ANADOLU ÜNİVERSİTESİ BİLİM VE TEKNOLOJİ DERGİSİ –B
Teorik Bilimler

ANADOLU UNIVERSITY JOURNAL OF SCIENCE AND TECHNOLOGY –B
Theoretical Sciences
Cilt/Vol.:1-Sayı/No: 2 : 105-116 (2011)

REGRESYONDA BİR ETKİLİ GÖZLEMİN SAPTANMASI İÇİN KULLANILAN TANI YÖNTEMLERİNİN KARŞILAŞTIRILMASI

Irmak ACARLAR [1]

ÖZ

Regresyonda etkili gözlem ve gözlem grupları, tahmin değerlerinde önemli derecede farklılaşmalara neden olabilir. Bu farklılaşmalar modelin açıklanabilirliğini azalttığı için verideki etkili gözlem veya gözlem gruplarının saptanması regresyon analizinin verimliliği açısından önemlidir. Bu çalışmada etkili gözlem ve gözlem gruplarının saptanması için kullanılan DFFITS, DFBETAS, COVRATIO, Cook Uzaklığı, S tanı istatistikleri ve grafik yöntemi incelenmiştir. Bu yöntemler etkili bir gözlem içeren veride bu gözlemi etkili gözlem olarak saptama oranı bakımından karşılaştırılmıştır.

Anahtar Kelimeler : Etkili gözlem, Tanı istatistikleri, Tanı grafikleri, Simülasyon.

COMPARISON OF DIAGNOSTIC METHODS FOR DETECTING AN INFLUENTIAL OBSERVATION IN REGRESSION

ABSTRACT

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

Keywords: Influential observation, Diagnostics, Diagnostic graphs, Simulation.

[1,] Gazi Üniversitesi, Fen Edebiyat Fakültesi, İstatistik Bölümü, Pk:06500 Teknikokullar, Ankara, Türkiye.
E-mail: irmakacarlar@gazi.edu.tr.

Geliş: 27 Ekim 2009; **Düzeltme:** 14 Haziran 2010; **Kabul:** 24 Eylül 2010

1. GİRİŞ

Regresyonda, veri kümesindeki gözlemlerden biri veya birkaçı verinin 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. Regresyon parametrelerinin en küçük kareler (EKK) tahminlerinde önemli derecede farklılaşmalara neden olan gözlemler, etkili gözlemler olarak tanımlanır(Montgomery vd., 2001; Cook, 1977a).

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

- Şapka (projeksiyon) matrisine dayalı tanı istatistikleri,
- Artıklara dayalı tanı istatistikleri,
- Güven elipsoitlerinin hacmine dayalı tanı istatistikleri,
- Etki eğrisine dayalı tanı istatistikleri ve
- Kısmi etkililiğe dayalı tanı istatistikleri

biçiminde ifade edilmiştir(Chatterjee ve Hadi, 1986).

Verideki 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 tasarım 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 doğrusal regresyon modeli,

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

biçiminde yazılır. Hata değişkenine ilişkin varsayımlar altında, bu modele ilişkin β parametre vektörünün EKK tahmin edicisi, $\hat{\beta} = \left(X^T X\right)^{-1} X^T Y$ ile bulunur. Tahmin değerlerinin vektörü $\hat{Y}$ olmak üzere, artık vektörü,

$$e = (I - H)Y \tag{2}$$

ile verilir. Eşitlik (2)' de H matrisi, Şapka (Hat) matrisi olarak tanımlanır ve

$$H = X(X^T X)^{-1} X^T \tag{3}$$

ile verilir. Hoaglin ve Welsch (1978) x-yönünde aykırı gözlemlerin belirticisi olan yüksek dereceli kaldıraç noktalarını (high leverage points) saptamak için projeksiyon matrisi olarak da bilinen Şapka matrisinin köşegen elemanlarının kullanılabileceğini belirtmişlerdir. Kaldıraç değeri (leverage value) olarak bilinen Şapka matrisin köşegen elemanları h_{ii} ile gösterilir ve

$$h_{ii} = x_i (X^T X)^{-1} x_i^T \quad , \qquad (i = 1, 2, ..., n) \tag{4}$$

olarak verilir. Ayrıca Hoaglin ve Welsch (1978), tahmin edilecek parametre sayısı p ve örnek çapı n olmak üzere, h_{ii} değeri $2p/n$'den büyük olan gözlemleri yüksek dereceli kaldıraç noktası olarak tanımlamışlardır. Buna ek olarak 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 incelene-

bildiği kısmi artık grafiği (partial residuals plot) de kullanılmaktadır(Hoaglin ve Welsch, 1978; Chatterjee ve Hadi, 1986).

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 artıklara sahip gözlemler etkili gözlemler olarak değerlendirilebilir. Dahili (Internal) ve Harici (External) olarak ikiye ayrılan student türü artıklar Margolin (1977) ve David (1981) tarafından tartışılmıştır.

Güven elipsoitlerinin hacmine dayalı tanı istatistiklerine Andrews ve Pregibon (1976) tarafından önerilen Andrews-Pregibon istatistiği örnek verilebilir. Ayrıca Belsley vd. (1980) tarafından önerilen kovaryans oranlarına dayalı *COVRATIO* istatistiği oldukça kullanışlıdır. *COVRATIO* istatistiği,

$$COVRATIO_i = \frac{\det\left\{\hat{\sigma}^2_{(i)}\left(X^T_{(i)}X_{(i)}\right)^{-1}\right\}}{\det\left\{\hat{\sigma}^2\left(X^TX\right)^{-1}\right\}} \qquad (i=1,2,...,n) \tag{5}$$

ile verilir. Yukarıdaki eşitlikte $\hat{\sigma}^2_{(i)}$, veriden i. gözlem çıkartıldığında geri kalan gözlemlerden hesaplanan ortalama artık karedir ve $X_{(i)}$ matrisi i. gözlemin silinmesiyle elde edilen tasarım matrisidir. Bu istatistikle regresyon parametrelerinin tahminleri üzerinde hem tek başına etkili olan gözlemler hem de ortak bir etkililiğe sahip gözlemler incelenebilir. 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 İstatistiği etki eğrisi (influence curve/function) kavramının örnek versiyonu olan örnek etki eğrisi (sample influence curve/function) kavramına dayalı bir istatistiktir. Gözlem silme tekniğine dayalı olan Cook Uzaklığı hem tek başına etkili olan bir gözlemi hem de ortak etkililiğe sahip gözlem kümelerini saptamada kullanılır. Tahmin edilecek parametre sayısı p olmak üzere Cook Uzaklığı İstatistiği,

$$D_i\left(X^TX, p\hat{\sigma}^2\right) = \frac{\left(\hat{\beta}_{(i)}-\hat{\beta}\right)^T X^TX\left(\hat{\beta}_{(i)}-\hat{\beta}\right)}{p\hat{\sigma}^2} \qquad i=1,2,...,n \tag{6}$$

ile verilir. 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 regresyon tahminlerini değiştirme eğiliminin olduğunu göstermektedir(Cook, 1977a; Cook ve Weisberg, 1982).

Belsley vd. (1980) tarafından önerilen gözlem silmeye dayalı olan $DFFITS_i$ istatistiği, i. gözlemin silinmesiyle bu gözlemin tahmin değerleri üzerinde yaptığı etkiyi inceleyen bir tanı yöntemidir. Verinin tümünden elde edilen regresyon katsayılarıyla hesaplanan i. gözleme ilişkin tahmin değeri $\hat{y}_i$ ve veriden i. gözlemin silinmesiyle elde edilen regresyon katsayılarıyla hesaplanan i. gözleme ilişkin tahmin değeri de $\hat{y}_{(i)}$ ile gösterilsin. Buna göre $DFFITS_i$,

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

olarak tanımlanır. Belsley vd. (1980), bu istatistik için kritik değer (cutoff value) olarak $2\sqrt{p/n}$ ' i önermiştir. Bu durumda $\left|DFFITS_i\right| > 2\sqrt{p/n}$ koşulunu sağlayan gözlemler etkili gözlemlerdir.

Belsley vd. (1980) tarafından önerilen gözlem silmeye dayalı diğer bir istatistik olan $DFBETAS_{ij}$, *i*. gözlemin silinmesi durumunda standart sapmaya bağlı olarak regresyon katsayılarının ne kadar değişeceğini gösteren bir tanı yöntemidir. *i*. gözlemin silinmesiyle elde edilen β vektörünün EKK tahmin edicisi $\hat{\beta}_{(i)}$ ile gösterilsin. $\hat{\beta}$ vektörünün *j*. elemanı $\hat{\beta}_j$ ve $\hat{\beta}_{(i)}$ vektörünün *j*. elemanı da $\hat{\beta}_{j(i)}$ ile ifade edilirse $DFBETAS_{ij}$ istatistiği,

$$DFBETAS_{ij} = \frac{\hat{\beta}_j - \hat{\beta}_{j(i)}}{\sqrt{\hat{\sigma}^2_{(i)}(X^T X)^{-1}_{jj}}} \qquad i = 1,2,...,n; \quad j = 1,...,p \tag{8}$$

olarak tanımlanır. Belsley vd. (1980)' in $DFBETAS_{ij}$ istatistiği için önerdiği kritik değer $2/\sqrt{n}$ olarak bilinir. Buna göre $\left|DFBETAS_{ij}\right| > 2/\sqrt{n}$ koşulunun sağlanması durumunda *i*. gözlem etkili gözlemdir.

Tek başına etkili olan gözlemlerin saptanması için kullanışlı bir başka istatistik Pena (2005) tarafından önerilmiştir. Bu tanı istatistiği ise tahmin değerlerine dayalı Cook uzaklığının geliştirilmiş bir biçimidir ve verideki her bir gözlemin silinmesiyle *i.* gözleme ilişkin tahminin duyarlılığını ölçmektedir.

Altunkaynak (2003), çoklu doğrusal regresyonda etkili gözlemlerin saptanması için doğrusal sınırlamalar, izdüşüm teorisi ve genelleştirilmiş Cook Uzaklığına dayalı üç aşamalı bir yöntem geliştirmiştir. Lojistik regresyonda aykırı gözlemlerin incelenmesiyle ilgili bir çalışma Vupa (2009) tarafından yapılmıştır.

Etkili gözlemlerin saptanması için bir başka yöntem Li vd. (2001) tarafından önerilen grafiksel yöntemdir. 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 benzer yöntemlere göre daha az uğraşmak amacıyla geliştirmişlerdir.

Çalışmanın ikinci bölümünde etkili gözlemlerin saptanması için iki yeni yöntem olan Pena' nın S_i tanı istatistiği ve grafik yöntemi hakkında bilgi verilmiştir. *DFFITS, DFBETAS,* Cook Uzaklığı, *COVRATIO,* S_i tanı istatistikleri ve grafik yönteminin simülasyon çalışmasıyla karşılaştırılması üçüncü bölümde verilmiştir. Son olarak dördüncü bölümde de sonuç ve öneriler sunulmuştur.

2. PENA' NIN S_i İSTATİSTİĞİ VE GRAFİK TEKNİĞİ

Etkili gözlemlerin saptanması için son yıllarda önerilen yöntemlerden biri gözlem silme tekniğine dayalı Pena'nın S_i istatistiğidir(Pena, 2005). Bu istatistik Cook Uzaklığı istatistiğinin geliştirilmiş bir biçimidir. Li vd. (2001) tarafından önerilen grafik tekniği ise yüksek boyutlu bir regresyon probleminin iki boyutlu tanı grafiklerinin bir setine indirgenmesine dayalıdır. Bu bölümde bu iki yöntem tanıtılmıştır.

2.1 Pena' nın S_i İstatistiği

Gözlem silme tekniğine dayalı olarak tahmin değerlerindeki farklılaşmanın incelendiği istatistiklerden biri Pena (2005) tarafından önerilen $S_{(i)}$ istatistiğidir. Bu istatistik *i*. gözlemin tahmininin her bir gözlemin tek tek silinmesiyle nasıl değişeceğini ölçen alternatif bir yöntemdir. Böylece etkili gözlemler verideki diğer gözlemlerin yardımıyla belirlenir.

Eşitlik (1)' deki model dikkate alınırsa tüm veriden elde edilen $\hat{\beta}_j$ istatistikleriyle hesaplanan $i.$ gözlemin tahmin değeri $\hat{y}_i$ ile veriden bir gözlemin çıkartılmasıyla elde edilen $\hat{\beta}_{j(k)}$ istatistikleriyle hesaplanan $i.$ gözlemin tahmin değeri $\hat{y}_{i(k)}$ arasındaki farka ilişkin vektör,

$$s_i = \left(\hat{y}_i - \hat{y}_{i(1)}, ..., \hat{y}_i - \hat{y}_{i(n)} \right)^T \tag{9}$$

biçiminde tanımlansın (Pena, 2005). Bu $i.$ gözlemin tahmin değerinin verideki her bir gözlemin silinmesine karşı duyarlılığını göstermektedir. Böylece $S_{(i)}$ istatistiği, s_i vektörünün standartlaştırılmış karesel normu olacak biçimde,

$$S_{(i)} = \frac{s_i^T s_i}{p \cdot \hat{\sigma}_{(\hat{y}_i)}} \quad , \qquad (i = 1, ..., n) \tag{10}$$

ile verilir. Burada $\hat{\sigma}_{(\hat{y}_i)} = \sqrt{s^2 h_{ii}}$ ile hesaplanır (Pena, 2005).

Hiçbir aykırı gözlem olmaması ve h_{ii} değerlerinin tümünün küçük olması durumunda $S_{(i)}$ istatistiğinin beklenen değeri yaklaşık olarak $1/p$ olur. Diğer bir deyişle yüksek dereceli kaldıraç noktalarının olmadığı bir veride gözlemlerin tümünün aynı duyarlılığa sahip olması beklenir. Bu beklenen değer kaldıraç noktalarına oldukça bağlı olan Cook uzaklığına göre önemli bir avantajdır(Pena, 2005).

Pena (2005) etkili gözlemlerin saptanması için önerdiği S_i istatistiğine ilişkin,

$$\left| S_{(i)} - \text{med}\left(S\right) \right| \geq 4.5 \times \text{MAD}\left(S_{(i)}\right) \tag{11}$$

karar kuralını önermiştir. Eğer bu eşitsizlik sağlanırsa $i.$ gözlem etkili gözlemdir. (11) eşitsizliğinde $\text{med}\left(S\right)$ değeri $S_{(i)}$ değerlerinin medyanı, $\text{MAD}\left(S_{(i)}\right)$ ise $S_{(i)}$ değerlerinin medyandan sapmalarının mutlak değerlerinin medyanıdır(Pena, 2005).

2.2 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 $l = 1, 2$, $j = 1, 2, ..., p$ ve $j.$ özdeğer a_j olmak üzere aşağıdaki gibidir:

Adım 1: X matrisinin faktöriyel QR ayrıştırması, $n \times n$ boyutlu $Q = \left[Q_1, Q_2 \right]$ 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.

Adım 2: $p \times p$ 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.

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

Adım4: 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.

Adım5: $w_l^{(j)} = (\phi_0 + u_j) / 2^{1/2}$ 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 (RSF) olarak tanımlanır ve

$$\lambda^{(j)} = \frac{a_j^{1/2}}{\sum_{j=1}^{p} a_j^{1/2}} \quad , \qquad (j = 1,2,...,p) \tag{12}$$

ile verilir.

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

$$\left(w - w_0^{(j)}\right)^T \left[M^{(j)}\right]^{-1} \left(w - w_0^{(j)}\right) = \rho \tag{13}$$

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. SİMÜLASYON

Bu bölümde önce etkili gözlemlerin saptanması için kullanılan *DFBETAS*, *DFFITS*, Cook Uzaklığı, *COVRATIO*, $S_{(i)}$ istatistikleri ve grafik yöntemi, örnek hacminin ve bağımsız değişken sayısının farklı durumları için simülasyon kullanılarak etkili gözlem içeren verideki etkili gözlemi saptama oranı bakımından karşılaştırılmıştır. Sonra örnek hacmi ve bağımsız değişken sayısı sabit iken, verideki etkili gözlemin hata terimi mutlak değer olarak daha da büyültülerek bu noktanın verinin merkezinden uzaklaştırıldığı durumda, bu yöntemler etkili gözlem içeren verideki etkili gözlemi saptama oranı bakımından karşılaştırılmıştır.

DFBETAS, *DFFITS* ve *COVRATIO* tanı istatistiklerine ilişkin karar kuralında yer alan kritik değerler tahmin edilecek parametre sayısı olan p ve örnek hacmi olan n değerlerinin bir fonksiyonu olup örnek hacmi arttıkça bu kritik değerler küçülür. Böylece ilgili tanı istatistiği, verinin geneline uymamasıyla birlikte EKK tahminlerini değiştirme eğilimi düşük olan bir gözlemi bile etkili gözlem olarak saptayabilir. Bu sorunu dikkate alan Belsley vd. (1980) *DFBETAS*, *DFFITS* ve *COVRATIO* tanı istatistiklerinin hacmi 100' den büyük örnekler için kullanışlı olmadığını belirtmişlerdir. Bu nedenle simülasyon çalışmasında örnek hacimleri 20, 30 ve 50 olarak alınmıştır. Bağımsız değişken sayısı da 2, 3 ve 4 olarak belirlenmiştir.

Bu bölümde yapılan simülasyon çalışmalarında veri üretmek için Hadi ve Simonoff (1993) tarafından yapılan çalışmadaki veri üretme yönteminden yararlanılmıştır. Hadi ve Simonoff (1993) tarafından yapılan çalışmada veri aykırı gözlem içermek amacıyla üretildiği için bu çalışmada aynı yöntemle veri etkili gözlem içerecek biçimde MATLAB2008a programı kullanılarak üretilmiştir.

Etkili bir gözlemin kaldıraç (leverage) değeri etkili olmayan gözlemlerinkine göre daha büyüktür. Bu tanımdan yararlanarak 1 indisi ile gösterilecek olan etkili gözleme ilişkin kaldıraç değerini büyütmek amacıyla bu gözlemin $p-1$ sayıda bağımsız değişkenlerin değerleri, verinin geneline uyan gözlemlere ilişkin bağımsız değişkenlerin değerlerinin türetildiği $[0,15]$ aralığının en uç değeri olan 15 olarak belirlenmiştir. Bu gözlemin hata (error) değişken değeri ise Hadi ve Simonoff (1993) tarafından yapılan çalışmadaki veri üretme yönteminin doğrultusunda $\varepsilon_1 = -5$ olarak alınmıştır. Buradaki amaç bu gözlemin hata değişken değerini mutlak değerce arttırarak, bu gözlemin artık değerini mutlak değerce büyültmektir. Sonra parametrelerinin değerleri 1 olan,

$$Y_1 = 1 + X_{11} + ... + X_{1\,p-1} + \varepsilon_1 \tag{14}$$

modeline göre etkili gözlemin bağımlı değişken değeri türetilmiştir. Simülasyon boyunca etkili gözleme ilişkin bağımsız değişkenlerin değerleri ve bağımlı değişken değerleri sabit kalmıştır.

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

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

<u>*Adım* 2</u>: Verinin geneline uyan $n-1$ sayıda gözlem için hata değişkenlerinin değerleri, $\varepsilon_i \sim N(0,1)$ dağılımından üretilir.

<u>*Adım* 3</u>: Verinin geneline uyan $n-1$ sayıda gözlem için $p-1$ sayıda bağımsız değişkenlerin değerleri ve hata değişkenlerinin değerleri kullanılarak ilgili modele yani

$$Y_i = \beta_0 + \beta_1 X_{i1} + ... + \beta_{p-1} X_{i\,p-1} + \varepsilon_i \qquad (i = 2,3,...,n) \tag{15}$$

modeline göre bağımlı değişken değerleri türetilir.

<u>*Adım* 4</u>: Türetilen verideki her bir gözlem için *DFBETAS* , *DFFITS* , *COVRATIO* , D_i ve S_i istatistiklerinin değerleri hesaplanır ve her bir tanı istatistiğine ilişkin karar kuralına göre 1 indisi ile gösterilen gözlemin etkili bir gözlem olup olmadığına karar verilir.

<u>*Adım* 5</u>: Aynı veri için grafik yöntemine göre tanı grafikleri oluşturulur. Eşitlik (12)' de verilen ve göreli duyarlılık faktörü olarak tanımlanan karar değişkenine göre etkili gözlemi belirleme gücü en yüksek olan grafik alınır. Bu grafikte etkili gözlem haricindeki diğer gözlemlere ilişkin noktaların merkeze uzaklıkları,

$$r = \left(w - w_0^{(j)}\right)^T \left[M^{(j)}\right]^{-1} \left(w - w_0^{(j)}\right) \tag{16}$$

ile hesaplanıp en büyük uzaklık deneysel güven elipsinin yarıçap uzunluğu olan ρ olarak alınır. Eğer 1 indisi ile gösterilen gözleme ilişkin noktanın elipsin merkezine uzaklığı, ρ değerinden büyük ise bu nokta deneysel güven elipsinin dışındadır ve grafik yöntemine göre bu gözlem etkili bir gözlemdir.

Her 1000 tekrarda verinin geneline uyan gözlemler için bağımsız değişkenlerin değerleri yeniden üretilmek üzere bu deneme 100000 kez tekrarlanmıştır. Sonra etkili gözlem içeren veride, bir etkili gözlemi saptamak için kullanılan bu tanı yöntemlerinin etkili bir gözlemi tespit etme oranları hesaplanmıştır.

3.1 Bağımsız Değişken Sayısı ve Örnek Hacminin Farklı Değerleri için Tanı Yöntemlerinin Karşılaştırılması

İki bağımsız değişkenin olduğu bir model, yani $p = 3$ durumu, için bölümün başında belirtilen adımlar doğrultusunda yapılan simülasyonla elde edilen sonuçlar Tablo 1' de verilmiştir.

Tablo 1. $p = 3$ iken bir etkili gözlem için tanı yöntemlerinin etkili bir gözlem içeren veride etkili gözlemi saptama oranları

$p = 3$			
	$n = 20$	$n = 30$	$n = 50$
$DFFITS_{(1)}$	0,9995	0,9997	0,9999
$DFBETAS_{1(1)}$	0,9994	0,9997	0,9998
$DFBETAS_{2(1)}$	0,9947	0,9994	0,9998
$DFBETAS_{3(1)}$	0,9948	0,9996	0,9997
$D_{(1)}$	0,8759	0,7744	0,3889
$COVRATIO_{(1)}$	0,9641	0,9955	0,9996
$S_{(1)}$	0,9614	0,9820	0,9900
Grafik Yöntemi	0,9440	0,9852	0,9963

Tablo 1' deki sonuçlar incelendiğinde, farklı örnek hacimleri altında $DFFITS_{(i)}$ ve $DFBETAS_{j(i)}$ istatistiklerinin etkili gözlemi saptama oranlarının büyük olduğu görülmektedir. Bununla birlikte, Cook Uzaklığı istatistiğinin etkili gözlemi saptama oranı örnek hacmi arttıkça düşmektedir. $COVRATIO_{(i)}$, $S_{(i)}$ istatistiklerinin ve grafik yönteminin etkili gözlemi saptama oranları ise örnek hacmi arttıkça büyümektedir.

Modelde üç bağımsız değişkenin olduğu durum yani $p = 4$ iken bir etkili gözlem için tanı istatistikleri ve grafik yöntemini karşılaştırmak amacıyla yapılan simülasyon bölümün başında belirtilen adımlar doğrultusunda oluşturulmuştur. Elde edilen sonuçlar Tablo 2' de verilmiştir.

Tablo 2. $p = 4$ iken bir etkili gözlem için tanı yöntemlerinin etkili bir gözlem içeren veride etkili gözlemi saptama oranları

$p = 4$			
	$n = 20$	$n = 30$	$n = 50$
$DFFITS_{(1)}$	0,9993	0,9995	0,9998
$DFBETAS_{1(1)}$	0,9994	0,9995	0,9998
$DFBETAS_{2(1)}$	0,9623	0,9922	0,9994
$DFBETAS_{3(1)}$	0,9647	0,9921	0,9994
$DFBETAS_{4(1)}$	0,9645	0,9926	0,9995
$D_{(1)}$	0,8431	0,7228	0,3668
$COVRATIO_{(1)}$	0,8978	0,9839	0,9995
$S_{(1)}$	0,9935	0,9988	0,9998
Grafik Yöntemi	0,8372	0,9486	0,9901

Yukarıdaki tabloda verilen sonuçlara göre örnek hacminin tüm durumları için $DFFITS_{(i)}$, $DFBETAS_{j(i)}$ ve $S_{(i)}$ istatistiklerinin etkili gözlemi saptama oranları yüksektir. Fakat örnek hacmi 20 iken $DFBETAS_{j(i)}$ istatistiğinin etkili gözlemi saptama oranı $DFFITS_{(i)}$ ve $S_{(i)}$ istatistiklerine ilişkin oranlara göre daha düşüktür. Ayrıca bu sonuçlara göre örnek hacmi arttıkça etkili gözlemi saptama oranı artan yöntemler $COVRATIO_{(i)}$ istatistiği ve grafik yöntemidir. Cook Uzaklığı istatistiği için sonuçlar incelendiğinde örnek hacmi arttıkça bu istatistiğe ilişkin etkili gözlemi saptama oranının düştüğü görülmektedir. Bir etkili gözlemi saptamak için kullanılan tanı yöntemlerine ilişkin son simülasyon bağımsız değişken sayısı 4 iken, bölümün başında belirtilen adımlar doğrultusunda yapılmıştır. Simülasyon sonuçları Tablo 3' te verilmiştir.

Tablo 3. $p = 5$ iken bir etkili gözlem için tanı yöntemlerinin etkili bir gözlem içeren veride etkili gözlemi saptama oranları

$p = 5$			
	$n = 20$	$n = 30$	$n = 50$
$DFFITS_{(1)}$	0,9994	0,9996	0,9998
$DFBETAS_{1(1)}$	0,9998	0,9995	0,9998
$DFBETAS_{2(1)}$	0,9221	0,9709	0,9959
$DFBETAS_{3(1)}$	0,9171	0,9678	0,9960
$DFBETAS_{4(1)}$	0,9206	0,9702	0,9961
$DFBETAS_{5(1)}$	0,9168	0,9714	0,9958
$D_{(1)}$	0,7831	0,6843	0,3390
$COVRATIO_{(1)}$	0,7926	0,9593	0,9979
$S_{(1)}$	0,9967	0,9998	0,9997
Grafik Yöntemi	0,7139	0,8851	0,9747

Tablo 3' te verilen sonuçlar incelendiğinde, $p=4$ durumunda olduğu gibi $DFFITS_{(i)}$, $DFBETAS_{j(i)}$ ve $S_{(i)}$ istatistiklerinin etkili olarak üretilen gözlemi saptama oranlarının ele alınan tüm örnek hacimlerinde büyük olduğu görülmektedir. Fakat örnek hacmi 20 iken $DFBETAS_{j(i)}$ istatistiğinin bu etkili gözlemi saptama oranı $DFFITS_{(i)}$ ve $S_{(i)}$ istatistiği için elde edilen oranlara göre daha düşüktür. $COVRATIO_{(i)}$ istatistiği ve grafik yöntemine ilişkin etkili gözlemi saptama oranları incelendiğinde $n=20$ iken bu yöntemlere ilişkin etkili gözlemi saptama oranlarının küçük olmasıyla birlikte örnek hacmi arttıkça bu yöntemlere ilişkin oranların arttığı görülmektedir. Son olarak örnek hacmi arttıkça Cook Uzaklığı istatistiğinin etkili gözlem olan ve 1 indisi ile gösterilen gözlemi saptama oranının düştüğü gözlenmektedir.

Örnek hacmi artarken $DFFITS_{(i)}$, $DFBETAS_{j(i)}$ ve $COVRATIO_{(i)}$ istatistiklerinin etkili bir gözlemi saptama oranlarının artmasının nedeni bu istatistiklere ilişkin karar kuralında yer alan kritik değerlerin örnek hacminin azalan bir fonksiyonu olmasıdır. Özellikle bu artış $COVRATIO_{(i)}$ istatistiği için daha açık bir şekilde görülmektedir.

Bağımsız değişken sayısının tüm durumlarında örnek hacmi arttıkça Cook Uzaklığı istatistiğinin etkili gözlemi saptama oranı düşmektedir. Çünkü gözlem silme tekniğine dayalı olarak etkili bir gözlem için elde edilen $\hat{\beta}_{(i)}$ tahmini, örnek hacmi arttıkça $\hat{\beta}_{(i)}$ tahminleri için oluşturulan ve bir elipsoide karşılık gelen güven bölgesine yaklaşır. Bu durumda etkili gözlem için elde edilen $\hat{\beta}_{(i)}$ tahmininin güven elipsoidinin merkezine uzaklığına karşılık gelen, bu gözleme ilişkin Cook Uzaklığı değeri küçülür. Böylece Cook Uzaklığı yöntemiyle etkili gözlemin saptanma oranı düşer.

Pena' nın $S_{(i)}$ istatistiğine ilişkin sonuçlar incelendiğinde bağımsız değişken sayısının ve örnek hacminin tüm durumlarında bu istatistiğe ilişkin saptama oranlarının büyük olduğu gözlenmektedir. Bunun nedeni $S_{(i)}$ istatistiğine ilişkin karar kuralından kaynaklanmaktadır. Bu istatistiğe ilişkin karar kuralı,

$$\left|S_{(i)}-\text{med}\left(S\right)\right| \geq 4.5\times \text{MAD}\left(S_{(i)}\right) \tag{17}$$

ile verilir. Burada $\text{MAD}\left(S_{(i)}\right)$ ifadesi $S_{(i)}$ değerlerinin medyandan sapmalarının mutlak değerlerinin medyanıdır. Yukarıdaki karar kuralına bağlı olarak etkili gözlem için hesaplanan $S_{(i)}$ değerinin medyandan mutlak değer bakımından sapması, verinin geneline uyan gözlemlerinkine göre büyüktür. Bağımsız değişken sayısının tüm durumlarında örnek hacmi artarken $\text{med}\left(S\right)$ ve $\text{MAD}\left(S_{(i)}\right)$ değerlerinde büyük bir değişim olmamaktadır. Böylece örnek hacmi arttıkça etkili gözlem için hesaplanan $S_{(i)}$ değerinin medyandan mutlak değerce sapması büyük bir değişim göstermediği için bu istatistiğin etkili gözlemi saptama oranı yüksek olacaktır.

Grafik yöntemine ilişkin sonuçlar incelendiğinde bağımsız değişken sayısının tüm durumlarında örnek hacmi arttıkça bu yönteme ilişkin etkili gözlemi saptama oranının arttığı görülmektedir. Eğer veride etkili gözlem yoksa tüm grafiklerde gözlemlere ilişkin noktalar geniş bir yayılım gösterecektir. Fakat veride en az bir etkili gözlem varsa göreli duyarlılık faktörüne bağlı olarak belirlenen grafikte verinin geneline uyan noktalar grafiğin merkezi olan $\left(0,0\right)$ noktası etrafında kümelenecektir ve etkili gözlem ilişkin nokta da verinin geneline uyan noktaların oluşturduğu bu kümeden uzak bir konumda bulunacaktır. Veride en az bir etkili gözlem varken bağımsız değişken sayısının tüm durumları için örnek hacmi arttıkça verinin geneline uyan gözlemler grafiğin merkezine daha da yaklaşmakta ve buna bağlı

olarak güven elipsleriyle belirlenen güven bölgesi daha da daralmaktadır. Böylece grafik yönteminin bir etkili gözlemi saptama oranı örnek hacmi arttıkça artmaktadır.

Örnek hacmi sabit tutulduğunda ele alınan tanı yöntemlerinin Tablo 1, Tablo 2 ve Tablo 3' teki sonuçlara göre incelenmesi bu tanı yöntemlerinin bağımsız değişken sayısındaki artışa göre değerlendirilmesi açısından önemlidir. Örnek hacmi sabit iken $DFFITS_{(i)}$ ve $S_{(i)}$ tanı istatistiklerinin etkili gözlemi saptama oranlarının büyük değişimler göstermediği açıkça görülmektedir. Bununla birlikte bağımsız değişken sayısı arttıkça $DFBETAS_{j(i)}$, $COVRATIO_{(i)}$, Cook Uzaklığı istatistikleri ve grafik yöntemine ilişkin etkili gözlemi saptama oranları azalmaktadır. Özellikle bu azalma grafik yöntemi için diğer yöntemlerinkine nazaran büyüktür. Bu da regresyon probleminin boyutu arttıkça grafik yönteminin etkili gözlemi saptama duyarlılığının azaldığını göstermektedir.

3.2 Bağımsız Değişken Sayısı ve Örnek Hacmi Sabitken Tanı Yöntemlerinin Karşılaştırılması

Bu simülasyonda, bağımsız değişken sayısı ve örnek hacmi sabit iken etkili gözlemin hata değişken değeri mutlak değerce arttırılarak, bunun verinin merkezinden uzaklaştırıldığı durumlarda, tanı yöntemlerinin etkili gözlem içeren veride etkili gözlemi saptama oranları elde edilmiştir. Buradaki amaç etkili gözleme ilişkin hata değişken değerinin daha da arttığı durumlarda tanı yöntemlerini karşılaştırmaktır.

Bir etkili gözlemi saptamak için kullanılan tanı yöntemlerine ilişkin simülasyon örnek hacmi 20 ve bağımsız değişken sayısı 2 iken bölümün başında belirtilen adımlar doğrultusunda yapılmıştır. Bu simülasyonda önce etkili gözlem olarak belirlenen ve 1 indisi ile gösterilen gözleminin hata değişken değeri $\varepsilon_1 = -5$ alınmıştır. Döngü,

$$Y_i = \beta_0 + \beta_1 X_{i1} + ... + \beta_{p-1} X_{i\,p-1} + \varepsilon_i \qquad (i = 2,3,...,n) \tag{18}$$

modeli dikkate alınarak bölümün başında verilen adımlar doğrultusunda, her $1\,000$ tekrara da bir verinin geneline uyan gözlemlerin bağımsız değişkenlerinin değerleri yeniden üretilmek üzere $100\,000$ kez tekrar edilmiştir ve tanı yöntemlerinin etkili olan bu gözlemi saptama oranları elde edilmiştir. Sonra etkili gözlemin hata değişken değeri önce $\varepsilon_1 = -7,5$ sonra da $\varepsilon_1 = -10$ alınıp bu değerler için aynı simülasyon tekrarlanmıştır. Sonuçlar Tablo 4' te verilmiştir.

Tablo 4. $n = 20$ ve $p = 3$ iken bir etkili gözlem için tanı yöntemlerinin farklı hata değişken değerlerine göre üretilen etkili gözlemi saptama oranları

$n = 20$, $p = 3$			
	$\varepsilon_1 = -5$	$\varepsilon_1 = -7,5$	$\varepsilon_1 = -10$
$DFFITS_{(1)}$	0,9995	1	1
$DFBETAS_{1(1)}$	0,9994	1	1
$DFBETAS_{2(1)}$	0,9947	0,9985	0,9989
$DFBETAS_{3(1)}$	0,9948	0,9985	0,9990
$D_{(1)}$	0,8759	0,9958	0,9997
$COVRATIO_{(1)}$	0,9641	0,9996	1
$S_{(1)}$	0,9614	0,9886	0,9932
Grafik Yöntemi	0,9440	0,9943	0,9989

Bağımsız değişken sayısı ve örnek hacmi sabitken, veride kaldıraç değeri diğer gözlemlerinkine göre büyük olan herhangi bir gözlemin hata değişken değeri arttıkça, bu gözlemin regresyon tahminlerini değiştirme eğiliminin artması beklenir. Tablo 4' teki sonuçlar da bunu doğrulamaktadır. Çünkü etkili gözlemin hata değişkeni değeri arttıkça etkili bir gözlemin saptanması için kullanılan bu altı tanı yönteminin verideki etkili gözlemi saptama oranları artmaktadır. Özellikle bu artış Cook Uzaklığı istatistiği için daha açık bir şekilde görülmektedir.

4. SONUÇ

Sonuç olarak bir etkili gözlemin saptanmak için kullanılan yöntemlerden verinin geneline uymayan bu gözlemi saptama bakımından en duyarlı tanı yöntemleri $DFFITS_{(i)}$ ve Pena' nın $S_{(i)}$ istatistikleridir. Bu iki yöntemden sonra $DFBETAS_{(i)}$ ve $COVRATIO_{(i)}$ yöntemlerinin de etkili gözlemi saptama bakımından duyarlılığı yüksektir. Cook Uzaklığı istatistiği ise örnek hacmi arttıkça etkili gözlemi daha düşük bir oranla saptamaktadır. Son olarak yüksek boyutlu bir regresyon probleminin iki boyutlu bir probleme indirgenmesi amacıyla veriye bir dönüşüm uygulandığı grafik yöntemi ise reg-resyon probleminin boyutu arttıkça etkili gözlemi saptama bakımından duyarlılığı azalmaktadır.

KAYNAKLAR

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

Andrews, D.F. ve Pregibon, D. (1976). "Finding Outliers That Matter", *J. Roy. Statist. Soc., Ser. B.* 40, 85-93.

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

Chatterjee, S. ve 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. ve 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.", *Willey*, New York, 110-150.

Hadi, A.S. ve 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. ve Welsch, R.E. (1978). "The Hat Matrix in Regression and ANOVA", *The American Statistician* 32(1), 17-22.

Li, B., Martin, E.B. ve 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.. ve Vining, G.G. (2001). "Introduction to Linear Regression Analysis", *Willey 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.

Vupa, Ö. (2009). "Investigation of Influence Observation and Outliers in Logistic Regression Model", *VI. İstatistik Günleri Sempozyumu Bildiriler Kitabı* 453-457.