G.Ü. Fen Bilimleri Dergisi,
16(3):457-466, 2003
ISSN 1303-9709
G.U. Journal of Science
16(3):457-466, 2003

DOĞRUSAL SINIRLAMALAR VE İZDÜŞÜM TEORİSİ YARDIMIYLA ÇOKLU DOĞRUSAL REGRESYONDA ETKİLİ GÖZLEMLERİN BELİRLENMESİ

Bülent ALTUNKAYNAK
Gazi Üniversitesi, Fen Edebiyat Fakültesi, İstatistik Bölümü,06500, Ankara, TÜRKİYE,
bulenta@gazi.edu.tr,

ÖZET

Bu çalışmada, doğrusal sınırlamalar, izdüşüm teorisi ve Cook uzaklığından oluşan üç aşamalı bir yöntem kullanılarak etkin gözlemlerin tespit edilmesine çalışılmıştır. Önerilen bu üç aşamalı yöntem etkin gözlemlerin belirlenmesinde hesaplama kolaylığı getirmektedir. Özellikle, izdüşüm teorisinin kullanımının hesaplamalardaki matris boyutlarını nasıl küçülttüğü gösterilerek bir örnek üzerinde uygulaması yapılmıştır.

Anahtar Kelimeler : Etkili gözlem, izdüşüm teorisi, cook uzaklığı, çoklu doğrusal regresyon

DETERMINING INFLUENTIAL OBSERVATIONS ON MULTIPLE LINEAR REGRESSION BY LINEAR RESTRICTIONS AND PROJECTION THEORY

ABSTRACT

In this paper, the assessment of influential observations through the use of a three-step method made up of linear restrictions, the projection theory, and the Cook distance was studied. The proposed method facilitates the calculation in the determination of influential observations. Specifically, how the projection theory used here minimizes the size of the matrices was clearly demonstrated and shown in an example as an application.

Key Words : Influential observation, projection theory, cook distance, multiple linear regression

1. GİRİŞ

Regresyon çözümlemeleri yapılırken katsayı tahminlerinde her bir gözlemin ayrıntılı olarak incelenmesi ve olağan olmayan gözlemlerin belirlenmesi gerekir. Çünkü, tek bir gözlem bile regresyon modelinin katsayı tahminleri üzerinde büyük bir etkiye sahip olabilir. Dolayısıyla ilgili gözlemin veri kümesinden çıkartılması regresyon denklemini tamamen değiştirebilir.

Olağan dışı gözlemlerin belirlenmesi ihtiyacı, aykırı ve etkili olan gözlemlerin bulunması için bazı yöntemlerin geliştirilmesini sağlamıştır. Aykırı değer, en büyük artıklı değer yada veri kümesindeki

gözlemlere uzak olan gözlem değeri olarak verilir (1). Aykırı değerin varlığı, model yapısının yanlış olmasından, gerekli bazı dönüşümlerin yapılmamış olmasından, ölçüm, tartım, kaydetme hatalarından örnekleme rasgeleliğinden kaynaklanabilir. Etkin gözlem (influential observation); veri kümesinden çıkartıldığında en küçük kareler kestirimlerini büyük ölçüde değiştiren gözlem olarak tanımlanır (2). Çoğu zaman aykırı bir gözlem etkili bir gözlem olabilir. Ancak, etkili gözlemler çoğu zaman aykırı gözlem değildir (3).

Bu nedenle etkili gözlemlerin belirlenmesi için artıkların incelenmesi yeterli değildir ve ilgili gözlemin parametre tahminleri üzerinde etkili olup olmadığını test etmek için farklı yöntemler geliştirilmiştir. Bu yöntemlerden yaygın olarak kullanılanlarından biri, gözlem uzaklığı ve student türü artıklara dayalı Cook uzaklığı (Cook's distance) yöntemidir (4). Belsley ve arkadaşları ise, gözlemlerin tahmin edilen regresyon modelindeki etkilerini geliştirdikleri daha farklı ölçütlerle incelemişlerdir (5).

Bir gözlem veya gözlem kümesinin parametre tahminlerinde meydana getirdiği farkın hesaplanması karmaşık matris işlemleri gerektirmektedir. Bu nedenle, gözlemlerin doğrusal sınırlamalarının dikkate alındığı ve parametre tahminlerinin iki aşamalı Sıradan En Küçük Kareler (SEK) yöntemiyle elde edildiği bir çalışma Pino tarafından ele alınmıştır (6). Pino çalışmasında, regresyon modeli üzerine doğrusal bir sınırlama getirmiş ve elde edilen sınırlı regresyon modeline izdüşüm teorisi uygulayarak En Küçük Kareler çözümleri elde etmiştir. Bu yöntemin faydası hesaplamalarda karmaşık matris işlemleri gerektirmemesidir. Bu çalışma Gupta ve Kabe tarafından çok değişkenli duruma genişletilmiştir (7).

Bu çalışmada, doğrusal sınırlamalar birinci aşama, izdüşüm teorisi ikinci aşama ve Cook uzaklığı üçüncü aşama olarak ele alınmış ve etkili gözlemlerin tespiti için bu üç aşamalı yöntem önerilmiştir. Burada önerilen yöntemde $\hat{\beta}-\hat{\beta}_{\psi}$ değerleri hesaplanırken izdüşüm teorisinden yararlanılmıştır. Bu sayede daha küçük boyutlu X matrisi ile çalışma imkanı doğmuş bu da büyük bir hesaplama kolaylığı getirmiştir. Ayrıca etkili gözlemlerin tespit edilmesi için $\hat{\beta}-\hat{\beta}_{\psi}$ değerlerini kullanan Cook uzaklığı üçüncü aşama olarak önerilmiştir.

Bu çalışma altı bölümden oluşmaktadır. İkinci Bölümünde doğrusal sınırlamalar altında parametre tahminleri elde edilmiştir. Tam ranklı regresyon modelinin parametre tahminleri $\hat{\beta}$, eksik gözleme sahip regresyon modelindeki parametre tahminleri, eksik gözlemlere ait alt küme ψ olmak üzere $\hat{\beta}_{\psi}$ notasyonu ile gösterilmiştir. Üçüncü Bölümde, $\hat{\beta}-\hat{\beta}_{\psi}$, klasik yöntemle ve izdüşüm teorisi yöntemiyle cebirsel olarak hesaplanmış, bu formülün izdüşüm teorisi kullanılarak kolayca nasıl hesaplandığı gösterilmiştir. Dördüncü Bölümde, Çoklu Doğrusal Regresyon modellerinde bir ve birden fazla gözlemin parametre tahminleri üzerindeki etkisi, $\hat{\beta}-\hat{\beta}_{\psi}$ kullanılarak Cook uzaklığı yöntemi ile ifade edilmiştir. Beşinci Bölümde sayısal bir örnek verilerek hesaplamaların nasıl yapılacağı gösterilmiş ve gözlem etkileri ile ilgili sonuçlar yorumlanmıştır. Son bölümde ise sonuç ve önerilere yer verilmiştir.

2. DOĞRUSAL SINIRLAMA ALTINDA PARAMETRE TAHMİNLERİ

n gözlem sayısı, m bağımsız değişken sayısı olmak üzere tam ranklı çoklu doğrusal regresyon modeli aşağıdaki gibi yazılabilir.

$$Y = X\beta + \varepsilon \qquad [1]$$

Burada, Y, gözlemlerin $n\times1$ boyutlu vektörü, X, bilinen değerlerin $n\times(m+1)$ boyutlu matrisi ve $rank(X)=(m+1)<n$, β, $(m+1)\times1$ boyutlu bilinmeyen katsayı vektörü ve ε, rasgele hata terimlerinin $n\times1$ boyutlu vektörüdür. Burada $E(\varepsilon)=0$ ve $V(\varepsilon)=I$ alınmıştır. Bu durumda, [1]'de verilen regresyon modelindeki β parametresinin doğrusal en iyi sapmasız tahmin edicisi (DESTE) SEK yöntemi ile şu şekilde elde edilebilir.

$\varepsilon'\varepsilon$ ifadesi $\varepsilon'\varepsilon = (Y - X\beta)'(Y - X\beta)$ şeklinde yazılır (8). β ya göre türev alınır sıfıra eşitlenirse, $\partial\varepsilon'\varepsilon / \partial\beta = -X'(Y - X\hat{\beta}) = 0$ dan,

$$\hat{\beta}=(X'X)^{-1}X'Y \qquad [2]$$

bulunur. Gözlemler üzerinde doğrusal bir sınırlama aşağıdaki gibi ifade edilebilir.

$$T=A'Y \qquad [3]$$

Burada A, $n\times r$ boyutlu bir matristir ve $r \leq n$ olmak üzere rank(A)=r'dir (6). Dolayısıyla [3] de verilen kısıt altında [1] de verilen doğrusal regresyon modeli sınırlandırılmış doğrusal regresyon modeline dönüşür.

$$T=A'Y=A'X\beta+A'\varepsilon \qquad [4]$$

Bu model için $E(T) = A'X\beta$ ve $Var(T) = A'A$ olduğu görülür. $u = A'\varepsilon$ olmak üzere, [4] deki β parametresi için DESTE,

$$\hat{\beta}_{\psi}=(X'A(A'A)^{-1}A'X)^{-1}(X'A(A'A)^{-1}A'Y) \qquad [5]$$

olarak elde edilir.

Tanım 2.1. $q_1\times q_2$ boyutlu herhangi bir M matrisinin sütunlarına dik olan izdüşüm matrisi P_M, rank(M)=$q_2 \leq q_1$ olmak üzere, $P_M = M(M'M)^{-1}M'$ şeklinde tanımlanır (9).

Bu tanım yardımıyla $P_A = A(A'A)^{-1}A'$ olmak üzere [5] nolu eşitlik aşağıdaki gibi ifade edilebilir.

$$\hat{\beta}_{\psi}=(X'P_AX)^{-1}(X'P_AY) \qquad [6]$$

ya da $< a, b >= a'b$ şeklinde iç çarpımlar yardımıyla [6],

$$\hat{\beta}_{\psi}=<X,P_AX>^{-1}<X,P_AY> \qquad [7]$$

yazılabilir. İzdüşüm matrisinin genel özelliklerinden;

$$<a,P_A b>=<P_A a,b>=<P_A a,P_A b> \quad [8]$$

olduğu bilinir (10). [8] eşitliği kullanılırsa $\hat{\beta}_\psi$,

$$\hat{\beta}_\psi=<P_A X,P_A X>^{-1}<P_A X,Y> \quad [9]$$

olarak ifade edilebilir. Bu eşitliğin bir sonucu olarak aşağıdaki teorem elde edilir.

Teorem 2.1.

$$Y=P_A X\beta+\varepsilon \quad [10]$$

şeklindeki regresyon modelinden $\hat{\beta}_\psi$ DESTE olarak elde edilir (7). Burada $E(\varepsilon)=0$ ve $V(\varepsilon)=I$ dır. Bu teoremin ispatı Pino tarafından ele alınmıştır (6).

3. $\hat{\beta}-\hat{\beta}_\psi$'NİN ELDE EDİLMESİ

Bu bölümde, incelenen gözlem ya da gözlemlerin katsayı tahminlerinde meydana getirdiği farkı hesaplamak için gerekli olan, $\hat{\beta}-\hat{\beta}_\psi$ farkının klasik yoldan ve izdüşüm teorisi yöntemiyle cebirsel olarak nasıl elde edildiği incelenecektir. Ancak bu yöntemlere geçmeden aşağıdaki teoremin verilmesinde fayda vardır.

Teorem 3.1. Bir doğrusal regresyon modeli aşağıdaki gibi verilsin.

$$Y=X\beta+Z\gamma+\varepsilon \quad [11]$$

Burada $E(\varepsilon)=0$ ve $V(\varepsilon)=I$ dır. β^*, [11] nolu doğrusal regresyon modelindeki β'nın DESTE'si olsun. Bu durumda, $\hat{\beta}_\psi$ tahmin edicisi aşağıdaki iki koşulun sağlanmasıyla β^*'ya denktir.

(i) rank(A)+rank(Z)=n
(ii) $Z'A=0$

İspat: Koşul (i) ve (ii) den ve Teorem 2.1'den $P_A+P_Z=I$ elde edilir (7).

$$Y=(I-P_Z)X\beta+\varepsilon=P_A X\beta+\varepsilon \quad [12]$$

yazılır. Burada $E(\varepsilon)=0$ ve $V(\varepsilon)=I$ dır. Buradan β'nın DESTE tahmin edicisi β^* dır.

3.1. Klasik Yöntem

Tanım 3.2. M ve N tekil olmayan matrisler olmak üzere,

$$(M+NDN')^{-1}=M^{-1}-M^{-1}N(N'M^{-1}N+D^{-1})^{-1}N'M^{-1} \quad [13]$$

dır (10).

$\hat{\beta}_{\psi} = (X'P_A X)^{-1}(X'P_A Y)$ idi. [12] yardımıyla $P_A = I - Z(Z'Z)^{-1}Z'$ yazılabilir. Bu durumda,

$$\hat{\beta}_{\psi}=(X'X-X'Z(Z'Z)^{-1}Z'X)^{-1}(X'Y-X'P_Z Y) \quad [14]$$

olur. [13] de, $M = X'X$, $N = X'Z$ ve $D = -(Z'Z)^{-1}$ olarak tanımlanırsa $P_X = X(X'X)^{-1}X'$ eşitliği de kullanılarak,

$$\hat{\beta}_{\psi}=[(X'X)^{-1}-(X'X)^{-1}X'Z(Z'P_X Z-Z'Z)^{-1}Z'X(X'X)^{-1}](X'Y-X'P_Z Y) \quad [15]$$

şeklinde ifade edilebilir. Parantez çarpımları yapılırsa ve $R = Z'(I-P_X)Z$ olarak alınırsa (rank(A) + rank (Z) =n olduğu hatırlanırsa rank(Z)=n-r=k' dır. Dolayısıyla R matrisi k$\times$ k tam ranklı bir matristir)

$$\hat{\beta}-\hat{\beta}_{\psi}=(X'X)^{-1}X'(P_Z-ZR^{-1}Z'P_X+ZR^{-1}Z'P_X P_Z)Y \quad [16]$$

elde edilir. Bu denklemde, $W = (P_Z - ZR^{-1}Z'P_X + ZR^{-1}Z'P_X P_Z)$ olsun.

$$R^{-1}R=I \quad [17]$$

olduğunda [17] eşitliğinin her iki tarafı Z matrisi ile soldan, $(Z'Z)^{-1}Z$ ile sağdan çarpılırsa,

$$ZR^{-1}R(Z'Z)^{-1}Z=Z(Z'Z)^{-1}Z \quad [18]$$

elde edilir. $R = Z'(I-P_X)Z$ idi. Öyleyse $ZR^{-1}Z'(I-P_X)P_Z = P_Z$ eşitliği kullanılarak

$$W = (ZR^{-1}Z(I-P_X)P_Z - ZR^{-1}Z'P_X + ZR^{-1}Z'P_X P_Z = Z(R^{-1}Z'(I-P_X)) \quad [19]$$

yazılabilir. Bu eşitlik [16] de yerine yazılırsa

$$\hat{\beta}-\hat{\beta}_{\psi} = (X'X)^{-1}X'Z(Z'(I-P_X)Z)^{-1}Z'(I-P_X)Y \quad [20]$$

olur.

3.2. İzdüşüm Teorisi Yöntemi

Teorem 3.1'den $\hat{\beta}_{\psi} = \hat{\beta}^{*}$ olduğu görülür. Bu sonuç kullanılarak

$$\hat{\beta} - \hat{\beta}_{\psi} = (X'X)^{-1}X'Z\hat{\gamma} \qquad [21]$$

yazılabilir. $P_X = X(X'X)X'$ olmak üzere $\hat{\gamma}$,

$$Y = (I - P_X)Z\gamma + \varepsilon \qquad [22]$$

modelindeki γ 'nın DESTE'sidir. [22] de verilen model için,

$$\varepsilon'\varepsilon = (Y - (I - P_X)Z\gamma)'(Y - (I - P_X)Z\gamma) \quad \text{ve} \quad \partial\varepsilon'\varepsilon / \partial\gamma = Z'(I - P_X)(Y - (I - P_X)Z\hat{\gamma}) = 0$$

izdüşüm matrisinin özelliklerinden $(I - P_X)^2 = (I - P_X)$ dir. Bu durumda,

$$\hat{\gamma} = (Z'(I - P_X)Z)^{-1}Z'(I - P_X)Y \qquad [23]$$

olarak düzenlenebilir. Sonuç olarak, [3] sınırlamasıyla,

$$\hat{\beta} - \hat{\beta}_{\psi} = (X'X)^{-1}X'Z(Z'(I - P_X)Z)^{-1}Z'(I - P_X)Y \qquad [24]$$

şeklinde de ifade edilebilir.

Yukarıda verilen her iki ispata bakıldığında, izdüşüm teorisi ile $\hat{\beta} - \hat{\beta}_{\psi}$ ifadesinin karmaşık matris işlemleri gerektirmeden daha kolay bir şekilde elde edildiği açıkça görülür. Aşağıda verilen tanım yardımıyla, [24] daha farklı bir şekilde ifade edilebilir.

Tanım 3.3. Bir $G_{a\times b}$ matrisi için, $(i_1, i_2, ..., i_r)$ satırları ve $(j_1, j_2, ..., j_s)$ sütunları tarafından seçilen alt matris formu G_H^F notasyonu ile gösterilebilir. Burada $H = \{i_1, i_2, ..., i_r\}$ ve $F = \{j_1, j_2, ..., j_s\}$ dir. Eğer, $H = \{1, ..., a\}$ veya $F = \{1, ..., b\}$ ise bunların dışında kalan satır ve sütunlar, G matrisinden silinir (11).

$Q = (I - P_X)$ ve $e = (I - P_X)Y = (Y - X\hat{\beta})$ olmak üzere [24]'nolu denklem

$$\hat{\beta} - \hat{\beta}_{\psi} = (X'X)^{-1}X'T^{D}(Q_D^D)^{-1}I_D(Y - X\hat{\beta}) \qquad [25]$$

dır.

4. ETKİLİ GÖZLEMLERİN BELİRLENMESİ

Parametre tahminleri üzerinde önemli değişikliklere sebep olan gözlemler etkili gözlemler olarak tanımlanmıştır (5). Bu etkinliğin araştırılması için bir çok etkinlik ölçütü türetilmiştir. Bu ölçütlerden birisi de Cook uzaklığıdır. Cook uzaklığının 1'den büyük değerleri için, ilgili gözlem ya da gözlem kümesinin parametre tahminleri üzerinde önemli bir etkiye sahip olduğunu söylenebilir (8, 11).

Cook uzaklığı;

$$C = \left[(\hat{\beta} - \hat{\beta}_{\psi})'X'X(\hat{\beta} - \hat{\beta}_{\psi})\right] / \left[(m+1)HKO\right] \qquad [26]$$

şeklinde tanımlanır. Burada HKO, hata kareler ortalamasıdır ve $HKO = e'e/(n-m-1)$ 'dir. $(\hat{\beta} - \hat{\beta}_{\psi})$ değeri ise [25] dan kolayca hesaplanır. C>1 olursa veri kümesinden çıkarılan gözlem etkin gözlemdir. Bu durumda bu gözlem veri kümesinden ya atılır yada model yeniden ele alınır veya modele yeni değişkenler eklenir (12).

5. SAYISAL BİR ÖRNEK

Bu bölümde, [25] verilen formüle ilişkin hesaplamaların nasıl yapıldığı gösterilmiş ve Cook uzaklığı yardımıyla gözlemlerin etkinlikleri incelenmiştir. Bu işlem yapılırken aşağıda verilen veri kümesinden yararlanılmıştır.

Çizelge 1. 10 gözleme ait veri kümesi

Gözlem No	Y	X_1	X_2	Gözlem No	Y	X_1	X_2
1	3	70	60	6	2	90	80
2	6	136,5	97	7	4	67	42
3	3	99	80	8	2	78	53
4	5	58	35	9	3	103	62
5	7	135	95	10	4	82	35

Bu veri kümesinden yararlanılarak bazı matris ve vektörler aşağıdaki gibi yazılabilir.

$$Y' = \begin{pmatrix} 3 & 6 & 3 & 5 & 7 & 2 & 4 & 2 & 3 & 4 \end{pmatrix} \text{ ve } X' = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 70 & 136.5 & 99 & 58 & 135 & 90 & 67 & 78 & 103 & 82 \\ 60 & 97 & 80 & 35 & 95 & 80 & 42 & 53 & 62 & 35 \end{pmatrix}$$

$Q = (I - P_X)$ ve $P_X = X(X'X)X'$ olmak üzere

$$Q = \begin{bmatrix} 0,694 & 0,053 & -0,191 & -0,175 & 0,057 & -0,293 & -0,148 & -0,141 & 0,047 & 0,097 \\ 0,053 & 0,596 & -0,145 & 0,129 & -0,394 & -0,081 & 0,068 & -0,006 & -0,180 & -0,041 \\ -0,191 & -0,145 & 0,794 & -0,031 & -0,135 & -0,258 & -0,043 & -0,080 & -0,018 & 0,108 \\ -0,175 & 0,129 & -0,031 & 0,715 & 0,119 & -0,057 & -0,238 & -0,173 & -0,073 & -0,215 \\ 0,057 & -0,394 & -0,135 & 0,119 & 0,615 & -0,069 & 0,060 & -0,010 & -0,185 & -0,059 \\ -0,293 & -0,081 & -0,258 & -0,057 & -0,069 & 0,638 & -0,058 & -0,096 & 0,057 & 0,219 \\ -0,148 & 0,068 & -0,043 & -0,238 & 0,060 & -0,058 & 0,797 & -0,154 & -0,086 & -0,197 \\ -0,141 & -0,006 & -0,080 & -0,173 & -0,010 & -0,096 & -0,154 & 0,871 & -0,081 & -0,130 \\ 0,047 & -0,180 & -0,018 & -0,073 & -0,185 & 0,057 & -0,086 & -0,081 & 0,791 & -0,272 \\ 0,097 & -0,041 & 0,108 & -0,215 & -0,059 & 0,219 & -0,197 & -0,130 & -0,272 & 0,490 \end{bmatrix}$$

dir. Daha önceki bölümlerde $e = (I - P_X)Y = (Y - X\hat{\beta})$ ifade edilmişti. Buradan,

$$e = QY = \begin{bmatrix} 0,694 & 0,053 & -0,191 & -0,175 & 0,057 & -0,293 & -0,148 & -0,141 & 0,047 & 0,097 \\ 0,053 & 0,596 & -0,145 & 0,129 & -0,394 & -0,081 & 0,068 & -0,006 & -0,180 & -0,041 \\ -0,191 & -0,145 & 0,794 & -0,031 & -0,135 & -0,258 & -0,043 & -0,080 & -0,018 & 0,108 \\ -0,175 & 0,129 & -0,031 & 0,715 & 0,119 & -0,057 & -0,238 & -0,173 & -0,073 & -0,215 \\ 0,057 & -0,394 & -0,135 & 0,119 & 0,615 & -0,069 & 0,060 & -0,010 & -0,185 & -0,059 \\ -0,293 & -0,081 & -0,258 & -0,057 & -0,069 & 0,638 & -0,058 & -0,096 & 0,057 & 0,219 \\ -0,148 & 0,068 & -0,043 & -0,238 & 0,060 & -0,058 & 0,797 & -0,154 & -0,086 & -0,197 \\ -0,141 & -0,006 & -0,080 & -0,173 & -0,010 & -0,096 & -0,154 & 0,871 & -0,081 & -0,130 \\ 0,047 & -0,180 & -0,018 & -0,073 & -0,185 & 0,057 & -0,086 & -0,081 & 0,791 & -0,272 \\ 0,097 & -0,041 & 0,108 & -0,215 & -0,059 & 0,219 & -0,197 & -0,130 & -0,272 & 0,490 \end{bmatrix} \begin{bmatrix} 3 \\ 6 \\ 3 \\ 5 \\ 7 \\ 2 \\ 4 \\ 2 \\ 3 \\ 4 \end{bmatrix}$$

$$e' = (0,421 \quad 0,585 \quad -0,634 \quad 2,069 \quad 1,594 \quad -1,014 \quad 0,779 \quad -1,460 \quad -1,757 \quad -0,583)$$

olarak bulunur.

Öncelikle D={i} durumunu inceleyelim (yani veri kümesinden tek bir gözlem noktasının çıkarıldığı durum), Örneğin birinci gözlem veri kümesinden çıkarılsın yani D={1} olsun, Bu durumda, Q matrisinin birinci köşegen elemanı, $Q_D^D = 0,694$ olarak alınır, D={1}olduğundan, I matrisinin birinci sütunu ve birinci satırı aşağıdaki gibi seçilir,

$$I^D = (1 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0)' \text{ ve } I_D = (1 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0)$$

şeklinde yazılır. Bu değerler [25]'da yerine yazılırsa,

$$\hat{\beta} - \hat{\beta}_{\psi} = (0,279 \quad -0,007 \quad 0,006)'$$

bulunur. Bu değerler [26]'da yerine yazılırsa Cook uzaklığı $C = 0,018$ olarak hesaplanır. C>1 için, gözlem kümesinden çıkarılan gözlemin parametre tahminleri üzerinde önemli bir etkiye sahip olduğunu söylemişti. Sonuç olarak 1-inci gözlemin etkili bir gözlem olmadığı söylenebilir. Benzer şekilde diğer gözlemler için hesaplanan değerler Çizelge 2'de verilmiştir.

Çizelge 2. Tek bir gözlem çıkartıldığında elde edilen Cook uzaklığı değerleri

Çıkarılan Gözlem No	C_i	Çıkarılan Gözlem No	C_i
1	0,018	6	0,143
2	0,061	7	0,030
3	0,021	8	0,057
4	0,373	9	0,161
5	0,404	10	0,113

Çizelgeye bakıldığında, gözlemlerin tek tek ele alındığı durumda hiçbir gözlemin parametre tahminleri üzerinde önemli bir etkiye sahip olmadığı görülür. Çünkü hiçbir C_i değeri 1'den büyük değildir.

Şimdi D{i,j}, $i \neq j$ durumunu ele alalım (yani veri kümesinden iki gözlem noktasının çıkarıldığı durum). Örneğin 1. ve 2. gözlemler veri kümesinden çıkarılsın. Bu durumda,

$$Q_D^D = \begin{bmatrix} 0,694 & 0,053 \\ 0,053 & 0,596 \end{bmatrix}$$

olarak seçilir. D{1,2} olduğundan birim matrisin 1 ve 2. satırları ile sütunları sırasıyla I^D ve I_D matrislerini oluşturur.

$$I^D = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}' \quad \text{ve} \quad I_D = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

dir. Bu değerler [26] nolu denklemde yerine yazılırsa Cook uzaklığı, C=0.060 olarak hesap edilir. Dolayısıyla 1 ve 2 nolu gözlemler bir arada düşünüldüğünde regresyon katsayıları üzerinde önemli bir etkiye sahip olmadığı söylenebilir (C=0.060<1) .Benzer şekilde farklı D{i,j}, i≠j durumları için Q_D^D, I^D ve I_D 'nin uygun seçimiyle Çizelge 3'de verilen sonuçlara ulaşılır.

Çizelge 2'deki gibi gözlemler tek tek ele alındığında etkili olmamasına rağmen, Çizelge 3'de de görüldüğü gibi iki gözlem aynı anda ele alındığında parametre tahminleri üzerinde etkili olabilmektedir. Bu durum maskeleme (masking) problemi olarak adlandırılır (13). Çizelge 3'e bakıldığında, {2,5}, {4,7} ve {9,10} gözlem alt kümelerine ait $C_{\{i,j\},i>j}$ değerlerinin 1'den büyük olduğu görülür dolayısıyla {2,5}, {4,7} ve {9,10} gözlem alt kümelerinin regresyon modelinin katsayı tahminleri üzerinde önemli bir etkiye sahip olduğu söylenebilir. Bunun anlamı, bu gözlem alt kümelerinin veri kümesinden çıkarılmasıyla regresyon doğrusunun önemli derecede değişeceğidir. Benzer şekilde üçlü, dörtlü v.b gözlem alt kümelerinin etkileri de incelenebilir.

Çizelge 3. İkili gözlem alt kümeleri çıkartıldığında elde edilen Cook uzaklığı değerleri

Durum	Çıkarılan Gözlemler	$C_{\{i,j\},i>j}$	Durum	Çıkarılan Gözlemler	$C_{\{i,j\},i>j}$	Durum	Çıkarılan Gözlemler	$C_{\{i,j\},i>j}$
1	1,2	0,060	16	2,9	0,117	31	5,6	0,389
2	1,3	0,007	17	2,10	0,138	32	5,7	0,363
3	1,4	0,818	18	3,4	0,361	33	5,8	0,438
4	1,5	0,382	19	3,5	0,344	34	5,9	0,149
5	1,6	0,162	20	3,6	0,617	35	5,10	0,400
6	1,7	0,120	21	3,7	0,036	36	6,7	0,129
7	1,8	0,038	22	3,8	0,137	37	6,8	0,376
8	1,9	0,221	23	3,9	0,198	38	6,9	0,205
9	1,10	0,207	24	3,10	0,078	39	6,10	0,082
10	2,3	0,032	25	4,5	0,355	40	7,8	0,013
11	2,4	0,322	26	4,6	0,380	41	7,9	0,124
12	2,5	6,469*	27	4,7	1,245*	42	7,10	0,039
13	2,6	0,134	28	4,8	0,168	43	8,9	0,380
14	2,7	0,058	29	4,9	0,328	44	8,10	0,440
15	2,8	0,112	30	4,10	0,399	45	9,10	1,888*

*Etkili durumlar

5. SONUÇ ve ÖNERİLER

Bu çalışmada, doğrusal sınırlama ve izdüşüm teorisi yardımıyla Pino tarafından iki aşamada elde edilen parametre tahminleri arasındaki fark formülünden yararlanılmıştır. İkinci bölümde de değinildiği gibi, uygun A matrisinin seçimiyle bir veya birden fazla gözlem veri kümesinden çıkarılarak katsayı tahminleri elde edilebilmektedir. Ayrıca bu doğrusal sınırlama yardımıyla model denklemi, P_A izdüşüm matrisini de içeren basit bir forma indirgenebilmektedir. Bu modelden yararlanılarak, izdüşüm teorisinin de kullanımıyla parametre tahminleri arasındaki fark formülünün nasıl elde edildiği üçüncü bölümde gösterilmiştir. Bu formül karmaşık matris işlemleri gerektirmemekte ve hesaplama kolaylığı

getirmektedir. Bu formül gerek çok değişkenli doğrusal regresyon modelinde gerekse çoklu doğrusal regresyon modelinde gözlem ya da gözlem vektörlerinin etkilerini incelemek amacıyla kullanılabilir.

Bu çalışmada, doğrusal sınırlamalar birinci aşama, izdüşüm teorisi ikinci aşama ve Cook uzaklığı üçüncü aşama olarak düşünülmüştür. Bu üç aşamalı yöntemin uygulanmasıyla Çoklu Doğrusal Regresyon modelinde bir tek gözlemin ve ikişerli gözlem kümelerinin parametre tahminleri üzerindeki etkisi incelenmiştir. Sayısal örnekte de görüldüğü gibi tek tek incelendiğinde etkili olmayan gözlemler ikişerli ele alındığında etkili olabilmektedir. Bu nedenle verilerin ileri derecede analizi için d gözlem sayısını göstermek üzere d=2, 3, 4, ..., r ($r<n/2$) gözlem gruplarının etkililikleri incelenmelidir. Ancak bu incelemelerde bir çok durumla karşılaşılacağından, tüm kombinasyonların hızlı bir şekilde incelenmesi için sezgisel optimizasyon yöntemlerinden yararlanılabilinir.

KAYNAKLAR

1. Draper, N.R., and Smith, H., "Applied Regression Analysis", ***Wile***y, New York, 152-153 (1980).
2. Weisberg, S., "Applied Linear Regression", ***Wiley***, New York, 107-124 (1980).
3. Barnett, V., and Lewis, T., "Outliers in Statistical data 3rd Edition", ***Wiley***, Chichester, 206-220 (1994).
4. Cook, R.D., "Detection of influential observation in linear regression", ***Technometrics***, 19(1): 15-18 (1977).
5. Belsley, D.A., Kuh E., and Welch, R.E., "Regression Diagnostics: Identifying Influential Data and Surces of Collinearity", ***Wiley***, New York, 42-51 (1980).
6. Pino, G.E., "Linear restrictions and two step least squares with applications", ***Statistics&Probability Letters***, 2: 245-248 (1984).
7. Gupta, A.K. and Kabe, D.G., "Linear restrictions and two step multivariate least squares with applications", ***Statistics&Probability Letters***, 32: 413-416 (1997).
8. Cook, R.D., and Weisberg, S., "Residual and Influence in Regression", ***Chapman and Hall***, New York (1982).
9. Belsley, D.A., "Conditioning Diagnostics:Collinearity and Weak Data in Regression", ***Wiley***, New York, 247-248 (1990).
10. Rao, C.R., "Linear Statistical Inference and Its Applications", ***Wiley***, New York, 186-187 (1973).
11. Fung, W.K., "A cautionary note on the use of generalized Cook-type measures", ***Computational Statistics&Data Analysis***, 19: 321-326 (1995).
12. Christensen, R., "Plane Answers to Complex Questions", ***Springer***, New York, 245-248 (1987).
13. Barrett, B.E., and Gray, J.B., "Leverage, residual, and interaction diagnostics for subsets of cases im least squares regression", ***Computational Statistics&Data Analysis***, 26: 39-52 (1997).

Geliş Tarihi:07.08.2002 *Kabul Tarihi:03.02.2003*