T.C.
ONDOKUZ MAYIS ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

GRAFİK YÖNTEMLERLE ETKİN GÖZLEMLERİN VE AYKIRI DEĞERLERİN TESPİTİ

YEŞİM AYDIN

YÜKSEK LİSANS TEZİ
İSTATİSTİK ANABİLİMİ DALI

DANIŞMAN
Yrd. Doç. Dr. V. REZAN USLU

SAMSUN - 2006

GRAFİK YÖNTEMLERLE ETKİN GÖZLEMLERİN VE AYKIRI DEĞERLERİN TESPİTİ

ÖZET

Regresyon analizlerinde güvenilir sonuçlara ulaşabilmek tamamiyle model varsayımlarının sağlanmasına bağlıdır. Özellikle aykırı değer, ardışık bağımlılık ve değişen varyans gibi sorunların mevcudiyeti elde edilen sonuçlar üzerinde şüphe uyandırmayı gerektirir.

Toplanan veri için uygun bir regresyon modeli öngörüldükten sonra başlıca adım sonuçları çok kuvvetli etkileyen aykırı değer ve çekim gücü yüksek gözlemleri tespit etmektir. Veri kümesi içinde bir aykırı değerin varlığı bile verilerin çoğunun vermek istediği bilgiye engel olmakta ve En Küçük Kareler (EKK) tahminleri üzerinde yanıltıcı bir etki yaratmaktadır. Bu yüzden bu çalışmada etkili gözlemlerin belirlenmesi için regresyon sonuçlarına farklı taraftan bakılmış, artıklara (klasik, normalleştirilmiş, standartlaştırılmış, jackknife ve silinmiş) ve İz Düşüm Matrisi H' nin köşegen elemanlarına dayalı tanısal grafikler incelenmiştir.

Anahtar Kelimeler: Aykırı değer, çekim gücü yüksek gözlem, tanısal grafikler, tanısal ölçüler.

IDENTIFYING OF INFLUENTIAL OBSERVATIONS AND OUTLIERS WITH DIAGNOSTIC PLOTS

ABSTRACT

To obtain the reliable results from a regression analysis depends on whether the model assumptions are violated or not. The presence of problems in the data such as outliers, heterescadasticity and serial correlations is needed for us to have a doubt on the regression results.

The main step after assuming that a model for the colected data set is appropriate, is to identify the high leverage points and outliers which have a significant impact on the results of Standard methodology. The presence of even a single outlier in data set can hide the information which can be obtained from the rest of the data and have an misleading effect on Least Squares estimates. Therefore we lok the regression results from the another points of view and diagnostic plots based on residuals (ordinary, normalized, standardized, jackknife and deleted) and the diagonal elements of the projection matrix are discussed in this study.

Key Words : Outliers, High Leverage Observations, Diagnostic Plots, Diagnostic Measure

İÇİNDEKİLER

SİMGE VE KISALTMALAR

ε_i	Hata terimi
e_i	Artıklar
$\hat{\beta}$	EKK tahmin edicisi
H	İz düşüm matrisi
h_{ii}	İz düşüm matrisinin köşegen elemanları
$e_{S,i}$	Standartlaştırılmış artık
$e_{N,i}$	Normalleştirilmiş artık
$e_{J,i}$	Jackknife artık
$e_{P,i}$	Silinmiş artık
EKK	En Küçük Kareler
AKT	Artık Kareler Toplamı
AKO	Artık Kareler Ortalaması

TABLOLARIN LİSTESİ

Sayfa

ŞEKİLLERİN LİSTESİ

Sayfa

EKLERİN LİSTESİ

1. GİRİŞ

Regresyon analizi, aralarında sebep-sonuç ilişkisi bulunan iki veya daha fazla değişken arasındaki ilişkiyi, ilgilenilen kitleye yönelik tahminler ya da kestirimler yapabilmek amacıyla regresyon modeli olarak adlandırılan matematiksel bir model ile karakterize edilebilen istatistiksel analiz tekniğidir. Doğrusal regresyon analizi, hatalara ilişkin standart varsayımlar dediğimiz varsayımların sağlanması durumunda En Küçük Kareler (EKK) tahmin yöntemleri ile en iyi yansız tahminleri verir. Ancak veride aykırı değerlerin bulunması durumunda bu varsayımlarda bozulmalar meydana gelebilir. Örneğin; hata dağılımının normalliği bozulabilir; değişen varyans problemi ortaya çıkabilir ve bu yüzden tahminler yanlı ve büyük varyanslı olabilirler (Rousseeuw ve Zomeron, 1990). Demek ki güvenilir tahminlere ulaşmak için veri içerisinde bu tür problemlere sebep olabilecek gözlemlerin tespit edilmesi gerekmektedir. Üzerinde çalışılan probleme yönelik veri içerisinde bu tür gözlemler tespit edildikten sonra ya veri setinden atılmaları önerilebilir ya da bunların etkilerinin azaltılabileceği diğer bazı yöntemlere başvurulabilir ki bu tezin kapsamı dışında bırakılmıştır.

Literatürde veri seti içerisinde sadece bir tane aykırı değer bulunması durumunda bunları tespit etmek için güvenilir ve kolay uygulanabilir teknikler mevcuttur. Fakat birden çok aykırı değer mevcut olması durumunda aykırı değerler bazen birbirlerinin varlıklarını gizleyebilmekte ve hatta bu aykırı değerler klasik tahmin yöntemlerinde herhangi bir sorun teşkil etmeyen gözlemlerin bile aykırı değerler olarak görünmesine neden olabilmektedir (Hadi ve Simonoff, 1993).

Aykırı değerlerin varlığını teşhis etmede çeşitli tanısal yöntemler geliştirilmiş. Bunların içerisinde standartlaştırılmış ve jackknife artıklara dayalı tanısal yöntemler aykırı değerlerin tespiti için sıkça başvurulan yöntemlerdir (Cook ve Weisberg, 1982). Çekim gücü yüksek gözlemlerin belirlenmesi için Dodge ve Hadi (1999) tarafından iz düşüm matrisi H'nin köşegen ve köşegen üzerinde olmayan elemanları için üst ve alt sınırları esas alan basit bir yöntem geliştirilmiştir. Regresyon tanısal araçları içerisinde yer alan grafiksel tanı yöntemleri, analizciye kolay ve daha hızlı bir şekilde, doğru model belirlemede ve/veya veri içerisinde

olabilecek sorunları ortaya çıkarmada fayda sağladıklarından etkili gözlemlerin tespitinde de oldukça önemli rol oynarlar.

Bu yüzden bu çalışmanın amacı, etkili gözlemlerin belirlenmesinde literatürdeki tanısal yöntemleri tanıtmak ve bunlara ilaveten artıklara ve iz düşüm matrisinin köşegen üzerindeki elemanlarına dayalı tanısal grafik yöntemlerini araştırmaktır.

İkinci bölümde basit ve çoklu doğrusal regresyon modeli, EKK varyasımları ve bunların kontrolü, parametre tahmini ve bu tahmin edicilerin özellikleri daha sonra regresyon uyumunun değerlendirilmesi anlatılmıştır. Üçüncü bölümde etkili gözlemlerin incelenmesi, dördüncü bölümde etkili gözlemlerin belirlenmesi için tanısal yöntemler anlatılmış, artıklara ve iz düşüm matrisinin köşegen elemanlarına dayalı tanısal grafikler verilmiştir. Beşinci bölümde etkili gözlemlerin tespiti için kullanılan tanısal istatistiklerin yanı sıra artıklara ve iz düşüm matrisinin köşegen elemanlarına dayalı tanısal grafikler ile bu noktaların daha hızlı ve kolay tespit edildiğini ortaya koymak amacıyla bir uygulama verilmiştir.

2. DOĞRUSAL REGRESYON MODELİ

2.1. Basit Doğrusal Regresyon Modeli

Basit doğrusal regresyon modeli sadece bir bağımsız değişken içeren modeldir ve aşağıdaki gibi ifade edilir.

$$Y = \beta_0 + \beta_1 X + \varepsilon \qquad (2.1)$$

Bu modelde β_0 kesişimi, β_1 eğimi temsil eden bilinmeyen regresyon katsayılarıdır. ε terimi modelin rasgele hata bileşenini temsil eder. Hataların sıfır ortalamaya ve bilinmeyen σ^2 varyansına sahip olduğu varsayılır.

2.2. Çoklu Doğrusal Regresyon Modeli

Bir bağımlı değişkenin kendisini etkileyen birden çok bağımsız değişken ile olan ilişkisi çoklu regresyon analizi başlığı altında incelenmektedir.

O halde, p bağımsız değişkenli bir çoklu doğrusal regresyon modeli

$$Y_i = \beta_0 + \beta_1 X_{1i} + ... + \beta_p X_{pi} + \varepsilon_i \qquad i = 1, 2, ..., n \qquad (2.2)$$

şeklinde tanımlanır. Burada n: gözlem sayısını, p: değişken sayısını, $X_{1i}, ..., X_{pi}$: açıklayıcı değişkenleri, Y_i: bağımlı değişkene ait i. nci gözlemleri, $\beta_j : j = 0, 1, ..., p$ j. nci açıklayıcı değişkenin bilinmeyen regresyon katsayısı olan parametreyi, ε_i : hata terimlerini temsil eder.

Parametreler modele bağımsız değişkenlerin veya bağımsız değişkenlerin fonksiyonlarının birer basit katsayıları olarak girerler.

(2.2) eşitliğinin matris notasyonu ile ifadesi

$$Y = X\beta + \varepsilon \qquad (2.3)$$

olarak yazılabilir. Burada $Y : n \times 1$ boyutlu bağımlı değişkenlerin vektörü, $X : n \times p$ boyutlu bağımsız değişkenler matrisi $(n \rangle m)$ ve $m = p + 1$, $\beta : m \times 1$ boyutlu bilinmeyen regresyon parametreleri vektörü ve $\varepsilon : n \times 1$ boyutlu hata vektörüdür.

Hata vektörünün, ortalaması sıfır $\left(E(\varepsilon)=0\right)$ ve varyansı $\left(Var(\varepsilon)=\sigma^2 I\right)$ olan ve birbirinden bağımsız rasgele değişkenler vektörü olduğu varsayılır.

Bağımlı ve bağımsız değişkenler arasındaki ilişkiyi gösteren parametrelerin tahmininde en iyi sonuçlara ulaşabilmek için sağlanması gereken bazı varsayımlar vardır.

2.3. En Küçük Kareler Varsayımları

1. Hata terimi, ortalaması sıfır, varyansı σ^2 olan normal dağılıma sahiptir (Normallik ve Sabit varyans varsayımı).

$$E(\varepsilon_i)=0 \qquad Var(\varepsilon_i)=\sigma^2 \qquad \varepsilon_i \sim N(0,\sigma^2)$$

2. Hata terimleri birbirleriyle ilişkisizdir (Hataların ilişkisizlik varsayımı).

$$Cov(\varepsilon_i,\varepsilon_j)=0 \qquad i \neq j$$

3. Bağımsız değişkenler arasında güçlü veya tam bir ilişki yoktur (Çoklu bağlantı varsayımı).

2.4. En Küçük Kareler Varsayımlarının Kontrolü

Regresyon analizinde hata teriminin bağımsız, ortalaması sıfır, varyansı sabit olan normal dağılım gösterdiği varsayılır (Chatterjee ve Hadi, 1988). Uydurulan model için bu varsayımlar tutmadığı takdirde o model ile ilgili her türlü yorum şüphe ile karşılanır ve tartışmaya açıktır. Çünkü tahminler minimum varyanslı olma özelliğini yitirirler. Bu nedenle model kullanılmadan önce varsayımlar kontrol edilmelidir.

2.4.1. Normallik Varsayımı

Normal dağılım varsayımı özellikle hipotez testlerinin yapılabilmesi ve güven aralıklarının oluşturulabilmesi için oldukça önemlidir. Bu varsayımın tutup tutmadığının belirlenmesi için en pratik yöntemlerden birisi normal olasılık grafikleridir (Kleinbaum ve Kupper, 1978; Draper ve Smith, 1981; Montgomery ve Peck; 1992). Grafik çizildiğinde noktalar bir doğru üzerinde olmalıdır. Elde edilen grafikteki doğrudan sapmalar, ilgili gözlemlerin normallikten sapmalarını ifade

eder. Normalden sapan gözlemlerin olması uygun olmayan bir regresyon modeli kullanıldığını veya varyansın homojen olmadığını ifade eder. Grafiksel yöntemler yanında normallik için Kolmogorov-Smirnov, Shapiro-Wilks istatistiği, çarpıklık ve basıklık testleri gibi analitik testler kullanılarak kapsamlı bir incelemeden sonra artıkların normal dağılım gösterip göstermediği test edilebilir (Kleinbaum ve Kupper, 1978; Shimek, 1999). Sonuçta artıkların normal dağılım göstermediğine karar verilirse bu durumda EKK tekniği uygulanamaz.

2.4.2. Sabit Varyans Varsayımı

Sabit varyans varsayımı, regresyon analizinde temel varsayımlardan birisidir. Bu nedenle, artıkların sabit varyansa sahip olup olmadıklarının belirlenmesi ve doğrulanması önemlidir. Eğer bu problem çözülemezse, EKK tahmin edicileri yine sapmasızdırlar fakat minimum varyans özelliğine sahip olmayabilirler. Diğer bir ifade ile katsayılar gerekenden büyük standart hatalara sahip olacaklardır.

Problemin varlığının incelenmesi için en çok kullanılan yöntem artıkların, tahmin değerlerine $\left(\hat{Y}\right)$ karşı grafiklerinin oluşturulmasıdır. Bu problemin çözümü için önerilen en etkili yöntem Y değişkeninde bir dönüşüm uygulamaktır (Montgomery ve Peck, 1992).

2.4.3. Hataların İlişkisizlik Varsayımı

EKK regresyon analizinde hata teriminin ortalaması sıfır, varyansı sabit olduğu ve birbirleriyle ilişkisiz olduğu varsayılır. Bağımsızlık varsayımı çoğu zaman mevcut olmayabilir. Özellikle zaman serisi verilerinde ardışık bağımlılık mevcuttur.

Hatanın ardışık bağımlı olması, EKK regresyonu üzerinde bazı etkilere sahiptir. Regresyon katsayıları hala yansızdırlar fakat minimum varyansa sahip değildir.

Hata teriminin ardışık bağımlı olup olmadığının araştırılması için en basit yöntem artık değerlerinin zamana karşı grafiğinin oluşturulmasıdır. Diğer bir yöntem ise Durbin Watson istatistiğidir (Draper ve Smith, 1981).

2.4.4. Çoklu Bağlantı

Çoklu bağlantı terimi, X matrisinin kolonlarının doğrusal bağımlılığı şeklinde tanımlanabilir. Bağımsız değişkenler arasında çoklu bağlantı problemi,

1. Veri toplama yönteminin yanlış olması

2. Örneklem veya modelde yapılan kısıtlamalar

3. Model seçiminde yapılan hatalar

gibi nedenlerden kaynaklanabilir. Verideki bu problem, regresyon katsayılarına ait EKK tahmin edicilerinin varyans ve kovaryans değerlerinin gerçekte olduğundan daha büyük olmasına, buna bağlı olarak da regresyon modeline dayanan yorumların hatalı olmasına neden olabilir (Montgomery ve Peck, 1992).

Bağımsız değişkenler arasında çoklu bağlantı problemi olup olmadığının belirlenmesi için birçok yöntem önerilmektedir. Bu problem aslında bir veri problemi olduğundan tespiti için önerilen yöntemlerin çoğunluğu X bağımsız değişken matrisi ile ilgilidir. Bunlar arasında en çok kullanılan istatistikler Varyans Şişme Faktörü $\left(VIF_j\right)$, Durum İndeksi $\left(\eta_j\right)$, Durum Sayısı $\left(\kappa_j\right)$ istatistikleridir.

Bağımsız değişkenler arasında tespit edilen çoklu bağlantı probleminin çözümü için modeli yeniden tanımlamak, konu ile ilgili ek veriler toplamak ve EKK yöntemi yerine Ridge regreyon yöntemini uygulamak önerilmektedir.

2.5. Model Parametrelerinin Tahmini

Doğrusal regresyonda bilinmeyen parametrelerin tahminlerini yapmak için kullanılan en yaygın yöntem EKK yöntemidir.

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

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} \qquad X = \begin{bmatrix} 1 & X_{11} & \cdots & X_{p1} \\ 1 & X_{12} & \cdots & X_{p2} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{1n} & \cdots & X_{pn} \end{bmatrix} \qquad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} \qquad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

şeklinde tanımlanır. EKK varsayımları sağlandığı takdirde parametre tahmin edicileri, hata kareler toplamının minimizasyonu ile bulunur. EKK fonksiyonu

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

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

$(\beta'X'Y)' = Y'X\beta$ olduğu için

$$s(\beta) = \varepsilon'\varepsilon = Y'Y - 2\beta'X'Y + \beta'X'X\beta$$

şeklinde bulunur. β' ya göre minimizasyon, β' ya göre türevinin alınıp, sıfıra eşitlenmesi olduğu için

$$\left.\frac{\partial s(\beta)}{\partial \beta}\right|_{\hat{\beta}} = -2X'Y + 2(X'X)\hat{\beta} = 0$$

$$X'X\hat{\beta} = X'Y \qquad (2.4)$$

normal denklemleri bulunur. Bu denklem sisteminin çözümü için $(X'X)^{-1}$ matrisinin mevcut olması gereklidir. Bunun içinse şart, açıklayıcı değişkenlerin lineer bağımsız olması yani X açıklayıcı değişkenler matrisinin sütunları arasında biri diğerinin doğrusal kombinasyonu olmamalıdır (Montgomery ve Peck, 1992).

(2.3) denkleminin çözümünden β parametresinin EKK tahmin edicisi;

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

$$\hat{\beta} = (X'X)^{-1}X'Y \qquad (2.5)$$

şeklinde elde edilir. Elde edilen bu tahmin ediciler, regresyon parametrelerinin en iyi, doğrusal, yansız tahminleridir.

2.6. En Küçük Kareler Tahmin Edicilerinin Özellikleri

1. $\hat{\beta}$, regresyon parametresi β' nın yansız tahmin edicisidir.

$$E(\hat{\beta}) = \beta \qquad (2.6)$$

2. Diğer tüm doğrusal yansız tahminler arasında $\hat{\beta}$ minimum varyansa sahiptir.

3. EKK tahmin metodu kullanılarak elde edilen Y bağımlı değişkeninin tahmin değerleri vektörü;

$$\hat{Y} = X\hat{\beta} \qquad (2.7)$$

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

$$\hat{Y} = HY \qquad (2.8)$$

şeklinde elde edilir. Burada $X(X'X)^{-1}X' = H$ iz düşüm matrisi olarak tanımlanır. $\hat{Y}$ tahmin vektörü, Y vektörünün doğrusal bir kombinasyonudur.

$$E(\hat{Y}) = X\beta \qquad Var(\hat{Y}) = \sigma^2 H$$

4. Artık vektörü e, Y'nin doğrusal bir dönüşümüdür.

$$e = Y - \hat{Y} \qquad (2.9)$$

$$e = Y - X(X'X)^{-1}X'Y$$

$$e = [I - H]Y \qquad (2.10)$$

elde edilir. Burada I birim matristir.

$$E(e) = 0 \qquad Var(e) = \sigma^2[I - H]$$

5. Artık kareler ortalaması σ^2'nin EKK tahmin edicisi olup

$$AKO = \hat{\sigma}^2 = \frac{e'e}{n - m} \qquad (2.11)$$

eşitliği ile elde edilir ve σ^2'nin yansız bir tahmin edicisidir.

2.7. Regresyon Uyumunun Değerlendirilmesi

EKK tekniği uygulanarak uydurulan regresyonun istatistiksel olarak önemli olup olmadığını test etmek amacıyla varyans analizi tekniği, modeldeki katsayıların ayrı ayrı önem kontrolleri için ise t testi kullanılır. Ancak bu yöntemler tek başlarına uydurulan modelin değişkenler arasındaki ilişkiyi ortaya en iyi model olduğunun veya verileri gerçekten temsil eden bir model olduğunu göstermez. Çünkü, model, en iyi model olmasa bile regresyon ve katsayılar önemli olabilirler.

Belirtme Katsayısı (R^2): Kullanılan X değişkenlerinin Y'deki toplam değişimi açıklayabilme oranını verir ve $0 < R^2 < 1$ dir.

Bu katsayı

$$R^2 = RKT/GKT \qquad (2.12)$$

formülü ile hesaplanmaktadır. Burada RKT: varyans analiz tablosundaki regresyon kareler toplamını, GKT: genel kareler toplamını göstermektedir. R^2'nin büyük çıkması her zaman modelin iyi olduğu sonucunu göstermez. Çünkü, modele konu ile ilgili veya ilgisiz bir değişkenin eklenmesi R^2'nin değerini artıracaktır. Dolayısıyla da büyük R^2'si olan modeller her zaman tahmin yapmada en iyi model olmayabilir (Montgomery ve Peck, 1992).

Modelin uyum iyiliğini ölçmede son yıllarda yaygın olarak kullanılmaya başlayan bir istatistik MEP (silinmiş artık kareler ortalaması) aşağıdaki gibi tanımlanır.

$$MEP = \frac{\sum_{i=1}^{n}\left(y_i - x_i'\hat{\beta}_{(i)}\right)^2}{n} \qquad (2.13)$$

Burada; $\hat{\beta}_{(i)}$: i. nci gözlem çıkarıldığında geriye kalan diğer gözlemler ile hesaplanan regresyon parametrelerinin tahmini, x_i': X matrisinin i. nci satırıdır. Silinmiş artıklar

$$e_{P,i} = y_i - x_i'\hat{\beta}_{(i)}$$

$$e_{P,i} = \frac{e_i}{1-h_{ii}} \qquad (2.14)$$

olarak tanımlanır. Burada $h_{ii} = x_i\left(X'X\right)^{-1}x_i'$, H iz düşüm matrisinin köşegen elemanlarıdır. Bu silinmiş artıkları kullanarak MEP istatistiğinin diğer bir ifadesi

$$MEP = \sum_{i=1}^{n}\frac{e_i^2}{\left(1-h_{ii}\right)^2.n} \qquad (2.15)$$

şeklinde yazılabilir. Büyük örneklemler için h_{ii} elemanları sıfıra yaklaşacağından

$$MEP = \frac{\sum_{i=1}^{n}e_i^2}{n} = \frac{AKT}{n} \qquad (2.16)$$

olur.

MEP istatistiği ile hesaplanan yeni belirtme katsayısı R_p^2 diye ifade olunup

$$R_p^2 = 1 - \frac{n.MEP}{GKT} \qquad (2.17)$$

ile hesaplanır.

Karşılaştırılan modeller arasında, MEP değeri en küçük, R_p^2 değeri de en büyük değere sahip model, en uygun model olarak tespit edilebilir (Meloun ve Militky, 2001).

3. ETKİLİ GÖZLEMLERİN İNCELENMESİ

Regresyon analizinde gözlemler,

1. Normal gözlem

2. Aykırı değer

3. Çekim gücü yüksek gözlem

şeklinde incelenebilir.

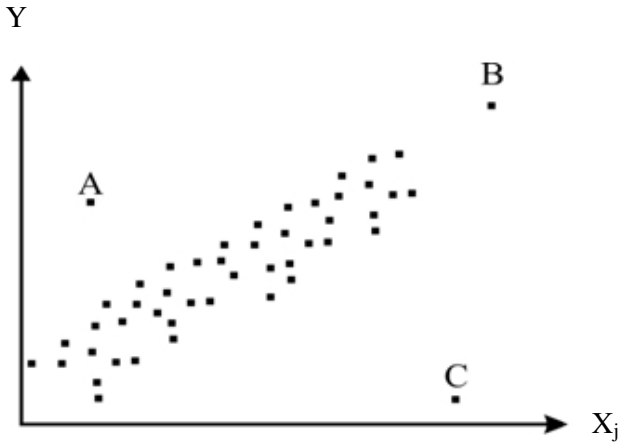
Normal Gözlem: Tüm parametrelere eşit etki yaptığı düşünülen gözlemlerdir. Bu nedenle bir gözlemin normal gözlemler grubuna girmesinden çok, diğer gruplardan birisinde bulunması daha önemli sonuçlar doğurabilir.

Aykırı Değer: Artıklar içerisinde mutlak değerce diğerlerinden oldukça büyük değere sahip veya artık ortalamasından 3 ya da 4 standart sapma kadar büyük değere aykırı değer denilmektedir. Model yapısının yanlış olmasından, ölçüm, kaydetme, örnekleme hatalarından kaynaklanabilir.

Aykırı değerler genellikle bağımlı değişken değerleri arasında diğerlerine göre farklı davranış gösteren gözlem değerleridir. Bu tip gözlemler EKK tahminlerini kendilerine doğru çekerler ve yanıltıcı sonuçların elde edilmesine neden olur.

Aykırı değerlerin belirlenmesindeki amaç; modelin yeterli olup-olmadığını, varsayımlardan sapmaları ve veride bir düzeltmeye gerek olup-olmadığını belirlemektir. Aykırı değerlerin tespiti artıkların incelenmesiyle yapılır.

Çekim Gücü Yüksek Gözlem: Bu tip gözlemler genellikle X açıklayıcı değişkenlerinin yapısındaki farklılıktan kaynaklanır. Çekim gücü yüksek gözlemlerin belirlenmesindeki amaç; X değişkeninin merkezinden uzak noktaları belirlemek. Bu gözlemleri belirlemek için kullanılan istatistik $H = X(X'X)^{-1}X'$ iz düşüm matrisinin köşegen elemanları h_{ii}' dir.



Şekil 3.1. Bağımlı ve bağımsız değişken arasındaki serpme grafiği

Şekil 3.1' de verilen A grafikteki verilere EKK regresyon doğrusu uydurulduğunda noktasının artık değeri çok büyük çıkacağından aykırı değer ve analizden çıkarıldığında uydurulan doğrunun Y eksenini kestiği nokta değişeceğinden etkili gözlem olabileceği, ancak açıklayıcı değişkenlerin merkezine yakın bir yerde olduğundan çekim gücü yüksek bir gözlem olmadığı düşünülür. Aynı zamanda A noktasının analizde tutulması hata ve katsayıların varyans tahminlerini değiştirecektir (Catterjee ve Hadi, 1988).

B noktası ise küçük artık değerine sahiptir, fakat çekim gücü yüksek bir noktadır. Çünkü açıklayıcı değişkenlerin merkezinden oldukça uzaktadır. B noktası aynı zamanda Y' nin merkezinden de uzaktadır. Bu durumda regresyon katsayılarının tahminlerine dolayısıyla uydurulan regresyon doğrusuna önemli etkisi olmayacaktır fakat noktalar arasındaki varyansı arttıracağından regresyon katsayılarının standart hatalarında etkili olabilir.

Şekil 3.1' de verilen C noktası ise hem aykırı değer hem de çekim gücü yüksek bir gözlemdir, çünkü büyük artık değerine sahiptir, çekim gücü yüksek noktadır, çünkü açıklayıcı değişkenlerin merkezinden uzaktadır. Yani bu gözlem etkili gözlemdir. Bu gözlemin analizde olması regresyon eşitliğinin bazı özelliklerini önemli ölçüde değiştirecektir.

4. Etkili Gözlemlerin Belirlenmesi İçin Tanısal Yöntemler

X ' de oluşan çekim gücü yüksek gözlemler, Y ' de oluşan aykırı değerler EKK tahmin edicileri üzerinde olumsuz sonuçlar yaratır. Doğru ve en iyi tahmin edicilere ulaşmak için veri setinin incelenmesi ve varsa sorunların teşhis edilmesi gerekmektedir. Bunun için geliştirilmiş bazı tanısal araçlar mevcuttur. Bu yöntemlerle verilerin çoğundan farklı olan gözlemler sayısal ve grafiksel olarak ortaya çıkarılıp incelenerek sağlıklı sonuçlara ulaşılabilir. Bunlardan belli başlıcaları aşağıdaki gibi belirtilmiştir.

1. Doğrudan yöntemler
2. Kümeleme yöntemleri
3. Artıkları temel alan tanısal araçlar
4. Artıkların grafiklerine dayalı tanısal araçlar
5. İz düşüm matrisini temel alan tanısal araçlar
6. Artıkları ve iz düşüm matrisinin köşegen elemanlarını temel alan tanısal grafikler
7. Diğer tanısal yöntemler

4.1. Doğrudan Yöntemler

Doğrusal model için doğrudan yöntemler veri miktarı ile oynayarak aykırı değerleri tespit prensibine dayanır. Bunlardan ileri tarama yönteminde aykırı değersiz bir veri seti tespit edilir ve bu altküme ile uyumlu veriler alt kümeye katılarak altküme büyütülür. Bu süreç bazı veriler altkümeye katılmak için uygun kritere ulaşamayıncaya kadar devam eder ve bu sayede aykırı değere ulaşılabilir. Bu yöntemde öne çıkanlar şunlardır: EKK' den elde edilen standartlaştırılmış artıkların mutlak değeri en küçük olanlardan, modeldeki parametre sayısından bir fazlası adedinde, temel bir altküme ele alınır. Verilerin yarısı kadarını içerecek şekilde EKK' e göre standartlaştırılmış artıkların en küçük değerini kullanarak bir altküme oluşturulur. Bu verilere ilaveten t dağılımına göre tespit edilmiş bir kriter değerine $\left(t_{0,95}\left(n-m-1\right)\right)$ kadar olanlar dışındakiler aykırı değer olarak tespit edilir (Hadi ve Simonoff, 1993).

Bir başka yöntem geri tarama yöntemidir. Bu yöntemde ilk önce bütün veri seti ele alınır ve kullanılan yöntemdeki kritere göre aykırı değer olarak tespit edilen veriler elenir ve aykırı değersiz bir küme elde edilmeye çalışılır. Genellikle geri tarama yöntemleri tercih edilmektedir. Bunun en basit açıklaması çok sayıda aykırı değer bulunması durumunda bunların kendilerini gizleyebilecekleri ve aykırı değer olmayan verileri de aykırı değer gibi göstermesine neden olacağı sebebi sayılabilir.

4.2. Kümeleme Yöntemleri

Aykırı değer tespiti için verilerin birbirlerine yakınlıklarını ölçüp kümeler oluşturarak verilerin çoğunun oluşturduğu kümeden uzak ve azınlıkta olan verileri aykırı değer olarak belirleme ilkesine dayalı kümeleme yöntemleri mevcuttur. Bunu başarmak için ilk ortaya atılan yöntemlerden birisi Mahalanobis Uzaklığı olmuştur. Klasik fakat güvenilir olmayan bir yöntemdir. Bütün veriler kullanılarak verilerin merkez noktası ve sapması tespit edilir, fakat birden çok aykırı değer bulunması bu istatistikleri kendi yönlerine doğru çekerek etkilemekte ve kendilerini gizlemektedir. Hesaplaması kolay olmasına rağmen verilerin tamamını dikkate alarak bir uzaklık ölçütü ortaya koyduğu için aykırı değerleri tespit etmesi zaman zaman mümkün olmayacaktır (Rousseeuw ve Leroy, 1987).

En Küçük Varyans Kovaryans Determinantı (MCD) ilk ve halen yaygınlıkla kullanılmakta olan bir sağlam regresyon kümeleme yöntemidir. Bu yöntemde amaç verilerin yarısına yakın bir altküme ile bu verilerin varyans kovaryans matrisini hesaplayıp bu matrisi en küçük yapacak temel veri alt kümesini tespit etmeye çalışmaktır. Daha sonra bu temel altküme etrafında kikare dağılımına göre kritik mesafe tespit edilir ve bu alanın dışında kalan veriler aykırı değer olarak tespit belirlenir (Rousseeuw, 1984).

En Küçük Hacimli Elipsoid (MVE) tahmin yönteminde amaç, verilerin yarısına yakınını içeren bir elips tespit edip bu elipsin belli bir sınırından sonra kalan noktaları aykırı değer olarak değerlendirmektir.

4.3. Artıkları Temel Alan Tanısal Araçlar

Regresyon artıklarının çeşitli tiplerinin analizi veya artıkların birkaç dönüşümünün analizi, modelin yetersizliğini belirlemede veya veride sorun yaratabilecek noktaların belirlenmesinde çok yararlıdır. Regresyon modelinde hataların normal dağılımlı olduğu varsayılır. $\varepsilon \sim N\left(0,\sigma^2 I\right)$

4.3.1. Klasik Artıklar $\left(e_i\right)$

Klasik artıklar, $e_i = y_i - x_i'\hat{\beta}$ şeklinde tanımlanır. Burada; $x_i' : X$ matrisinin i. nci satırıdır.

$$e = Y - \hat{Y}$$

$$e = Y - X\left(X'X\right)^{-1}X'Y$$

$$e = \left[I - H\right]Y$$

$$e = \left[I - H\right]\left(X\beta + \varepsilon\right)$$

$$e = \left[I - H\right]\varepsilon \qquad (4.1)$$

Artıklar bağımlı değişkenin veya hatanın doğrusal kombinasyonu olarak ifade edilir.

Bundan dolayı i. nci artık için

$$e_i = \left(1 - h_{ii}\right)Y_i - \sum_{j \neq i}^{n} h_{ii}Y_i \qquad (4.2)$$

$$e_i = \left(1 - h_{ii}\right)\varepsilon_i - \sum_{j \neq i}^{n} h_{ii}\varepsilon_i \qquad (4.3)$$

yazılabilir. Artıkların dağılımı aşağıdaki durumlara bağlıdır.

i. Hataların dağılımına

ii. H matrisinin elemanlarına

iii. n boyutlu örnekleme

Büyük örneklemlerde, $\frac{1}{n} \approx 0$ için, $e_i \approx \varepsilon_i$ olur ve artık dağılımının analizi hataların dağılımı hakkında doğrudan bilgiler verir. Klasik artıklar sabit varyansa sahip değildir. Küçük ve orta örneklemler için klasik artık, etkili noktaların tespiti veya teşhisi için iyi sonuç veremeyebilirler.

4.3.2. Standartlaştırılmış Artıklar $\left(e_{S,i}\right)$

$$e_{S,i} = \frac{e_i}{\hat{\sigma}\sqrt{1-h_{ii}}} \qquad i = 1,2,...,n \qquad (4.4)$$

ile elde edilir. Burada;

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} e_i^2}{(n-m)}$$

artık kareler ortalamasıdır.

Standartlaştırılmış artıklar birim varyansa sahip olur. t rasgele değişkenine benzer davranış gösterir. Bu yüzden çoğu tanısal grafikte klasik artıklar yerine standartlaştırılmış artıklar kullanılmaktadır.

4.3.3. Normalleştirilmiş Artıklar $\left(e_{N,i}\right)$

Zaman zaman başvurulan bir artık türü de

$$e_{N,i} = \frac{e_i}{\hat{\sigma}} \qquad (4.5)$$

ile hesaplanır. Normalleştirilmiş artıklar olarak isimlendirilen bu artıkların sıfır ortalamalı birim varyanslı normal dağıldığı varsayılmasına karşılık bu durum gerçekte böyle değildir. Yine de uygulamada, normalleştirilmiş artıklar arasında $\pm 3\hat{\sigma}$ değerini aşan gözlemler aykırı değer olarak belirlenir. Bu yaklaşım oldukça yanıltıcıdır ve veri ile ilgili yanlış kararlar almaya neden olur.

4.3.4. Jackknife Artıklar $\left(e_{J,i}\right)$

Jackknife artıklar, $\hat{\sigma}^2$'nin e_i'lerden bağımsız bir tahminini gerektirir. Böyle bir tahmin şu ifade kullanılarak elde edilir. Hataların normalliği altında i. nci noktanın etkisi araştırıldığında regresyon modeli $Y = X\beta + U_i\theta + \varepsilon$ şeklinde tanımlanır. Burada U_i : i. nci satırı bir, diğer satırları sıfır olan $(n \times 1)$ boyutlu vektördür. Bu regresyon modeli için artık kareler toplamı $Y'(I-H)Y - Y'(I-H)U_i\left(U_i'PU_i\right)^{-1}U_i'(I-H)Y$ ifadesinden elde edilir. Bu ifade

kullanılarak $\hat{\sigma}^2$'nin e_i'lerden bağımsız bir tahminini elde edilir. Bu artık kareler toplamı kullanılarak, i. nci gözlem olmaksızın hesaplanan artık kareler ortalaması $\left(\hat{\sigma}^2_{(i)}\right)$ aşağıdaki gibi elde edilir.

$$\hat{\sigma}^2_{(i)} = \frac{(n-m).\hat{\sigma}^2 - e_i^2/(1-h_{ii})}{n-m-1}$$

$$\hat{\sigma}^2_{(i)} = \hat{\sigma}^2\left(\frac{n-m-e^2_{S,i}}{n-m-1}\right) \qquad (4.6)$$

Normallik varsayımı altında, $\hat{\sigma}^2_{(i)}$ ve e_i bağımsızdır ve jackknife artıklar aşağıdaki gibi tanımlanır.

$$e_{J,i} = \frac{e_i}{\hat{\sigma}^2_{(i)}.(1-h_{ii})^{1/2}} \qquad (4.7)$$

$e_{J,i}$'nin dağılımı, $(n-m-1)$ serbestlik dereceli Student t dağılımıdır. (4.7) eşitliğinde $\hat{\sigma}^2_{(i)}$'nin yerine (4.6) ifadesi yazılırsa, $e_{J,i}$'nin $e_{S,i}$ cinsinden ifadesini verir.

$$e_{J,i} = e_{S,i}\left(\frac{n-m-1}{n-m-e^2_{S,i}}\right) \qquad (4.8)$$

Böylelikle $\hat{\sigma}^2_{(i)}$ ve jackknife artıkların, gözlemin çıkarılmasıyla regresyon analizinin tekrarlanmasına gerek kalmaksızın tek bir analiz sonucu ile elde edilebileceği görülmektedir. Jackknife artıklar, aykırı değerlerin tespiti için sık sık kullanılır. $e^2_{j,i} \le F_{1-\alpha/n}(1; n-m-1; 0.5)$ şartı geçerli olduğu zaman ilgili gözlemin etkili gözlem olmadığını belirtir. Çekim gücü yüksek noktaların durumu hakkında bu artıklar hiçbir belirti vermez.

4.3.5. Silinmiş Artıklar $\left(e_{P,i}\right)$

Klasik ve standartlaştırılmış artıklar bütün veriyi kullanarak hesaplanan bir tahmini temel alırken, silinmiş artıklar $\left(e_{P,i}\right)$, veriden i. nci gözlemi attıktan sonra geriye kalan gözlemlerle hesaplanan tahmini temel alır.

$\hat{\beta}$: Bütün veri kullanarak hesaplanan β'nın EKK tahminidir.

$\hat{\beta}_{(i)}$: i. nci gözlem çıkarıldıktan sonra hesaplanan β'nın EKK tahminidir.

Böylece i. nci silinmiş artık aşağıdaki gibi tanımlanır.

$$e_{P,i} = y_i - x_i'\hat{\beta}_{(i)} \qquad i = 1,2,...,n \tag{4.9}$$

$$e_{P,i} = y_i - x_i'\left(\hat{\beta} - \frac{e_i (X'X)^{-1} x_i}{1-h_{ii}}\right) \tag{4.10}$$

$$e_{P,i} = \frac{e_i}{1-h_{ii}} \tag{4.11}$$

Böylece silinmiş artıklarda tek bir analiz ile elde edilebilir. Anderson, Allen ve Cady (1972) ve Allen (1974), model seçimi için silinmiş artık kareler toplamını (PRESS) bir kriter olarak kullanmıştır. $PRESS = \sum_{i=1}^{n} e_{P,i}^2$. Buna göre küçük PRESS değerine sahip model iyi model olarak belirtilmiş.

4.4. H Matrisinin Köşegen Elemanlarını Temel Alan Tanısal Araçlar

$H = X(X'X)^{-1}X'$ şeklinde tanımlanan H matrisi simetrik $(H' = H)$ ve idempotenttir $(H^2 = H)$. Bu matrisin köşegen elemanları çekim gücü yüksek noktaların belirlenmesi için kullanılır. Hoaglin ve Welsch (1978) tarafından, $h_{ii} \rangle 2m/n$ olan gözlemler çekim gücü yüksek noktalar olarak belirtilmiştir. X matrisinin sütunları lineer bağımsız olduğunda ve regresyon modeli sabit terim içerdiğinde $\sum_{i=1}^{n} h_{ii} = m$ olduğu gösterilebilir. Bundan dolayı H matrisinin köşegen elemanlarının ortalaması m/n olur. $0 \le h_{ii} \le 1$ olduğundan $n \le 2m$ ise bu kriter geçerli olamaz.

4.5. Artıklara Dayalı Tanısal Grafikler

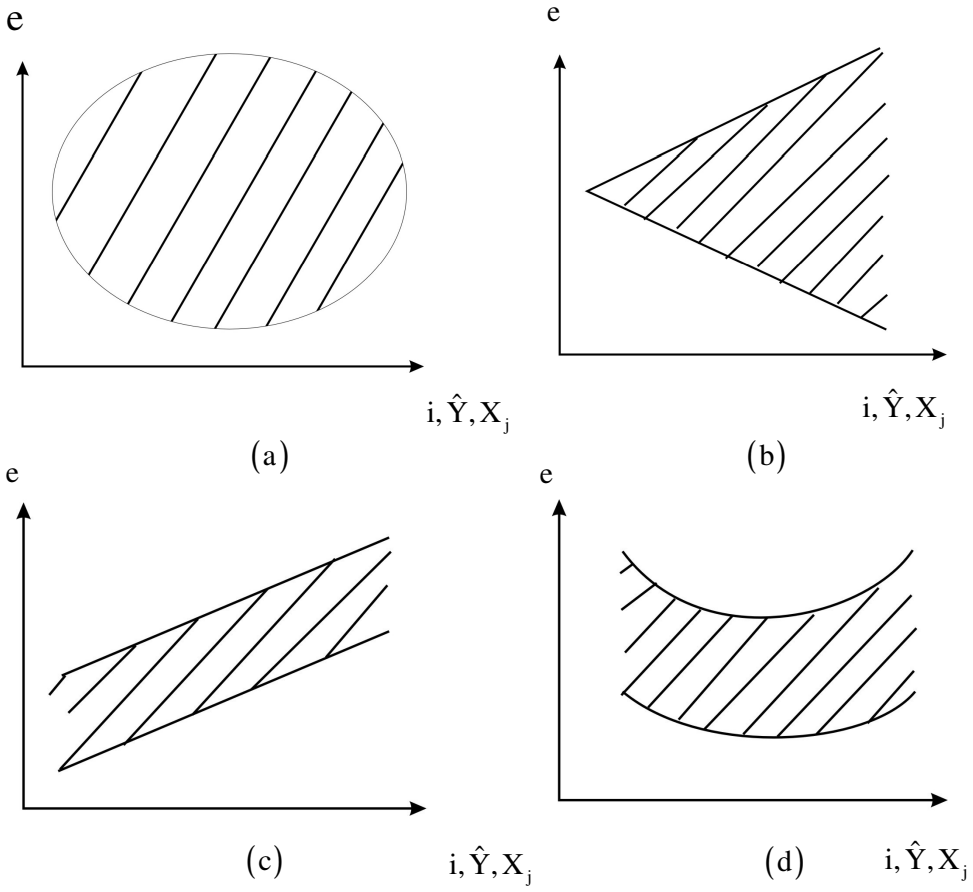
Tanısal grafikler, regresyon model yapılandırma ve kontrol aşamasında oldukça sık kullanılırlar. Modeldeki yetersizlikleri, veri içerisinde yer alabilecek olan aykırı değerleri veya çekim gücü yüksek olan gözlemleri ortaya çıkarmada oldukça kullanışlı olan grafikler:

1. Artıkların i indeksine karşı grafiği

2. Artıkların $\hat{Y}$ tahmin değerlerine karşı grafiği

3. Artıkların X_J açıklayıcı değişkenlerine karşı grafiği

şeklinde sıralanabilir.



Şekil 4.1. Artık grafiklerindeki olası sonuçlar

Artık grafiklerinde noktaların dağılımı eğer şekil 4.1.(a)' da ki gibi bir tesadüfi görünüme sahip ise EKK varsayımları doğrudur.

Artıkların i indeksine veya $\hat{Y}$ tahmin değerlerine karşı çizilen grafiklerinde noktaların dağılımı eğer şekil 4.1.(b)' de ki gibi ise EKK varsayımlarından bazıları bozulmaktadır ve bu görünüm veride değişen varyans sorunu olduğunu gösterir. Yani sabit varyansa sahip değildir, buna getirilen çözüm ağırlıklı EKK yöntemi analizi en uygundur veya analize başlamadan önce Y_i gözlemleri üzerinde bir dönüşüm uygulamak gerekmektedir.

Artıkların i indeksine veya X_J açıklayıcı değişkenlerine karşı çizilen grafiklerinde eğer noktaların dağılımı şekil 4.1.(c)' de ki bant şeklinde ise yine EKK varsayımları bozuluyor demektir. Hesaplamalarda bazı hatalar olduğunu veya modelde X_j' nin yokluğunu gösterir. X açıklayıcı değişkeninin lineer etkisi kaybolmuştur.

Artıkların i indeksine, X_J açıklayıcı değişkenlerine ve $\hat{Y}$ tahmin değerlerine karşı çizilen grafiklerinde eğer noktaların dağılımı şekil 4.1.(d)' de ki gibi bir lineer olmayan görünüm sergiliyorsa ise ileri sürülen modelin yanlış olduğunu gösterir. Modele karesel veya çarpım halindeki bir terimin eklenmesine veya Y_i gözlemleri üzerinde bir dönüşümün yapılmasına gerek duyulmaktadır.

Artıkların Y tahmin değerlerine karşı çizilen grafikleri kullanışlı değildir. Çünkü bu iki nicelikte birbirleriyle güçlü ilişkilidir.

4.6. Artıkları ve H Matrisinin Köşegen Elemanlarını Temel Alan Tanısal Grafikler

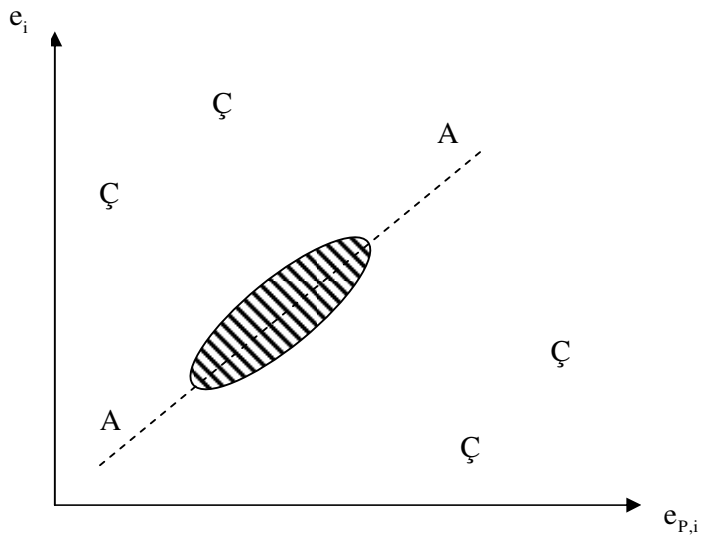
Regresyon tanısal araçları içerisinde yer alan grafiksel tanı yöntemleri, analizciye kolay ve daha hızlı bir şekilde, veri içerisinde olabilecek sorunları ortaya çıkarmada fayda sağladıklarından etkili noktaların tespitinde oldukça önemli rol oynarlar. Bu çalışmada da veri içerisinde aykırı değer veya çekim gücü yüksek gözlem olarak etkili olabilecek noktaların belirlenmesi için çeşitli

regresyon sonuçlarına farklı taraftan bakılmış, Artıklara ve İz Düşüm Matrisi H' nin köşegen elemanlarına dayalı tanısal grafikler incelenmiştir.

1) Silinmiş Artık Grafiği
2) Williams Grafiği
3) Pregibon Grafiği
4) İndeks Grafiği
5) Rankit Grafiği (Q-Q grafiği)

4.6.1. Silinmiş Artıkların Grafiği

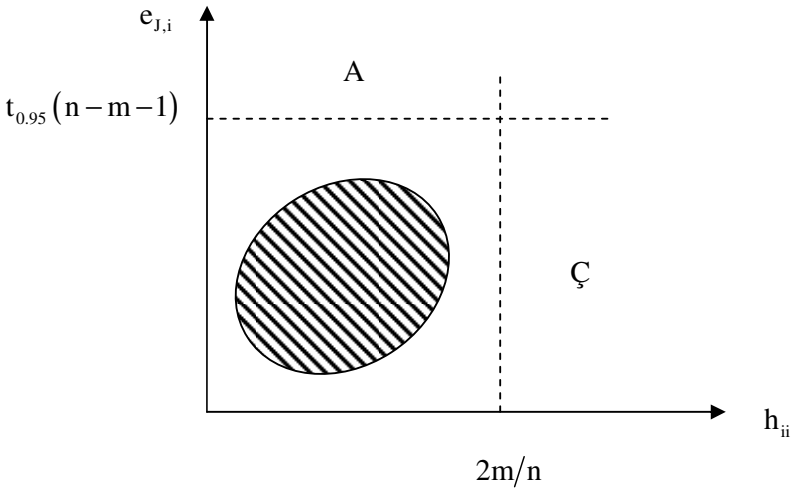
Bu grafikte X ekseninde silinmiş artıklar $e_{P,i}$, Y ekseninde de klasik artıklar e_i yer alır (Williams, 1973). Çekim gücü yüksek noktalar yerlerinden dolayı kolaylıkla tespit edilir ve $y = x$ doğrusu dışında uzanmışlardır. Aykırı değerler $y = x$ doğrusu üzerinde yerleşmişler fakat merkezi örneklemden uzaktadırlar.



Şekil 4.2. Silinmiş artıkların grafiği. Ç bir çekim gücü yüksek nokta, A bir aykırı değerdir.

4.6.2. Williams Grafiği

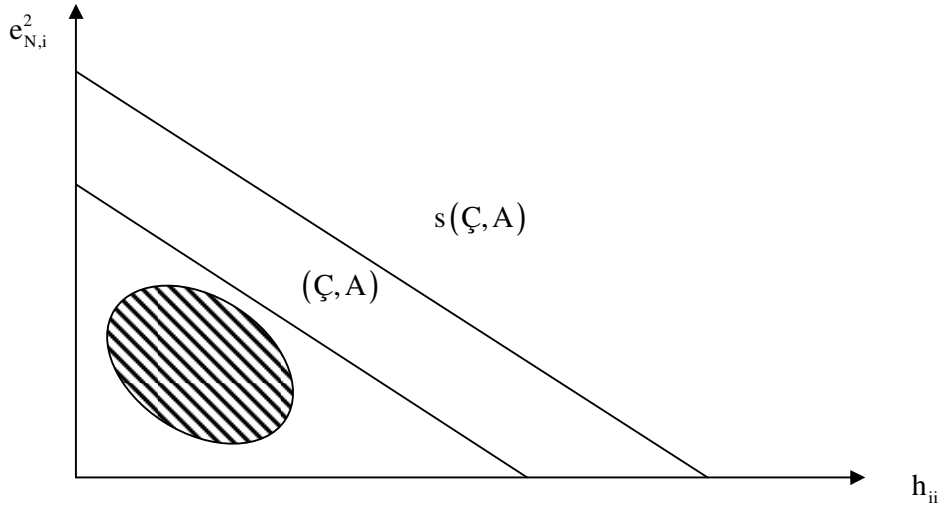
Bu grafikte X ekseninde h_{ii} köşegen elemanları, Y ekseninde de jackknife artıklar $e_{J,i}$ bulunur. 2 sınırlayıcı çizgi çizilir (Williams, 1973). İlki artıklar için $Y = t_{0.95}(n-m-1)$ ve ikincisi çekim gücü yüksek noktalar için $X = 2m/n$ dir. Burada $t_{0.95}(n-m-1)$; $(n-m-1)$ serbestlik dereceli student t dağılımının %95 lik t tablo değeridir.



Şekil 4.3. Williams grafiği. Ç çekim gücü yüksek nokta, A aykırı değerdir.

4.6.3. Pregibon Grafiği

Bu grafikte X ekseninde köşegen elemanları h_{ii}, Y ekseninde de normalleştirilmiş artıkların karesi $e^2_{N,i}$ bulunur (Pregibon, 1981). Bu grafik için $E(h_{ii} + e^2_{N,i}) = (m+1)/n$ ifadesi geçerli olduğundan, 2 farklı birbirine benzeyen doğrular çizilir. $y = -x + 2(m+1)/n$ ve $y = -x + 3(m+1)/n$.



Şekil 4.4. Pregibon grafiği. $(Ç, A)$ etkili nokta, $s(Ç, A)$ güçlü etkili noktadır.

İçindeki etkili noktaları ayırt etmek için aşağıdaki kurallar kullanılır.

1. Eğer bir nokta en üstteki doğrunun üstünde ise, bu nokta çok güçlü etkilidir.

2. Eğer bir nokta iki doğru arasında ise, bu nokta etkilidir.

Etkili nokta, bir aykırı değer veya bir çekim gücü yüksek noktalardan birisi olabilir.

4.6.4. İndeks Grafiği

X ekseninde sıralı i indeksi, Y ekseninde de artıkların çeşitleri $(e_{S,i}, e_{P,i}, e_{J,i}, e_i)$ veya köşegen elemanları h_{ii} veya diğer tahminler bulunur. Bu grafik şüpheli noktaları belirtir. Bu şüpheli noktalar aykırı değer veya çekim gücü yüksek gözlemler olarak etkili olabilir.

4.6.5. Rankit Grafiği (Q-Q Çizimi)

Bu grafik türünde Y eksenine artıkların herhangi bir çeşidi, X eksenine seçilen artıkların artan sıralı değerlerine bağlı hesaplanan olasılıklarına

$\left(p_i = i/(n+1)\right)$ karşılık gelen standart normal değerleri $U_{p,i}$' ler yerleştirilir. Bu grafik yardımı ile verideki aykırı değerleri tespit edebiliriz.

4.7. Diğer Tanısal Ölçüler

Bu ölçüler parametre tahminlerinde verilen bir noktanın gerçek etkisini belirtir. i. nci gözlemin etkili olup olmadığını görmek için analizden çıkarılıp, tekrar analiz edilmesi gerekir. i. nci gözlemin silinmesiyle ortaya çıkan etkiyi ölçen ölçütler aşağıdaki gibi tanımlanır.

4.7.1. Cook Uzaklığı (D_i)

Bu ölçü, i. nci gözlemin veri setinden çıkartılmasının parametre tahminler üzerindeki etkisini gösterir (Cook ve Weisberg, 1982).

$$D_i = \frac{\left(\hat{\beta} - \hat{\beta}_{(i)}\right)' X'X \left(\hat{\beta} - \hat{\beta}_{(i)}\right)}{m\hat{\sigma}^2} \qquad (4.12)$$

Yukarıdaki formülde $\hat{\beta}_{(i)}$' yi hesaplamak için gözlem sayısı kadar analizi tekrar etmek gerekir. Bunun yerine Cook Uzaklığı;

$$D_i = \frac{e_{S,i}}{m} \frac{h_{ii}}{1-h_{ii}} \qquad (4.13)$$

D_i değeri $4/n$ değerinden büyük olursa i. nci gözlem etkili bir gözlemdir.

4.7.2. Atkinson Ölçüsü (A_i)

Atkinson tarafından ifade edilen, Cook Uzaklığının düzeltilmiş hali olan bu ölçü aşağıdaki gibi ifade edilir (Atkinson, 1985).

$$A_i = \left|e_{J,i}\right| \sqrt{\frac{n-m}{m} \frac{h_{ii}}{1-h_{ii}}} \qquad (4.14)$$

Bu ölçü ayrıca grafik yorumlaması için uygundur. Atkinson, A_i' nin mutlak değerlerinin artıklara karşı grafiğinin çizilebileceğini önerir. $h_{ii} = m/n$ ise Atkinson ölçüsü sayısal olarak Jackknife artığına eşit olur.

4.7.3. DFFITS Ölçüsü (DFFITS)

Veri setinden i. nci gözlemin silinmesiyle $\hat{Y}_i$ tahmin değerleri üzerindeki etkisini ölçen ölçüttür (Belsey, Kuh ve Welsch, 1980). Aşağıdaki gibi ifade edilir;

$$DFFIT_i = e_i \frac{h_{ii}}{1-h_{ii}} \qquad (4.15)$$

DFFIT ' nın ölçeklenmiş biçimi ise

$$DFFITS_i = \sqrt{\frac{h_{ii}}{1-h_{ii}}} \frac{e_i}{\hat{\sigma}_{(i)}\sqrt{1-h_{ii}}} \qquad (4.16)$$

şeklinde tanımlanır. Burada $\hat{\sigma}_{(i)}$ i. nci gözlem çıkarıldıktan sonra elde edilen regresyon standart hatasıdır. Bu ölçütte $|DFFITS_i| > 2\sqrt{m/n}$ ise i . nci noktanın $\hat{Y}$ üzerinde anlamlı bir etki yapabileceği düşünülebilir (İpek, 2002).

4.7.4. DFBETA Ölçüsü

i. nci gözlemin veri setinden çıkartılmasının katsayı tahminleri üzerindeki etkisini gösterir.

$$DFBETA_i = \hat{\beta} - \hat{\beta}_{(i)} \qquad (4.17)$$

biçiminde verilir. DFBETA ' nın ölçeklenmiş biçimi olarak, bu vektörün j. nci elemanı için bu istatistik

$$DFBETAS_{j(i)} = \frac{\hat{\beta}_j - \hat{\beta}_{j(i)}}{\hat{\sigma}_{(i)}\sqrt{(X'X)^{-1}_{jj}}} \qquad (4.18)$$

eşitliği ile elde edilir. Bu ölçüt $|DFBETAS| > 2/\sqrt{n}$ ise ilgili gözlem j. değişkenin katsayısı üzerinde etkilidir (İpek, 2002).

5. UYGULAMA

Bu bölümde EKK yöntemi kullanılarak veri setindeki etkili noktaların tespiti için tanısal ölçülerin ve indeks grafiklerinin yanı sıra artıkların ve iz düşüm matrisinin köşegen elemanlarına dayalı tanısal grafikler (Silinmiş artık grafiği, Williams grafiği, Pregibon grafiği, ve Rankit Q-Q grafiği) yardımı ile etkili noktaların kolaylıkla tespit edildiğini göstermek amacıyla gerçek veri seti üzerinde bir uygulama yapılmıştır. Verilerin analizi için Minitab programı kullanılmıştır.

Uygulamada kullanılan veri seti, nitrik asit yükseltgenmesi sırasında sızan amonyağın nelerden etkilendiği düşünülerek bir fabrikadan 21 günlük gözlemler elde edilmiştir (Analitica Chimica Acta 439, 2001). Nitrik asit amonyağının yükseltgenmesi sırasında sızan amonyak gazının yüzdelik miktarları bağımlı değişken olarak alınmış, hava akışıyla ölçülen fabrikadaki reaksiyon hızı, nitrik asit en yüksek karşı takım emiliminde sarmal içindeki soğutulmuş su döngüsünün giriş sıcaklığı ve 10 kat olarak belirlenmiş en üst düzeyde emilimi sağlayan sıvıdaki nitrik asit değişim oran değerleri bağımsız değişken olarak tanımlanmıştır.

Regresyon modeli

$$Y_i = \beta_0 + \beta_1 X_{1i} + ... + \beta_p X_{pi} + \varepsilon_i$$

olmak üzere,

Y_i : Nitrik asit amonyağının yükseltgenmesi sırasında sızan amonyak gazının yüzdelik miktarları

X_1 : Hava akışıyla ölçülen fabrikadaki reaksiyon hızı

X_2 : Nitrik asit en yüksek karşı takım emiliminde sarmal içindeki soğutulmuş su döngüsünün giriş sıcaklığı

X_3 : 10 kat olarak belirlenmiş en üst düzeyde emilimi sağlayan sıvıdaki nitrik asit değişim oranları

En Küçük Kareler Yöntemi uygulanarak elde edilen model denklemi;

$$\hat{Y} = -35,1 + 0,740X_1 + 1.43X_2 - 0.262X_3$$

şeklindedir. Varyans analizi tablosu ve regresyon katsayıları anlamlılık testi sonuçları aşağıda verilmiştir.

Tablo 5.1. Tüm veriler için EKK yöntemi varyans analizi sonuçları

	Serbestlik Derecesi	Kareler Toplamı	Kareler Ortalaması	F	p Değeri
Regresyon	3	1897,74	632,58	62,71	0,000
Artık	17	171,50	10,09		
Toplam	20	2069,24			

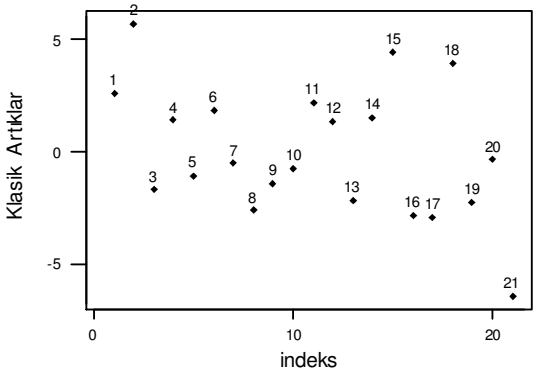
Tablo 5.2. Tüm veriler için regresyon katsayıları ve anlamlılık testi sonuçları

	Katsayılar	Std. Hata	t	p	VIF
Sabit	-35,07	11,60	-3,02	0,008	
X_1	0,7403	0,1337	5,54	0,000	2,8
X_2	1,4284	0,3482	4,10	0,001	2,4
X_3	-0,2623	0,1594	-1,65	0,118	1,5

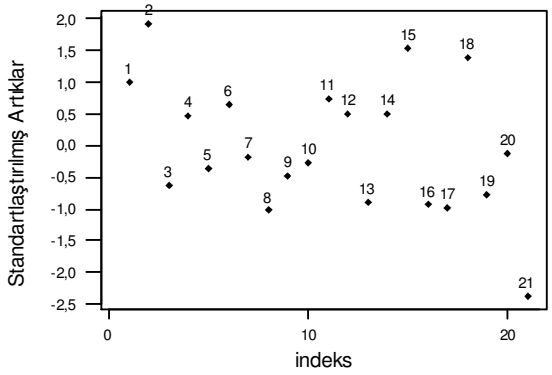
$\alpha = 0,05$ anlamlılık seviyesinde her bir parametrenin ayrı ayrı anlamlılık testleri yapıldığında sadece X_3 değişkeninin modele olan katkısı istatistiksel olarak anlamlı değildir.

Belirtme katsayısı $R^2 = \%91,7$ yani verideki değişimin $\%91,7$'i regresyon modeli ile açıklanır. Standart sapma değeri $\hat{\sigma} = 3,176$, silinmiş artık kareler toplamı olan $PRESS = 278,128$, silinmiş artık karelerle hesaplanan belirtme katsayısı $R_p^2 = \%86,6$, düzeltilmiş belirtme katsayısı $R_d^2 = \%90,2$, silinmiş artık kareler ortalaması $MEP = 13,24$'dır. Bu istatistikler modelin uyumunu kontrol etmede kullanılacaktır.

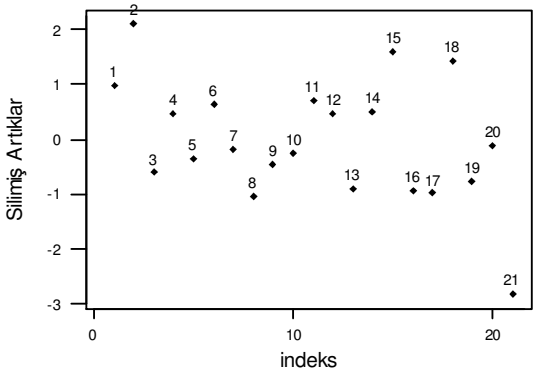
Veriye EKK yöntemi uygulandıktan sonra etkili gözlemleri tespit etmek için kullanılan tanısal istatistikler Ek.2.'de verilmiştir. Bu tanısal istatistik değerlerine karşı çizilen indeks grafikleri ise aşağıdaki elde edilir ve yorumlanır.



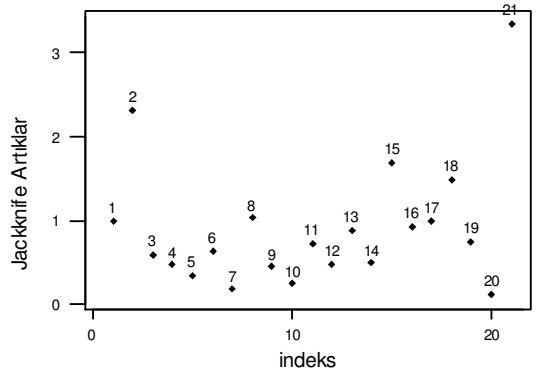
Şekil 5.1. Klasik Artıkların İndeks Grafiği



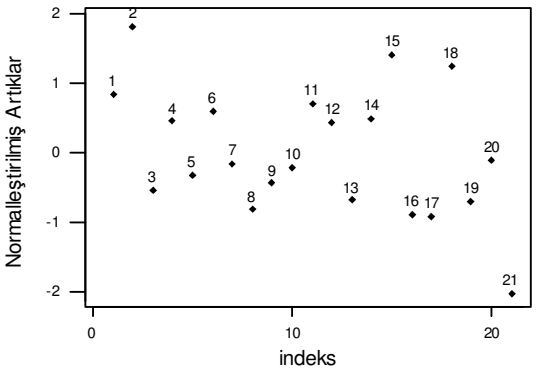
Şekil 5.2. Standartlaştırılmış Artıkların İndeks Grafiği



Şekil 5.3. Silinmiş Artıkların İndeks Grafiği

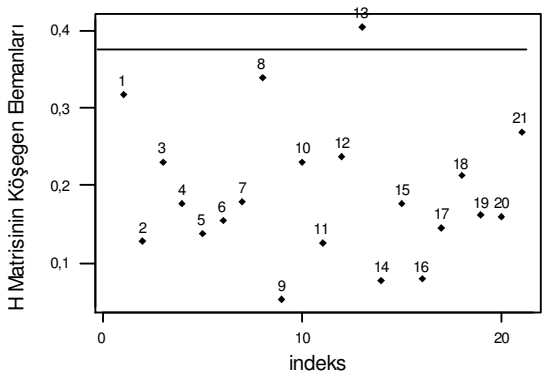


Şekil 5.4. Jackknife Artıkların İndeks Grafiği



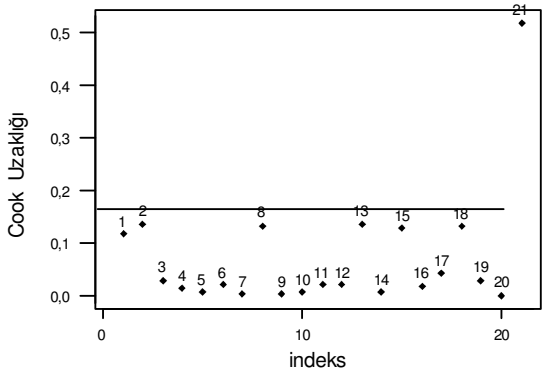
Şekil 5.5. Normalleştirilmiş Artıkların İndeks Grafiği

Artıkların (Klasik, Standartlaştırılmış, Silinmiş, Jackknife, Normalleştirilmiş) indeks grafikleri incelendiğinde diğerlerine göre farklılık gösteren şüpheli gözlemlerin 2 ve 21. gözlemler olabileceği görülmektedir.



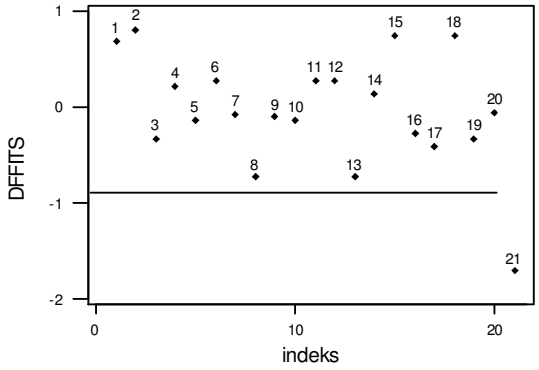
Şekil 5.6. H Matrisinin Köşegen Elemanlarının İndeks Grafiği

H matrisinin köşegen elemanları olan h_{ii} değerleri $2m/n = 0,38$ değerinden büyük olursa i. gözlem çekim gücü yüksek gözlem olarak ifade edilir. Şekil 5.6. incelendiğinde bu kriter değerini aşan gözlemin 13. gözlem olduğu görünmektedir.



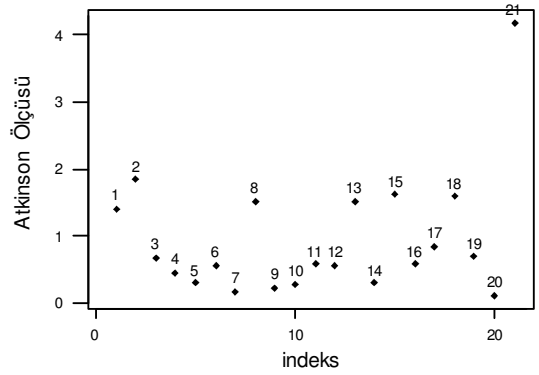
Şekil 5.7. Cook Uzaklığının İndeks Grafiği

Cook uzaklığı D_i değerleri, $4/n = 0,19$ değerinden büyük olursa i. gözlem parametre tahminleri üzerinde etkili gözlemdir. Şekil 5.7.' de verilen Cook uzaklığının indeks grafiği incelendiğinde kriter değerini aşan gözlemin 21. gözlem olduğu görülmektedir.



Şekil 5.8. DFFITS Ölçüsünün İndeks Grafiği

DFFITS değerleri gözlemlerin tahmin değerine etkisini gösterir, Şekil 5.8. incelendiğinde kriter olarak karşılaştırılan $2\sqrt{m/n} = 0,87$ değerinden büyük olan 21. gözlemin etkili gözlem olabileceği sonucuna varılabilir.

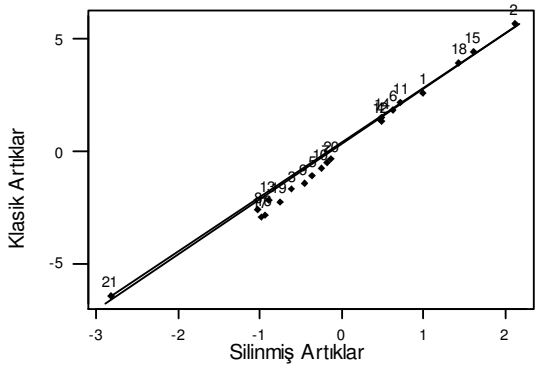


Şekil 5.9. Atkinson Ölçüsünün İndeks Grafiği

Şekil 5.9.' da verilen Atkinson ölçüsünün indeks grafiği incelendiğinde A_i değerleri arasında $A_i^2 > 10$ kriter değerini aşan 21. gözlemin etkili gözlem olabileceği söylenir.

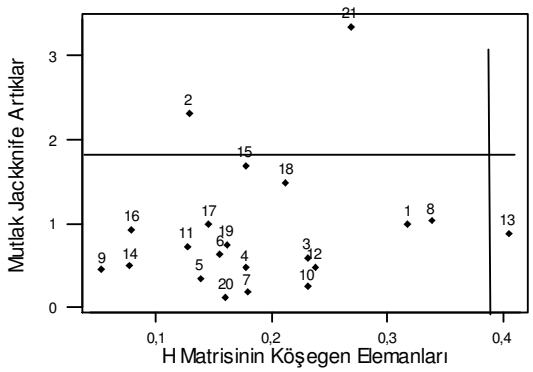
İncelediğimiz indeks grafikleri yardımı ile veri seti içerisinde 2, 13 ve 21. gözlemlerin etkili gözlemler olabileceği ifade edilebilir.

Bu çalışmanın asıl amacı etkili gözlemleri yani aykırı değer veya çekim gücü yüksek gözlem olabilecek noktaları, artıkların ve iz düşüm matrisinin köşegen üzerindeki elemanlarına dayalı tanısal grafikler (Silinmiş artık grafiği, Williams grafiği, Pregibon grafiği, Jackknife artıkların Rankit Q-Q grafiği) yardımı ile kısa sürede ve kolaylıkla elde etmek olduğundan, bu amaç için çizilen grafikler aşağıdaki gibi elde edilir.



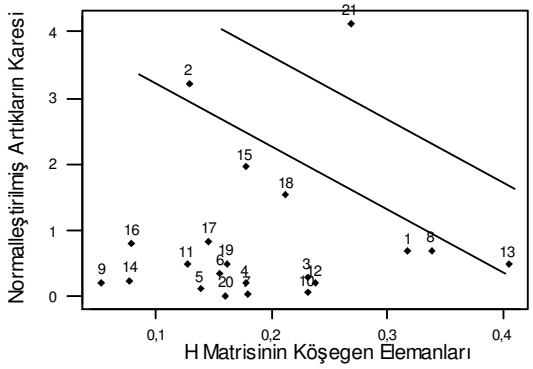
Şekil 5.10. Silinmiş Artık Grafiği

Şekil 5.10.' da verilen silinmiş artık grafiğini incelediğimizde bütün gözlemlerin $y = x$ doğrusu üzerinde yer aldığını, ancak merkezi örneklemden uzakta bulunan 2. ve 21. gözlemlerin veri içerisinde aykırı değer olarak etki yapabileceği söylenebilir.



Şekil 5.11. Williams Grafiği

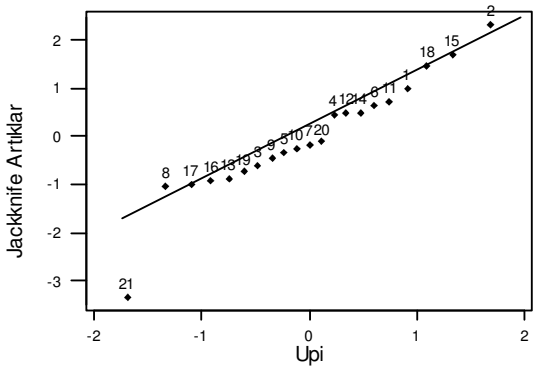
Şekil 5.11.' da verilen Williams grafiği yardımı ile h_{ii} değerleri içerisinde $2m/n = 0,38$ kriter değerini aşan 13. gözlemin çekim gücü yüksek gözlem, Jackknife artıkları arasında ise $t_{0,95}(21-4-1) = 1,746$ tablo değerini aşan 2. ve 21. gözlemlerin aykırı değer olabileceği söylenebilir.



Şekil 5.12. Pregibon Grafiği

Şekil 5.12.' de ki Pregibon grafiği içinse en üstteki doğrunun üzerinde yer alan 21. gözlemin veri üzerinde çok güçlü etkili olduğu, iki doğru arasında yer alan 2 ve 13. gözlemin ise daha az etkili olduğu ancak bu gözlemlerin aykırı

değer mi? yoksa çekim gücü yüksek gözlem olarak mı? etkili olduğu hakkında bilgi vermez.



Şekil 5.13. Jackknife Artıkların Rankit Q-Q Grafiği

Jackknife artıklarının Rankit Q-Q grafiğinde, Y ekseninde jackknife artıklar, X ekseninde jackknife artıklarının artan sıralı değerlerine bağlı hesaplanan olasılıklarına karşılık gelen standart normal değerleri yer alıyor. Grafiğe göre bütün gözlemler $y = x$ üzerinde iken 21. gözlem doğru dışında olduğundan bu gözlemin aykırı değer olarak etki yapabileceği söylenebilir.

Bilinen tanısal istatistikler, indeks grafikleri, artıkları ve iz düşüm matrisinin köşegen elemanlarını temel alan tanısal grafiklerin incelenmesi sonucunda aykırı değer olduğu tespit edilen 2. ve 21. gözlemler, çekim gücü yüksek gözlem olduğu düşünülen 13. gözlem veri setinden çıkarılarak analiz tekrarlanmış ve sonuçlar aşağıdaki gibi elde edilmiştir.

Tablo 5.3. 2, 13, ve 21. gözlemler çıkarıldıktan sonra EKK yöntemi varyans analizi sonuçları

	Serbestlik Derecesi	Kareler Toplamı	Kareler Ortalaması	F	p Değeri
Regresyon	3	1792,93	597,64	120,66	0,000
Artık	14	69,34	4,95		
Toplam	17	1862,28			

Tablo 5.4. 2, 13, ve 21. gözlemler çıkarıldıktan sonra regresyon katsayıları ve anlamlılık testi sonuçları

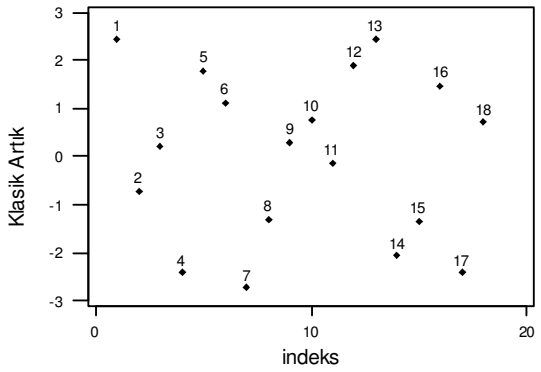
	Katsayılar	Std. Hata	t	p	VIF
Sabit	-35,08	10,44	-3,36	0,005	
X_1	0,9393	0,1054	8,91	0,000	3,1
X_2	0,8289	0,2881	2,88	0,012	3,1
X_3	-0,2539	0,1378	-1,84	0,087	1,3

Tablo 5.5. Etkili gözlemler çıkarılmadan ve çıkarıldıktan sonraki regresyon sonuçları

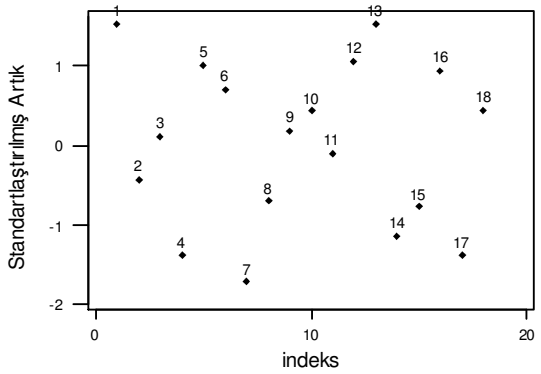
	Tüm Gözlemler Varken	2, 13, ve 21. Gözlemler Çıkarıldıktan Sonra
AKO	10,09	4,95
PRESS	278,128	118,514
R^2	% 91,7	% 96,3
R_p^2	% 86,6	% 93,64
R_d^2	% 90,2	% 95,5

Veri setinden, etkili gözlemler olduğu düşünülen 3 gözlem çıkarıldıktan sonra elde edilen EKK sonuçları incelendiğinde, belirtme katsayısında $\%4,6$'lik bir artış, artık kareler ortalamasında 5,14'lük bir azalma, silinmiş artık kareler toplamında 159,614'lük azalma, silinmiş artıklarla hesaplanan belirtme katsayısında $\%7,04$'lık bir artış ve düzeltilmiş belirtme katsayısında da $\%5,3$'lük bir artış görülmektedir. Bu da istediğimiz bir durumdur. Buna göre sildiğimiz gözlemler veri üzerinde anlamlı bir etki yapmıyormuş deriz.

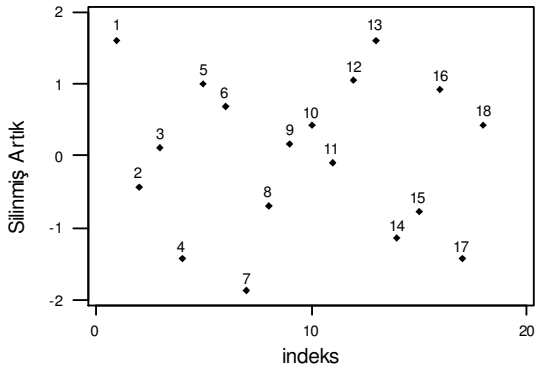
Etkili gözlemler silindikten sonra elde edilen 18 gözlemli veri seti için hesaplanan tanısal istatistikler Ek.3' de verilmiştir. Tanısal istatistiklere karşı çizilen indeks grafikleri ise aşağıdaki gibi elde edilir.



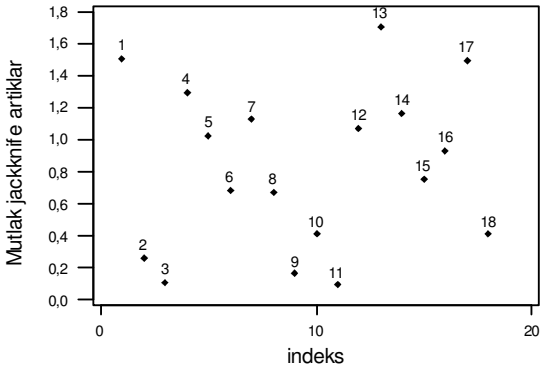
Şekil 5.14. Klasik Artıkların İndeks Grafiği



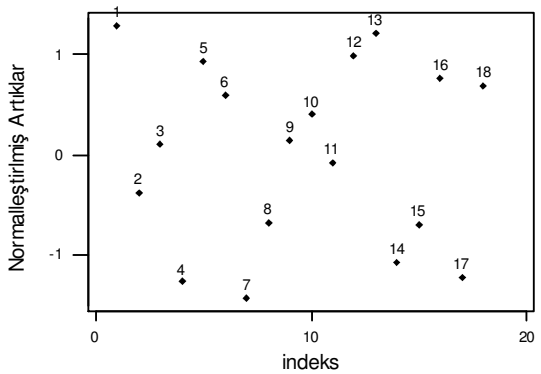
Şekil 5.15. Standartlaştırılmış Artıkların İndeks Grafiği



Şekil 5.16. Silinmiş Artıkların İndeks Grafiği

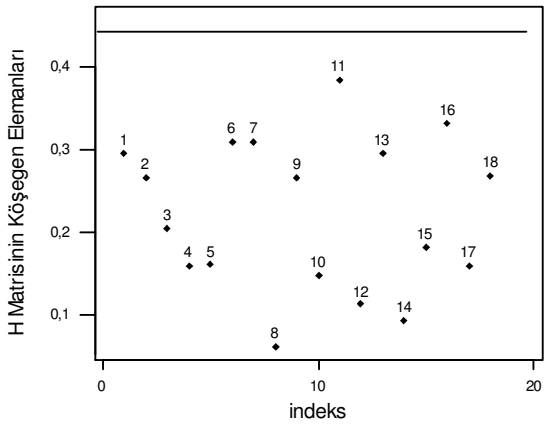


Şekil 5.17. Jackknife Artıkların İndeks Grafiği



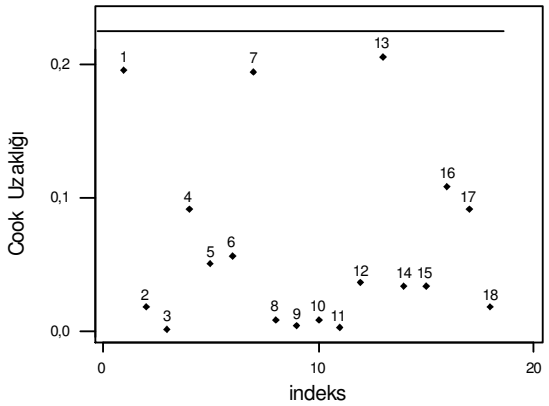
Şekil 5.18. Normalleştirilmiş Artıkların İndeks Grafiği

Etkili gözlemler çıkarıldıktan sonra artıkların indeks grafikleri incelendiğinde diğerlerine göre farklılık gösteren yani veri içerisinde etkili olabilecek gözlemin olmadığı görünmektedir.



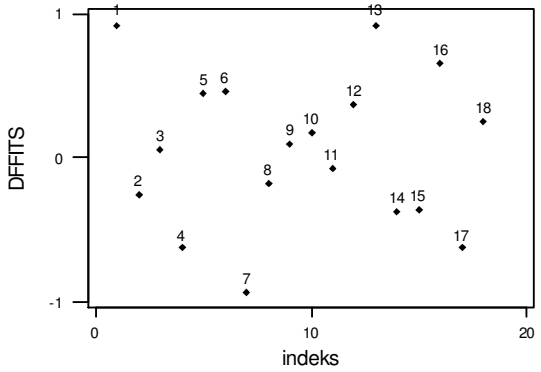
Şekil 5.19. H Matrisinin Köşegen Elemanlarının İndeks Grafiği

Şekil 5.19.' da verilen H matrisinin köşegen elemanlarının indeks grafiğine baktığımızda h_{ii} değerleri arasında $2m/n = 0,44$ kriter değerini aşan çekim gücü yüksek gözlemin olmadığı görünmektedir.



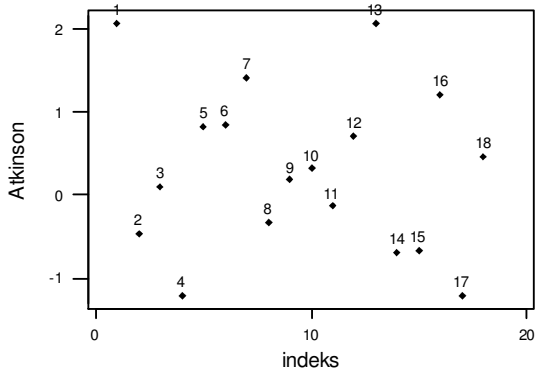
Şekil 5.20. Cook Uzaklığının İndeks Grafiği

Şekil 5.20. incelendiğinde Cook uzaklığı D_i değerleri arasında $4/n = 0,22$ kriter değerini aşan gözlemin bulunmadığı söylenebilir.



Şekil 5.21. DFFITS Ölçüsünün İndeks Grafiği

Şekil 5.21. incelendiğinde DFFITS değerleri arasında mutlak değerce $2\sqrt{m/n} = 0,94$ kriter değerini aşan gözlemin bulunmadığı görülmektedir.

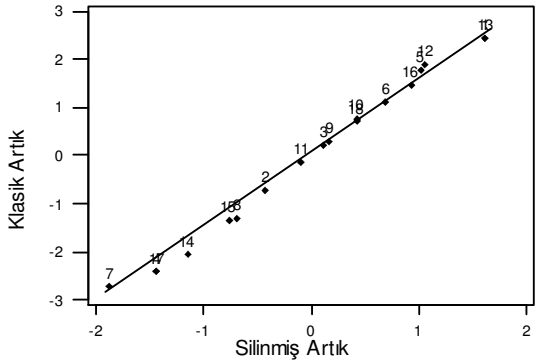


Şekil 5.22. Atkinson Ölçüsünün İndeks Grafiği

Atkinson değerleri içerisinde diğerlerine göre farklılık gösteren ve $A_i^2 > 10$ kriter değerini aşan etkili gözlemin olmadığı şekil 5.22.' de görülmektedir.

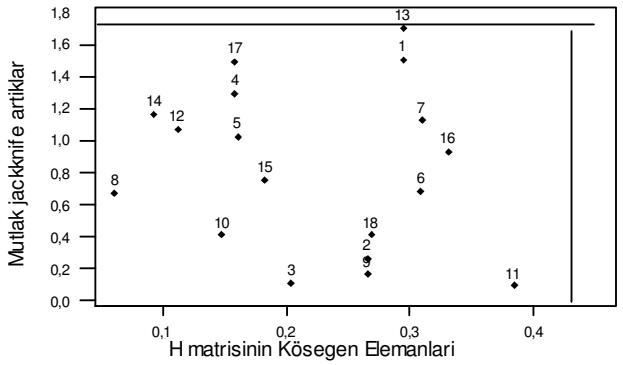
Tanısal istatistikler ve bu istatistik değerlerine karşı çizilen indeks grafikleri incelendiğinde veri üzerinde etkili olabilecek gözlemlerin bulunmadığı söylenebilir.

Veri içerisinde etkili gözlemlerin olup olmadığını artıkların ve iz düşüm matrisinin köşegen elemanlarına dayalı tanısal grafikler ile araştırıldığında aşağıdaki grafikler elde edilir..



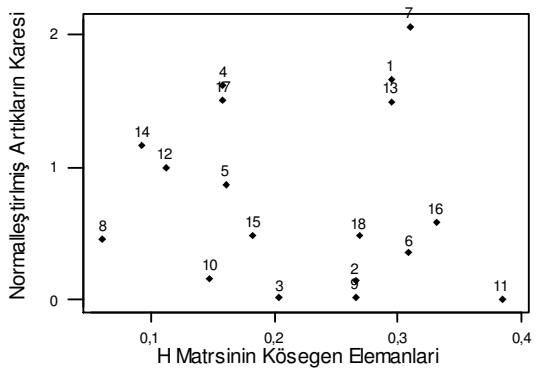
Şekil 5.23. Silinmiş Artık Grafiği

Silinmiş Artık Grafiği incelendiğinde bütün gözlemler $y = x$ doğrusu üzerinde yer alıp, merkezi örneklemden uzakta herhangi bir gözlemin var olmadığı yani veri içerisinde aykırı değerlerin bulunmadığı söylenebilir.



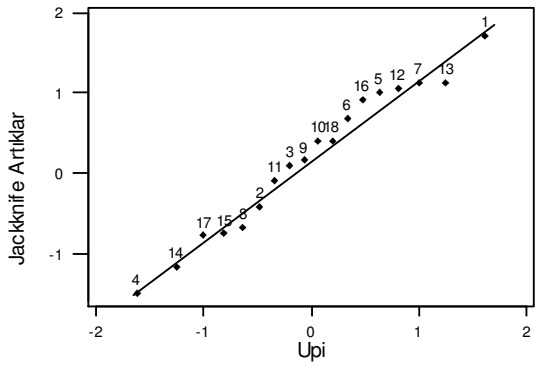
Şekil 5.24. Williams Grafiği

Williams Grafiğine göre h_{ii} değerlerinden, kriter değeri $2m/n = 0,44$ ' u aşan gözlemin olmadığı yani veride çekim gücü yüksek gözlemin bulunmadığı, Jackknife artıkları arasında ise $t_{0,95}(18-4-1) = 1,771$ tablo değerini geçen gözlemin bulunmadığı böylelikle veride aykırı değerin var olmadığı söylenebilir.



Şekil 5.25. Pregibon Grafiği

Aykırı değerler ve çekim gücü yüksek gözlemi analizden çıkardıktan sonra Pregibon grafiğini incelediğimizde şüpheli gözlemlerin veri içerisinde yer almadığı, bütün gözlemlerin rasgele dağıldığı görülmektedir.



Şekil 5.26. Jackknife Artıkların Rankit Q-Q Grafiği

Şekil 5.26.' da ki Jackknife artıklarının Rankit Q-Q grafiği incelendiğinde bütün gözlemler doğru üzerinde yer alıp, bu doğru dışında herhangi bir gözlem bulunmadığından veri içerisinde aykırı değerin varlığından söz edilemeyeceği söylenebilir.

3 gözlem çıkarıldıktan sonra elde edilen 18 gözlemli veri setine uygulanan EKK sonuçları, bilinen tanısal istatistikler, indeks grafikleri, artıkları ve iz düşüm matrisinin köşegen elemanlarını temel alan tanısal grafikler incelendiğinde veri seti içerisinde aykırı değer veya çekim gücü yüksek gözlem tespit edilmemiştir. Böylece analiz tamamlanmıştır.

6. SONUÇ VE ÖNERİLER

Bu çalışma, etkili gözlemlerin tahmin ediciler üzerindeki etkisini dikkate alarak, bu değerlerin tespiti üzerinde durumaktadır. Bir veri setinde etkili gözlemlerin EKK tahmin yöntemine göre hesaplanan tanısal istatistikler ile elde edilebildiği gibi, artıklara ve iz düşüm matrisinin köşegen elemanlarına dayalı tanısal grafikler yardımı ile bu gözlemlerin görsel olarak daha kolay ve daha hızlı tespit edildiği incelenmiştir.

İkinci bölümde doğrusal regresyon modeli, model varsayımları, veride karşılaşılabilecek sorunlara genel olarak değinilmiş, model parametrelerinin tahmini ve regresyon uyumunun değerlendirilmesi incelenmiştir. Üçüncü bölümde etkili gözlemlerin belirlenmesi üzerinde durulmuş, 4. bölümde ise etkili gözlemlerin tespiti amacıyla kullanılan yöntemler ve bu yöntemlerin dışında Silinmiş artık grafiği, Williams grafiği, Pregibon grafiği ve Rankit Q-Q tanısal grafikleri ile bu gözlemlerin daha net tespit edilebileceği anlatılmıştır.

Uygulamada bir fabrikada nitrik asit amonyağının yükseltgenmesi sırasında amonyağın sızmasını etkileyen nedenlerin ölçümlerinden oluşan bir veri seti kullanılmıştır. Bu veri setine öncelikle EKK yöntemi uygulanmış ve bilinen tanısal istatistikler, indeks grafikleri ile 3 etkili gözlem tespit edilmiştir. Artıklara ve iz düşüm matrisinin köşegen elemanlarına dayalı tanısal grafiklerde (Silinmiş artık grafiği, Williams grafiği, Pregibon grafiği ve Rankit Q-Q grafiği) incelendiğinde aynı etkili gözlemlerin daha net ve daha kolay tespit edildiği görülmüştür. Bulunan 3 etkili gözlem veriden çıkarılarak analiz tekrarlanmış ve elde edilen sonuçlar incelendiğinde çıkardığımız gözlemlerin veride anlamlı bir etki yapmadığı söylenmiştir. Gözlemleri çıkardıktan sonra veri içersinde etkili gözlem olup olmadığını araştırmak için yine tanısal istatistiklere ve tanısal grafiklere bakıldığında tahminleri etkileyebileceği düşünülen gözlemlerin bulunmadığı görülmüştür.

Sonuç olarak, veride etkili gözlemlerin bulunması durumunda hata dağılımının normalliği bozulabilir, tahminler yanlı ve büyük varyanslı olabilir. Bu sorunları ortadan kaldırmak için etkili gözlemlerin tespit edilerek veriden çıkarılması gereklidir. Bu gözlemlerin tespiti içinse bilinen tanısal istatistikler yerine tanısal grafikler (Silinmiş artık grafiği, Williams grafiği, Pregibon grafiği ve Rankit Q-Q grafiği) yardımı ile etkili gözlemleri daha net gösteren bir sunum elde edilebilmektedir.

Bu çalışmada, veri içerisinde problem yaratan etkili gözlemler tespit edildikten sonra gözlemlerin veri setinden çıkartılarak analizin yapılması önerilmiştir. Ancak bu tip gözlemlerin analizden çıkartılması her zaman doğru sonuçlar vermeyebilir. Çünkü bu gözlemler veri içerisinde önemli bilgi taşıyabilirler ve bunların veriden atılması bilgi eksikliğine neden olabilir. Bu yüzden bu tip gözlemleri veriden çıkarmak yerine, bunların parametre tahminleri üzerindeki etkilerini azaltan yöntemlere başvurulabilir. Bunlardan en bilineni Ağırlıklı En Küçük Kareler yöntemidir. Eğer etkili gözlemlerin, ölçüm veya kaydetme hatalarından dolayı ortaya çıktığı belirlenmiş ise, bu tip gözlemlerin analizden çıkarılması yerine hataların düzeltilerek, yani doğru veriler girilerek, analize devam edilmesi önerilebilir.

KAYNAKLAR

Atkinson, A.C., 1985. Atkinson plots, Transformations and Regression: An Introduction to Graphical Methods of Diagnostic Regression Analysis, Claredon Press, Oxford.

Cook. R.D.; Weisberg S., 1982. Residuals and Influence in Regression. School of Statistics University of Minnesota, New York-London.

Chatterjee, S.; Hadi, A.S., 1988. Sensitivity Analysis in Linear Regression, Wiley, New York.

Draper, N.R.; Smith, H., 1981. Applied Regression Analysis. 2.ed. New York; John Wiley&Sons.

Dodge, Y.; Hadi, A.S., 1999. Simple Graphs and Bounds for the Elements of the Hat Matrix, J. Appl. Statist. 26, 817-823.

Hadi, A.S.; Simonoff, J.S., 1993. Procedures for the Identification of Multiple Outliers in Linear Models, J. Am. Statist. Assoc, 88, 1264-1272.

Hoaglin, D.C.; Welsch, R.E., 1978. The Matrix in Regression and ANOVA, Am. Statist. 32, 17-22.

İpek, O., 2002. Regresyonda Etkin Gözlemlerin Ortaya Çıkartılmasında Kullanılan İstatistiksel Ölçütler.

Kleinbaum, D.G.; Kupper, L.L., 1978. Applied Regression Analysis and Other Multivariable Methods. University of North California, Massachusetts.

Montgomery, D.C.; Peck, E.A., 1992. Introduction to Linear Regression Analysis. John Wiley and Sons, Inc. Canada.

Meloun, M.; Militky, J., 2001. Detection of Single Influential Points in OLS Regression Model Building. Analitica Chemica Acta, April, 169-191.

Pregibon, D., 1981. Logistic Regression Diagnostic, Ann. Statist. 9, 45-52.

Rousseeuw, P.J., 1984. " Least Median of Squares Regression", Journal of the American Statistical Association, 79, 871-880.

Rousseeuw, P.J.; Leroy, A.M., 1987. Robust Regression and Outliers Detection, John Wiley, New York.

Rousseeuw, P.J.; Zomeron van B.C., 1990. Unmasking Multivariate Outliers and Leverage Points, J. Am. Statist. Assoc. 85, 871-880.

Schimek, Michael, G., 1999. Smoothing and Regression: Approaches, Computation, and Application. John Wiley and Sons, Inc. New York.

Williams, D.X., 1973. Letter to the editor, Appl. Statist. 22, 407-408.

Weisberg, S., 1980. Applied Linear Regression, John Wiley&Sons, New York.

EKLER

Ek 1. 21 gözlemli nitrik asit veri seti

No	X1	X2	X3	Y
1	80	27	89	42
2	62	24	87	28
3	62	24	93	19
4	58	18	80	14
5	58	18	82	11
6	50	18	86	7
7	50	20	80	9
8	80	27	88	37
9	62	22	87	18
10	62	24	93	20
11	58	18	83	14
12	58	19	93	12
13	50	19	72	8
14	56	20	82	15
15	75	25	90	37
16	62	23	87	18
17	58	23	87	15
18	58	17	88	13
19	58	18	89	8
20	50	19	79	8
21	70	20	91	15

Ek 2. Tüm veriler olduğunda EKK Yöntemi gözlem analiz sonuçları

No	Klasik Artık	Standartlaştırılmş Artık	Silinmiş Artık	Jackknife Artık	Normalleştirlmiş Artık
1	**2,62132**	**0,99947**	**0,99944**	**0,99940**	**0,82535**
***2**	**5,70785**	**1,92542**	**2,11240**	**2,31755**	**1,79718**
3	**-1,71857**	**-0,61710**	**-0,60550**	**0,59411**	**-0,54111**
4	**1,40391**	**0,48735**	**0,47614**	**0,46518**	**0,44204**
5	**-1,07156**	**-0,36348**	**-0,35401**	**0,34478**	**-0,33739**
6	**1,90004**	**0,65102**	**0,63961**	**0,62839**	**0,59825**
7	**-0,53041**	**-0,18425**	**-0,17893**	**0,17376**	**-0,16700**
8	**-2,64095**	**-1,02222**	**-1,02366**	**1,02510**	**-0,83153**
9	**-1,43527**	**-0,46445**	**-0,45346**	**0,44274**	**-0,45191**
10	**-0,71857**	**-0,25802**	**-0,25081**	**0,24380**	**-0,22625**
11	**2,19070**	**0,73803**	**0,72775**	**0,71761**	**0,68977**
12	**1,38489**	**0,49927**	**0,48795**	**0,47689**	**0,43605**
13	**-2,20007**	**-0,89757**	**-0,89217**	**0,88680**	**-0,69272**
14	**1,55220**	**0,50877**	**0,49739**	**0,48625**	**0,48873**
15	**4,44205**	**1,54164**	**1,61257**	**1,68677**	**1,39863**
16	**-2,86371**	**-0,93954**	**-0,93611**	**0,93270**	**-0,90167**
17	**-2,90244**	**-0,98863**	**-0,98793**	**0,98723**	**-0,91386**
18	**3,93045**	**1,39399**	**1,43698**	**1,48131**	**1,23755**
19	**-2,23572**	**-0,76847**	**-0,75882**	**0,74929**	**-0,70394**
20	**-0,36423**	**-0,12512**	**-0,12144**	**0,11787**	**-0,11468**
***21**	**-6,45190**	**-2,37518**	**-2,81900**	**3,34576**	**-2,03146**

* : Tanısal araçlar ile etkili gözlem olabileceği belirlenen gözlemler

Ek 2(devamı). Tüm veriler olduğunda EKK Yöntemi gözlem analiz sonuçları

No	hii	Di	DFFITS	Ai
1	0,318140	0,116520	0,68268	1,40733
2	0,128855	0,137088	0,81242	1,83750
3	0,231194	0,028630	-0,33204	0,67165
4	0,177400	0,012805	0,22111	0,44535
5	0,138494	0,005310	-0,14194	0,28499
6	0,155631	0,019530	0,27460	0,55617
7	0,178556	0,001845	-0,08342	0,16701
8	0,338352	0,133588	-0,73202	1,51123
9	0,053337	0,003038	-0,10764	0,21665
10	0,231194	0,005005	-0,13754	0,27562
11	0,126596	0,019737	0,27706	0,56323
12	0,237289	0,019387	0,27217	0,54837
13	*0,404434	0,136772	-0,73520	1,50653
14	0,077343	0,005425	0,14401	0,29023
15	0,177004	0,127787	0,74784	1,61265
16	0,079076	0,018949	-0,27431	0,56344
17	0,145617	0,041645	-0,40786	0,84022
18	0,211933	0,130645	0,74520	1,58365
19	0,160968	0,028324	-0,33237	0,67659
20	0,160026	0,000746	-0,05301	0,10606
21	0,268562	*0,517845	*-1,70816	*4,17948

Ek 3. 2, 13 ve 21. gözlemler çıkarıldığında EKK yöntemi gözlem analiz sonuçları

No	Klasik Artık	Standartlaştırılmş Artık	Silinmiş Artık	Jackknife Artık	Normalleştirlmiş Artık
1	2,45411	1,53247	1,61872	1,50982	1,28690
2	-0,71529	-0,43764	-0,42463	0,26201	-0,37509
3	0,20343	0,11956	0,11527	0,11114	0,10668
4	-2,42456	-1,38609	-1,43798	1,29181	-1,27140
5	1,77855	1,01780	1,01921	1,02062	0,93264
6	1,12679	0,71068	0,69753	0,68462	0,59087
7	-2,73189	-1,72415	-1,87203	1,12463	-1,43256
8	-1,29557	-0,70117	-0,68785	0,67478	-0,67938
9	0,28471	0,17419	0,16804	0,16210	0,14929
10	0,76144	0,43261	0,41969	0,40716	0,39929
11	-0,14640	-0,09782	-0,09429	0,09090	-0,07677
12	1,90448	1,05992	1,06499	1,07008	0,99868
13	2,45411	1,53247	1,61872	1,70982	1,22066
14	-2,06344	-1,13620	-1,14913	1,16221	-1,08204
15	-1,33389	-0,77361	-0,76193	0,75043	-0,69947
16	1,45933	0,93563	0,93118	0,92675	0,76525
17	-2,42456	-1,38609	-1,43798	1,49181	-1,22474
18	0,70866	0,43455	0,42160	0,40903	0,69282

Ek 3(devamı). 2, 13 ve 21. gözlemler çıkarıldığında EKK yöntemi gözlem analiz sonuçları

No	hii	Di	DFFITS	Ai
1	**0,294744**	**0,195370**	**0,926450**	**2,06792**
2	**0,265341**	**0,017293**	**-0,255194**	**-0,46324**
3	**0,203879**	**0,000915**	**0,058335**	**0,10522**
4	**0,158561**	**0,090510**	**-0,624221**	**-1,21153**
5	**0,160260**	**0,049425**	**0,445251**	**0,83414**
6	**0,308681**	**0,056379**	**0,466098**	**0,85585**
7	**0,309574**	**0,194578**	**-0,933530**	**1,40886**
8	**0,061096**	**0,007998**	**-0,175463**	**-0,32203**
9	**0,265341**	**0,002740**	**0,100986**	**0,18225**
10	**0,148055**	**0,008131**	**0,174960**	**0,31754**
11	**0,384034**	**0,001491**	**-0,074454**	**-0,13427**
12	**0,112140**	**0,035473**	**0,378489**	**0,71147**
13	**0,294744**	**0,205370**	**0,926450**	**2,06792**
14	**0,092975**	**0,033082**	**-0,367911**	**-0,69613**
15	**0,182402**	**0,033379**	**-0,359882**	**-0,66311**
16	**0,330982**	**0,108271**	**0,654961**	**1,21949**
17	**0,158561**	**0,090510**	**-0,624221**	**-1,21153**
18	**0,268631**	**0,017340**	**0,255509**	**0,46376**

ÖZGEÇMİŞ

Adı Soyadı : Yeşim AYDIN
Doğum Yeri : Samsun
Doğum Yılı : 1980
Medeni Hali : Bekar

Eğitim ve Akademik Durumu:

Lise : 1994 - 1998 Tülay Başaran Anadolu Lisesi
Lisans : 1999 – 2003 Ondokuzmayıs Üni. Fen Edebiyat Fak. İstatistik Bölümü
Yabancı Dil : İngilizce

İş Tecrübesi:

2006 - Bugün Türkiye İstatistik Kurumu