Volume: 07 Number: 01 Category: 02 Page: 83-99 ISSN No: 1303-6319 Cilt: 07 Sayı: 01 Kategori: 02 Sayfa: 83-99

ÇOKLU DOĞRUSAL REGRESYONDA ETKİLİ GÖZLEM GRUPLARININ SAPTANMASI İÇİN KULLANILAN TANI YÖNTEMLERİNİN KARŞILAŞTIRILMASI

Irmak ACARLAR* **Hamza GAMGAM****

ÖZET

Regresyonda etkili gözlem ve gözlem grupları, tahmin değerlerinde önemli derecede farklılaşmalara neden olabilir. Bu farklılaşmalar modelin açıklanabilirliğini azalttığı için veri kümesindeki etkili gözlem veya gözlem gruplarının saptanması regresyon analizinin verimliliği açısından önemlidir. Bu çalışmada etkili gözlem gruplarının saptanması için kullanılan COVRATIO, Cook Uzaklığı istatistikleri ve grafik yöntemi incelenmiştir. Bu yöntemler, iki gözlemden oluşan etkili bir gözlem grubu içeren veri kümesinde bu gözlem grubunu etkili olarak saptama oranı bakımından karşılaştırılmıştır.

Anahtar Kelimeler: Etkili gözlem, Simülasyon, Tanı istatistikleri.

1. GİRİŞ

Regresyonda, gözlemlerden biri veya birkaçı veri kümesinin geneline uymayabilir. Bu tip gözlemler aykırı gözlemler (outliers) olarak adlandırılır. Bazı aykırı gözlemler ise mutlak değerce anormal büyüklükte artıklara sahip olabilir ve bunlar regresyon sonuçlarını olumsuz yönde etkileyebilir. Bir veri kümesinde, regresyon parametrelerinin En Küçük Kareler (EKK) tahminlerinde önemli derecede farklılaşmalara neden olan gözlemler, etkili gözlemler olarak tanımlanır. Etkili gözlemler için bir başka tanım ise, veri kümesinde bir etkili gözlem mevcutken ilgili gözlemin veri kümesinden çıkartılması sonucu regresyon tahminleri farklılaşıyorsa bu gözlem etkili gözlemdir biçiminde verilebilir (Cook, 1977; Montgomery vd., 2001).

Etkili gözlemlerin artıkları mutlak değerce oldukça büyük oldukları için, bunlar artık kareler toplamı değerinin büyümesine neden olur. Bundan dolayı bu gözlemler, belirleme katsayısı ve regresyon katsayılarının standart hataları gibi modele ilişkin istatistikler üzerinde olumsuz yönde etkide bulunur. Bu yüzden etkili gözlemlerin saptanması ve saptandıktan sonra, bu gözlemlerin etkisinin azaltılması için gerekli ağırlıklandırmaların yapılması, yapılan regresyon analizinin daha verimli olması açısından önemlidir.

Etkili gözlemler üzerinde ilk kez Cook (1977) tarafından çalışılmıştır. Son otuz yılda bu alanda birçok çalışma yapılmıştır. Bu süreç içerisinde etkili gözlemlerin saptanması için birimlerin tek tek incelenmesinin yanı sıra birimlerin gruplar halinde incelenmesinin de

* Gazi Üniversitesi, Fen Edebiyat Fakültesi, İstatistik Bölümü, Teknikokullar, Ankara,
e-posta: irmakacarlar@gazi.edu.tr
** Prof. Dr., Gazi Üniversitesi, Fen Edebiyat Fakültesi, İstatistik Bölümü, Teknikokullar, Ankara,
e-posta: gamgam@gazi.edu.tr

önemi ortaya çıkmıştır. Literatürde etkili gözlemlerin saptanması için önerilen tanı istatistikleri beş başlık altında toplanabilir. Bunlar

- Artıklara dayalı,
- Şapka (Projeksiyon, Hat) matrisine dayalı,
- Güven elipsoitlerinin hacmine dayalı,
- Etki eğrisine dayalı ve
- Kısmi etkililiğe dayalı tanı istatistikleridir (Chatterjee ve Hadi, 1986).

Bu tanı istatistiklerinin çoğunda gözlem silme tekniği kullanılmaktadır. Gözlem silme tekniğinde, bir gözlem veya gözlem kümesi veri kümesinden çıkartıldıktan sonra regresyon modeline ilişkin istatistiksel sonuçların nasıl etkilendiği incelenir.

Bilinen e_i artıklarına dayalı olan ve aykırı gözlemlerin saptanmasında kullanılan tanı istatistiklerinden biri Student-Türü Artıklar'dır. Student-Türü Artıklar, aykırı gözlemleri belirlemenin yanında, etkili gözlemlerin belirlenmesi için de kullanılır. Bu yöntemde önemli derecede büyük değerli Student-Türü Artıklar'a sahip gözlemler etkili gözlemler olarak değerlendirilebilir. Dahili ve R-Student Türü olarak ikiye ayrılan Student-Türü Artıklar, Margolin (1977) ve David (1981) tarafından tartışılmıştır.

Hoaglin ve Welsch (1978) aykırı gözlemlerin ve etkili gözlemlerin belirticisi olan yüksek dereceli kaldıraç noktalarını (high leverage points) saptamak için projeksiyon matrisi olarak da bilinen, Şapka Matrisi'nin köşegen elemanlarının kullanılabileceğini belirtmişlerdir. Bu matrisin köşegen elemanları h_{ii} ile gösterilir ve kaldıraç değeri olarak bilinir.

Güven elipsoitlerinin hacmine dayalı tanı istatistiklerine Andrews-Pregibon (1976) tarafından önerilen Andrews-Pregibon istatistiği örnek verilebilir. Ayrıca Belsley vd. (1980) tarafından önerilen kovaryans oranlarına dayalı *COVRATIO* istatistiği de oldukça kullanışlıdır. Cook ve Weisberg (1982) tarafından geliştirilen iki tanı istatistiği olan; Ençok Olabilirlik Uzaklığı ve Cook-Weisberg İstatistiği de güven elipsoitlerin hacmine dayalı istatistiklerdendir.

Uygulamada sıkça kullanılan ve Cook (1977) tarafından önerilen Cook Uzaklığı, etki eğrisi (influence curve/function) kavramının örnek uyarlaması olan örnek etki eğrisi (sample influence curve/function) kavramına dayalı bir istatistiktir. Gözlem silme tekniğine dayalı olan Cook Uzaklığı hem tek başına etkili olan gözlemleri, hem de ortak etkililiğe sahip gözlem kümelerini saptamada kullanılır. Ayrıca bu tanı istatistiğinin değiştirilmiş bir biçimi olan Düzeltilmiş Cook Uzaklığı da etkili gözlemlerin tespit edilmesinde kullanılan diğer bir istatistiktir (Cook ve Weisberg, 1982). Bunlara ek olarak etkili gözlemlerin saptanmasında Belsley vd. (1980) tarafından geliştirilen ve gözlem silme tekniğine dayalı iki kullanışlı tanı istatistiği olan *DFBETAS* ve *DFFITS* istatistikleri de etki eğrisinden türetilmiş istatistiklerdir.

Yüksek dereceli kaldıraç noktalarının ve etkili gözlemlerin incelenmesinde, şapka matrisinin ayrıştırılmasıyla elde edilen *j*. değişkenin, *i*. gözlemin h_{ii} değerine katkısını ölçmeye yarayan kısmi kaldıraç değeri (partial leverage) ve bu katkının görsel olarak incelenebildiği kısmi artık grafiği (partial residuals plot) de kullanılmaktadır. Hoaglin ve Welsch (1978) ve Chatterjee ve Hadi (1986) kısmi kaldıraç değerleri ve kısmi artık grafiği ile etkili gözlemlerin bulunması konusu üzerine çalışmışlardır.

Altunkaynak ve Ekni (2002), çok değişkenli doğrusal regresyonda etkili gözlem vektörlerinin saptanması için, doğrusal sınırlamalar, izdüşüm teorisi ve genelleştirilmiş Cook Uzaklığı'ndan oluşan üç aşamalı bir yöntem önermişlerdir. Bu yönteme dayalı olarak Altunkaynak (2003) çoklu doğrusal regresyonda etkili gözlemlerin saptanması için yeni ve hesaplama kolaylığı sağlayan bir yöntem geliştirmiştir.

Etkili gözlemlerin saptanması için bir başka yöntem de bağımlı değişkene karşı bağımsız değişken için oluşturulan serpme diyagramının incelenmesidir. Basit doğrusal regresyonda bu serpme diyagramı iki boyutlu olduğu için etkili gözlemler açık bir şekilde belirlenir. Fakat bağımsız değişken sayısı iki olduğunda, bağımlı değişkene karşı bağımsız değişken için oluşturulan üç boyutlu serpme diyagramından, gözlemlerin genel eğilimine uymayan noktalar görsel olarak saptanamayabilir. Çünkü bu üç boyutlu diyagramda bir hacim söz konusudur. Bağımsız değişken sayısının ikiden fazla olduğu durumlarda ise serpme diyagramı boyut sorunundan dolayı oluşturulamaz. Bu zorluluğu gidermek için Li vd. (2001) çoklu doğrusal regresyonda etkili gözlemlerin saptanması için bir grafiksel yöntem geliştirmişlerdir. Bu yöntemdeki ana fikir yüksek boyutlu bir regresyon problemini iki boyutlu tanı grafiklerinin bir setine indirgeyerek, bu grafiklerin görsel olarak incelenmesine dayanır. Li vd. (2001) bu metodolojiyi hem daha kolay bir yorumlamayı elde etmek, hem de hesaplamalarla diğer yöntemlere göre daha az uğraşmak amacıyla geliştirmişlerdir.

Etkili gözlem grupları incelenirken veri kümesinde regresyon tahminleri üzerinde ortak etkisi olan gözlemler mevcut olabileceği gibi bu tahminler üzerinde koşullu olarak etkili olan gözlemler de bulunabilir. Son yıllarda ortak etkililik (joint influence) ve koşullu etkililik (conditional influence) kavramlarına bağlı olarak maskeleme (masking) ve sürükleme (yanılgıya-düşürme, swamping) sorunları üzerinde durulmaktadır. Maskeleme, iki ya da daha fazla etkili gözlemin bulunduğu bir veri kümesinde, bu etkili gözlemlerden birinin, diğer etkili gözlem veya gözlemleri veri kümesinden tamamen atmadan etkili olarak saptanamaması durumudur. Sürükleme ise etkili bir gözlemle, verinin geneline uyan bir başka gözlemin etkili gözlem grubu olarak saptanmasıdır. Lawrance (1995) etkili gözlemlerin analizinde maskelemeyi ve sürükleme sorunlarını incelemiştir ve etkili gözlemler için bu sorunların aykırı gözlemlerdeki orijinal tanımlarıyla aynı olmadığını vurgulamıştır.

Çalışmanın ikinci bölümünde regresyonla ilgili temel kavramlara yer verilmiştir. Üçüncü bölümde etkili gözlem gruplarının saptanması için kullanılan Cook uzaklığı, *COVRATIO* tanı istatistikleri ve yeni bir yöntem olan grafik yöntemi tanıtılmıştır. Dördüncü bölümde ise bu yöntemler, literatürde tanı yöntemlerinin incelenmesi için sıklıkla kullanılan bir veri kümesi üzerinde incelenmiştir. Cook Uzaklığı, *COVRATIO* istatistikleri ve grafik yönteminin simülasyon çalışmasıyla karşılaştırılması beşinci bölümde verilmiştir. Son olarak altıncı bölümde de sonuç ve öneriler sunulmuştur.

2. YÖNTEM

Bu bölümde çoklu doğrusal regresyon modeli, tanı yöntemlerine ilişkin temel kavramlar ve bazı tanı yöntemleri verilmiştir.

2.1 Çoklu Doğrusal Regresyon Modeli

Hata terimi ile ilgili bilinen varsayımlar altında regresyon modeli Eşitlik (1)'de verilen biçimde tanımlanır.

$$Y = X\beta + \varepsilon \tag{1}$$

Bu modelde $n \times 1$ boyutlu yanıt vektörü Y, $n \times p$ boyutlu ve p ranklı tasarım matrisi X, $p \times 1$ boyutlu parametre vektörü β ve $n \times 1$ boyutlu 0 ortalamalı ve σ^2 varyanslı hata vektörü de ε ile gösterilir. β parametre vektörünün EKK tahmin edicisi olan $\hat{\beta}$ için,

$$\hat{\beta} = \left(X^T X\right)^{-1} X^T Y \tag{2}$$

olduğu bilinir. Tahmin değerlerinin vektörü $\hat{Y}$ olmak üzere, artık vektörü olan e aşağıdaki gibi tanımlanır.

$$\begin{aligned} e = Y - \hat{Y} &= Y - X\hat{\beta} \\ &= (I - X(X^T X)^{-1} X^T)Y \\ &= (I - H)Y \end{aligned} \tag{3}$$

Burada H, şapka matrisi olarak bilinir. $n \times n$ boyutlu olan şapka matrisinin köşegen elemanları (h_{ii}) aykırı ve etkili gözlemleri saptamada kullanılır (Hoaglin ve Welsch, 1978). Böylece şapka matrisi,

$$H = X(X^T X)^{-1} X^T \tag{4}$$

ile verilir. Simetrik ve eşgüçlü bir matris olan H matrisinin $i.$ köşegen elemanına karşılık gelen değer h_{ii} ile gösterilir ve kaldıraç değeri olarak adlandırılır. Hoaglin ve Welsch (1978) $i.$ gözleme karşılık gelen kaldıraç değerinin $2p/n$'den büyük olması durumunda, bu gözlemi yüksek dereceli kaldıraç noktası olarak tanımlamışlardır.

$i.$ gözleme karşılık gelen kaldıraç değerinin hesaplanması için alternatif bir formül,

$$h_{ii} = x_i (X^T X)^{-1} x_i^T \quad , \qquad (i = 1, 2, ..., n) \tag{5}$$

olarak verilir. Burada x_i vektörü, X matrisinin $i.$ satırıdır. h_{ii} değerlerinin özelliği toplamlarının tahmin edilecek parametre sayısı olan p 'ye eşit ve $0 \le h_{ii} \le 1$ olmasıdır (Hoaglin ve Welsch, 1978).

Veri kümesinin geneline uymayan gözlemlerin belirlenmesinde kullanılan diğer istatistiklerden ikisi Dahili ve R-Student Türü Artıklar'dır. Bunlardan Dahili Student-Türü Artık, r_i ile gösterilir ve

$$r_i = \frac{e_i}{\hat{\sigma}\sqrt{(1-h_{ii})}} \qquad , \qquad (i = 1,2,...,n) \tag{6}$$

ile verilir. Doğrusal bir modelde, Dahili Student-Türü Artıklar ortalaması 0 ve varyansı 1 olan normal dağılıma sahiptir (Margolin, 1977).

R-Student Türü Artıklar ise gözlem silme tekniğine dayalıdır. Hataların normalliği varsayımı altında *i.* gözlemin silinmesiyle hesaplanan varyansın tahmin edicisi $\hat{\sigma}^2_{(i)}$ olmak üzere bu gözleme ilişkin R-Student Türü Artık değeri,

$$t_i = \frac{e_i}{\sqrt{\hat{\sigma}^2_{(i)}(1-h_{ii})}} \qquad , \qquad (i = 1,2,...,n) \tag{7}$$

ile verilir. t_i istatistiği $n-p$ serbestlik dereceli Student-t dağılır. Bu durumda bu istatistik için kritik değer olarak $t_{\alpha,n-p}$ kullanılır (David, 1981).

2.2 Etkili Gözlem Gruplarının Saptanması için Tanı Yöntemleri

Etkili gözlemlerin saptanması için kullanılan tanı yöntemlerinden üçü, Cook Uzaklığı, *COVRATIO* istatistikleri ve yeni bir yöntem olan grafik tekniğidir. Bu tanı yöntemlerinden Cook Uzaklığı ve *COVRATIO* istatistikleri gözlem silme tekniğine dayalıdır. Ayrıca bu istatistikler, veri kümesindeki gözlemleri teker teker inceleme imkanı sağlamasının yanı sıra veri kümesindeki gözlemleri gruplar halinde de inceleme imkanı sağlamaktadır. Bu bölümde bu üç yöntem tanıtılmıştır.

2.2.1 Cook Uzaklığı İstatistiği

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 olan Cook Uzaklığı istatistiği, Cook (1977) tarafından önerilmiştir. Cook Uzaklığı, D_i ile gösterilir ve

$$D_i\left(X^TX, p\hat{\sigma}^2\right) = \frac{\left(\hat{\beta}_{(i)} - \hat{\beta}\right)^T X^TX\left(\hat{\beta}_{(i)} - \hat{\beta}\right)}{p\hat{\sigma}^2} \qquad i = 1,2,...,n \tag{8}$$

ile verilir. Veri kümesinin *m* büyüklüğünde bir alt kümesi olan ve *I* ile gösterilen gözlem grupları incelenmek istenirse bu istatistik,

$$D_I\left(X^TX, p\hat{\sigma}^2\right) = \frac{\left(\hat{\beta}_{(I)} - \hat{\beta}\right)^T X^TX\left(\hat{\beta}_{(I)} - \hat{\beta}\right)}{p\hat{\sigma}^2} \tag{9}$$

ile verilir. Literatürde D_i istatistiğine ilişkin kritik değer $F_{0.50,p,n-p}$ olarak bilinir. Bu durumda $D_i > F_{0.50,p,n-p}$ koşulunun sağlanması, *i.* gözlemin veya *I* ile gösterilen gözlem grubunun etkili olduğunu işaret eder.

2.2.2 COVRATIO İstatistiği

Etkili gözlemlerin saptanmasında kullanışlı olan bir diğer tanı istatistiği, *i.* gözlemin silinmesiyle geri kalan veriden elde edilen varyans-kovaryans matrisinin determinantının, tüm veriden elde edilen varyans-kovaryans matrisinin determinantına oranı olan $COVRATIO_i$ istatistiğidir. $X_{(i)}$, *i.* gözlemin silinmesiyle geri kalan veriyi temsil eden tasarım matrisi olmak üzere $COVRATIO_i$,

$$COVRATIO_i = \frac{\det\left\{\hat{\sigma}^2_{(i)}\left(X^T_{(i)}X_{(i)}\right)^{-1}\right\}}{\det\left\{\hat{\sigma}^2(X^TX)^{-1}\right\}} \qquad (i=1,2,...,n) \qquad (10)$$

olarak tanımlanır. $\det\left(X^T_{(i)}X_{(i)}\right) = \left(1-h_{ii}\right)\det\left(X^TX\right)$ olduğundan dolayı $COVRATIO_i$ istatistiği, Student-Türü artıklar cinsinden de yazılabilir (Belsley vd., 1980).

$$\begin{aligned} COVRATIO_i &= \left(\frac{\hat{\sigma}^2_{(i)}}{\hat{\sigma}^2}\right)^p \left(\frac{1}{1-h_{ii}}\right) \\ &= \frac{1}{\left(\frac{n-p-1}{n-p}+\frac{t_i^2}{n-p}\right)^p\left(1-h_{ii}\right)} \qquad (i=1,2,...,n) \end{aligned} \qquad (11)$$

Veri kümesinin *m* büyüklüğünde bir alt kümesi olan ve *I* ile gösterilen gözlem grupları incelenmek istenirse bu istatistik,

$$COVRATIO_I = \frac{\det\left\{\hat{\sigma}^2_{(I)}\left(X^T_{(I)}X_{(I)}\right)^{-1}\right\}}{\det\left\{\hat{\sigma}^2(X^TX)^{-1}\right\}} \qquad (12)$$

ile verilir. Belsley vd. (1980), $\left|COVRATIO_i - 1\right| > 3p/n$ olması durumunda, *i.* gözlemin etkili olduğunu öne sürmüşlerdir.

2.2.3 Grafik Tekniği

Etkili gözlem veya gözlem gruplarının saptanması için kullanılan bir diğer yöntem Li vd. (2001) tarafından önerilen grafik tekniğidir. Bu yöntemin benzer grafiksel yöntemlere göre iki avantajı; hem daha kolay yorumlayabilmeyi sağlamak, hem de

hesaplamalarla daha az uğraşmaktır. Bu yöntem adımsal bir yöntemdir ve her bir öz değere karşılık gelen tanı grafiğinin belli bir algoritmaya göre oluşturulup, ayrı ayrı incelenmesine dayanır. Tanı grafiklerini elde etmek için önerilen algoritma $l = 1, 2$ ve $j = 1, 2, ..., p$ için, j. özdeğer a_j olmak üzere aşağıdaki gibidir.

<u>*Adım 1*</u>: X matrisinin faktöriyel QR ayrıştırması, $n \times n$ boyutlu $Q = [Q_1, Q_2]$ matrisi için $X = QR$ biçiminde elde edilir. Burada tüm elemanları sıfır olan matris O olmak üzere $R = \left[R_1^T, O_{(n-p)\times p}^T\right]^T$ matrisi $n \times n$ boyutlu bir üst üçgen matrisi, R_1 tekil olmayan $p \times p$ boyutlu bir üst üçgen matrisi ve Q_1 matrisi $n \times p$ boyutlu bir dik matristir.

<u>*Adım 2*</u>: $p \times p$ boyutlu R_1 matrisinin tekil değer ayrıştırması,

$$R_1^T = P_1 diag\left\{a_1^{-1/2}, ..., a_p^{-1/2}\right\} P_2^T$$

ile hesaplanır. Burada $p \times p$ boyutlu P_1 ve P_2 matrisleri dik matrislerdir.

<u>*Adım 3*</u>: $\phi = Q diag\left\{O_{p\times p}, I_{(n-p)\times(n-p)}\right\} Q^T Y$ ve $\phi_0 = \phi / (\phi^T \phi)^{1/2}$ hesaplanır.

<u>*Adım4*</u>: Keyfi olarak belirlenen $s_j \times 1$ boyutlu r_l vektörü için, $u_j = Q_1 \left[G_j^T, O^R\right]^T r_l$ hesaplanır. Burada G_j vektörü P_1 matrisinin j. sütunudur.

<u>*Adım 5*</u>: $w_l^{(j)} = (\phi_0 + u_j) / 2^{1/2}$ hesaplanır.

j. tanı grafiği, bu adımlar doğrultusunda elde edilen $w_1^{(j)}$ ve $w_2^{(j)}$ vektörleri için serpme diyagramı oluşturularak elde edilir. Özdeğer sayısı kadar olan tanı grafiklerinden etkili gözlemi veya etkili gözlem gruplarını en açık bir şekilde sunan grafiği belirlemek için Li vd. (2001) tarafından önerilen bir karar değişkeni Göreli Duyarlılık Faktörü (Relative Sensitivity Factor, RSF) olarak tanımlanır ve

$$\lambda^{(j)} = \frac{a_j^{1/2}}{\sum_{j=1}^{p} a_j^{1/2}} \quad , \qquad (j = 1, 2, ..., p) \tag{13}$$

ile verilir.

Li vd. (2001) tanı grafiklerinde veri kümesinin geneline uymayan gözlemlerin tespit edilebilmesi için ρ yarıçaplı deneysel güven elipslerinin oluşturulabileceğini belirtmişlerdir. Deneysel güven elipsleri,

$$\left(w - w_0^{(j)}\right)^T \left[M^{(j)}\right]^{-1} \left(w - w_0^{(j)}\right) = \rho \tag{14}$$

ile elde edilir. Burada 2×1 boyutlu olan $w_0^{(j)}$ vektörünün elemanları $w_1^{(j)}$ ve $w_2^{(j)}$ vektörlerinin elemanlarının konum parametrelerinden oluşmaktadır. $M^{(j)}$ ise bu iki vektörün elemanları için oluşturulan kovaryans matrisidir.

3. BULGULAR

Bu bölümde etkili gözlem gruplarının saptanması için kullanılan Cook Uzaklığı, COVRATIO istatistikleri ve grafik tekniğine ilişkin bir uygulamaya yer verilmiştir. Sonra bu üç tanı yöntemi, bir simülasyon çalışmasıyla karşılaştırılmıştır.

3.1 Uygulama

Etkili gözlemlerin saptanması için önerilen bu üç yöntem küçük bir uygulama üzerinden gösterilmek istenirse, literatürde etkili gözlemlerin saptanması için kullanılan yöntemlerin değerlendirilmesi için yaygın olarak kullanılan ve etkili gözlem içeren Fare Verisi (Rat Data) ele alınabilir (Cook ve Weisberg, 1982). Veri kümesi, farelerin karaciğerine uygulanan bir ilacın miktarının incelenmesi suretiyle yapılan bir deneyden elde edilmiştir. 19 fare rastgele seçilmiştir ve yanıt değişkeni karaciğerdeki dozun oranı (y) ve üç bağımsız değişken vücut ağırlığı (x_1), karaciğer ağırlığı (x_2) ve vücuda verilen doz oranı (x_3) olarak belirlenmiştir. Cook ve Weisberg (1982) bu veri kümesi için aşağıdaki doğrusal modeli oluşturmuşlardır.

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \varepsilon_i \quad , \qquad i = 1,...,19 \tag{15}$$

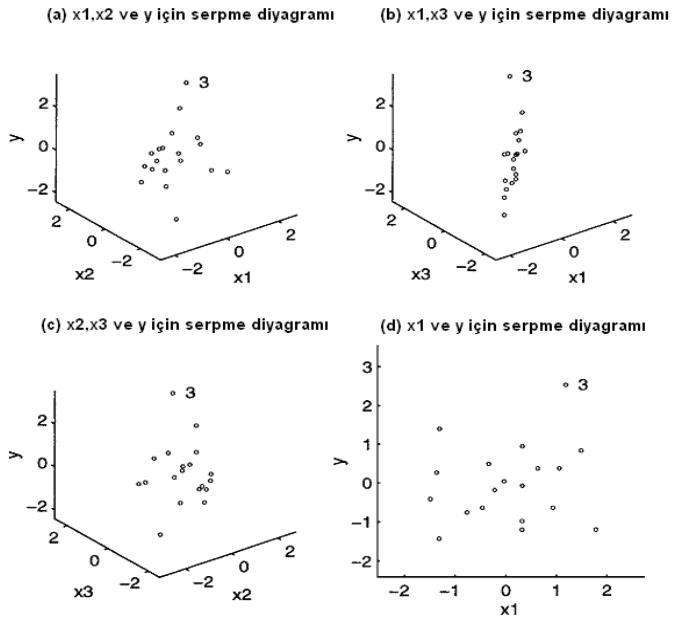
Gözlemlerin, Cook Uzaklığı ve COVRATIO değerleri Tablo 1'de verilmiştir. Tablo 1 incelendiğinde üçüncü gözleme ilişkin Cook Uzaklığı değeri olan D_3 değerinin, diğer gözlemlere ilişkin Cook Uzaklığı değerlerinden oldukça büyük olduğu gözlenmektedir. Ayrıca $D_3 > F_{0.50,3,16}$ olduğundan dolayı ilgili gözlem etkili gözlem olarak saptanır. Bu veri kümesindeki gözlemlerin COVRATIO değerleri, ilgili karar kuralına göre incelendiğinde ise üçüncü gözlemin etkili gözlem olduğu saptanır.

Tablo 1. Fare verisi için hesaplanan Cook Uzaklıkları ve *COVRATIO* değerleri

Gözlem No	D_i	$COVRATIO_i$
1	0,16883	0,63100
2	0,08854	1,01641
3	**0,92962**	**7,40080**
4	0,05718	0,8599
5	0,20292	1,52416
6	0,00049	1,56674
7	0,02462	1,28928
8	0,04686	1,52005
9	0,00049	1,40225
10	0,00005	1,49636
11	0,04144	1,06564
12	0,01890	1,44373
13	0,27260	0,97225
14	0,00537	1,46055
15	0,00373	1,35882
16	0,05099	1,37492
17	0,00425	1,60711
18	0,03163	1,27008
19	0,19994	0,51736

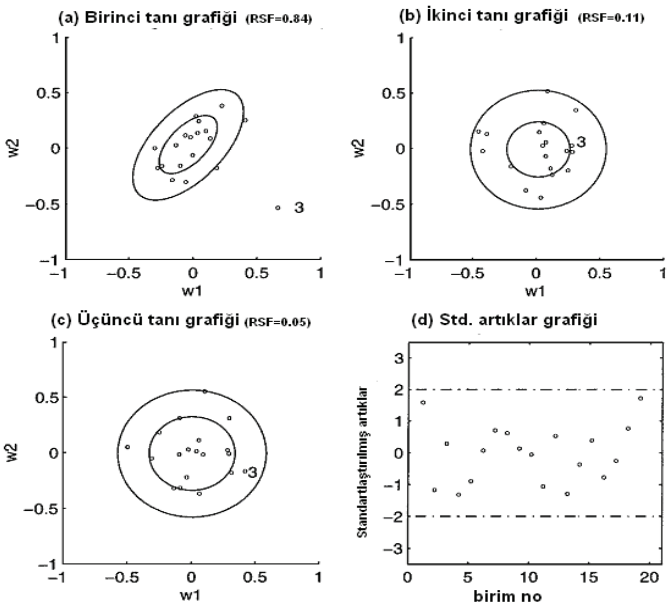
m hacimli etkili bir gözlem grubu, o grubu oluşturan etkili gözlemlerin genelleştirilmiş bir haline karşılık geldiği için bu gözlem grubunu saptamada aynı işlemler dizisi uygulanır.

3. gözlemi etkili olan bu veri kümesi için serpme diyagramları Şekil 1'de verilmiştir. İlk üç diyagram, bağımsız değişkenlerin ikili kombinasyonları ile yanıt değişkeni dikkate alınarak oluşturulan üç boyutlu diyagramlardır. Bu diyagramların üçünde de üçüncü gözlem etkili gözlem olmasından ziyade, aykırı gözlem olarak görülmektedir. İki boyutlu serpme diyagramında ise bu gözlemin etkili veya aykırı gözlem olduğuna kesin olarak karar vermek zordur.



Şekil 1. Fare Verisi için Serpme Diyagramları

Tanı grafikleri oluşturulmadan önce ham veri standartlaştırılmıştır. Li vd. (2001) tarafından önerilen algoritmaya göre oluşturulan tanı grafikleri Şekil 2'deki gibi elde edilmiştir. Grafikler incelendiğinde RSF değeri en büyük olan grafik birinci grafiktir. Bu durumda etkili gözlemleri en açık sunan grafiğin, bu grafik olduğu sonucuna varılır.



Şekil 2. Fare Verisi için Tanı Grafikleri

Birinci grafik incelendiğinde $w_1^{(1)}$ ve $w_2^{(1)}$ arasında doğrusal ilişki olduğu açık bir şekilde görülebilir. Ayrıca bu grafikteki noktalar incelendiğinde 3. gözlem veri kümesinin geneline uymadığı gözlenmektedir. Li vd. (2001) tarafından önerilen karar kuralına göre, bu nokta oluşturulan deneysel güven elipsinin dışında olduğu için regresyon tahminlerini değiştirme potansiyeline sahiptir. Diğer grafiklerde ise gözlemler düzgün bir yayılım göstermektedir. Bir başka ifade ile diğer grafiklerde 3. gözlem dışında veri kümesinin geneline uymayan başka gözlem yoktur. Bu durumda 3. gözlemin veri kümesindeki tek etkili gözlem olduğu sonucuna varılır.

3.2 Simülasyon Çalışması

Bu bölümde önce etkili gözlemlerin saptanması için kullanılan Cook Uzaklığı, *COVRATIO* istatistikleri ve grafik yöntemi, örnek hacminin ve bağımsız değişken sayısının farklı durumları için simülasyon kullanılarak, etkili gözlem grubu içeren veri kümesindeki etkili gözlem grubunu saptama oranı bakımından karşılaştırılmıştır. Sonra örnek hacmi ve bağımsız değişken sayısı sabit iken bu gözlemlere ilişkin hata terimleri mutlak değer olarak daha da büyütülerek bunların verinin merkezinden uzaklaştırıldığı durumda, bu yöntemler etkili gözlem grubu içeren veri kümesindeki etkili gözlem grubunu saptama oranı bakımından karşılaştırılmıştır.

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 aykırı gözlem içermek amacıyla üretildiği için bu çalışmada, aynı yöntemle veri kümesi etkili gözlem içerecek biçimde MATLAB2008a programı kullanılarak üretilmiştir.

Birden fazla etkili gözlemin bulunduğu bir veri kümesinde etkili gözlem veya gözlem grupları tanı istatistikleri kullanılarak saptanırken karşılaşılabilecek sorunlardan ikisi maskeleme ve sürüklemedir. Bu çalışmada, simülasyon maskeleme ve sürüklemeyle karşılaşılmayacak biçimde tasarlanmıştır.

Etkili bir gözlemin kaldıraç değeri etkili olmayan gözlemlerinkine göre daha büyüktür. Bu tanımdan yararlanarak 1 ve 2 indisleri ile gösterilen ve veri kümesindeki etkili gözlem grubunu oluşturan iki etkili gözlem için kaldıraç değerlerini büyütmek amacıyla bu gözlemlerin $p-1$ sayıda bağımsız değişkenlerin değerleri, veri kümesinin geneline uyan gözlemlere ilişkin bağımsız değişkenlerin değerlerinin tekdüze dağılımdan türetildiği $[0,15]$ aralığının en uç değeri olan 15 olarak belirlenmiştir. Bu gözlemlerin hata değişken değerleri ise sırasıyla $\varepsilon_1 = -5$ ve $\varepsilon_2 = -5.5$ olarak alınmıştır. Buradaki amaç bu gözlemlerin hata değişken değerini mutlak değerce arttırarak, bunların artık değerlerini de mutlak değerce büyütmektir. Sonra parametrelerinin değerleri 1 olan,

$$Y_i = 1 + X_{i1} + \ldots + X_{i\,p-1} + \varepsilon_i \qquad (16)$$

modeline göre etkili grubunu oluşturan gözlemlerin bağımlı değişken değeri türetilmiştir. Simülasyon boyunca bu gözlemlere ilişkin bağımsız değişkenlerin değerleri ve bağımlı değişken değerleri sabit kalmıştır.

Veri kümesinin geneline uyan $n-2$ sayıda gözlem için $p-1$ sayıda bağımsız değişken değerleri $[0,15]$ aralığında tekdüze dağılımdan türetilmiştir. Sonra simülasyon aşağıdaki adımlar doğrultusunda yapılmıştır.

***Adım 1*:** Bağımsız değişken sayısı $p-1$ olmak üzere, $p\times 1$ boyutlu β parametre vektörünün tüm elemanlarına 1 değeri atanır.

Adım 2: Veri kümesinin geneline uyan gözlemler için hata değişkenlerinin değerleri, $\varepsilon_i \quad N(0,1)$ dağılımından üretilir.

***Adım 3*:** $p-1$ sayıda bağımsız değişkenlerin değerleri ve hata değişkenlerinin değerleri kullanılarak,

$$Y_i = \beta_0 + \beta_1 X_{i1} + ... + \beta_{p-1} X_{i\,p-1} + \varepsilon_i \qquad (i = 3,4,...,n) \qquad (17)$$

modeline göre veri kümesinin geneline uyan gözlemler için bağımlı değişken değerleri türetilir.

***Adım 4*:** Türetilen veri kümesindeki her bir gözlem için $COVRATIO$ ve D_i istatistiklerinin değerleri hesaplanır ve her bir tanı istatistiğine ilişkin karar kuralına göre 1 ve 2 indisi ile gösterilen gözlemlerin etkili bir gözlem grubu olup olmadığına karar verilir.

***Adım 5*:** Aynı veri kümesi için grafik yöntemine göre tanı grafikleri oluşturulur. (13) eşitliğinde verilen ve göreli duyarlılık faktörü olarak tanımlanan karar değişkenine göre etkili gözlem grubunu belirleme gücü en yüksek olan grafik alınır. Bu grafikte etkili gözlem grubunu oluşturan gözlemler haricindeki diğer gözlemlere ilişkin noktaların merkeze uzaklıkları,

$$r = \left(w - w_0^{(j)}\right)^T \left[M^{(j)}\right]^{-1} \left(w - w_0^{(j)}\right)$$

ile hesaplanıp en büyük uzaklık deneysel güven elipsinin yarıçap uzunluğu olan ρ olarak alınır. Eğer 1 ve 2 indisleri ile gösterilen gözlemlere ilişkin noktaların elipsin merkezine uzaklıkları, ρ değerinden büyük ise bu noktalar deneysel güven elipsinin dışındadır ve grafik yöntemine göre bu gözlem grubu etkili bir gözlem grubudur.

Her 1 000 tekrarda veri kümesinin geneline uyan gözlemler için bağımsız değişkenlerin değerleri yeniden üretilmek üzere bu deneme 100 000 kez tekrarlanmıştır. Sonra etkili gözlem grubu içeren veri kümesinde, bu tanı yöntemlerinin etkili bir gözlem grubunu tespit etme oranları hesaplanmıştır.

3.2.1 Bağımsız Değişken Sayısı ve Örnek Hacminin Farklı Değerleri için Etkili Gözlem Gruplarını Saptamaya Dayalı Tanı Yöntemlerinin Karşılaştırılması

Bağımsız değişken sayısının iki olduğu, $p = 3$ olan durum için etkili gözlem grubu içeren bir veri kümesinde etkili gözlem gruplarını saptamak için kullanılan üç tanı yöntemine ilişkin simülasyon sonuçları Tablo 2'de verilmiştir.

Tablo 2. $p = 3$ iken Tanı Yöntemlerinin Veri Kümesinde Etkili Gözlem Grubunu Saptama Oranları

$p = 3$	$n = 20$	$n = 30$	$n = 40$
$D_{\{1,2\}}$	0,9989	0,9997	0,9998
$COVRATIO_{\{1,2\}}$	0,9790	0,9991	0,9995
Grafik Yöntemi	0,7450	0,8913	0,9253

Tablo 2'deki sonuçlar incelendiğinde Cook Uzaklığı ve $COVRATIO_{\{I\}}$ istatistiklerinin bu etkili gözlem grubunu saptama oranlarının ele alınan tüm örnek hacimlerinde oldukça büyük olduğu görülmektedir. Bununla birlikte tüm örnek hacimleri için diğer yöntemlere nazaran grafik yöntemin etkili gözlem grubunu saptama oranının düşük olduğu gözlenmektedir. Ayrıca örnek hacmi arttıkça grafik yönteminin, bu etkili gözlem grubunu saptama oranı da artmaktadır.

Bağımsız değişken sayısının üç olduğu durumda, $p = 4$ durumu için etkili gözlem gruplarını saptamaya dayalı tanı yöntemlerine ilişkin simülasyon sonuçları Tablo 3'te verilmiştir.

Tablo 3. $p = 4$ iken Tanı Yöntemlerinin Veri Kümesinde Etkili Gözlem Grubunu Saptama Oranları

$p = 4$	$n = 20$	$n = 30$	$n = 40$
$D_{\{1,2\}}$	0,9992	0,9996	0,9998
$COVRATIO_{\{1,2\}}$	0,9389	0,9937	0,9993
Grafik Yöntemi	0,3130	0,5917	0,7232

Tablo 3'teki sonuçlar incelendiğinde Cook Uzaklığı istatistiğine ilişkin etkili gözlem grubunu saptama oranının örnek hacminin tüm durumları için yüksek olduğu açıkça görülmektedir. $COVRATIO_{\{I\}}$ istatistiği ve grafik yöntemine ilişkin etkili gözlem grubunu saptama oranları ise örnek hacmi arttıkça artmaktadır. Buna ek olarak grafik yöntemi $n = 20$ iken etkili gözlem grubunu oldukça küçük bir oranla saptamaktadır.

Bağımsız değişken sayısının dört olduğu, $p = 5$ durumu için etkili gözlem grupları için tanı yöntemlerine ilişkin simülasyon sonuçları Tablo 4'te verilmiştir.

Tablo 4. $p = 5$ iken Tanı Yöntemlerinin Veri Kümesinde Etkili Gözlem Grubunu Saptama Oranları

$p = 5$	$n = 20$	$n = 30$	$n = 40$
$D_{\{1,2\}}$	0,9986	0,9991	0,9996
$COVRATIO_{\{1,2\}}$	0,8733	0,9855	0,9980
Grafik Yöntemi	0,1826	0,3340	0,5654

Tablo 4'teki sonuçlar incelendiğinde, farklı örnek hacimleri altında Cook Uzaklığı istatistiğinin etkili gözlem grubunu saptama oranının dikkate değer biçimde büyük olduğu görülmektedir. Bununla birlikte $COVRATIO_{\{I\}}$ istatistiği ve grafik yöntemine ilişkin oranlar farklı örnek hacimleri altında Cook Uzaklığı istatistiği için elde edilen oranlara göre daha küçüktür. Ayrıca bu iki yöntem için elde edilen etkili gözlem grubunu saptama oranları örnek hacmi arttıkça sayısal olarak büyümektedir.

Örnek hacmi $n = 20$ iken etkili gözlem gruplarını saptamak için kullanılan tanı yöntemlerine ilişkin Tablo 2, Tablo 3 ve Tablo 4'teki sonuçlar incelendiğinde bağımsız değişken sayısı arttıkça, $COVRATIO_{\{I\}}$ istatistiği ve grafik yöntemi için bulunan etkili bir gözlem grubunu saptama oranlarının küçüldüğü gözlenmektedir. Özellikle bu küçülme grafik yöntemi için daha da açık bir şekilde görülmektedir. Bununla birlikte ele alınan tüm örnek hacimleri ve bağımsız değişken sayıları için Cook Uzaklığı istatistiğine ilişkin etkili gözlem grubunu saptama oranları diğer yöntemler göre büyüktür.

Cook Uzaklığı istatistiğinin etkili bir gözlem grubunu saptama oranının ele alınan tüm örnek hacimleri ve bağımsız değişken sayıları için yüksek olması, m hacimli etkili bir gözlem grubunun bulunduğu veri kümesinde bu etkili gözlem grubunun silinmesiyle geri kalan gözlemlerden elde edilen EKK tahmini $\hat{\beta}_{(I)}$ ile tüm gözlemlere dayalı olarak elde edilen $\hat{\beta}$ arasındaki karesel uzaklığın oldukça büyük olmasından kaynaklanmaktadır. Dolayısıyla bu istatistik, β parametre vektörünün EKK tahminindeki büyük değişimi sayısal olarak diğer iki yönteme nazaran açıkça göstermektedir. Bağımsız değişken sayısı sabitken, örnek hacmi arttıkça $COVRATIO_{\{I\}}$ istatistiği için bulunan etkili gözlem grubunu saptama oranlarının artmasının nedeni ise bu istatistiğe ilişkin karar kuralında yer alan kritik değerin örnek hacminin azalan bir fonksiyonu olmasıdır.

Grafik yöntemi, çok boyutlu bir regresyon problemini iki boyutlu bir probleme dönüştürmek suretiyle oluşturulan p sayıda tanı grafikleri setinin, her bir grafikteki noktalar için oluşturulan güven bölgesi kriterine bağlı olarak incelemesine dayalıdır. Veri kümesinde bir etkili gözlem grubu bulunduğunda bu etkili gözlem grubunu saptama oranı bakımından grafik yönteminin incelenmesi, bu yöntemin regresyon problemini nasıl dönüştürdüğünü değerlendirmek açısından önemlidir. Bağımsız değişken sayısı sabitken grafik yöntemiyle etkili gözlem grubunu saptama oranlarına ilişkin sonuçlar incelendiğinde örnek hacminin artmasıyla bu yöntemin etkili gözlem grubunu daha iyi saptadığı açıkça görülmektedir. Fakat örnek hacmi sabit iken bağımsız

değişken sayısı arttıkça, grafik yönteminin etkili gözlem grubunu saptama oranı düşmektedir. Buna bağlı olarak regresyon probleminin boyutu arttıkça, grafik yönteminin regresyon problemini dönüştürme eğiliminin azaldığı sonucuna varılır.

3.2.2 Bağımsız Değişken Sayısı ve Örnek Hacmi Sabitken Etkili Gözlem Gruplarını Saptamaya Dayalı Tanı Yöntemlerinin Karşılaştırılması

Bu simülasyonda, bağımsız değişken sayısı ve örnek hacmi sabit iken etkili gözlem grubunu oluşturan gözlemlerin hata değişkenlerinin değerleri mutlak değerce arttırılarak, bunların verinin merkezinden uzaklaştırıldığı durumlarda, tanı yöntemlerinin etkili gözlem grubu içeren veri kümesinde bu gözlem grubunu saptama oranları elde edilmiştir. Buradaki amaç etkili gözlem grubundaki gözlemlerin, hata değişken değerlerinin daha da arttığı durumlarda tanı yöntemlerini karşılaştırmaktır.

$n = 20$ ve $p = 3$ iken etkili gözlem gruplarını saptamak için kullanılan tanı yöntemlerine ilişkin simülasyon bir öncekine benzer biçimde bölümün başında belirtilen adımlar doğrultusunda yapılmıştır. Simülasyon yapılırken etkili gözlem grubunu oluşturan iki gözlemin hata değişkeni değerleri sırasıyla $\varepsilon_1 = -5$ ve $\varepsilon_2 = -5,5$ alınıp bu gözlemlerin bağımsız değişken değerleri de kullanılarak bağımlı değişken değerleri elde edilmiştir. Sonra bölümün başında verilen adımlar doğrultusunda, her 1 000 tekrara da verinin geneline uyan gözlemlerin bağımsız değişkenlerinin değerleri yeniden üretilmek üzere,

$$Y_i = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_{p-1} X_{i\,p-1} + \varepsilon_i \qquad (i = 3, 4, \ldots, n)$$

modeli dikkate alınıp, bu deneme 100 000 kez tekrar edilmiştir ve etkili bir gözlem grubu içeren veri kümesinde etkili gözlem gruplarını saptamak için kullanılan tanı yöntemlerinin bu etkili gözlem grubunu saptama oranları elde edilmiştir. Aynı simülasyon bu iki gözlemin hata değişkeni değerleri önce $\varepsilon_1 = -7$, $\varepsilon_2 = -7,5$ için sonra $\varepsilon_1 = -10$, $\varepsilon_2 = -10,5$ için yapılıp sonuçlar Tablo 5'te verilmiştir.

Tablo 5. $n = 20$ ve $p = 3$ iken Tanı Yöntemlerinin Farklı Hata Değişken Değerlerine göre Üretilen İki Etkili Gözlemin Oluşturduğu Etkili Gözlem Grubunu Saptama Oranları

$n = 20,\ p = 3$	$\varepsilon_1 = -5$ $\varepsilon_2 = -5,5$	$\varepsilon_1 = -7$ $\varepsilon_2 = -7,5$	$\varepsilon_1 = -10$ $\varepsilon_2 = -10,5$
$D_{\{1,2\}}$	0,9989	1	1
$COVRATIO_{\{1,2\}}$	0,9790	0,9987	1
Grafik Yöntemi	0,7450	0,8829	0,9094

Tablo 5'teki sonuçlar incelendiğinde veri kümesindeki etkili gözlem grubunu oluşturan gözlemlerin hata değişkeni değerlerinin artmasıyla $COVRATIO_{(I)}$, Cook Uzaklığı istatistikleri ve grafik yönteminin bu gözlem grubunu saptama oranlarının arttığı gözlenmektedir. $COVRATIO_{(I)}$ ve Cook Uzaklığı istatistiklerinin etkili gözlem grubunu saptama oranlarının ele alınan tüm örnek hacimlerinde yüksek olmasıyla birlikte grafik yöntemi için deneysel olarak elde edilen oran bu iki istatistiğe göre daha küçüktür.

4. TARTIŞMA VE SONUÇ

Sonuç olarak Cook Uzaklığı İstatistiği etkili gözlem gruplarını saptamak için kullanılan yöntemler içinde bu gözlem gruplarını saptama bakımından en iyi yöntemdir. Bundan başka *COVRATIO* istatistiği de etkili gözlem gruplarına duyarlıdır. Yüksek boyutlu bir regresyon probleminin iki boyutlu bir probleme indirgenmesi amacıyla veri kümesine bir dönüşümün uygulandığı grafik yöntemi ise regresyon probleminin boyutu arttıkça etkili gözlem gruplarını saptama bakımından duyarlılığı azalmaktadır.

5. KAYNAKLAR

Altunkaynak, B., Ekni, M., 2002. Detection of influential observation vectors for multivariate linear regression. Hacettepe Journal of Mathematics and Statistics, 31: 139-151.

Altunkaynak, B., 2003. Doğrusal sınırlamalar ve izdüşüm teorisi yardımıyla çoklu doğrusal regresyonda etkili gözlemlerin tespiti. Gazi Üniversitesi Fen Bilimleri Dergisi, 16(3): 457-466.

Andrews, D. F., Pregibon, D., 1976. Finding outliers that matter. J. Roy. Statist. Soc., Ser. B. 40: 85-93.

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.

Chatterjee, S., Hadi, A. S., 1986. Influential observations, high leverage points and outliers in linear regression. Statistical Science, 1(3): 379-416.

Cook, R. D., 1977a. Detection of influential observations in linear regression. Technometrics, 19 (1): 15-18.

Cook, R. D., Weisberg, S., 1982. Residuals and influence in regression. Chapman and Hall, New York, 10-20, 101-156.

David, H. A., 1981, Order statistics, 2nd Edn. Wiley, New York, 110-150.

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 .

Hoaglin, D. C., Welsch, R. E., 1978. The hat matrix in regression and ANOVA. The American Statistician, 32 (1): 17-22.

Lawrance, A. J., 1995. Deletion influence and masking in regression. Journal of the Royal Statistical Society Series B (Methodological), 57(1): 181-189.

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.

Margolin, B. H., 1977. The distribution of internally studentized statistics via laplace transform inversion. Biometrica, 64: 573-582.

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.

COMPARISON OF DIAGNOSTIC METHODS FOR DETECTING INFLUENTIAL SETS IN MULTIPLE LINEAR REGRESSION

ABSTRACT

In regression, an influential observation and influential sets would cause noticeable differentiations on fitted values. Since these differentiations decrease explicability of model, detecting the influential observation or the influential sets in data is important for efficiency of regression analysis. In this study COVRATIO, Cook Distance statistics and graphical technique used for detecting influential sets are examined. These methods are compared with regard to ratios of detecting influential set in data which includes two influential observations.

Keywords: Influential observation, Simulation, Diagnostics.