Süleyman Demirel Üniversitesi,
Fen Bilimleri Enstitüsü Dergisi,
15-1(2011),8-14

Varyansın Sağlam Tahmin Edicilerine Dayalı Cook Uzaklığı İstatistiği'nin İncelenmesi

Irmak ACARLAR*
Gazi Üniversitesi, Fen Fakültesi, İstatistik Bölümü, ANKARA
Alınış Tarihi:11.11.2010 Kabul Tarihi:14.03.2011

Özet: Cook Uzaklığı istatistiği regresyon analizinde etkili gözlemlerin saptanmasında, uygulama kolaylığından dolayı diğer yöntemlere göre kullanışlı bir yöntem sağlar. Veri kümesinde etkili gözlem(ler) mevcutken, Artık Kareler Toplamı (AKT) arttığından hata değişkenlerinin varyansı σ^2 'nin Ençok Olabilirlik (EÇOB) tahmini olan $\hat{\sigma}^2$ büyük değer alır. Bu durumda her bir gözlem için ayrı ayrı hesaplanan Cook Uzaklığı değerleri $\hat{\sigma}^2$ EÇOB tahmini değerinin büyük olmasından olumsuz bir biçimde etkilenebilir. Bu gibi durumlarda Cook ve Weisberg (1982), Cook Uzaklığı İstatistiği için varyansın sağlam tahminlerinin kullanılabileceğini önermişlerdir. Bu çalışmada σ^2 'nin sağlam tahmin edicilerine dayalı Cook Uzaklığı istatistiği incelenmiştir.

Anahtar kelimeler: Cook Uzaklığı, Varyansın Sağlam Tahmin Edicileri, Simülasyon.

Investigation of Cook's Distance Based on Robust Estimators of Variance

Abstract: In regression Cook's Distance is more useful technique than the other methods due to its simple application about detect influential observations. Since residual sum of squares (SS_{Res}) increases, $\hat{\sigma}^2$ known as least likelihood estimate of variance of errors σ^2 has great values when data contains influential observations. In this case Cook's Distance values calculated separately for each observation could be affected negatively from great values of the least likelihood estimate of variance. In that case of situations Cook and Weisberg (1982) proposed using robust estimations of variance for the Cook's Distance. In this paper, Cook's Distance has been investigated on robust estimators of σ^2.

Key words: Cooks Distance, Robust Estimators of Variance, Simulation.

Giriş

Model parametrelerinin tahminlerinde her bir gözlemin ayrıntılı olarak incelenmesi ve veri kümesinin geneline uymayan gözlemlerin belirlenmesi regresyon analizi için oldukça önemlidir. Çünkü tek bir gözlem bile regresyon parametrelerinin en küçük kareler tahminleri (EKK) üzerinde olumsuz yönde bir etkiye sahip olabilir. Dolayısıyla ilgili gözlem veya gözlemlerin veri kümesinden çıkartılması regresyon denklemini tamamen değiştirebilir. Bu tip bir gözlem etkili gözlem (influential observation) olarak tanımlanır.

Etkili gözlemlerin saptanmasının önemi ilk kez Cook(1977) tarafından ortaya atılmıştır ve Cook (1977,1979) etkili gözlemlerin saptanması için gözlem silme tekniğine dayalı olan Cook Uzaklığı İstatistiği'ni önermiştir. Belsley vd. (1980) tarafından önerilen ve gözlem silme tekniğine dayalı olan tanı istatistikleri DFBETAS, DFFITS ve COVRATIO istatistikleridir. Ayrıca tek gözlem silme tekniğine dayalı olan bir diğer istatistik Pena (2005) tarafından geliştirilmiştir. Altunkaynak (2003) etkili gözlemlerin saptanması için üç aşamalı bir tanı yöntemi önermiştir. Bu yöntem çoklu doğrusal regresyonda etkili gözlemlerin saptanması için doğrusal sınırlamalar, izdüşüm teorisi ve genelleştirilmiş Cook Uzaklığına dayalıdır.

Literatürde etkili gözlemlerin görsel olarak saptanması için grafiksel yöntemler de mevcuttur. Bu grafiklerden bazılarını Cook ve Weisberg (1982) vermiştir. Bununla birlikte yüksek boyutlu bir regresyon probleminin iki boyuta indirgenmesine dayalı bir grafiksel yöntem Li vd. (2001) tarafından önerilmiştir. Acarlar ve Gamgam (2010), etkili gözlem grubu içeren veri kümeleri için, bu grafiksel yöntemle Cook Uzaklığı ve COVRATIO istatistiklerini karşılaştırmış ve Cook Uzaklığı İstatistiği'nin etkili gözlem gruplarını daha iyi saptadığı sonucuna varmışlardır.

Veri kümesinde etkili gözlemler mevcutken Artık Kareler Toplamı arttığı için varyansın En Çok Olabilirlik Tahmini (EÇOB) olan Artık Kareler Ortalaması da büyük olur. Bu durum model parametrelerinin EKK tahminlerinin anlamlılığı için bir sorun oluşturmaktadır. Ayrıca Artık Kareler Ortalaması'nın büyük olması verinin geneline uyan gözlemlerle birlikte etkili gözlem için ayrı ayrı hesaplanan Cook Uzaklığı İstatistiği değerlerini de küçültür. Cook ve Weisberg (1982) bu sorunu gidermek için Cook Uzaklığı İstatistiği'nde varyansın EÇOB tahmini yerine sağlam tahminlerinin kullanılabileceğini belirtmiştir.

* *irmakacarlar@gazi.edu.tr*

Bu çalışmada bir etkili gözlem içeren veride bu etkili gözlemin saptanması için kullanılan ve varyansın EÇOB tahmin edicisi ve literatürde mevcut üç sağlam tahmin ediciye dayalı olan Cook Uzaklığı İstatistiği etkili gözlemi saptama oranına göre incelenmiştir. Sağlam tahmin edicilerin hesaplanması için gerekli olan ağırlık fonksiyonlarından Huber'in t Fonksiyonu, Ramsay'in E_a Fonksiyonu, Andrew'in Dalga Fonksiyonu ve Hampel'in 17A Fonksiyonu ele alınmıştır(Yılmaz,2004).

Çalışmanın ikinci bölümünde Cook Uzaklığı istatistiği tanıtılmıştır. Sonra üçüncü bölümde hata değişkenlerinin varyansının bazı sağlam tahmin edicileri ve bunlar için gerekli ağırlık fonksiyonları verilmiştir. Dördüncü bölümde ise varyansın EÇOB tahmin edicisi ve üçüncü bölümde tanıtılan üç sağlam tahmin ediciye dayalı Cook Uzaklığı İstatistiği etkili gözlemi saptama oranı bakımından simülasyon yardımıyla karşılaştırılmıştır. Son olarak dördüncü bölümde sonuç ve öneriler verilmiştir.

Cook Uzaklığı İstatistiği

Veri kümesindeki gözlem sayısı *n* ve regresyon modelindeki parametre sayısı da *p* olmak üzere, $n \times 1$ boyutlu yanıt vektörü *Y*, $n \times p$ boyutlu ve *p* ranklı açıklayıcı değişken matrisi *X*, $p \times 1$ boyutlu parametre vektörü β ve $n \times 1$ boyutlu 0 ortalamalı ve σ^2 varyanslı hata değişkenlerinin vektörü ε ile gösterilsin. Bu durumda çoklu doğrusal regresyon modeli,

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

biçiminde yazılır. Hata değişkenlerine ilişkin varsayımlar altında, bu modele ilişkin β parametre vektörünün EKK tahmini $\hat{\beta} = (X^T X)^{-1} X^T Y$ ile bulunur. Tahmin değerlerinin vektörü $\hat{Y}$ olmak üzere, artık vektörü $e = (I - H)Y$ ile verilir. Burada *H* matrisi şapka (hat) matrisidir ve $H = X(X^T X)^{-1} X^T$ ile elde edilir. Hata değişkenlerinin varyansı olan σ^2 'nin EÇOB tahmin edicisi ise

$$\hat{\sigma}^2 = e^T e / (n - p) \qquad (2)$$

ile verilir.

Tüm veri kümesine dayalı EKK tahmin vektörü $\hat{\beta}$ ile *i.* gözlemin veya veri kümesinin bir alt kümesinin veri kümesinden atılmasıyla elde edilen EKK tahmin vektörü $\hat{\beta}_{(i)}$ arasındaki karesel uzaklığın bir ölçüsü Cook (1977) tarafından önerilmiştir. Cook Uzaklığı İstatistiği olarak bilinen bu tanı istatistiği D_i ile gösterilir ve

$$D_i(X^T X, p\hat{\sigma}^2) = \frac{\left(\hat{\beta}_{(i)} - \hat{\beta}\right)^T (X^T X)\left(\hat{\beta}_{(i)} - \hat{\beta}\right)}{p\hat{\sigma}^2} \qquad (3)$$

$$i = 1,2, \dots, n$$

ile verilir.

Cook Uzaklığı $\hat{\beta}_{(i)}$ ile $\hat{\beta}$ arasındaki karesel uzaklığa dayalı olduğundan dolayı bu istatistik için kritik değer $\hat{\beta}_{(i)}$ ile $\hat{\beta}$ arasındaki karesel uzaklığı temel alarak çıkartılır. Veri kümesindeki tüm gözlemlerden elde edilen β parametre vektörünün EKK tahmin edicisi $\hat{\beta}$ vektöründen oluşturulan bir elipsoid, klasik anlamdaki bir güven elipsoidine karşılık gelmektedir. Eğer *i.* gözlemin veriden çıkartılmasıyla elde edilen β parametre vektörünün EKK tahmin edicisi $\hat{\beta}_{(i)}$ vektörü, $\hat{\beta}$ vektörü kullanılarak oluşturulan %50' lik güven elipsoidinin sınırında ya da bu sınırın dışında ise bu durumda oldukça büyük bir sapma söz konusudur ve bu gözlemin etkili olduğu söylenebilir. Çünkü normal şartlarda etkili olmayan bir gözlem veriden çıkartıldığında verinin geri kalanından elde edilen $\hat{\beta}_{(i)}$ vektörü %10 ile %20 civarında oluşturulan bir güven elipsoidi içinde olmalıdır. $\hat{\beta}_{(i)}$ vektörünün güven elipsoidinin dışında olduğu bu vektörün elemanlarıyla belirlenen noktanın güven elipsoidinin merkezine olan uzaklığı ile belirlenir. Bu uzaklık *i.* gözlem için hesaplanan D_i değerine karşılık gelmektedir.

Eğer *i.* gözlemin veriden silinmesiyle elde edilen $\hat{\beta}_{(i)}$ vektörünün grafikteki koordinatı %50' lik güven elipsoidinin sınırında ise $D_i \cong F_{0.50,p,n-p}$ olacaktır. Bu durumda $D_i > F_{0.50,p,n-p}$ ise grafikte $\hat{\beta}_{(i)}$ vektörüne ilişkin nokta güven elipsoidinin dışında olacaktır. Bundan dolayı verideki herhangi bir gözlem için hesaplanan Cook Uzaklığı $F_{0.50,p,n-p}$ kritik değerinden büyükse bu gözlem etkili bir gözlemdir. Burada dikkat edilmesi gereken nokta Cook Uzaklığı istatistiğinin *F* dağılımına yakınsadığıdır(Cook ve Weisberg,1982).

Hata Değişkenlerinin Varyansının Sağlam Tahmin Edicileri

Veri kümesinde aykırı gözlemler mevcutken hata değişkenlerinin varyansının EÇOB tahmini iyi bir dağılış ölçüsü değildir. Bu nedenle sağlam dağılış ölçülerine gerek duyulmuştur. Literatürde hata değişkenlerinin varyansı için önerilen sağlam tahmin edicilerden biri artık değerlerinin medyanından mutlak sapmaların medyanına dayalı olan

$$\hat{\sigma}^2 = \left[\text{medyan}|r_i - \text{medyan}(r_i)| / 0.6745\right]^2 \qquad (4)$$

ile verilir(Montgomery vd.,2001; Kashid ve Kulkarni, 2002). Burada r_i , β_j parametrelerinin sağlam tahminlerine dayalı olarak elde edilen artıklardır. Örnek hacmi büyük iken ve hata terimlerinin dağılımının normal olduğu varsayımı altında 0.6745 sabiti $\hat{\sigma}^2$ tahminini, σ^2 parametresinin yaklaşık olarak yansız bir tahmin edicisi yapar.

σ^2 için diğer bir tahmin edici ise artık değerlerinin medyanına dayalı olan

$$\hat{\sigma}^2 = [\text{medyan}\{|r_i|\}]^2 \quad (5)$$

ile verilir(Huber,2004).

β_j parametreleri için M-tahmin ediciler oldukça iyi birer tahmin edici oldukları için bu tahminlerle elde edilen etkili gözleme ilişkin artık değerleri diğer gözlemler için artık değerlerinden mutlak değerce daha büyük olacaktır. Dolayısıyla (2) numaralı eşitlikte tanımlanan varyansın EÇOB tahmin edicisi etkili gözlemlere ilişkin artık değerlerinden dolayı büyük olabilir. Bu sorunu gidermek için ağırlıklandırılmış sapmalara dayalı σ^2 parametresinin bir tahmin edicisi önerilmiştir. Bu tahmin edici

$$\hat{\sigma}^2 = \frac{1}{n-p}\sum_{i=1}^{n}(w_i^2 r_i^2) \quad (6)$$

ile verilir(Kashid ve Kulkarni, 2002). Burada w_i ağırlıklı en küçük kareler yöntemiyle elde edilen *i.* gözlem için ağırlığı temsil etmektedir.

Literatürde önerilen birçok ağırlık fonksiyonu vardır ve bunların temeli Huber (1964) tarafından önerilen artıkların simetrik bir fonksiyonu olarak tanımlanan $\rho(\cdot)$ fonksiyonunun en küçüklenmesine dayalıdır. Bilinen EKK yönteminde w_i ağırlıkları 1 değerini alır. e_i^*, *i.* gözlemin standartlaştırılmış artık değeri olmak üzere bilinen EKK kriter fonksiyonu ve bazı sağlam kriter fonksiyonları Yılmaz (2004) tarafından incelenmiştir. Bu çalışmada Yılmaz (2004) tarafından yapılan çalışmadan farklı olarak etkili bir gözlem içeren veride sağlam kriter fonksiyonlarıyla elde edilen ağırlıklardan hesaplanan varyansın sağlam tahmin edicilerine dayalı Cook Uzaklığı İstatistikleri bu etkili gözlemi saptama oranı bakımından incelenmiştir. Bu fonksiyonlar Çizelge1'deki gibidir.

Çizelge1. *Sağlam Kriter Fonksiyonları (Yılmaz, 2004, s.52)*

Kriter	$\rho(e^*)$	$\psi(e^*)$	$w(e^*)$	Aralık
EKK		e^*	1	$\lvert e^*\rvert < \infty$
Huber'in t Fonksiyonu $t = 1.345$		e^*	1	$\lvert e^*\rvert \le t$
	$\lvert e^*\rvert t - \frac{1}{2}t^2$	$t\ \text{sgn}(e^*)$	$\frac{t}{\lvert e^*\rvert}$	$\lvert e^*\rvert > t$
Ramsay'in E_a Fonksiyonu $a = 0.3$	$a^{-2}\left[1 - e^{-a\lvert e^*\rvert}\right][1 + a\lvert e^*\rvert]$	$e^* e^{(-a\lvert e^*\rvert)}$	$e^{(-a\lvert e^*\rvert)}$	$\lvert e^*\rvert < \infty$
Andrew'in Dalga Fonksiyonu $a = 1.339$	$a\left[1 - \cos\left(\frac{e^*}{a}\right)\right]$	$\sin\left(\frac{e^*}{a}\right)$	$\frac{\sin\left(\frac{e^*}{a}\right)}{\frac{e^*}{a}}$	$\lvert e^*\rvert \le a\pi$
		0	0	$\lvert e^*\rvert > a\pi$
Hampel'in 17A Fonksiyonu $a = 1.7$ $b = 3.4$ $c = 8.5$		e^*	1	$\lvert e^*\rvert \le a$
	$a\lvert e^*\rvert - \frac{1}{2}a^2$	$a\ \text{sgn}(e^*)$	$\frac{a}{\lvert e^*\rvert}$	$a < \lvert e^*\rvert \le b$
	$\frac{a\left(c\lvert e^*\rvert - \frac{1}{2}(e^*)^2\right)}{c-b} - \frac{7}{6}a^2$	$\frac{a\ \text{sgn}(e^*)(c - \lvert e^*\rvert)}{c-b}$	$\frac{a(c - \lvert e^*\rvert)}{\lvert e^*\rvert(c-b)}$	$b < \lvert e^*\rvert \le c$
	$a(b + c - a)$	0	0	$\lvert e^*\rvert > c$

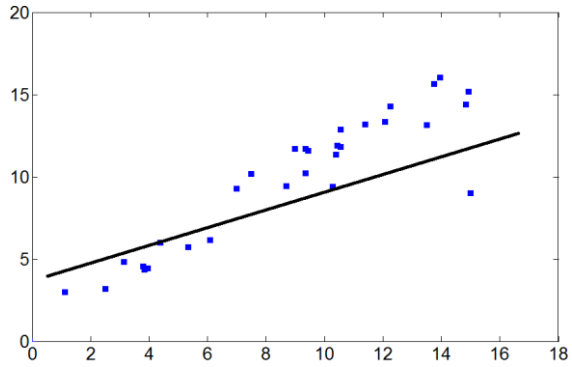
Simülasyon

Çalışmanın bu bölümünde (4), (5) ve (6) numaralı eşitliklerde verilen varyansın sağlam tahmin edicileri ve varyansın EÇOB tahmin edicisine dayalı Cook Uzaklığı İstatistikleri veri kümesindeki etkili gözlemi saptama oranı bakımından simülasyon yardımıyla karşılaştırılmıştır. Ayrıca varyansın sağlam tahmin edicileri için gerekli olan ağırlık fonksiyonlarından Huber'in t Fonksiyonu, Ramsay'in E_a Fonksiyonu, Andrew'in Dalga Fonksiyonu ve Hampel'in 17A Fonksiyonu ele alınmıştır. Simülasyonda örnek hacminin 20, 30 ve 50, tahmin edilecek parametre sayısının 2, 3 ve 5 olduğu durumlar ele alınmıştır.

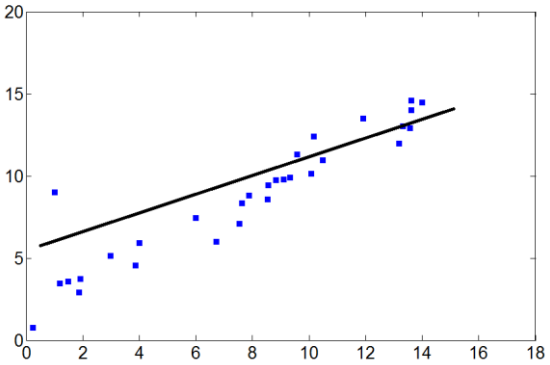
Simülasyon çalışmasında veri kümesini üretmek için Hadi ve Simonoff (1993) tarafından yapılan çalışmadaki veri üretme yönteminden yararlanılmıştır. Hadi ve Simonoff (1993) tarafından yapılan çalışmada veri kümesi sapan gözlem içermek amacıyla üretildiği için bu çalışmada, aynı yöntemle veri kümesi etkili gözlem içerecek biçimde MATLAB2010a programı kullanılarak üretilmiştir.

Simülasyon çalışmasında veri kümesinde bir etkili gözlemin bulunduğu durum ele alınmıştır. Buna dayalı

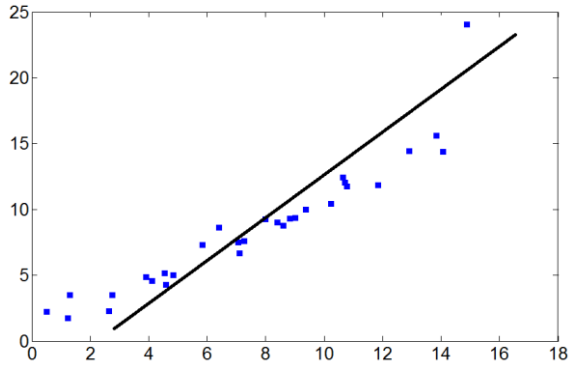
olarak bu veri için etkili gözlemin serpme diyagramındaki konumuna göre üç ayrı veri tipi oluşturulmuştur. Bu veri tipleri için birer örnek serpme diyagramı basit doğrusal regresyon durumu için Şekil1' deki gibidir.



Şekil1. *Simülasyon çalışmasında üretilen etkili gözlem içeren birinci tip veri için örnek durum*



Şekil2. *Simülasyon çalışmasında üretilen etkili gözlem içeren ikinci tip veri için örnek durum*



Şekil3. *Simülasyon çalışmasında üretilen etkili gözlem içeren üçüncü tip veri için örnek durum*

Bu çalışmada ele alınan (4), (5) ve (6) numaralı eşitliklerde verilen varyansın sağlam tahmin edicileri sırasıyla $\hat{\sigma}_1^2$, $\hat{\sigma}_2^2$ ve $\hat{\sigma}_3^2$ ile varyansın EÇOB tahmini ise $\hat{\sigma}_0^2$ ile gösterilmiştir.

Simülasyonun başlangıç aşamasında etkili gözlem içeren veri kümesi için önce bu verinin bağımsız değişkenlerinin değerleri türetilmiştir. *p* boyutlu parametre vektörünün elemanları 1 olacak şekilde bu vektör $\beta = [1\ 1 \dots 1]_{1\times p}$ olarak belirlenmiştir. Sonra veri kümesinin geneline uyan n-1 adet gözlem için bağımsız değişkenlerin değerleri [0, 15] aralığında tekdüze dağılımdan rastgele türetilmiştir. Sonra simülasyon aşağıdaki adımlar doğrultusunda yapılmıştır.

1.Adım: Veri kümesinin geneline uyan gözlemler için hata terimi ϵ_i (i = 2, 3, …, n) standart normal dağılımdan üretilmek üzere bu gözlemler için bağımlı değişken değerleri

$$Y_i = 1 + X_{i1} + X_{i2} + \cdots + X_{ip-1} + \epsilon_i \qquad i = 2,3, \dots, n$$

modeline göre türetilir.

2.Adım: Etkili gözlem 1. gözlem olarak gösterilmek üzere bu gözlem için bağımlı ve bağımsız değişkenlerin değerleri çalışmada ilgilenilen üç farklı veri tipine göre ayrı ayrı veri kümeleri için aşağıdaki gibi türetilir.

a) Birinci veri tipi için etkili gözlem olan 1. gözlemin bağımsız değişkenlerinin değerleri $x_1 = [15\ 15\ \dots\ 15]$ ve hata teriminin değeri $\epsilon_1 = -7$ alınarak bu gözlemin bağımlı değişken değeri

$$Y_1 = 1 + X_{11} + X_{12} + \cdots + X_{1\,p-1} - 7$$

modeline göre türetilir.

b) İkinci veri tipi için etkili gözlem olan 1. gözlemin bağımsız değişkenlerinin değerleri $x_1 = [1\ 1\ \dots\ 1]$ ve hata teriminin değeri $\epsilon_1 = +7$ alınarak bu gözlemin bağımlı değişken değeri

$$Y_1 = 1 + X_{11} + X_{12} + \cdots + X_{1\,p-1} + 7$$

modeline göre türetilir.

c) Üçüncü veri tipi için etkili gözlem olan 1. gözlemin bağımsız değişkenlerinin değerleri $x_1 = [15\ 15\ \dots\ 15]$ ve hata teriminin değeri $\epsilon_1 = +7$ alınarak bu gözlemin bağımlı değişken değeri

$$Y_1 = 1 + X_{11} + X_{12} + \cdots + X_{1\,p-1} + 7$$

modeline göre türetilir.

3.Adım: Veri kümesindeki her gözlem için $\hat{\sigma}_0^2$ ile gösterilen varyansın EÇOB tahminine dayalı olarak Cook Uzaklığı İstatistiği değerleri hesaplanıp bu istatistik için kritere göre etkili gözlem belirlenir.

4.Adım: w_i (i = 1, 2, …, n) ağrılıkları Huber'in t Fonksiyonu, Ramsay'in E_a Fonksiyonu, Andrew'in Dalga Fonksiyonu ve Hampel'in 17A Fonksiyonu ile elde edilir.

5.Adım: Veri kümesindeki her gözlem için 4. adımda elde edilen ağırlıklardan yararlanarak $\hat{\sigma}_1^2$, $\hat{\sigma}_2^2$ ve $\hat{\sigma}_3^2$ ile gösterilen varyansın sağlam tahminlerine dayalı Cook

Uzaklığı İstatistiği değerleri hesaplanıp bu istatistik için kritere göre etkili gözlem belirlenir.

Veri kümesinin geneline uyan n-1 adet gözlemin bağımsız değişkenlerinin değerleri her 1000 denemede yeniden türetilmek üzere bu deneme 50000 kez tekrarlanıp varyansın farklı tahminlerine dayalı Cook Uzaklığı İstatistiği'nin etkili gözlemi saptama oranları örnek hacminin ve tahmin edilecek parametre sayısının farklı durumları için elde edilmiştir. Sonuçlar Çizelge2, Çizelge3 ve Çizelge4'te sunulmuştur.

Çizelge2, Çizelge3 ve Çizelge4'teki sonuçlar incelendiğinde farklı örnek hacimleri ve parametre sayıları altında varyansın EÇOB tahminine dayalı Cook Uzaklığı İstatistiği'nin etkili gözlem olarak türetilen 1. gözlemi saptama oranlarının, varyansın sağlam tahminlerine dayalı Cook Uzaklığı İstatistiği'nin bu gözlemi saptama oranlarına göre daha düşük olduğu açıkça görülmektedir. Konum parametresi için örnek medyanı örnek ortalamasına göre daha sağlam bir tahmin edici olduğundan buna dayalı olan σ^2 parametresinin tahmini uç değerlerden doğal olarak etkilenmeyecektir. Etkili gözlem içeren bir veri kümesi için en uygun regresyon denklemi oluşturulduğunda bu etkili gözlemin artık değeri büyük olacağından artıkların medyanına dayalı olan varyansın sağlam tahminleri arzu edilen varyansa yakın olma eğilimindedir. Bu nedenle verinin geneline uymayan ve regresyon tahminlerini değiştirme eğiliminde olan bu gözlem için $\hat{\sigma}_1^2$ ve $\hat{\sigma}_2^2$ sağlam tahminlerine dayalı Cook Uzaklığı İstatistiği değeri diğer gözlemler için $\hat{\sigma}_1^2$ ve $\hat{\sigma}_2^2$ sağlam tahminlerine dayalı Cook Uzaklığı İstatistiği değerlerinden daha büyük olacaktır. Simülasyon sonuçları da bunu desteklemektedir. Ayrıca sağlam ağırlık fonksiyonlarıyla elde edilen w_i ağırlıkları da varyansın veri kümesindeki mevcut olan etkili gözlemden etkilenmesinin önüne geçmektedir. Çünkü etkili gözleme ilişkin ağırlık değeri varyansın büyük olmasını engeller.

Sonuç ve Öneriler

Bu çalışmada etkili gözlemlerin saptanması için kullanılan, varyansın EÇOB ve sağlam tahmin edicilerine dayalı Cook Uzaklığı İstatistiği, veri kümesindeki etkili gözlemi saptama oranı bakımından simülasyon yardımıyla incelenmiştir. Ayrıca çalışmada varyansın sağlam tahmin edicileri için gerekli olan ağırlık fonksiyonlarından Huber'in t Fonksiyonu, Ramsay'in E_a Fonksiyonu, Andrew'in Dalga Fonksiyonu ve Hampel'in 17A Fonksiyonu da ele alınmıştır. Simülasyonda türetilen etkili gözlem içeren veri kümesi için üç tip veri kümesi düşünülmüştür. Tüm simülasyon sonuçları genelleştirildiğinde farklı örnek hacimleri ve farklı parametre sayıları altında bir etkili gözlem içeren veri kümesindeki etkili gözlemi, varyansın sağlam tahminlerine dayalı Cook Uzaklığı İstatistiği varyansın EÇOB tahminine dayalı Cook Uzaklığı İstatistiği'ne göre saptama oranı daha büyüktür. Ağırlık fonksiyonları dikkate alındığında bu dört fonksiyonla elde edilen ağırlıklardan yararlanarak ayrı ayrı hesaplanan varyansın sağlam tahminine dayalı Cook Uzaklığı İstatistiği her dört fonksiyon için hemen hemen aynı sonuçları vermiştir. Her ne kadar varyansın sağlam tahmin edicilerine dayalı Cook Uzaklığı İstatistiği'nin varyansın EÇOB tahmin edicisine dayalı Cook Uzaklığı İstatistiği'ne göre veri kümesindeki etkili gözlemi saptama oranı daha büyük olsa da varyansın sağlam tahminleri küçük değerler aldığından dolayı bunlara dayalı olan Cook Uzaklığı İstatistiği etkili gözlem içeren veri kümesindeki verinin geneline uyan gözlemlerin birkaçını da etkili gözlem olarak saptama eğiliminde olabilir. Bu durum analiz için dikkat edilmesi gereken bir durumdur.

Çizelge2 *Örnek hacminin ve parametre sayısının farklı durumlarına göre, varyansın farklı tahmin edicilerine dayalı Cook Uzaklığı İstatistiği'nin birinci tip veride etkili gözlem olan 1. gözlemi saptama oranları*

			Huber			**Andrew**			**Ramsay**			**Hampel**		
p	**n**	F_{X_1}	F_{X_2}	F_{X_3}	F_{X_4}	F_{X_5}	F_{X_6}	F_{X_7}	F_{X_8}	F_{X_9}	$F_{X_{10}}$	$F_{X_{11}}$	$F_{X_{12}}$	$F_{X_{13}}$
	20	0.9049	0.9986	1	1	0.9989	1	1	0.9983	1	1	0.9988	1	1
2	**30**	0.8161	0.9957	1	0.9998	0.9965	1	1	0.9960	1	1	0.9963	1	1
	50	0.4558	0.9679	1	0.9996	0.9701	1	1	0.9985	1	1	0.9692	1	0.9999
	20	0.946	0.9989	1	1	0.9993	1	1	0.999	1	1	0.9989	1	1
3	**30**	0.9111	0.9979	1	1	0.9979	1	1	0.9975	1	0.9996	0.9980	1	1
	50	0.7057	0.9852	0.9999	0.9996	0.9866	1	0.9997	0.9850	1	0.9998	0.9860	1	1
	20	0.9476	0.9971	1	0.9993	0.9979	1	1	0.9973	1	1	0.9976	1	0.9996
5	**30**	0.9296	0.9974	1	0.9998	0.9981	1	1	0.9970	1	1	0.9977	0.9999	1
	50	0.7938	0.9894	0.9998	0.9995	0.9915	0.9999	0.9999	0.9892	0.9999	0.9999	0.9902	1	0.9999

Çizelge3. *Örnek hacminin ve parametre sayısının farklı durumlarına göre, varyansın farklı tahmin edicilerine dayalı Cook Uzaklığı İstatistiği'nin ikinci tip veride etkili gözlem olan 1. gözlemi saptama oranları*

			Huber			**Andrew**			**Ramsay**			**Hampel**		
p	**n**	F_{X_1}	F_{X_2}	F_{X_3}	F_{X_4}	F_{X_5}	F_{X_6}	F_{X_7}	F_{X_8}	F_{X_9}	$F_{X_{10}}$	$F_{X_{11}}$	$F_{X_{12}}$	$F_{X_{13}}$
	20	0.6819	0.9957	1	0.9998	0.9961	1	1	0.9952	1	1	0.9951	1	1
2	**30**	0.4967	0.9864	1	0.9997	0.9857	1	1	0.9859	1	1	0.9860	1	1
	50	0.1310	0.9010	0.9999	0.9944	0.9022	0.9999	1	0.9007	0.9999	1	0.9015	0.9999	0.9978
	20	0.7838	0.9970	1	0.9999	0.9976	1	1	0.9966	1	1	0.9973	1	1
3	**30**	0.6604	0.9914	1	0.9996	0.9930	1	0.9998	0.9918	1	0.9996	0.9918	1	1
	50	0.2924	0.9407	1	0.9971	0.9468	1	1	0.9418	1	0.9998	0.9403	1	0.9987
	20	0.8095	0.9954	1	0.9988	0.9978	1	1	0.9961	1	1	0.9961	1	0.9997
5	**30**	0.7251	0.9929	1	0.9995	0.9944	1	1	0.9927	0.9999	1	0.9933	1	0.9999
	50	0.4046	0.9601	0.9999	0.9972	0.9639	1	0.9999	0.9619	1	0.9999	0.9593	0.9999	0.9980

Çizelge4. *Örnek hacminin ve parametre sayısının farklı durumlarına göre, varyansın farklı tahmin edicilerine dayalı Cook Uzaklığı İstatistiği'nin üçüncü tip veride etkili gözlem olan 1. gözlemi saptama oranları*

			Huber			Andrew			Ramsay			Hampel		
p	**n**		X^TX	X^TX	X^TX	X^TX	X^TX	X^TX	X^TX	X^TX	X^TX	X^TX	X^TX	X^TX
2	**20**	0.9059	0.9983	1	1	0.9988	1	1	0.9984	1	1	0.9986	1	1
	30	0.8186	0.9956	1	1	0.9962	1	0.9999	0.9957	1	1	0.9964	1	1
	50	0.4569	0.9682	1	0.9995	0.9696	0.9998	1	0.9678	1	0.9999	0.9696	1	0.9999
3	**20**	0.9465	0.9988	1	0.9999	0.9992	1	1	0.9989	1	0.9998	0.9992	1	1
	30	0.9116	0.9974	1	1	0.9984	1	0.9998	0.9975	1	1	0.9976	1	1
	50	0.7057	0.9856	1	0.9998	0.9870	1	1	0.986	1	1	0.9859	0.9998	1
5	**20**	0.9473	0.9968	1	0.9993	0.9984	1	0.9999	0.9971	0.9999	0.9999	0.9974	1	0.9994
	30	0.9285	0.9971	0.9998	0.9999	0.9981	1	1	0.9974	1	0.9996	0.9975	1	0.9999
	50	0.7901	0.9891	1	0.9997	0.9907	0.9999	0.9998	0.9890	1	1	0.9883	1	1

Kaynaklar

Acarlar I., Gamgam H., 2010, Çoklu Doğrusal Regresyonda Etkili Gözlem Gruplarının Saptanması İçin Kullanılan Tanı Yöntemlerinin Karşılaştırılması. TÜİK İstatistik Araştırma Dergisi, 07(1), 83-99.

Altunkaynak, B., 2003. Doğrusal Sınırlamalar ve İzdüşüm Teorisi Yardımıyla Çoklu Doğrusal Regresyonda Etkili Gözlemlerin Tespiti. Gazi Üniversitesi Fen Bilimleri Dergisi, 16(3): 457-466.

Belsley, D.A., Kuh, E., Welsch, R.E., 1980. Regression Diagnostics: Identifying Influential Data and Sources of Collinearity. Wiley Series in Probability and Mathematical Statistics, New York, 6-84.

Cook, R.D., 1977a. Detection of Influential Observations in Linear Regression. Technometrics, 19 (1): 15-18.

Cook, R.D., 1979, Influential Observations in Linear Regression. Journal of the American Statistical Association, 74 (365): 169-174.

Cook, R.D., Weisberg, S., 1982. Residuals and Influence in Regression. Chapman and Hall, New York, 124s.

Hadi, A.S., Simonoff, J.S., 1993. Procedures for the Identification of Multiple Outliers in Linear Models. Journal of the American Statistical Association, 88(424): 1264-1272 .

Huber, P.J., 2004. Robust Statistics. Willey Series in Probability and Statistics, New Jersey, 153-195.

Kashid, D.N., Kulkarni, S.R., 2002. A More General Criterion for Subset Selection in Multiple Linear Regression. Communications in Statistics – Theory and Methods, 31(5), 795-811

Li, B., Martin, E.B., Morris, A.J., 2001. A Graphical Technique for Detecting Influential Cases in Regression Analysis. Communications in Statistics – Theory and Methods, 30(3): 463-483.

Montgomery, D.C., Peck, E.A., Vining, G.G., 2001. Introduction to Linear Regression Analysis. Wiley Series in Probability and Mathematical Statistics, New York, 207-219.

Pena, D., 2005. A New Statistics for Influence in Linear Regression. American Statistical Association and the American Society for Quality, 47(1): 1-12.

Yılmaz, S.S, 2004. Regresyonun M, L, R, Tahmin Edicileri ile Yanlı Tahmin Edicilerinin Kombinasyonu. Yüksek Lisans Tezi, Gazi Üniversitesi Fen Bilimleri Enstitüsü, Ankara, 52s.