAMA

Bu bölümde 3. ve 4. bölümlerde verdiğimiz çok değişkenli doğrusal regresyon modellerinde tanılama yöntemleri kullanılarak "Rohwer Data" kümesi incelenecektir.

Bu uygulama; *"Rohwer Data" (ROHWER, 1975)* kümesi üzerinde

- D_i, $DFFITS_i$, $COVRATIO_i$, AP_i, h_{ii}, e_i, w_i^2 ölçülerini elde etme,
- Leverage (etki) ,rezidü grafiklerinin *(Barrett ve Ling (1992))* oluşturulması,
- *Cook (1986)*'un yerel etki yöntemini uygulama,
- LD^* ölçüsü *(Billor ve Loynes,(1992))* ile ilgili işlemlerin yapılması ve sonuçların değerlendirilmesi olmak üzere dört alt bölümde ele alınacaktır. İşlemler MINITAB paket programınından yararlanılarak elde edilmiştir.

5.1 "Rohwer Data" Tanıtımı ve Tanılama Yöntemlerinin Uygulanışı

5.1.1 "Rohwer Data" Tanıtımı

Rohwer (1975) tarafından Amerika'nın herhangi bir bölgesinde yaşayan üst sınıfta okuyan öğrencilerinden rastgele *32* öğrenci seçilmiş ve bu öğrenciler üzerinde "Paired-Association Learning Proficency Tests" (PA) adını verdiğimiz test kümesi uygulanarak en iyi veri kümesi elde edilmek istenmiştir.

Araştırma;

x_1: Kelime gösterme

x_2: Şekil gösterme

x_3: Hem kelime hem de şekil gösterme

x_4: Hareketle kelimeyi söyletme

x_5: Öğretilen kelimelerle doğru cümle kurma

kavramlarının değerlendirilmesi amacıyla her öğrenciye toplam 20 soru sorularak yapılmıştır.

Kullanılan testler

- "Peabody Picture Vocabulary" testi,
- "Student Achievement" testi,
- "Ravin Progressive" testi

olup, test sonuçlarından elde edilen değerler sırasıyla y_1 , y_2 , y_3 değişkenlerine atanmıştır. İnceleme sonucu elde edilen değerler *Tablo 5.1* de görülmektedir. Bu verilere göre çok değişkenli doğrusal regresyon modeli E nin satırları $N_3(0,\Sigma)$ *(3 değişkenli, 0 ortalamalı,* σ^2 *varyanslı)* dağılımından gelmek üzere;

$$Y_{32X3} = X_{32X6}B_{6X3} + E_{32X3}$$

olarak yazılır.

Bu model için β_1,β_2,β_3 'ün kestirimleri $\hat{B}$ ve Σ nın yansız kestiricisi $\hat{\Sigma}$ olarak tanımlanırsa bu veri kümesi için elde edilen $\hat{B}$ ve $\hat{\Sigma}$ matrisleri sırasıyla

$$\hat{B} = \begin{bmatrix} 39.6972 & 13.2439 & -28.4674 \\ 0.0673 & 0.0593 & 3.2571 \\ 0.3700 & 0.4924 & 2.9966 \\ -0.3744 & -0.1640 & -5.8591 \\ 1.5230 & 0.1190 & 5.6662 \\ 0.4102 & -0.1212 & -0.6227 \end{bmatrix}$$

ve

$$\hat{\Sigma} = \begin{bmatrix} 121.844 & 8.793 & 39.988 \\ 8.793 & 5.532 & 19.493 \\ 39.988 & 19.493 & 535.441 \end{bmatrix}$$

olarak elde edilir.

y_1	y_2	y_3	x_1	x_2	x_3	x_4	x_5	x_6
68	15	24	1	0	10	8	21	22
82	11	8	1	7	3	21	28	21
82	13	88	1	7	9	17	31	30
91	18	82	1	6	11	16	27	25
82	13	90	1	20	7	21	28	16
100	15	77	1	4	11	18	32	29
100	13	58	1	6	7	17	26	23
96	12	14	1	5	2	11	22	23
63	10	1	1	3	5	14	24	20
91	18	98	1	16	12	16	27	30
87	10	8	1	5	3	17	25	24
105	21	88	1	2	11	10	26	22
87	14	4	1	1	4	14	25	19
76	16	14	1	11	5	18	27	22
66	14	38	1	0	0	3	16	11
74	15	4	1	5	8	11	12	15
68	13	64	1	1	6	10	28	23
98	16	88	1	1	9	12	30	18
63	15	14	1	0	13	13	19	16
94	16	99	1	4	6	14	27	19
82	18	50	1	4	5	16	21	24
89	15	36	1	1	6	15	23	28
80	19	88	1	5	8	14	25	24
61	11	14	1	4	5	11	16	22
102	20	24	1	5	7	17	26	15
71	12	24	1	0	4	8	16	14
102	16	24	1	4	17	21	27	31
96	13	50	1	5	8	20	28	26
55	16	8	1	4	7	19	20	13
96	18	98	1	4	7	10	23	19
74	15	98	1	2	6	14	25	17
78	19	50	1	5	10	18	27	26

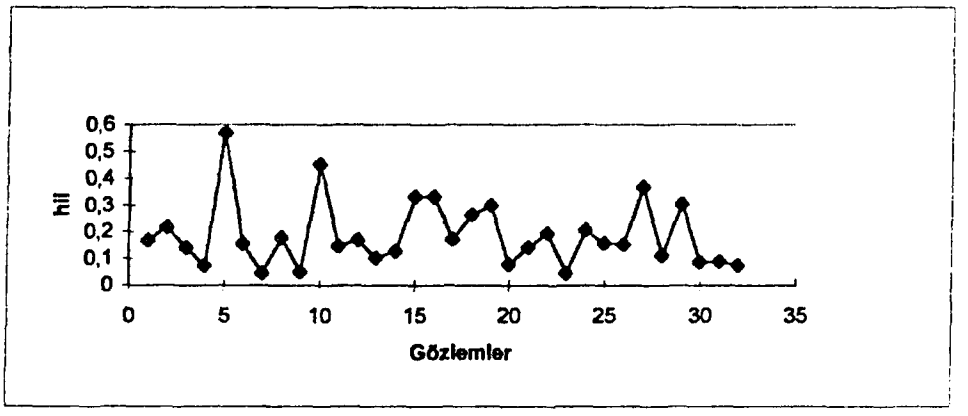
Tablo 5.1 Rohwer Data

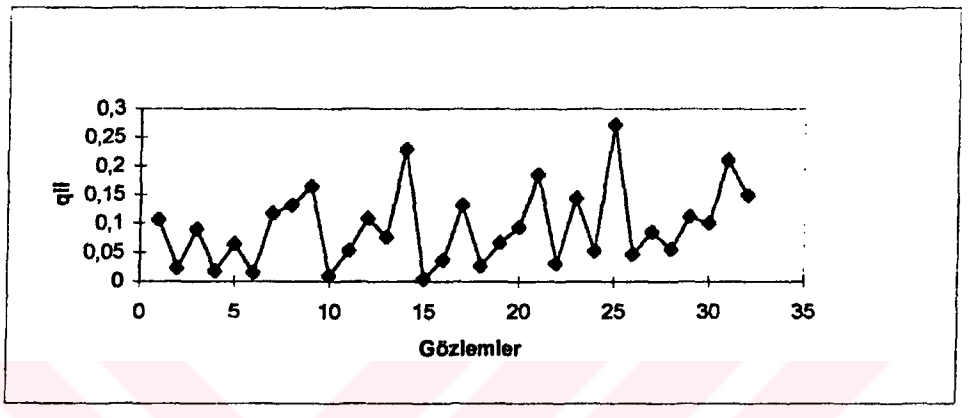
5.1.2 Tek Satır Etkisinin İncelenmesi

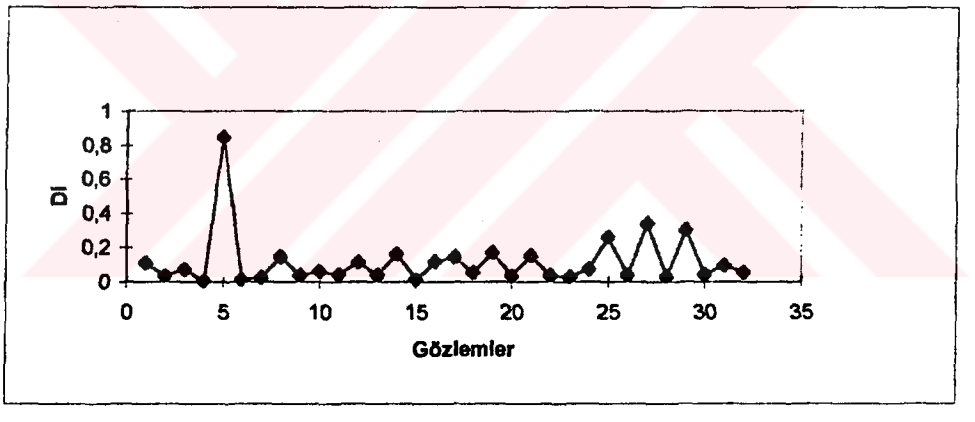
Tek satırın etkisinin araştırılması amacıyla "Rohwer Data" kullanılarak elde edilen ölçü değerleri *Tablo 5.2* de ve bu tablonun kullanımıyla elde edilen serpilme diagramları; *Şekil 5.1* de görülmektedir.

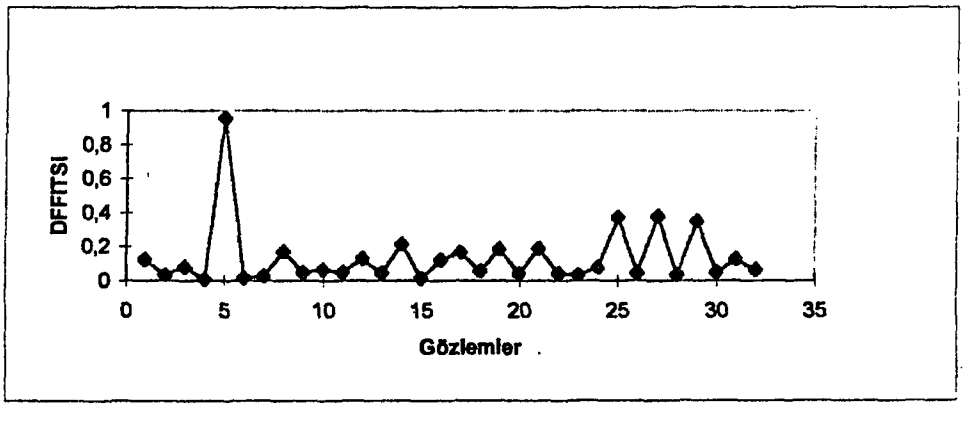
	Di	DFFITSi	COVRATIOi	FVARATIOi	Wi	AP_i	hii	qii
1	0,110671	0,121949	1,5475	1,6983	27,230	0,726882	0,167009	0,106109
2	0,03576	0,035431	3,5452	2,2867	8,432	0,758472	0,218453	0,023074
3	0,074113	0,079496	1,6628	1,595	17,228	0,769378	0,141735	0,088887
4	0,006454	0,006325	2,2694	1,3861	1,269	0,909363	0,073144	0,017493
5	0,846738	0,956152	9,5929	1,18984	411,878	0,367672	0,568215	0,064114
6	0,014585	0,014288	2,9955	1,8256	3,142	0,83008	0,154322	0,015598
7	0,025295	0,027734	1,0593	1,1337	5,403	0,837268	0,04531	0,117423
8	0,147674	0,168818	1,2851	1,6949	38,135	0,692569	0,176611	0,13082
9	0,040404	0,046943	0,7624	1,0903	9,204	0,785147	0,051313	0,16354
10	0,063391	0,062055	1,10315	6,6996	21,048	0,538647	0,451612	0,009741
11	0,045681	0,046825	2,2117	1,6908	10,191	0,801635	0,145428	0,052937
12	0,11629	0,12861	1,5332	1,7135	28,838	0,721199	0,170504	0,108297
13	0,042671	0,044845	1,6506	1,4295	9,307	0,820011	0,103746	0,076243
14	0,164274	0,213967	0,492	1,2459	45,561	0,644844	0,126499	0,228657
15	0,015191	0,01471	6,528	3,755	4,099	0,662834	0,332467	0,004698
16	0,11832	0,12039	4,8372	3,5636	33,513	0,63143	0,331835	0,036735
17	0,144486	0,165228	1,2667	1,6735	37,170	0,6952	0,173206	0,131594
18	0,056707	0,056596	4,0558	2,7131	14,294	0,709529	0,263539	0,026932
19	0,173212	0,183832	3,2435	2,9504	48,732	0,635687	0,298358	0,065955
20	0,037332	0,039917	1,3707	1,294	8,060	0,828424	0,078806	0,09277
21	0,151641	0,185636	0,7485	1,3902	40,160	0,675309	0,140238	0,184453
22	0,040243	0,040251	3,0522	2,0638	9,286	0,775052	0,193803	0,031145
23	0,030357	0,03435	0,8746	1,0959	6,687	0,81191	0,044553	0,143537
24	0,072947	0,074995	2,7132	2,1051	17,577	0,742223	0,206417	0,05136
25	0,26008	0,368825	0,3287	1,2737	81,389	0,571504	0,157126	0,27137
26	0,042609	0,043323	2,3879	1,7528	9,517	0,800694	0,153339	0,045967
27	0,338659	0,376304	3,3582	3,8426	110,618	0,547542	0,367265	0,085193
28	0,034224	0,035109	1,9612	1,5052	7,353	0,832434	0,111898	0,055669
29	0,3026	0,34625	2,1182	2,8069	92,568	0,584642	0,30427	0,111088
30	0,045051	0,048657	1,3235	1,3139	9,908	0,813225	0,086554	0,100221
31	0,097584	0,121842	0,5593	1,1446	24,882	0,701404	0,089219	0,209378
32	0,055032	0,063053	0,8891	1,1859	12,654	0,777789	0,07321	0,149001

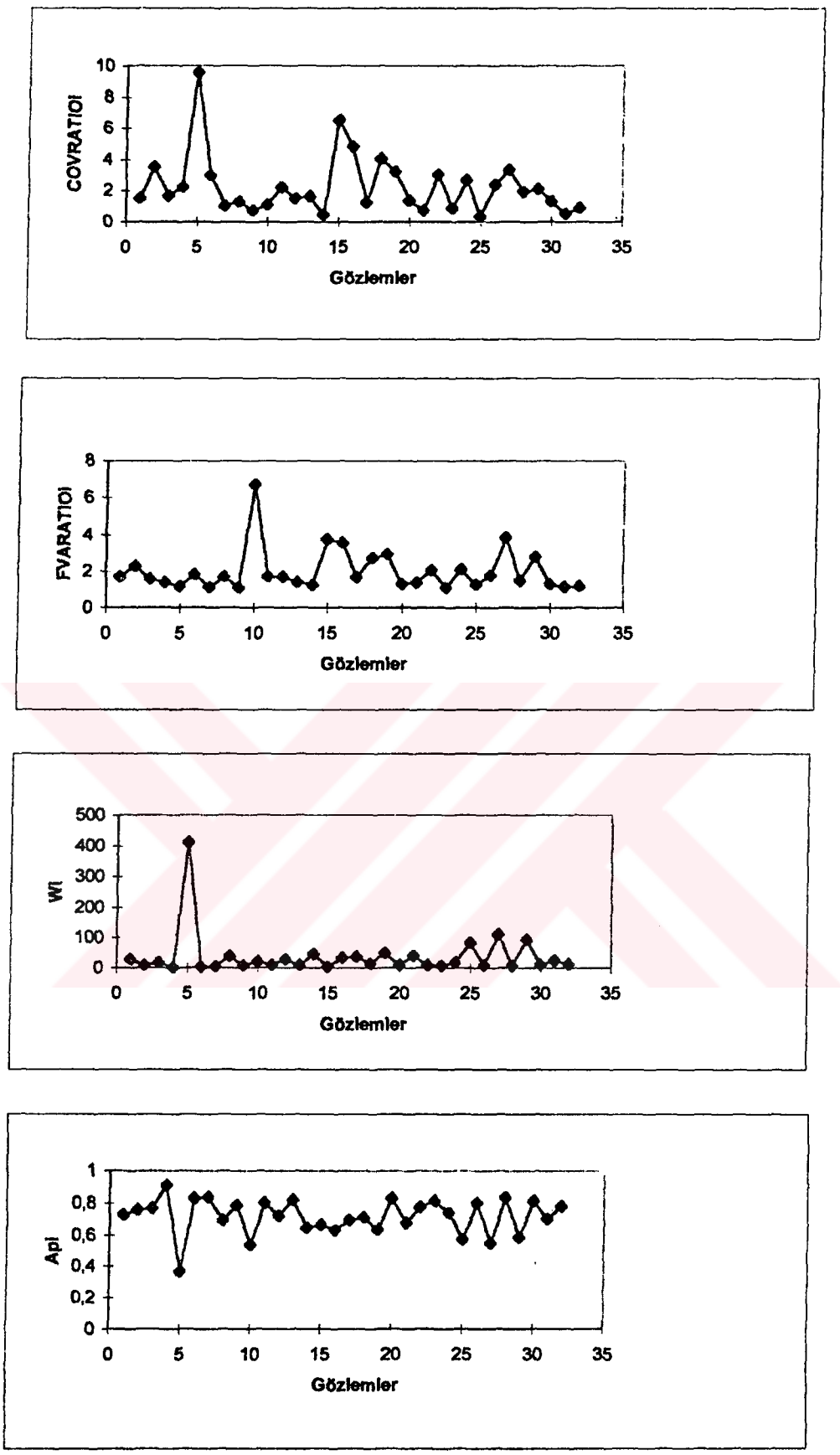
Tablo 5.2 "Rohwer Data" ya ilişkin ölçü değerleri


hii
0,6
0,5
0,4
0,3
0,2
0,1
0
0 5 10 15 20 25 30 35
Gözlemler


qii
0,3
0,25
0,2
0,15
0,1
0,05
0
0 5 10 15 20 25 30 35
Gözlemler


Di
1
0,8
0,6
0,4
0,2
0
0 5 10 15 20 25 30 35
Gözlemler


DFFITSI
1
0,8
0,6
0,4
0,2
0
0 5 10 15 20 25 30 35
Gözlemler



Şekil (5.1) "Rolwer Data"ya ilişkin ölçüm değerlerini gösteren grafikler.

Ölçü ve grafik sonuçlarına göre aşağıdaki yorumlar yapılabilir.

- 5. gözlem X uzayının en uzağındaki veri olup en büyük "leverage" (etki) özelliğini gösteririr.
- 25. gözlem uydurulmuş regresyon doğrusunun en uzağındaki nokta olup en büyük rezidüye sahiptir. Dolayısıyla bu gözlemin modele dahil edilmesi modelin eğiminin, parametre kestiriminin değişimine neden olabilir.
- 5. gözlemin uydurulmuş değerler üzerindeki etkisi büyüktür.
- 5. gözlem e.k.k. kestiricisi $\hat{B}$ üzerinde etkilidir.
- 5. ve 15. gözlemler güven elipsoidinin hacminde önemli değişikliklere neden olabilirler.
- 10. gözlem $\hat{y}_i$ nın varyansında önemli değişikliklere meydana getirebilirler.

Genel olarak tüm ölçüler üzerinden bir değerlendirme yapılırsa 5. gözlemin en büyük etkiye sahip olduğu açıktır.

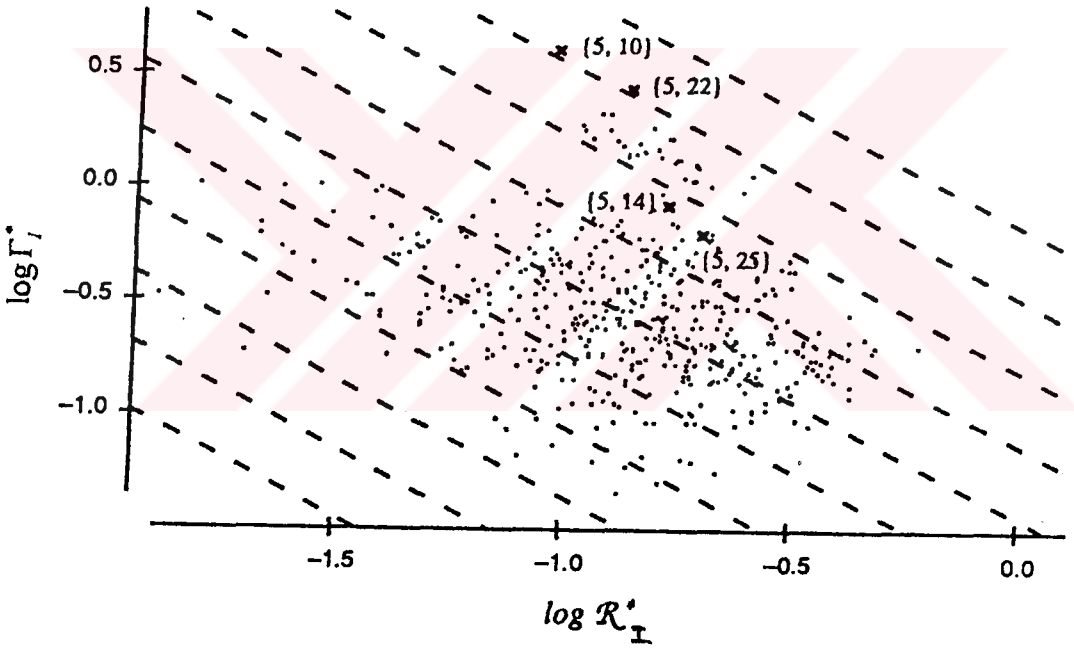
5.1.3 Çoklu Satırın Etkisinin Saptanması (Leverage ve Rezidü Grafiklerinin Kullanımı)

"Rohwer data" kullanılarak 3. bölümde verilen (ikili alt kümelerin ortak etkisini ortaya çıkarmayı sağlayan) leverage-rezidü grafiklerini çizmek için öncelikle *Tablo 3.4* de tanımlanmış leverage (etki) ve rezidü elemanlarının belirlenmesi gerekmektedir. *"Rohwer data"* nın eleman sayısı 32 olup, herbir ölçü için leverage ve rezidü elemanlarının belirlenmesi $\binom{32}{2}$ kombinasyonu kadar alt küme üzerinde ayrı ayrı işlem yapılmasını gerektirir. Bu işlemin yapılması çok zaman alacağından inceleme sadece $D_I(X'X, pS)$ Cook uzaklığı kullanılarak yapılmıştır. İncelemede seçilen her I ikili alt kümesi için aşağıdaki formların elde edilmesi gereklidir.

- Etki (L_I) ve rezidü (R_I) matrisleri,
- Bu matris formlarının özdeğerleri,
- $\Gamma_I = \|L_I\|$ ve $\Re_I = \|R_I\|$ formları,
- $\Gamma_I^* = \Gamma_I (Cos\theta_I)^{1/2}$ ve $\Re_I^* = \Re_I (Cos\theta_I)^{1/2}$ formları.

Tüm bu verileri elde etmekle etki ve rezidü matris formları bir skalere atanmış olur. Bu şekilde elde ettiğimiz değerleri tablo üzerinde yorumlamak çok zor olduğundan inceleme leverage-rezidü grafiğine bakılarak yapılacaktır.

$D_I(X'X, pS)$ ölçüsü J_I^{tr} sınıfının bir üyesi olduğundan 3. bölümde yapılan açıklamaya göre incelememizi $\log \Re_I^*$ ın $\log \Gamma_I^*$ a karşı grafiğini çizerek yapmalıyız.



Şekil 5.2 $D_I(X'X, pS)$ Ölçüsü Kullanılarak Elde Edilen Rohwer Data için Ortak Etkiyi Gösteren Grafik.

Yapılan işlemler sonucu elde edilen grafik *Şekil 5.2* de görülmektedir. Bu şekil incelendiğinde $\{5,14\}, \{5,25\}$ alt kümelerinin ilgi duyucağımız alt kümeler olduğunu,

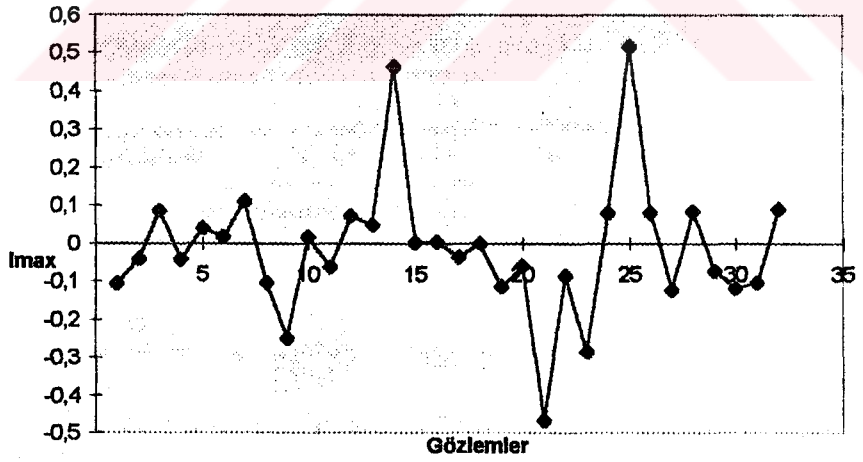
bunların ortak etkide birbirlerinin etkilerini iptal ettiğini ancak bunlardan daha da etkili olan $\{5,10\}$, $\{5,22\}$ alt kümelerinin bulunduğunu söyleyebiliriz.

5.1.4 LD nin Çok Değişkenli Doğrusal Regresyon Modellerine Uygulanışı

Bu yöntemi uygulamak için

$$\ddot{F} = -(E\hat{C}\hat{C}'E') * H - \frac{1}{n}\sum_{i=1}^{3} D_i E Q_i E' D_i$$

formu, $2\ddot{F}$ nın özdeğerlerini (C_i) ve bu özdeğerlere karşılık gelen özvektörleri (l_i) elde edildi. Hesaplamalar sonucunda mutlak değerce en büyük özdeğer C_{max}=14.5392 olarak elde edildi. Bu değer eşik değer 2 den çok büyük olduğundan hemen yerel etki ile ilgili bir problemin olduğunu düşünebiliriz. Problemin hangi satırdan kaynaklandığını incelemek için en büyük özdeğer C_{max} a karşılık gelen özvektörü incelemeliyiz. Bu özvektöre ilişkin serpilme diyagramı *Şekil 5.3* de görülmektedir. Şekil incelendiğinde 14 ve 25 nolu gözlemlerin en etkili gözlemler olduğu görülmektedir. Dolayısıyla bu satırlar en büyük yerel etkiye neden olabilecek satırlardır yorumu yapılabilir.



Şekil 5.3 Maksimum Eğrilikle İlişkili Özvektörün İndeks Grafiği

Görüldüğü gibi l_{max} a bakarak etkili olan satırlar kolaylıkla belirlenebilmektedir.

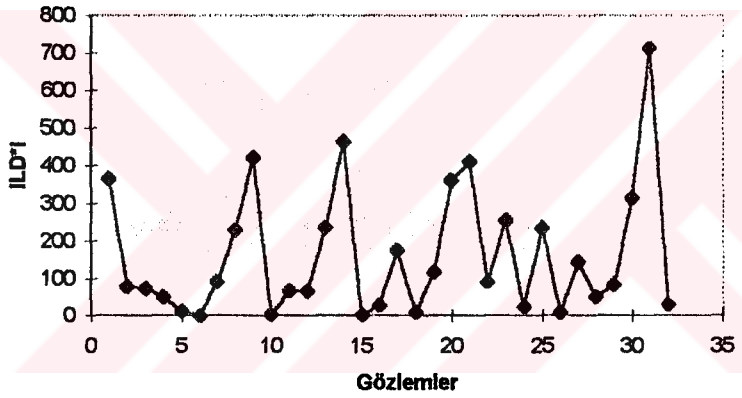
5.1.5 LD ın Çok Değişkenli Doğrusal Regresyon Modellerine Uyarlanması*

Bu ölçü *Cook (1986)* 'un yerel etki yöntemindeki eksik tanımlamaları ortadan kaldırdığı için daha doğru bilgi vermektedir.

LD* ın artışındaki maksimum oran $\nabla LD^*(W_0)$ ın doğrultusunda gerçekleştiğinden inceleme için öncelikle

$$\frac{\partial LD^*(W_0)}{\partial w_j} = (r - e_j^2 \sum_{i\ 1}^{3} (\hat{c}_i')^2), \quad j = 1,2,\ldots,32$$

değerlerini elde etmeliyiz.



Şekil 5.5 LD ın Çok Değişkenli Doğrusal Regresyona Uyarlanması İle Elde Edilen LD* ın Grafiği*

Bu değerlere göre $\frac{\partial LD^*(W_0)}{\partial w_i}$ $i=1,2,\ldots,32$ nin gradyenti (maksimum eğriliği) $S_{max} = |\nabla LD^*(W_0)| = 1331.49$ olarak elde edildi. Bu değer *Billor ve Loynes (1992)* tarafından tanımlanan eşik değer formunun kullanılması ile elde edilen değerden

(12.19) çok büyük olduğundan yerel etkiye neden olan gözlemlerin bulunduğunu hemen söyleyebiliriz.

Yerel olarak etkinin hangi satırdan kaynaklandığı konusundaki araştırma için $S_i = \frac{\partial LD^*(W_0)}{w_i}$, $i = 1,2,...,32$ değerlerinin incelenmesi gereklidir. Bu değerler için serpilme diagramı *Şekil 5.5* de görülmektedir. İnceleme sonucu 14 ve 31 nolu gözlemin en büyük yerel etkiye sahip olduğunu söyleyebiliriz.

31 nolu gözlem en büyük rezidüye sahip olan gözlemdir. Bu gözlemin rezidüsünün büyük oluşu bizi varyansın sabitliğini etkileyen bir gözlem olduğu sonucuna ulaştırır.

Dolayısıyla Cook istatistiği ile belirleyemediğimiz rezidüsü büyük olup, varyansın sabitlik özelliğini bozan gözlemler bu istatistik kullanılarak kolaylıkla belirlenebilir.

ÖZET

Doğrusal regresyon modelleri kendi içinde tek değişkenli ve çok değişkenli olmak üzere iki gruba ayrılırlar. Tek değişkenli doğrusal regresyon modellerinde *Y; nx1* tpinde bir vektör, *X; nxp* tipinde bir matris ve ε ; *nx1* tipinde bir vektör iken, çok değişkenli doğrusal regresyon modellerinde *Y; nxr* tipinde *(r>1)* bir matris, ε ; *nxr* tipinde *(r>1)* bir matris ve β ; *pxr* tipinde bir matris olmak zorundadır. Her iki model için de aynı varsayımlar geçerlidir. Bu varsayımlar *Bölüm 1* de ayrıntılı olarak verilmiştir.

Regresyon modelleri; değişkenler, gözlemler ve model varsayımları ile belirlenirler. Modeli tam olarak temsil edemeyen değişkenler regresyon katsayılarında büyük değişikliklerin meydana gelmesine neden olurlar. Modelin verileri sağlıklı bir şekilde temsil edebilmesi için bu değişkenlerin (gözlemlerin) belirlenmesi gereklidir. Bu amaca yönelik çok sayıda tanılama yöntemi bulunmaktadır. Bunlar içinden; *Cook (1977)*'un D_i uzaklığı, *Hoaglin ve Welsch (1978)*'in *H şapka matrisi, Andrews ve Pregibon (1978)*'un AP_i si, *Belsley ve ark. (1980)*'nın $DFBETAS_i$, $DFFITS_i$, $COVRATIO_i$, $FVARATIO_i$ sı sayılabilir.

Tanılama yöntemlerinde, etkili görülen satır (veya satırlar) veri kümesinden çıkarılarak ya da bu satırlara bozulma terimi etkilettirilerek etkinin belirlenmesi yolu izlenir.

Silme yolu ile, tek satırın etkisinin araştırılması işleminde (*i* etkili görülen satırı ifade etmek üzere) D_i, $DFFITS_i$, AP_i, h_{ii}, $COVRATIO_i$, $FVARATIO_i$, W_i gibi pek çok ölçüden yararlanılırken, birden fazla satırın etkisinin araştırılmasında; tek değişkenli doğrusal regresyon modelleri için *Jones ve Ling (1988)* tarafından tanımlanan

$$J_I(f;\, u,v,c)=\left[e_I'(I-H_I)^{-u}H_I^{v}e_I\right]f\big(m,\, n,\, p)\big)/c$$

J_I sınıfından ve çok değişkenli doğrusal regresyon modelleri içinde *Barrett ve Ling (1992)* tarafından tanımlanan

$$J_1^{tr}(f;a,b) = f(.)tr\left[H_1 Q_1 (I - H_1 - Q_1)^a (I - H_1)^b\right]$$

$$J_1^{det}(f;a,b) = f(.)det\left[(I - H_1 - Q_1)^a (I - H_1)^b\right]$$

J_1^{tr} ve J_1^{det} sınıflarından yararlanınız. Bu sınıflara ait ölçüler u, v ya da a, b nin seçimlerine göre kolaylıkla elde edilebileceğinden fazla işlem yapma problemini ortadan kaldırmaktadır.

Cook (1986) likelihooda dayalı modellerde varsayımlardan yerel sapmaların etkisini saptamak üzere genel bir yöntem verdi. Bu yöntem ile modele küçük bir bozulma etkilettirildiğinde analiz sonuçlarında önemli bir değişiklik ortaya çıkıyorsa bir problem olduğu sonucuna varılır.

Cook bu değişimi saptamak üzere likelihood displacement yüzeyinin normal eğriliğinin kullanılması gerektiğini ileri sürdü. Ancak bu yöntemde bir takım problemler, örneğin parametrelerin tanımlanmasındaki eksiklik ve işlem zorlukları söz konusudur. *Billor ve Loynes (1992)* bu eksik tanımlamaları ortaya çıkartıp, fazla işlem yapmamızı ve daha güvenilir sonuç elde etmemizi sağlayan alternatif yöntem vermişlerdir. LD^* ile tanımlanan bu ölçü ve bu ölçünün çok değişkenli doğrusal regresyon modellerine uyarlanmış formu hakkında detaylı bilgi 4. bölümde bulunmaktadır.

SUMMARY

Linear regression models are divided into two different groups, one is called univariate linear regression, the other is called multivariate linear regression. In the univariate linear regression models;

Y is a *nx1* vector,

X is an *nxp* matrix,

ε is a *nx1* vector.

However; in the multivariate linear regression models

Y is an *nxr* matrix

and

ε is an *nxr* matrix *(r>1)*.

For these two models the assumptions are the same. These assumptions are explained in details in part 1.

Regression models are assessed by the variates, observations and assumptions. The variates which do not represent the model exactly may cause gross changes on the regression coefficients and fitted model.

Therefore the assessment of the variates and these observations are very significant for the model adequacy. For this purpose, many diagnostic methods (or influence measures) have been proposed. Some of these are; D_i *(Cook's (1977))*, *Hat matrix (H) (Hoaglin and Welsch's (1978))*, $COVRATIO_i$, $FVARATIO_i$, $DFFITS_i$, $DFBETAS_i$ *(Belsley et al. 's (1980))*, AP_i *(Andrews and Pregibon's (1978))*.

In these diagnostic methods, influential case (or cases) is (are) omitted from the model or the perturbation is introduced to the model or the perturbation is introduced to the case (or cases) and then influential observations are assessed based on these measures.

While D_i, $DFFITS_i$, $DFBETAS_i$, $COVRATIO_i$, $FVARATIO_i$, AP_i are used in the case of assessment of single-case influence (i is the case or row of X and Y), J_I class

$$J_I(f;u,v,c)=\left[e_I'(I-H_I)^{-u}H_I^{v}e_I\right]f(m,n,p)/c$$

is used for the case of assessment of multiple-case influence. For the multivariate regression models, J_I^{tr} and J_I^{det} class

$$J_I^{tr}(f;a,b)=f(.)tr\left[H_IQ_I(I-H_I-Q_I)^a(I-H_I)^b\right]$$

$$J_I^{det}(f;a,b)=f(.)det\left[(I-H_I-Q_I)^a(I-H_I)^b\right]$$

defined by *Barrett and Ling (1992)* are used for the assessment of multiple case influence. These classes offer considerable computational savings.

Cook (1986) gave a general method for assessing the influence of local departures from assumptions in likelihood based models. With this method, if a minor perturbations in the model leads to a major change in the results of the analysis, then there is evidence of difficulty. In order to assess local influence *Cook* suggests using the normal curvature of the likelihood dislacement surface.

However in this method there are some problems such as lack of the definition of the parameters, computational difficulties. Therefore *Billor and Loynes (1992)* proposed an alternative method, LD^*.

This measure in univariate case is examined and applied to the multivariate case in order to assess local influence in part 4.

KAYNAKLAR

Andrews, D. F. and Pregibon, D. (1978). *Finding outliers that matter.* J. Roy. Statist. Soc., Ser. B., 40, 85-93.

Ancombe, F. J.(1961). *Examination of residuals.* Proc. Fourth Berkeley Symp., 1. 1-36.

Atkinson, A. C., (1985). *Plots, Transformations, and Regression.* Oxford University.

Barrett, B. E. and Ling, R. F., (1992). *General Classes of Influence Measures for Multivariate Regression.* Journal of the American Statistical Association, 87, 417,184-191.

Beckman, R. J. and Cook, R. D. (1983). *Outlier ...s.* Technometrics 25, 119-149, 161-163.

Beckman, R., Nachtsheİm, C. J., and Cook, R. D. (1987). *Diagnostics for mixed-model analysis of variance.* Technometrics, 29, 413-426.

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

Billor, N., Loynes and R. M., (1993). *Local Influence: A New Approach.* Commun. Statist., 22, (6), 1595-1611.

Box, G. E. P and Cox, D. R. (1964). *An Analysis of transformations (with discussion).* J. Roy. Statist. Soc., Ser. B., 26, 211-246.

Carroll, R. J. and Ruppert, D. (1981). *On Prediction and the power transformation family.* Biometrica, 68, 609-616.

Carroll, R. J. and Ruppert, D. (1988). *Transformation and Weighting in Regression.* Chapman and Hall, London.

Chatterjee, S. and Price, B. (1977). *Regression Analysis by Example.* John Wiley, New York.

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

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

Cook R. D. (1979). *Influential Observations in Linear Regression.* Journal of the American Statistical Association, 74, 169-174.

Cook, R. D., (1986). *Assessment of Local Influence.* J.R. Statist. Soc., 48, 2, 133-169.

Cook, R. D. (1987). *Influence assessment.* J. Appl. Statist., 14, 2, 117-131.

Cook, R. D. and Wang, P. C. (1983). Transformations and influential cases in regression. Technometrics, 25, 337-343.

Cook, R. D. and Weisberg (1982). R*esiduals and Influence in Regression.* Chapman and Hall, New York.

Cook, R. D., Pena, D., Weisberg, S., (1988). *The Likelihood Displacement: A Unifying Principle for Influence Measures.* Commun. Statist. -Theory Meth., 17, 3, 623-640.

Cox, D.R. and Hinkley, D. V. (1974). *Theoritical Statistics.* Chapman and Hall, London.

Daniel, C. (1978). *Patterns in residuals in the two-way layout.* Technometrics, 20, 385-395.

Donald, S. G. and Maddala, G. S., (1987). *Identifying Outliers and Influental Observations in Econometric models.* Journal of Applied Statistics, 14, 185, 663- 699.

Draper, N. and John, J. A. (1981). *Influential observations and outliers in regression.* Technometrics 23, 21-26.

Fung, W. K., (1995). G*raphical Summaries for Influence of Multiple Observations.* Commun. Statist.-Theory Meth., 24, 2, 415-427.

Gentleman, J. F. and Wilk, M. B., (1975). *Detecting Outliers II: Supplementing the Direct Analysis of Residuals.* Biometrics, 31, 387-410.

Gray, J. B. (1983). *"The L-R plot: a graphical tool for assessing influence," Proceedings of the Statistical Computing Section,* Amer. Statist. Asso. 159-164.

Gray, J. B. (1985). *"Graphics for regression diagnostics," Proceedings of the Statistical Computing Section,* Amer. Statist. Asso. 102-107.

Gray, J. B., (1986). *A Simple Graphic for Assessing Influence in Regression.* J. Statist. Comput., 24, 121-134.

Graybill, F. A. (1976). *Theory and Application of the linear model.* MA: Duxbury, North Scituate.

Graybill, F.A., (1983). *Matrices with Applications in Statistics.* Wadsworth, Inc., California.

Hadi, A.S. and Simonoff, J.S., (1993). *Procedures for the Identification of Multiple Outliers in Linear Models.* Journal of the American Statistical Association, 88, 424, 1264-1272.

Hoaglin, D. C. and Welsch, R. (1978). *The hat matrix in regression and ANOVA.* Amer. Statistican, 32, 17-22.

Hocking, R. R. (1984). *Discussion of "K-Clustering as a Detection Tool for Influential Subsets in Regression."* by J.B. Gray and R.F. Ling, Tehnometrics, 26, 321-323.

Hossain, A. and Naik, D. N., (1989). *Detection of Influential Observations in Multivariate Regression.* Journal of Applied Statistics, 16, 1, 25-37.

Huber, P. (1981). *Robust Statistics.* Wiley, New York.

Lawrence, A, J. (1988). *Regression transformation diagnostics using local influence.* Journal of the American Statistical Association, 83, 1067-1072.

Johnson, W. and Geisser, S. (1980). *A predictive view of the detection and characterization of influential observations in regression analysis.* University of Minnesota, School of Statistics Technical Report no. 365.

Johnson, W. and Geisser, S. (1983). *The detection and characterization of influential observations in regression analysis.* J. Amer. Statist. Assoc. 8, 137-144.

Jones, W. D. and Ling, R. F., (1988). *A New Unifying Class of Influence Measures for Regression Diagnostics.* American Statistical Association, 305-310.

Judge, G. G., Griffits, W. E., Hill, R. C., Lutkepohl, H., and Lee, T.C. (1985). *The Theory and Practice of Econometrics.* John Wiley and Sons, New York.

Kim, M.G., (1995). *Local Influence In Multivariate Regression.* Commun. Statist.-Theory Meth., 24, 5, 1271-1278.

Lehmann, E. (1982). *Theory of Point Estimation*, Newyork: John Willey and Sons.

Little, J. K. (1985). *Influence and a Quadratic Form in the Andrews-Pregibon Statistic.* Technometrics, 27, 13-15.

McCulloch, C. E. and Meeter, D. (1983). *In discussion of "outliers" by Beckman, R. J. and Cook, R. D.*. Technometrics, 25, 152-155.

Miller, R. (1966). *Simultaneous Inference.* McGraw Hill, New York.

Milliman, R. S. and Parker, G. D. (1977). *Elements of Differential Geometry.* Englewood Cliffs, NJ: Prentice Hall.

Montgomery, D.C. and Peck, E.A., (1992). *Introduction to Linear Regression Analysis.* John Wiley and Sons., Inc., Canada.

Naik, D.N., (1989). D*etection of Outliers In the Multivariate Linear Regression Model.* Commun. Statist.-Theory Meth., 18, 6, 2225-2232.

Nelder, J. and Wedderburn R. (1972). *Generalized linear models.* J. Roy. Statist. Soc., Ser. A, 135, 370-384.

Rao, C. R. (1973). *Linear Statistical Inference and its Applications.* 2nd edition, John Wiley and Sons, New York.

Rousseeuw, P. J. and Leroy, A. M., (1987). *Robust Regression and Outlier Detection.* John Wiley, New York.

Schall, R. and Dunne, T. T. (1992). *A note on the relationship between parameter collinearity and local influence.* Biometrika, 79, 399-404.

Schwarzmann, B. (1991). *Connection between local influence analysis and residual diagnostics.* Technometrics, 33, 103-104.

Searle, S. R. (1971). *Linear Models.* John Wiley and Sons, New York.

Srikantan, K. S. (1961). *Testing for a single outlier in a regression model.* Sankhya, A, 23, 251-260.

Storer, B. E. and Crowley, J. (1985). *A diagnostic for Cox regression and general conditional likelihoods.* Journal of the American Statistical Association, 80, 139-147.

Thomas, W. (1990). *Influence on confidence regions for regression coefficients in generalized linear models,* Journal of the American Statistical Association, 85, 393-397.

Thomas W. and Cook ,R. D. (1989). *Assessing influence on regression coefficients in generalized linear models.* Biometrica, 76, 741-749.

Thomas W. and Cook ,R. D. (1990). *Assessing influence on predictions from generalized linear models.* Technometrics, 32, 59-65.

Tsai, C. L. (1986). *Discussion of Assessment of Local Influence* by R.D. Cook. Journal of the Royal Statistical Society, Ser. B., 48, 165.

Tsai, C. and Wu, X. (1992). *Transformation-model diagnostics.* Technometrics, 34, 197-202.

Tukey, J. W. (1970). *Exploratory Data Analysis.* Limited preliminary edition. Reading Mass. Addison-Wesley.

Weisberg, S. (1980). *Applied Linear Regression.* Wiley, New York.

Welsch, R. E. (1982). *"Influence Functions and Regression Diagnostics," in Modern Data Analysis.* (R.L. Launer and F. Siegel, eds.), Academic press, New York.

TEŞEKKÜR

Bu çalışmayı yöneten ve değerli zamanlarını harcayarak yakın ilgi ve yardımlarını esirgemeyen sayın hocam Yrd.Doç.Dr. Nedret Billor'a teşekkürlerimi saygılarımla sunarım. Ayrıca tüm matematik bölümü akademik personeline ilgi ve alakalarından dolayı teşekkür ederim. Bu tezin yazımında yardımlarını esirgemeyen değerli eşim Araş. Gör. Ersin Kıral'a teşekkür ederim.

ÖZGEÇMİŞ

1973 yılında Adana'da doğdum. İlk öğrenimimi Celalettin Sayhan İlkokulunda, orta öğrenimimi Gazi Ortaokulunda ve lise öğrenimimi Abdülkadir Paksoy Kız Lisesinde tamamlayarak 1989 yılında Çukurova Üniversitesi Fen Edebiyat Fakültesi Matematik Bölümüne girdim. 1993 yılında mezun olup yine aynı yıl master programına başladım. 1994 yılında evlendim ve aynı yıl Çukurova Üniversitesi İktisadi ve İdari Bilimler Fakültesi Ekonometri Bölümüne Araştıma Görevlisi olarak girdim. Halen aynı görevde çalışmaktayım.