BÖLÜM 11

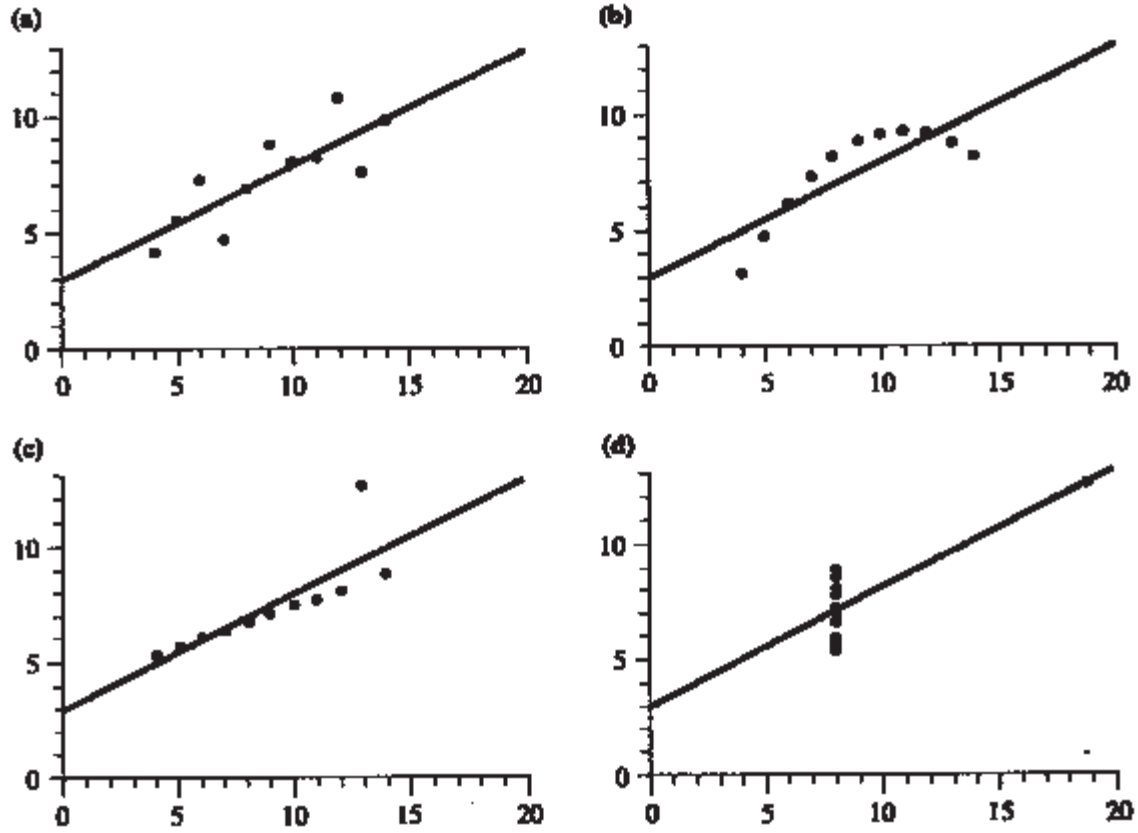
REGRESYON TANILARI

GİRİŞ

- Regresyon tanıları, regresyondaki problemleri belirlemeye dönük genel teknik sınıfı içermektedir.
- Bu problemler model ya da veri seti ile ilgilidir.
- Bu da konusu ile ilgili bir çok güncel yayın olan aktif bir araştırma alanıdır.
- Önerilen teknikler arasında en kullanışlı olanının belirli olmadığı görülmektedir.
- Bu Bölümde çalışmalarda tercih edilen bazı basit teknikler sunulmaktadır.
- Belsley, Kuh ve Welsch (1980) ve Cook ve Weisberg (1982) teori ile ilgili daha ayrıntılı bilgi ve tanı teknikleri metotları ile ilgili incelenebilir.

ARTIK ANALİZİ

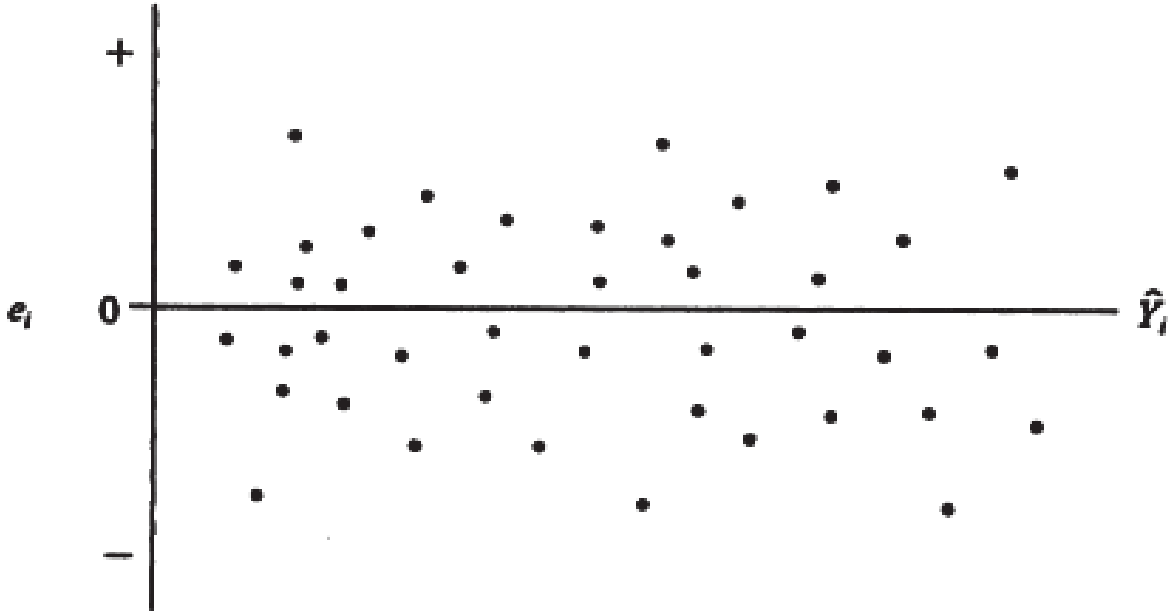
Regresyon artıkları analizi ya da bazı artık dönüşümleri, modeldeki ya da veri setindeki problemlerin yetersizliklerini belirlemede çok kullanışlıdır. Regresyon modelindeki doğru hatalar normal ve sıfır ortalamalı ve ortak varyansa sahip bağımsız dağılan rassal değişkenlerdir $\epsilon \sim N(\mathbf{0}, \boldsymbol{I}\sigma^2)$. Gözlemlenen artıklar, bağımsız değildir, $\boldsymbol{I}\sigma^2$ varsayımı geçerli olsa bile ortak varyansa sahip değildir. Sıradan en küçük kareler varsayımı altında, $\mathcal{E}(\boldsymbol{e}) = \mathbf{0}$ ve $\mathbf{Var}(\boldsymbol{e}) = (\boldsymbol{I} - \boldsymbol{P})\sigma^2$ ile $\boldsymbol{e} = (\boldsymbol{I} - \boldsymbol{P})\boldsymbol{Y}$ çok değişkenli normal dağılıma sahiptir. $\mathbf{Var}(\boldsymbol{e})$'nin köşegen elemanları eşit değildir, öyleyse gözlemlenen kalıntıların ortak varyansı yoktur, köşegen dışı elemanlar sıfırdan farklıdır, böylece bağımsız değildir.



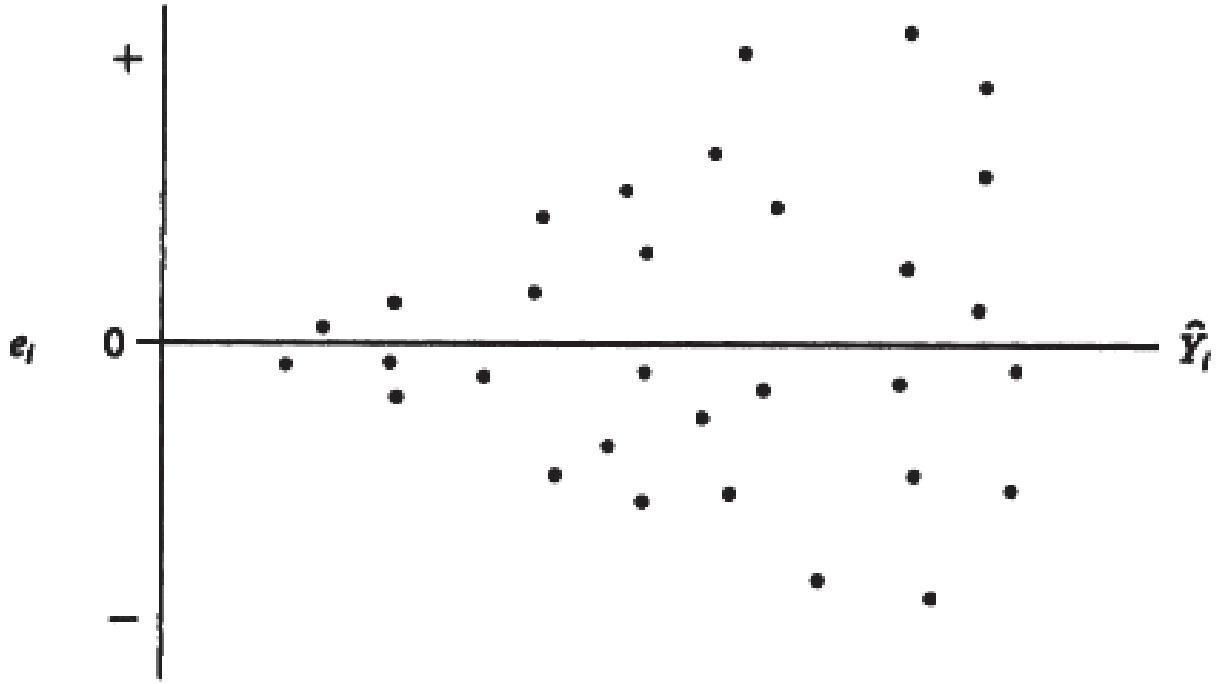
ŞEKİL 11.1. *Y nin X üzerindeki doğrusal regresyonları dört veri seti için benzer sayısal sonuçları vermektedir. [Anscombe (1973)'den uyarlanmıştır.]*

$\widehat{Y}$'ye karşı e Grafiği

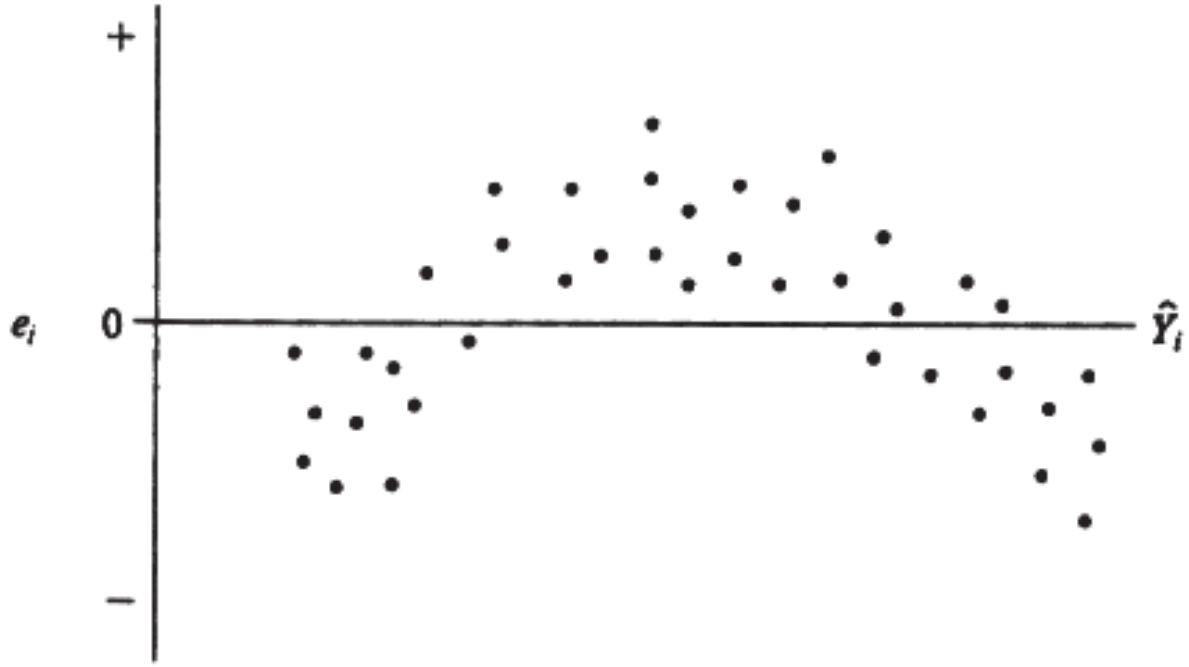
Bağımlı değişkenin kestirilmiş değerlerinin artıklara karşı grafiği önem taşımaktadır. Noktaların $e = 0$ çizgisinin üstündeki ve altındaki rassal dağılımı, eğer varsayımlar sağlanıyorsa hemen hemen tüm noktalar bant içinde olacak şekilde, $e = \pm 2s$ ile gösterilmektedir (Şekil 11.2). $\widehat{Y}$ burada Y yerine kullanılmaktadır çünkü e, $\widehat{Y}$'ye diktir fakat Y ye değildir. e nin Y karşısındaki grafiği bu diklik yokluğu yokluğu nedeniyle desen gösterecektir.



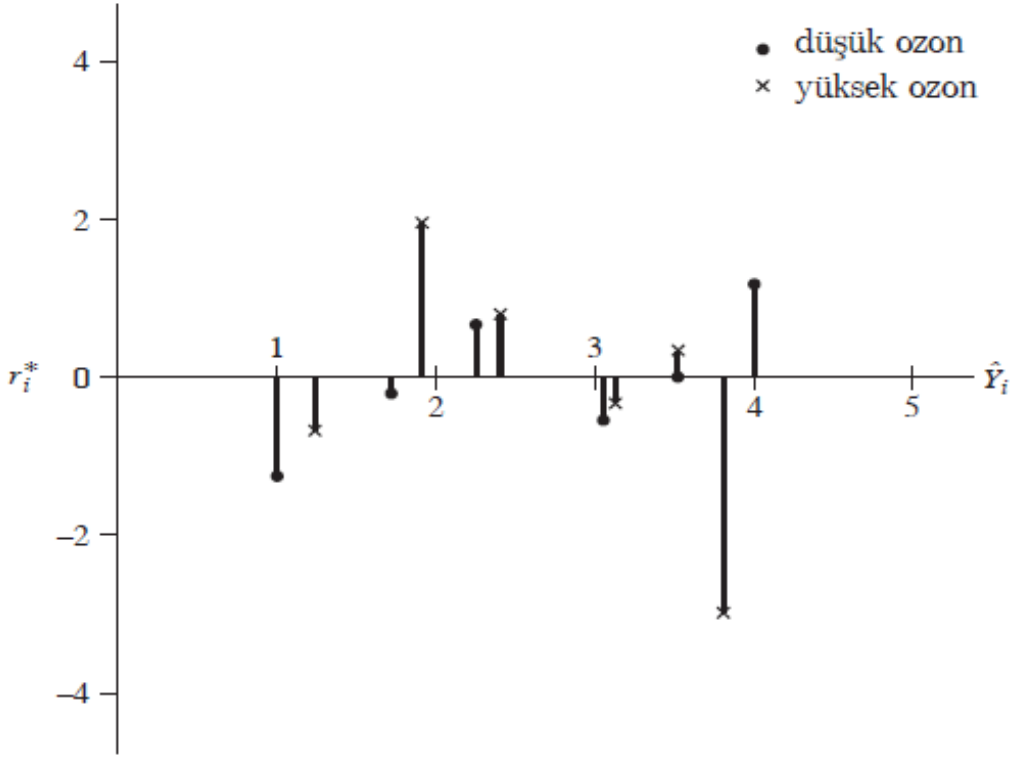
ŞEKİL 11.2. *Varsayımlar karşılığında e nin $\hat{Y}$ karşısında beklenen tipik davranışı*



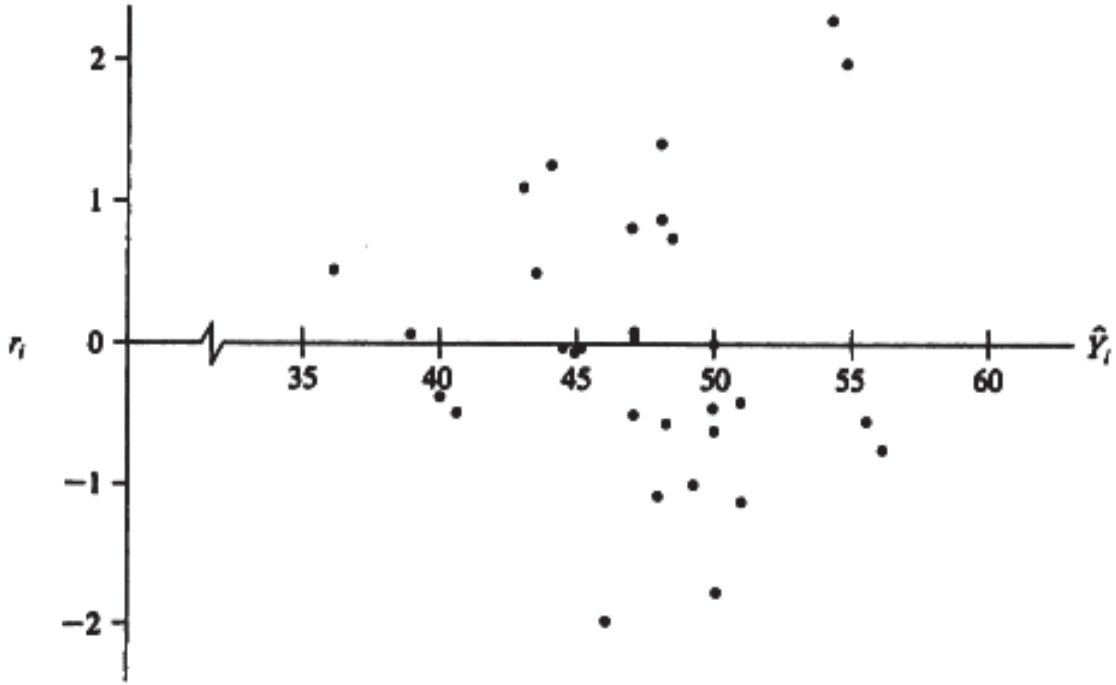
ŞEKİL 11.3. *e nin* $\hat{Y}$ *karşısında, artan yayılım (büyük varyans) gösterdiğindeki grafiği.*



ŞEKİL 11.4. *Artıkların asimetrik deseni $\hat{Y}$ karşısında çizilmektedir. Model burada önemli bir bağımsız değişkeni kaçırmaktadır. Karesel terim büyük ihtimalle modelde yer almamaktadır.*



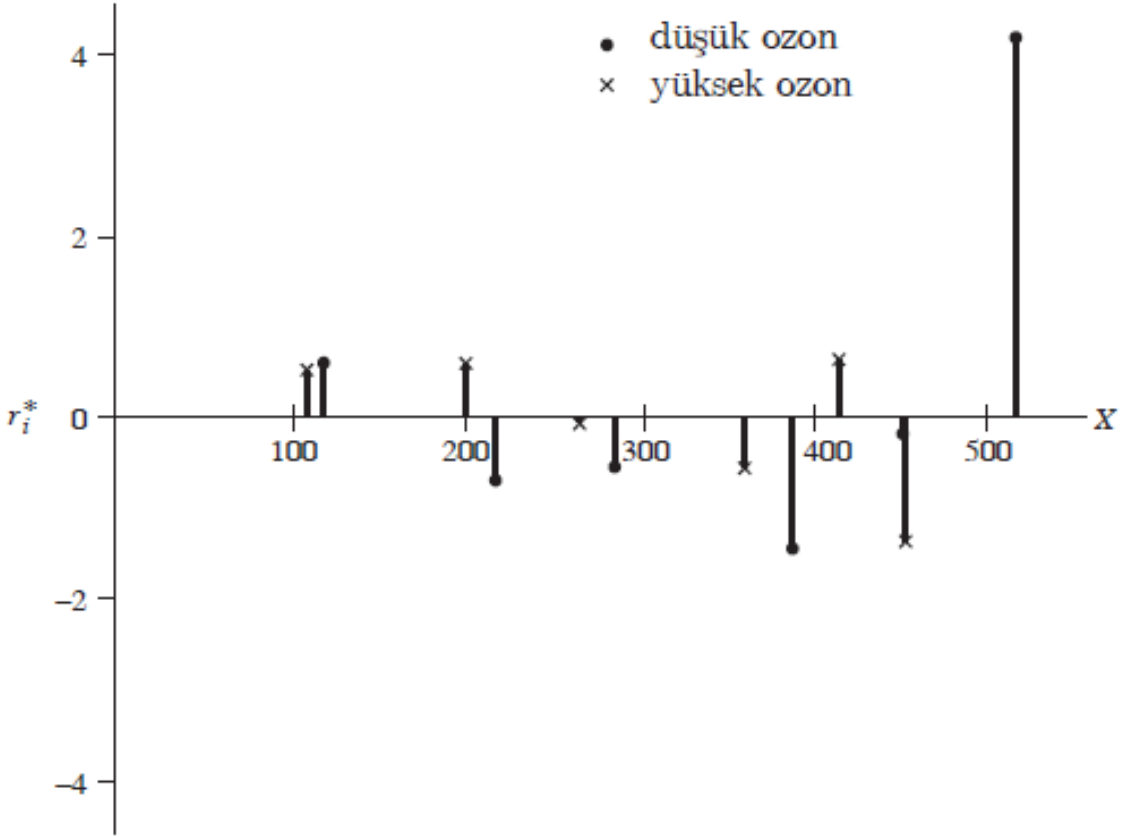
ŞEKİL 11.5. *Lesser–Unsworth veri seti için r_i^* ile $\hat{Y}_i$ grafiği (Alıştırma 1.19). İki farklı ozon düzeyine maruz kalan soyafasülyesi tohumu ağırlığının kümülatif güneş yenilemesi üzerindeki etkisi. Model (tohum ağırlığı)$^{1/2}$ nin ozon düzeyi ve güneş yenilemesi üzerindeki doğrusal regresyonunu içermektedir.*



ŞEKİL 11.6. *Zaman içinde alınan oksijen, dinlenme kalpatış hızı, koşma kalpatış hızı ve maksimum kalpatış hızı regresyonu için* r_i *ve* $\widehat{Y}_i$ *grafiği. Orjinal veri seti Tablo 4.3 te yer almaktadır.*

X_i'ye karşı e Grafiği

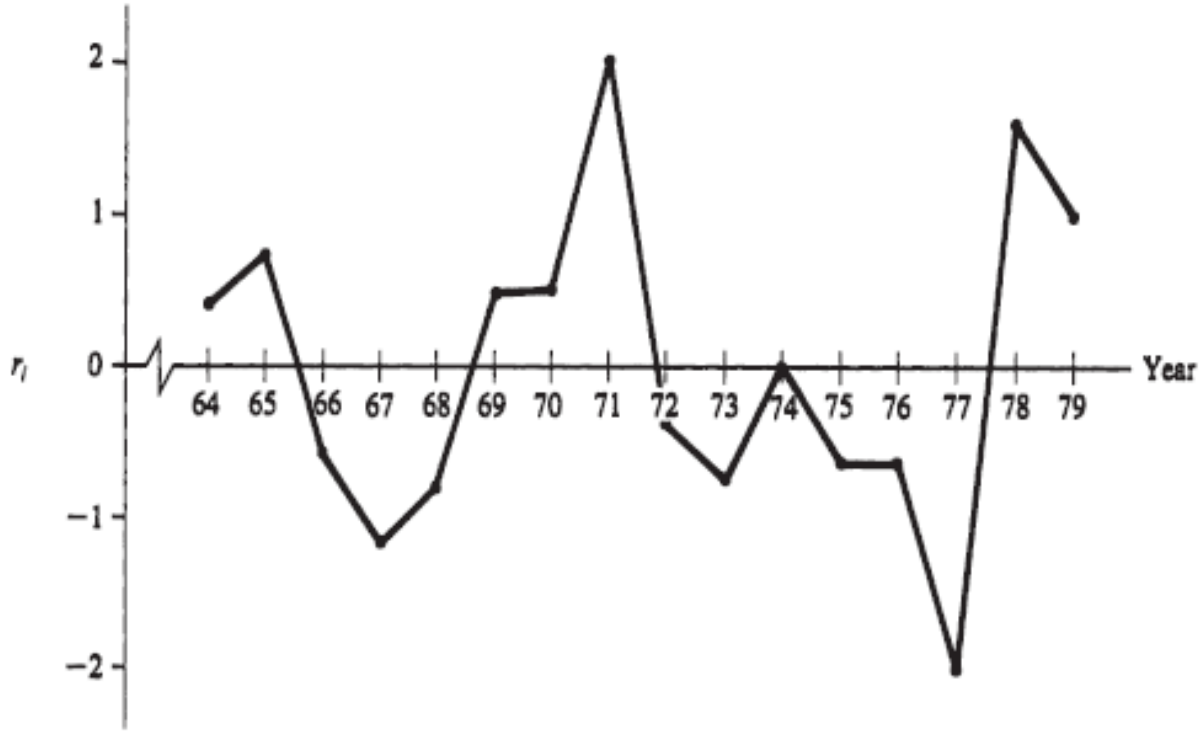
Bağımsız değişkenlere karşı artıkların çizimleri $\hat{Y}$ ye karşı çizimleriyle benzer yorumlara sahiptir. Dağılımdaki farklar heterojen varyanslara işaret etmektedir. Bu grafiklerde eksik yüksek dereceden polinom terimi bağımsız değişken için bu grafiklerde bellidir. Halbuki, bir değişkenle ilişkili modeldeki yüksek dereceden polinom terimi eksikliği gibi yetersizlikler diğer bağımsız değişkenlerin dağılımı ve etkileri ile belirsiz hale gelebilir. Kısmi regresyon kaldıraç grafikleri (Kısım 11.1.6'da tartışılmaktadır), bazı bağımsız değişkenler ilave edildiğinde daha açıklayıcı hale gelmektedir.



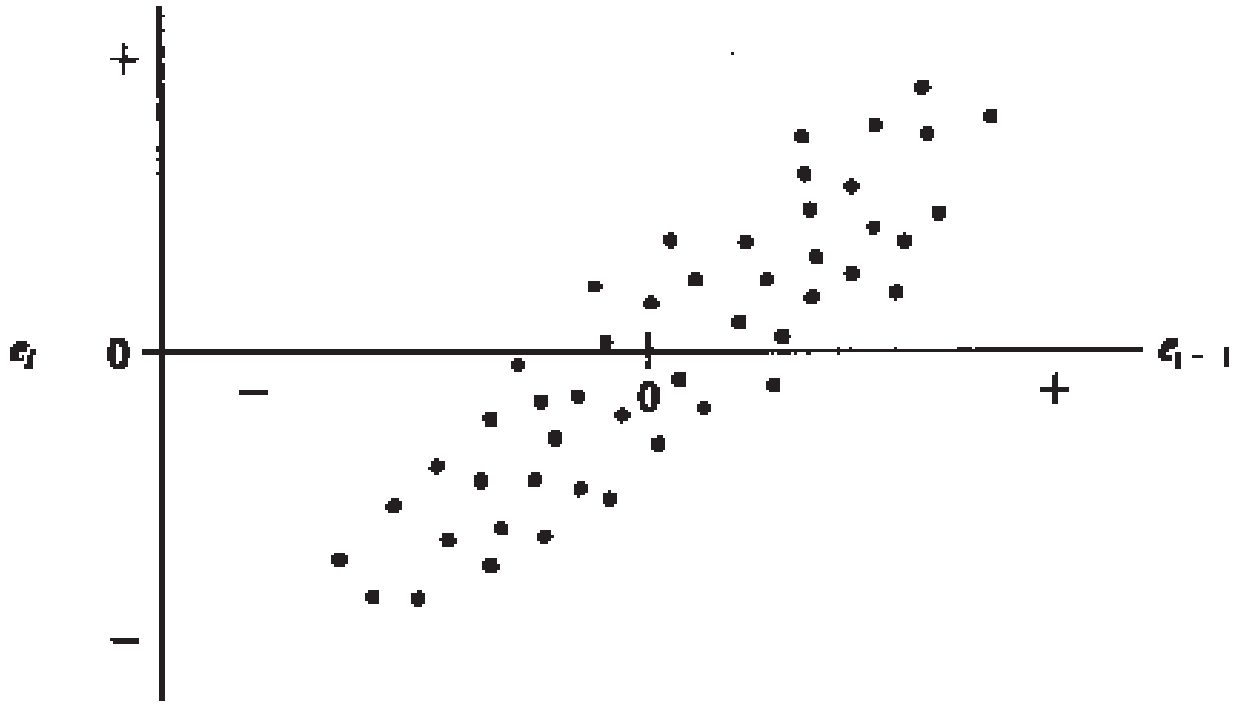
ŞEKİL 11.7. *Lesser–Unsworth veri seti için Studentize kalıntılarla ışınım (X) grafiği. Kalıntılar tohum ağırlığının ozon seviyesi ve kümülatif güneş ışıması üzerine regresyonundan elde edilmektedir.*

e‘nin ZAMANA GÖRE GRAFİKLERİ

- Bireysel gözlemsel birimler üzerine toplanan veri seti genellikle serisel korelasyon gösteren artıklara sahip olacaktır.
- Zaman içindeki bir noktadaki artık geçmiş artıklara bağlı olacaktır.
- Sürecin sürekli izlemesi ile elde edilen veri seti gibi klasik zaman serileri, halihazırda korelasyon gösteren ve korelasyon göstermesi beklenen artıklar olarak adlandırılır.
- Zaman serileri modelleri ve analizleri bu serisel korelasyonları göz önüne almalı ve bu durumlarda kullanılmalıdır (Fuller, 1996; Bloomfield, 1976).



ŞEKİL 11.8. *r_i İle yıllık ringabalığı yakalama regresyonu için yakalama yılı grafiği. [Veri seti Nelson ve Ahrenholz (1986) den elde edildi.]*

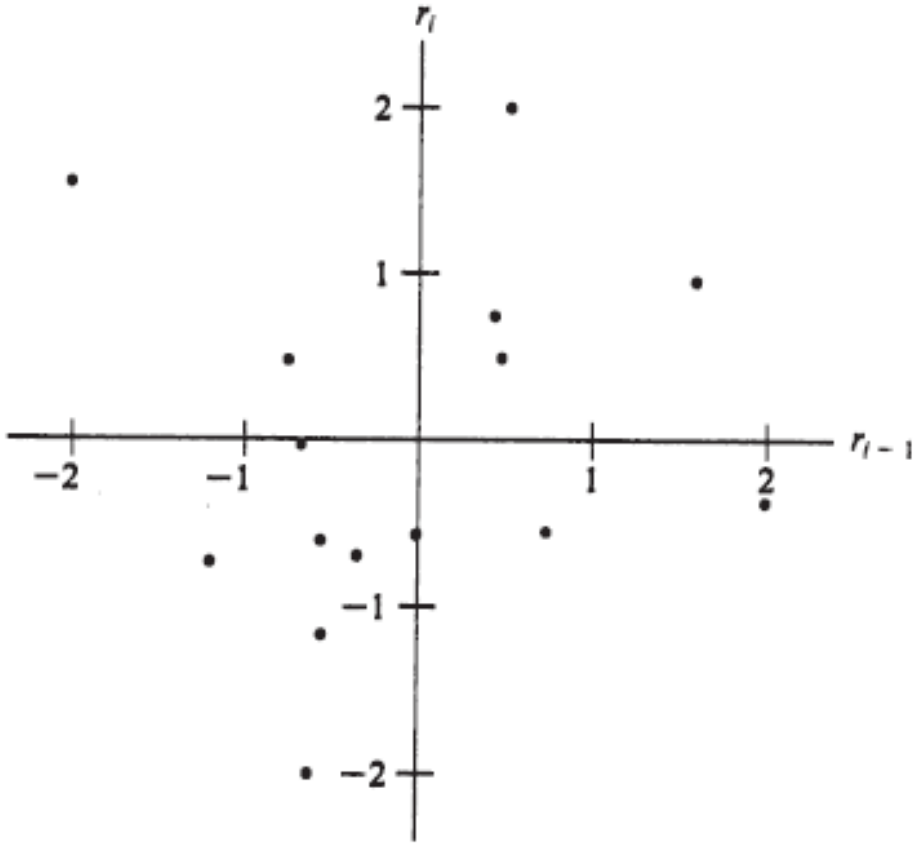


ŞEKİL 11.9. *Kalıntılar arasında pozitif serisel korelasyona işaret eden* e_i *ile* e_{i-1} *grafiği.*

e_i'ye karşı e_{i-1} Grafikleri

Veri setindeki serisel korelasyon, önceki kalıntılarla her bir kalıntının net bir şekilde çizimini ortaya koymaktadır. Pozitif serisel korelasyon Şekil 11.9'daki gibi net pozitif eğimle beraber dağılım noktaları üreteceklerdir.

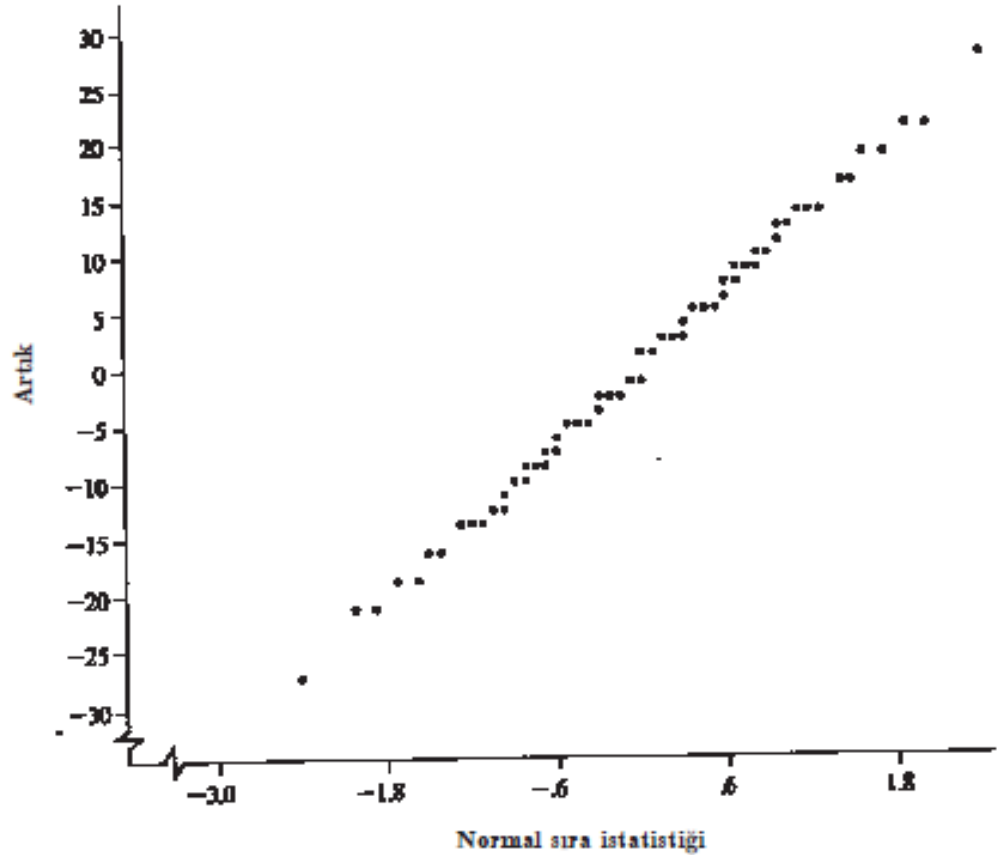
Şekil 11.10'da Menhaden veri seti için r_i nin r_{i-1} e karşı grafiği yer almaktadır. Yukarı sol tarafta yer alan kadranda yer alan uç nokta en büyük ikinci positif kalıntı (1978) ya karşı en büyük negatif kalıntıdır (1977). 1977'den 1978'e yakalamada ani kayış büyük ölçüde serisel korelasyondan sorumludur. Yine de pozitif serisel korelasyon belirgindir.



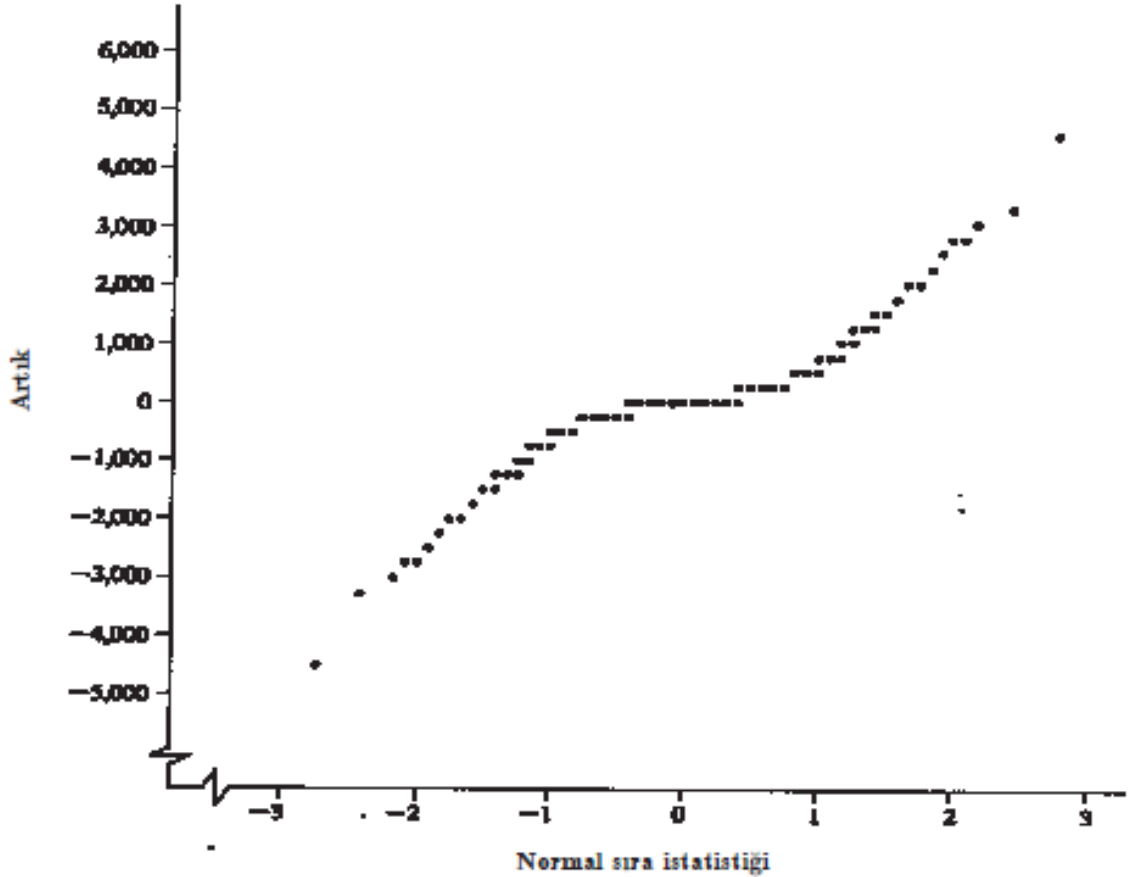
ŞEKİL 11.10. *Menhaden yakalama veri seti için r_i nin r_{i-1} ye karşı grafiği. Kalıntılar, yıllık yakalamanın yakalama yılına regresyonlarından elde edilmiştir.*

NORMAL OLASILIK GRAFİKLERİ

- Normal olasılık grafiği normal olmamayı yakalamak için dizayn edilmiştir.
- Sıralanmış kalıntıların uygun örneklem büyüklüğü için normal sıralama istatistiğine karşı grafiğidir.
- Normal sıralama istatistiği sıfır ortalama ve birim varyansa sahip normal dağılımdan elde edilen sıralanmış gözlemlerin beklenen değerleridir.



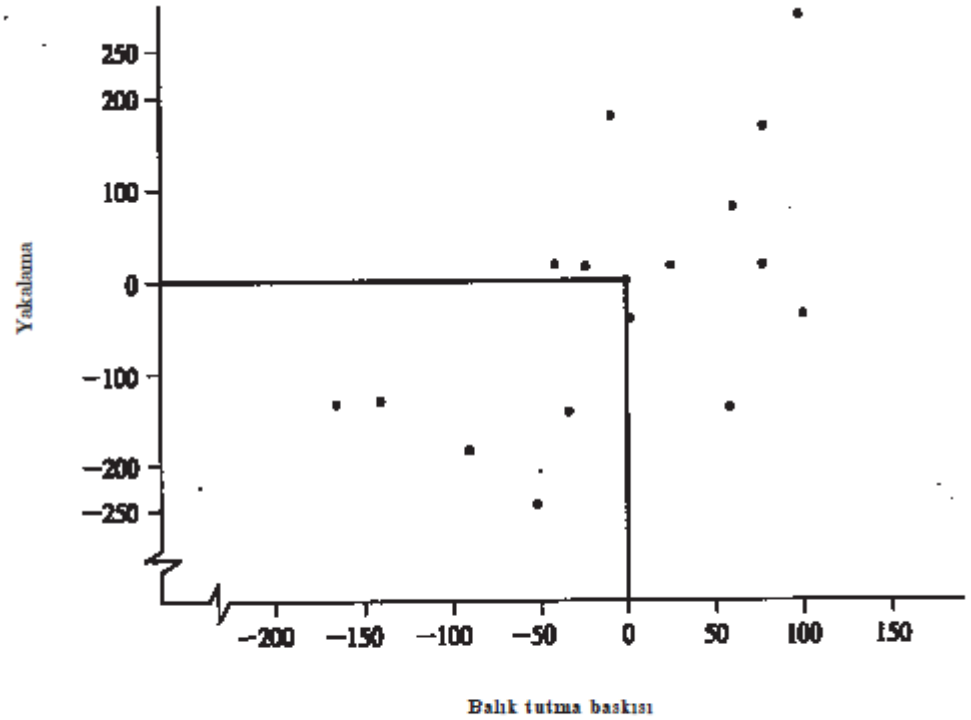
ŞEKİL 11.11. *Tütünde mavi küf hastalığı ile ilgili bir çalışmada, nihai bitki ağırlıklarının varyansı analizinde normal artık grafiği. (M. Moss ve C. C. Main, North Carolina State Üniversitesi.)*



ŞEKİL 11.12. *Tipik büyük kuyruklar normal olasılık grafiği. Bu durumda, S-şekli veri setindeki heterojen varyanotan ortaya çıkar.*

KISMİ REGRESYON KALDIRAÇ GRAFİKLERİ

Bazı bağımsız değişkenler dahil edildiğinde, kalıntıların bir bağımsız değişkenle ilişkisi diğer değişkenlerin etkileri ile belirsizleşebilir. **Kısmi regresyon kaldıraç grafikleri** diğer değişkenlerin şaşırtıcı etkilerini kaldırmak için bir teşebbüstür. $e(j)$ bağımlı değişkenin j'nci dışında diğer tüm bağımsız değişkenler üzerine regresyon eşitliği tahmin edilmesi sonucu elde edilen artıklar göstermektedir. $e(j)$ ile $u(j)$ ye karşı grafiği j'nci değişken için kısmi regresyon kaldıraç grafiğidir. $e(j)$ ve $u(j)$ nin her ikisi de digger tüm bağımsız değişkenler için modele uyarlanmıştır.



ŞEKİL 11.13. *Yıllık ringa balığı yakalamanın tekne sayısı ve balık tutma çabası üzerine regresyonundan elde edilen kısmi regresyon kaldıracı grafiği. [Veri seti Nelson ve Ahrenholz (1986)'dan alınmıştır.]*

ETKİ İSTATİSTİKLERİ

Potansiyel olarak etkili noktalar ya da yüksek kaldıraca sahip noktalar X-uzayında örneklem noktaları bulutunun kenarında yer alır. Projeksiyon matrisinin $\boldsymbol{P}$ (bazı kaynaklarda şapma matris olarak adlandırılır) i'nci diagonal elemanı v_{ii} X-uzayının merkezinden i'nci veri noktasının uzaklığı ile ilgilidir. Bu uzaklık ölçütü, örneklem noktalar bulutunun toplam şeklini göz önüne almaktadır. Örneğin veri setlerinin eliptik bulutu tarafındaki veri noktası, merkezi noktadan benzer uzaklıktaki fakat eliptik bulutun ana eksenindeki merkezi diğer veri noktasına göre daha büyük v_{ii} 'ye sahip olacaktır. $\boldsymbol{P}$ nin i'nci diagonal elemanı şu şekildedir:

$$v_{ii} = \boldsymbol{x}_i'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_i, \tag{11.10}$$

COOK'UN D'Sİ

Cook'un *D'si* (Cook, 1977; Cook ve Weisberg, 1982) belirli bir gözlem ihmal edildiğinde, $\widehat{\boldsymbol{\beta}}$'daki kaymayı ölçmek için oluşturulmuştur. O gözlemin tüm regresyon katsayıları üzerindeki etkisinin birleştirilmiş ölçütüdür. Cook'un D'si şu şekilde tanımlanmaktadır:

$$D_i = \frac{(\widehat{\boldsymbol{\beta}}_{(i)} - \widehat{\boldsymbol{\beta}})'(\boldsymbol{X}'\boldsymbol{X})(\widehat{\boldsymbol{\beta}}_{(i)} - \widehat{\boldsymbol{\beta}})}{p's^2}. \tag{11.12}$$

Sayısal olarak, D_i daha kolay bir şekilde elde edilebilir:

$$D_i = \frac{r_i^2}{p'}\left(\frac{v_{ii}}{1 - v_{ii}}\right), \tag{11.13}$$

r_i burada standartlaştırılmış kalıntılardır ve v_{ii} tam regresyondan hesaplanan $\boldsymbol{P}$'nin i'nci köşegen elemanıdır. Dikkat edilirse eğer standartlaştırılmış kalıntı geniş ve veri noktası $\boldsymbol{X}$-uzayının merkezine uzaksa D_i büyüktür.

DFFITS

Eşitlik 11.13, i'nci gözlem $\boldsymbol{\beta}$ tahmininde kullanılmadığında, Cook'un D'sinin $\widehat{\boldsymbol{Y}}$ de kayma ile ilgili bir ölçüt sağlamaktadır. Daha yakından ilgili bir ölçüt DFFITS (Belsley, Kuh ve Welsch 1980) tarafından önerilmiştir ve şu şekilde tanımlanmaktadır:

$$\begin{aligned} \text{DFFITS}_i &= \frac{\widehat{Y}_i - \widehat{Y}_{i(i)}}{s_{(i)}\sqrt{v_{ii}}} \\ &= \left(\frac{v_{ii}}{1-v_{ii}}\right) \frac{e_i}{s_{(i)}(1-v_{(ii)})^{1/2}}, \end{aligned} \tag{11.15}$$

$\widehat{Y}_{i(i)}$ burada i nci gözlem için tahmin edilen ortalamadır fakat i nci gözlem $\boldsymbol{\beta}$ tahmininde kullanılmamıştır. Dikkat edilirse σ burada $s_{(i)}$ ile beraber tahmin edilmiştir ve σ tahmini i nci gözlem haricinde elde edilmiştir. $s_{(i)}$ regresyonu yeniden yapmadan aşağıdaki ilişki kullanılarak elde edilmiştir

$$(n-p'-1)s^2_{(i)} = (n-p')s^2 - \frac{e_i^2}{1-v_{ii}}. \tag{11.16}$$

DFFITS ile Cook'un D si arasındaki ilişki şu şekildedir:

$$D_i = (\text{DFFITS}_i)^2 \left(\frac{s^2_{(i)}}{p's^2}\right). \tag{11.17}$$

DFBETAS

Cook'un D_i'si i nci gözlemin tahmin edilen regresyon katsayıları vektörü üzerindeki etkisini ortaya çıkarmaktadır. Bireysel regresyon katsayılar için etkili gözlemler $\text{DFBETAS}_{j(i)}$, $j = 0, 1, 2, \ldots, p$ (Belsley, Kuh ve Welsch, 1980) tarafından belirlenmiştir. Burada i nci gözlem analizden çıkartıldığında, her bir $\text{DFBETAS}_{j(i)}$ $\widehat{\beta}_j$ deki stardartlaşmış değişimdir. Böylece:

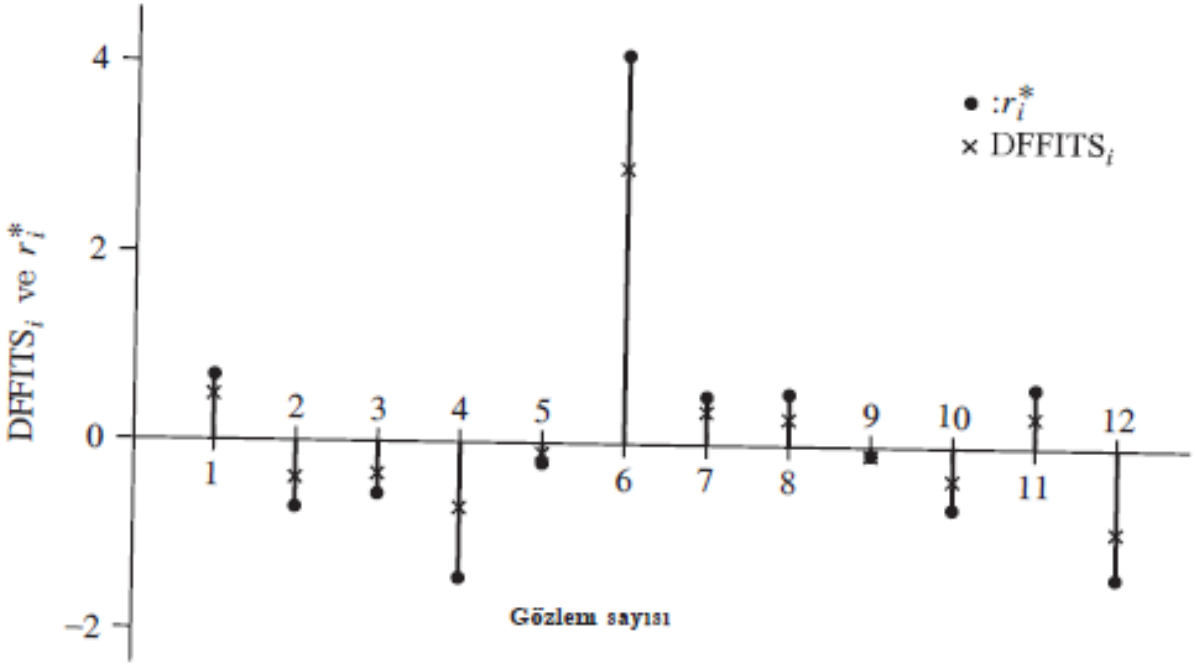
$$\text{DFBETAS}_{j(i)} = \frac{\widehat{\beta}_j - \widehat{\beta}_{j(i)}}{s_i\sqrt{c_{jj}}}, \tag{11.19}$$

c_{jj}, $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ dan elde edilen $(j+1)$ nci köşegen elemanlarıdır. Formül DFFITS_i kadar basit değilse de $\text{DFBETAS}_{j(i)}$ orjinal gregresyon sonuçlarından da hesaplanabilir. Okuyucu Belsley, Kuh ve Welsch (1980)'yi detaylar için inceleyebilir.

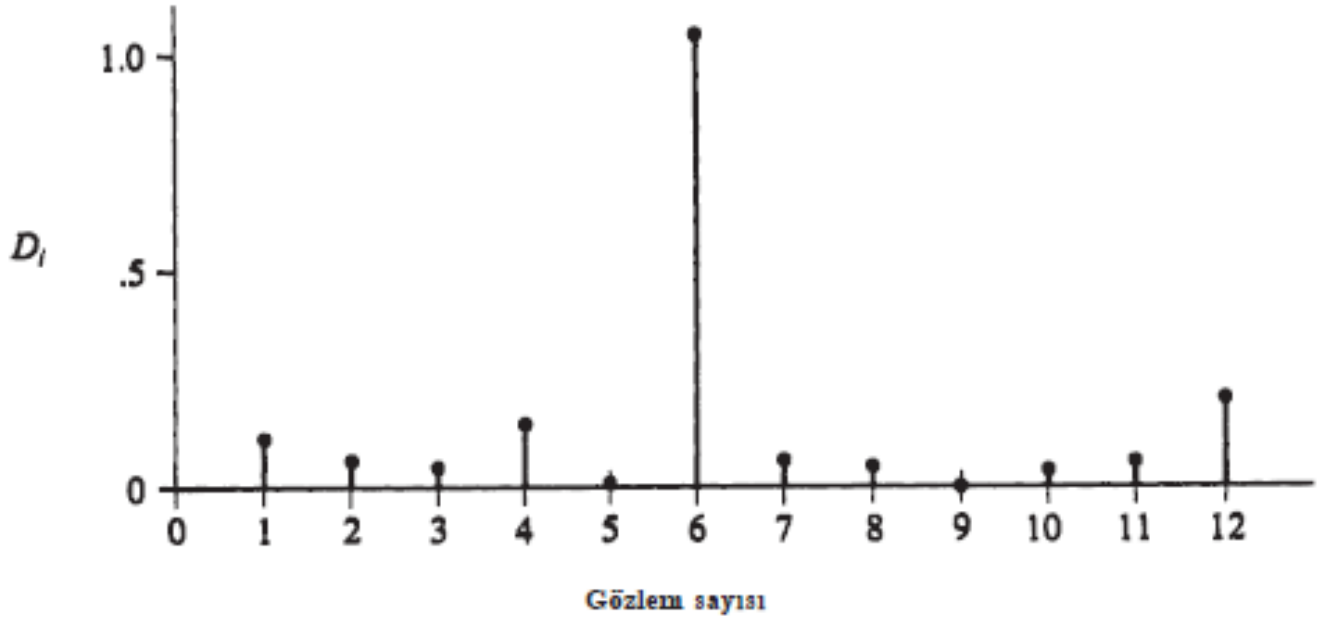
COVRATIO

i nci gözlemin tahmin edilen regresyon katsayılarından elde edilen varyans kovaryans matrisine etkisi, iki varyans kovaryans matrisinin belirleyicilerinin oranı tarafından ölçülmektedir. Belsley, Kuh ve Welsch (1980) bunu şu şekilde formüle etmektedir:

$$\begin{aligned} \text{COVRATIO} &= \frac{det(s^2_{(i)}[\boldsymbol{X}'_{(i)}\boldsymbol{X}_{(i)}]^{-1})}{det(s^2[\boldsymbol{X}'\boldsymbol{X}]^{-1})} \\ &= \left[\left(\frac{n-p'-1}{n-p'} + \frac{r_i^{*2}}{n-p'}\right)^p (1-v_{ii})\right]^{-1}. \quad (11.20) \end{aligned}$$



ŞEKİL 11.14. *Studentize kalıntılar ve $DFFITS_i$ ozon seviyesi üzerine tohum ağırlığı ve Lesser–Unsworth verisi kullanılarak kümülatif güneş ışığı regresyon eşitliği gözlem sayısına karşı çizilmiştir.*

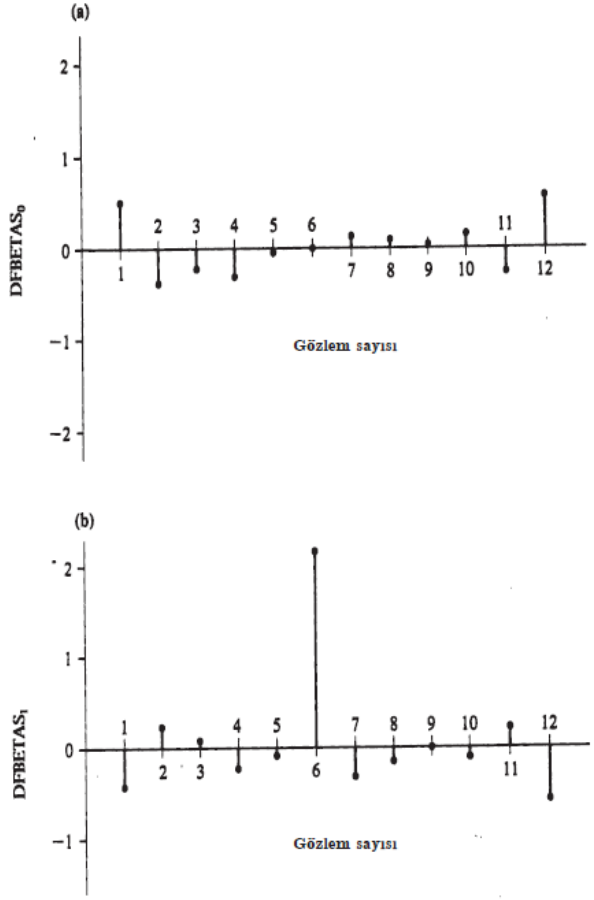


ŞEKİL 11.15. *Cook'un D_i, gözlem sayısına karşı çizilmiştir. Gözlem sayısı tohum ağırlığının ozon seviyesi ve kümülatif güneş ışıması regresyon eşitliğinden elde edilmiştir. Veri seti Lesser–Unsworth'dan alınmıştır.*

11.2.5 Etki Ölçütleri Özeti

Aşağıdaki tablo etki ölçütlerini özetlemektedir.

Etki Ölçütü	***Formül***	***i Gözlemi Etkili Olabilir Eğer:***
Cook's D_i	$\frac{(\widehat{\boldsymbol{\beta}}_{(i)}-\widehat{\boldsymbol{\beta}})'(\boldsymbol{X}'\boldsymbol{X})(\widehat{\boldsymbol{\beta}}_{(i)}-\widehat{\boldsymbol{\beta}})}{p's^2}$	$D_i > F_{(.5,p',n-p')}$
DFFITS$_i$	$\frac{\widehat{Y}_i-\widehat{Y}_{i(i)}}{s_{(i)}\sqrt{v_{ii}}}$	$\lvert\text{DFFITS}_i\rvert > 2\sqrt{p'/n}$
Atkinson's C_i	$\left(\frac{n-p'}{p'}\right)^{1/2}\lvert\text{DFFITS}_i\rvert$	$\lvert C_i\rvert > 2[(n-p')/n]^{1/2}$
DFBETAS$_{j(i)}$	$\frac{\widehat{\beta}_j-\widehat{\beta}_{j(i)}}{s_i\sqrt{c_{jj}}}$	$\lvert\text{DFBETAS}_{j(i)}\rvert > 2/\sqrt{n}$
COVRATIO$_i$	$\frac{\det(s^2_{(i)}[\boldsymbol{X}'_{(i)}\boldsymbol{X}_{(i)}]^{-1})}{\det(s^2[\boldsymbol{X}'\boldsymbol{X}]^{-1})}$	COVRATIO $\begin{cases} < 1-3p'/n \\ > 1+3p'/n \end{cases}$



ŞEKİL 11.16. *$DFBETAS_{0(i)}$ ve $DFBETAS_{1(i)}$ gözlem sayısına karşı çizilmiştir. Gözlem sayısı tohum ağırlığının ozon seviyesi üzerine ve kümülatif güneş ışıması regresyonu ile elde edilmiştir. Lesser–Unsworth veri seti kullanılmıştır.*

DOĞRUSAL BAĞIMLILIK TANILARI

- Regresyonda doğrusal bağımlılık sorunu, X matrisinin sütunları arasında yaklaşık tekillik (near-singularity) varsa, X sütunun doğrusal kombinasyonları sıfıra yakınsa yaratılan problem kümesine işaret etmektedir.

- Bu da bağımsız değişkenler arasında yakın fazlalıklara işaret etmektedir.

- Özellikle aynı bilgi birden çok biçimde sağlanmıştır.

- Geometrik olarak X uzayının en az bir boyutu veri noktaları arasında neredeyse hiç saçılma olmaması anlamında zayıf olarak tanımlandıysa ortaya çıkmaktadır.

KOŞUL SAYISI VE KOŞUL ENDEKSİ

$\boldsymbol{X}$ matrisinin koşul sayısı $K(\boldsymbol{X})$ en büyük tekil değerin en küçük tekil değere oranı olarak tanımlanmıştır (Belsley, Kuh ve Welsch, 1980).

$$K(\boldsymbol{X}) = \left[\frac{\lambda_{max}}{\lambda_{min}}\right]^{1/2}. \tag{11.22}$$

Koşul sayı $\boldsymbol{X}$ ya da $\boldsymbol{Y}$'deki küçük değişmelerin normal eşitliklerin çözümüne duyarlılığı ile ilgili bir ölçüt sunmaktadır. Büyük koşul sayı, tekilliğe yakınlığın matrisin zayıf bir şekilde koşullandığına işaret etmektedir. Eğer tüm sütünlar ikili dik ve birim uzunluğa sahip olacak şekilde ölçeklendirilmişlerse, matrisin koşul sayısı 1'dir (tüm λ_k bire eşittir).

Koşul sayı kavramı, X- uzayının her bir (temel bileşenler) boyutu için koşul sayısı sağlayacak şekilde genişletilmiştir. X uzayının k ncı temel bileşen boyutu için δ_k koşul endeksi şu şekilde tanımlanmaktadır

$$\delta_k = \left[\frac{\lambda_{max}}{\lambda_k}\right]^{1/2}. \tag{11.23}$$

TABLO 11.1. *Sayısal alıştırma için tekil değerler ve koşul endeksleri.*

Temel Bileşen	*Tekil Değerler*	*Keşul Endeksi*
1	1.7024	1.00
2	1.0033	1.70
3	0.3083	5.52
4	0.0062	273.60

VARYANS ŞİŞİRME FAKTÖRÜ

Doğrusal bağımlılığın bir başka yaygın ölçütü j nci regresyon katsayısı için **varyans şişirme faktörüdür**, VIF_j. Varyans şişirme faktörü bağımsız değişkenlerin korelasyon matrisinden, $\widehat{\boldsymbol{\rho}}$ elde edilmiştir. Böylece bağımsız değişkenler merkezileşmiş ve birim uzunluğa standartlaştırılmıştır. Köşegen elemanlar $\widehat{\boldsymbol{\rho}}^{-1}$, $\widehat{\boldsymbol{\rho}}$ nin tersi varyans şişirme faktörleridir. VIF_j ile doğrusal bağımlılık (standartlaştırılmış ve merkezileştirilmiş değişkenler) arasındaki bağlantı şu ilişki yoluyla oluşturulmaktadır:

$$VIF_j = \frac{1}{1 - R_j^2}, \tag{11.25}$$

VARYANS AYRIŞTIRMA ORANLARI

Her bir tahmin edilmiş regresyon katsayısı $\boldsymbol{X}'\boldsymbol{X}$ in özdeğerlerinin λ_k bir fonksiyonu ve özvektörlerin bir elemanı olarak ifade edilebilir.

u_{jk} k ncı özvektörün j nci elemanıdır. Böylece

$$\mathrm{Var}(\widehat{\beta}_j) = \sigma^2 \sum_k \left(\frac{u_{jk}^2}{\lambda_k} \right). \qquad (11.27)$$

$k = 1, \ldots, p'$ temel bileşen boyutları üzerine toplama yapılmıştır. Böylece her bir regresyon katsayısının varyansı, her bir temel bileşenden katkılara ayrıştırılabilir. Her bir katkının büyüklüğü (j nci regresyon katsayısının varyansı için) k ncı özvektörün u_{jk} j nci elemanının, $\lambda_k^{1/2}$ tekil değerine oranının karesi tarafından belirlenmektedir.

TABLO 11.2. *Alıştırma 11.11 için varyans ayrıştırma oranları. Tüm temel bileşenler (tablonun yukarı yarısı) ve dördüncü temel bileşen silinmiştir (alt yarısı)*

Temel	*Varyans Oranı*			
Bileşen	*Sabit*	X_1	X_2	X_3
1	.0000[a]	.0000[a]	.0000[a]	.0102
2	.0000[a]	.0000[a]	.0055	.0008
3	.0001	.0001	.0003	.6492
4	.9999	.9998	.9942	.3398
	1.0000	1.0000	1.0000	1.0000
1	.070	.060	.001	.015
2	.002	.002	.942	.001
3	.928	.939	.057	.983
	1.000	1.000	1.000	1.000

TABLO 11.3. $\boldsymbol{K}'\widehat{\boldsymbol{\beta}}$ *Doğrusal fonksiyonlar,* $\boldsymbol{K}' = (1\ 25\ 0\ 3)$ *için varyans bölmeleri ve varyans alanları.*

Temel Bileşen		*Varyans Bölmesi*	*Varyans Oranı*
1		.0451	.7542
2		.0003	.0050
3		.0142	.2375
4		.0002	.0033
	Toplam	.0597	1.0000

DOĞRUSAL BAĞIMLILIK TANILARININ ÖZETİ

Eşdoğrusallık Tanılaması	*Formül*	*Doğrusal Bağımlı Eğer*
Koşul Endeksi , δ_k	$\left[\frac{\lambda_{max}}{\lambda_k}\right]^{1/2}$	$\begin{cases} 30 \leq \delta_k \leq 100 \text{ (ılımlı)} \\ \delta_k > 100 \text{ (güçlü)} \end{cases}$
mci	$\sum_{j=1}^{p'} \left[\frac{\lambda_{p'}}{\lambda_j}\right]^2$	$\begin{cases} mci \leq 2 \\ mci \approx 1 \text{ (güçlü)} \end{cases}$
Varyans Şişirme Faktörü VIF	$\frac{1}{1-R_j^2}$	$VIF > 10$

LİNTHURST VERİ SETİ ÜZERİNE REGRESYON TANILARI

- Model oluşturma sürecinde değişkenlerin seçimini göstermek amacıyla Linthurst veri seti Bölüm 5'te kullanılmıştı.

- O örnekte, modelleme beş bağımsız değişkeni kullanıyordu: SALINITY, pH, K, Na ve Zn. Daha sonra iki değişkeni içeren model ile sonlandırılmıştı.

- En küçük kareler yönteminin sıradan varsayımları yapılmış ve tüm değişkenlerin bağımlı değişken BIOMASS ile doğrusal ilişki gösterdiği varsayılmıştır.

- Bu seçimde Linthurst veri seti için regresyon tanılamaları beş değişkenli regresyon modeli için sunulmuştur.

ARTIK GRAFİKLERİ

En küçük kareler artıkları ile tahmin edilmiş değerlerinin grafiğinin çizimi Şekil 11.17(a)'da sunulmaktadır. Buradaki şekil beş tahmin edilmiş değerin varlığına işaret etmektedir ve bunların değeri iki binden büyüktür ve diğerlerinden büyüktür. Bu noktalarla ilintili beş artıktan dördü dikkate değer değildir fakat beşinci nokta en büyük negatif artıktır, −748 ya da standartlaştırılmış artıktır, $r_{29} = 2.0804$. Şekil 11.17(a)'daki ikinci nokta dikkate değer biçimde pozitif artıklar arasında negative artıklara göre daha büyük saçılıma sahiptir. Bu da artıkların dağılımının çarpık olduğu anlamına gelmektedir. Çarpıklık kalıntıların frekans poligonundan net biçimde görülebilmektedir, Şekil 11.18 (sayfa 383). 2'den büyük dört kalıntı vardır fakat sadece bir tanesi –2'den küçüktür ve küçük negatif artıklar yüksek göreli frekansa sahiptir.

TABLO 11.4. *BIOMASS ın beş bağımsız değişken, SAL, pH, K, Na ve Zn üzerine regresyonundan elde edilen artık analizi * işareti etki ölçütünü göstermekte ve değerin referans değeri aştığına işaret eder. (SAS PROC REG den R opsiyonuyla.)*

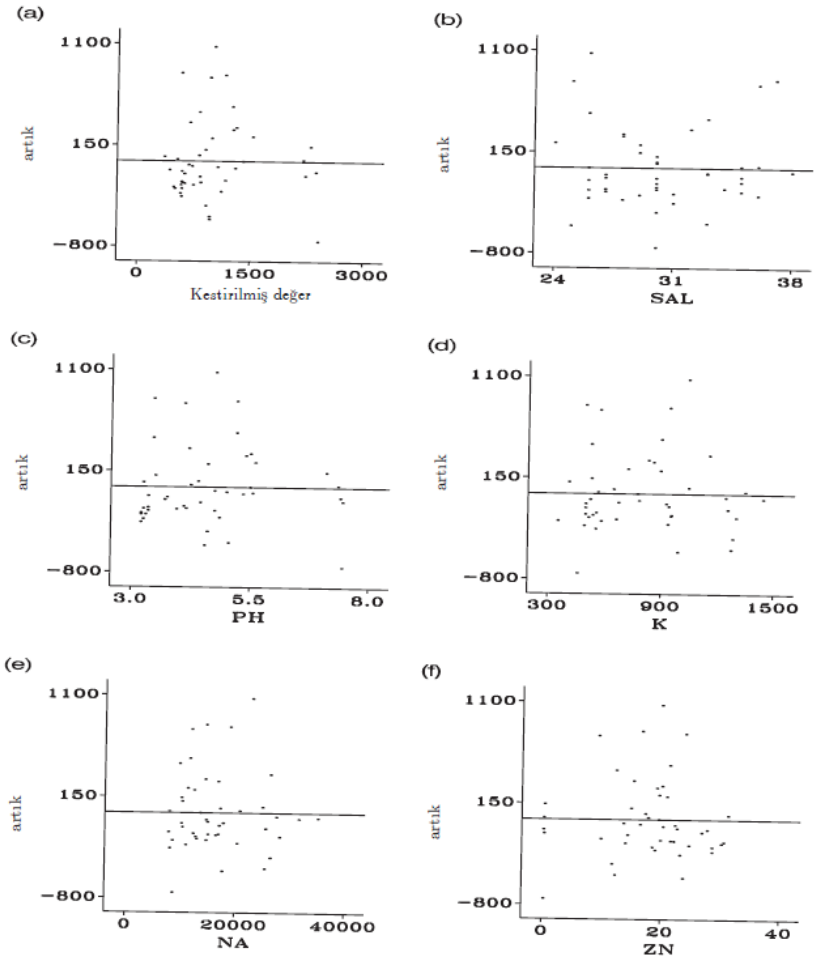
Gözlem	Y_i	$\hat{Y}_i$	$s(\hat{Y}_i)$	e_i	$s(e_i)$	r_i	*Cook's D*
1	676	724	176	−48	357	−.135	.001
2	516	740	142	−224	372	−.601	.009
3	1,052	691	127	361	378	.956	.017
4	868	815	114	53	382	.140	.000
5	1,008	1,063	321	−56	235	−.236	.017
6	436	958	126	−522	378	−1.381	.035
7	544	527	214	17	336	.050	.000
8	680	827	141	−147	373	−.394	.004
9	640	676	174	−36	358	−.101	.000
10	492	911	165	−419	362	−1.155	.046
11	984	1,166	167	−182	362	−.503	.009
12	1,400	573	147	827	370	2.232	.130*
13	1,276	816	153	460	368	1.252	.045
14	1,736	953	137	783	374	2.093	.099*
15	1,004	898	166	106	362	.293	.003
16	396	355	135	41	375	.109	.000
17	352	577	127	−225	377	−.595	.007
18	328	586	139	−258	373	−.691	.011
19	392	586	118	−194	380	−.511	.004
20	236	494	131	−258	376	−.687	.010
21	392	596	122	−204	379	−.537	.005
22	268	570	120	−302	380	−.795	.010
23	252	584	124	−332	378	−.877	.014
24	236	479	100	−243	386	−.631	.004
25	340	425	131	−85	376	−.226	.001
26	2,436	2,296	170	140	360	.388	.006
27	2,216	2,202	196	14	347	.040	.000
28	2,096	2,230	187	−134	351	−.381	.007
29	1,660	2,408	171	−748	360	−2.080	.163*
30	2,272	2,369	168	−97	361	−.270	.003
31	824	1,110	115	−286	381	−.750	.008
32	1,196	982	118	214	381	.562	.005
33	1,960	1,155	120	805	380	2.120	.075
34	2,080	1,008	124	1072	378	2.834	.145*
35	1,764	1,254	136	510	374	1.363	.041
36	412	959	111	−547	383	−1.431	.029
37	416	626	133	−210	376	−.558	.006
38	504	624	107	−120	384	−.313	.001
39	492	588	99	−96	386	−.250	.001
40	636	837	95	−201	387	−.521	.003
41	1,756	1,526	129	230	377	.610	.007
42	1,232	1,298	97	−66	386	−.171	.000
43	1,400	1,401	106	−1	384	−.004	.000
44	1,620	1,306	113	314	382	.822	.010
45	1,560	1,265	90	295	388	.759	.005

TABLO 11.5. *BIOMASS ın beş bağımsız değişken, SAL, pH, K, Na ve Zn üzerine regresyonundan elde edilen artıklar ve etki istatistikleri (SAS'ın PROC REG, INFLUENCE ile). * etki ölçütünün değerinin referans değeri üstünde olduğunu gösterir.*

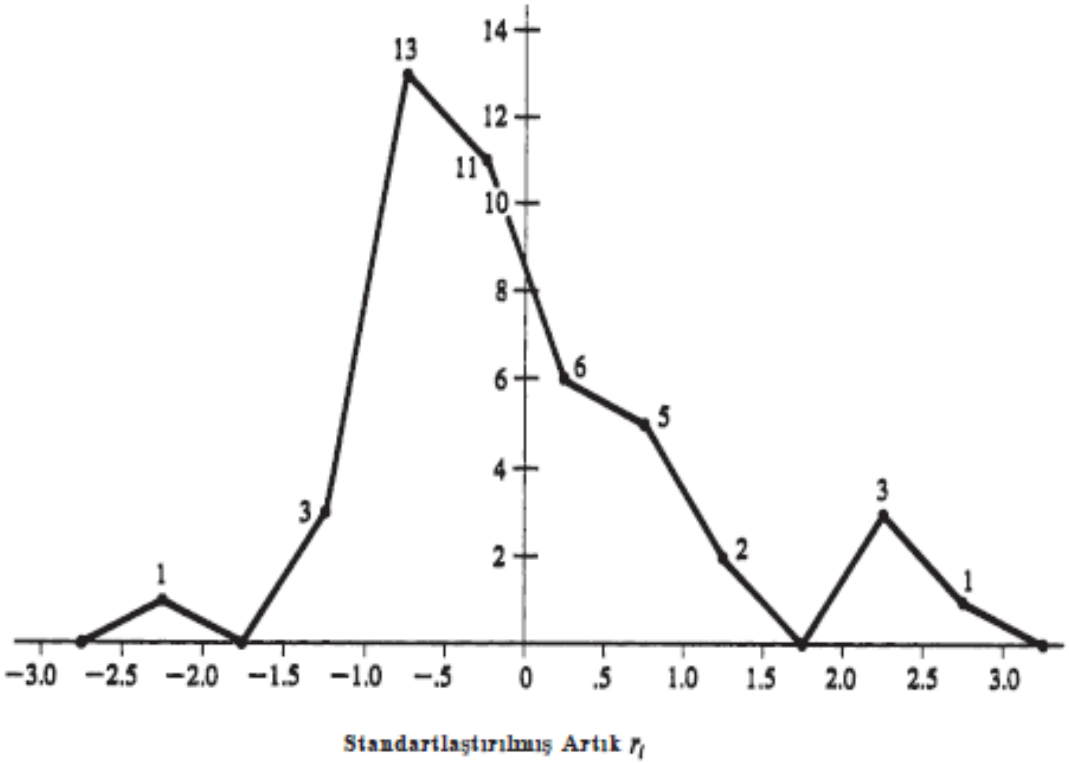
Gözlem	e_i	r_i^*	v_{ii}	*COV-RATIO*	*DF-FITS*
1	−48	−.133	.195	1.447∗	−.065
2	−224	−.596	.127	1.266	−.228
3	361	.955	.101	1.128	.321
4	53	.138	.082	1.269	.041
5	−55	−.233	.651∗	3.318∗	−.318
6	−522	−1.398	.100	.961	−.466
7	17	.050	.289∗	1.642∗	.032
8	−147	−.390	.125	1.304	−.147
9	−36	−.100	.191	1.443∗	−.049
10	−419	−1.160	.172	1.146	−.529
11	−182	−.498	.175	1.362	−.229
12	827	2.359	.135	.595∗	.934∗
13	460	1.261	.148	1.073	.526
14	783	2.193	.119	.649	.806∗
15	106	.289	.173	1.395	.132
16	41	.107	.115	1.317	.039
17	−225	−.590	.102	1.232	−.199
18	−258	−.687	.121	1.235	−.255
19	−194	−.506	.088	1.230	−.157
20	−258	−.682	.108	1.218	−.238
21	−204	−.532	.094	1.234	−.172
22	−302	−.791	.090	1.165	−.249
23	−332	−.874	.097	1.149	−.287
24	−243	−.626	.063	1.173	−.162
25	−85	−.224	.108	1.300	−.078
26	140	.384	.181	1.395	.181
27	14	.039	.243	1.543∗	.022
28	−134	−.376	.222	1.468∗	−.201
29	−748	−2.177	.184	.708	−1.034∗
30	−97	−.267	.178	1.406∗	−.124
31	−286	−.745	.083	1.168	−.224
32	214	.557	.087	1.219	.172
33	805	2.225	.091	.617	.704
34	1,072	3.140	.098	.325∗	1.032∗
35	510	1.379	.117	.988	.502
36	−547	−1.451	.078	.917	−.421
37	−210	−.553	.111	1.253	−.196
38	−120	−.309	.072	1.241	−.086
39	−96	−.247	.062	1.235	−.064
40	−201	−.516	.057	1.188	−.127
41	230	.605	.106	1.233	.208
42	−66	−.168	.060	1.237	−.043
43	−1	−.004	.070	1.257	−.001
44	314	.819	.081	1.144	.242
45	295	.755	.051	1.127	.176

TABLO 11.6. *BIOMASS ın beş bağımsız değişken, SAL, pH, K, Na ve Zn üzerine regresyonundan elde edilen etki istatistikleri (DFBETAS). (SAS PROC REG, INFLUENCE yoluyla). * işareti etki ölçütünün değerinin referans değeri üzerinde olduğunu gösterir.*

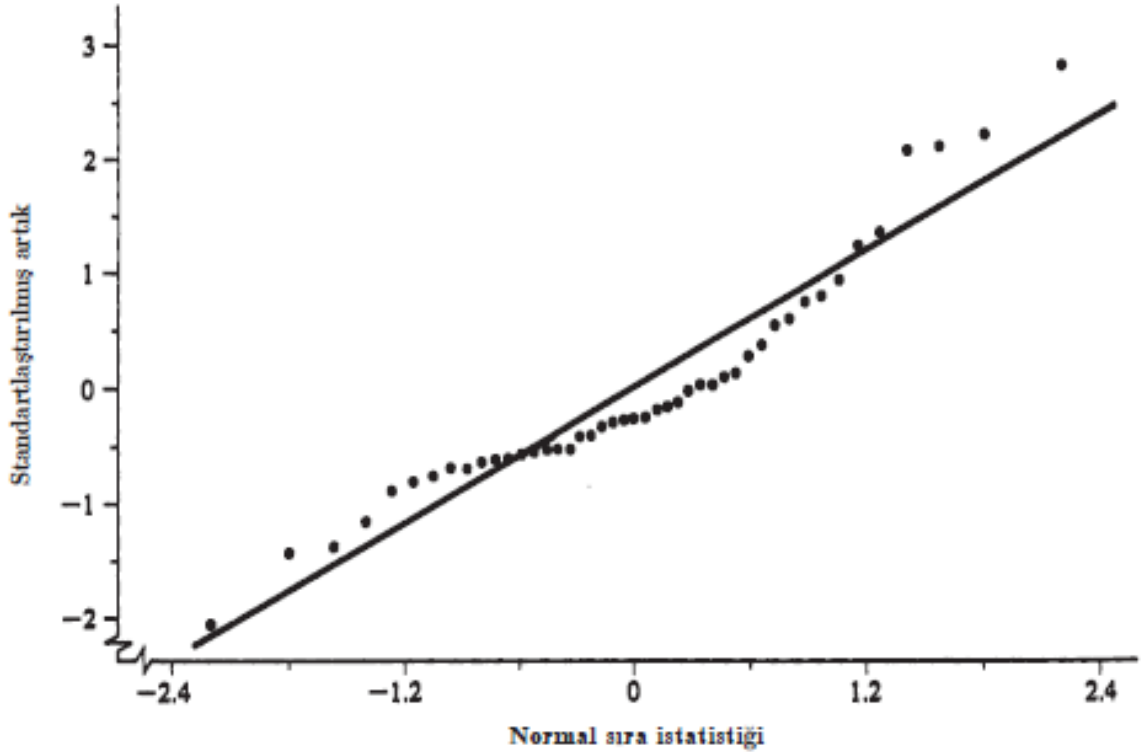
	DFBETAS					
Gözlem	X_0	*SAL*	*pH*	*K*	*Na*	*Zn*
1	.010	−.004	−.004	−.002	−.032	.001
2	.074	−.086	−.014	−.081	−.016	−.007
3	.123	−.094	−.166	−.005	.152	−.171
4	.020	−.020	−.019	−.010	.027	−.021
5	.065	−.030	−.108	.245	−.244	−.083
6	.054	−.069	.022	−.220	.007	.078
7	−.019	.022	.009	.026	−.021	.013
8	−.075	.069	.810	−.030	−.041	.091
9	.029	−.034	−.014	−.017	.004	−.014
10	−.310*	.285	.317*	−.068	−.177	.378*
11	−.174	.116	.172	.004	.022	.180
12	−.151	.442*	−.150	−.294	.092	.020
13	.307*	−.126	−.398*	−.052	−.023	−.351*
14	.133	.165	−.346*	−.041	−.090	−.331*
15	.107	−.076	−.104	−.062	.042	−.098
16	−.014	.013	.010	−.011	.005	.024
17	−.020	.027	.000	.081	−.028	−.061
18	.013	−.032	−.010	.084	.024	−.093
19	.008	−.056	.036	.041	.006	−.007
20	.043	−.118	.046	.039	.006	−.014
21	−.100	.070	.104	.106	−.084	.069
22	−.022	.012	.017	.074	.008	−.069
23	.010	−.075	.054	−.069	.163	−.044
24	.011	−.043	.030	−.014	.050	−.037
25	.041	−.057	−.012	−.007	.022	−.037
26	.074	−.074	−.006	−.047	.025	−.091
27	−.011	.012	.013	.005	−.010	.006
28	.090	−.094	−.118	.011	.037	−.042
29	−.130	.154	−.250	.235	−.010	.247
30	−.023	.026	−.024	.033	−.012	.038
31	−.141	.174	.069	.052	−.108	.097
32	−.066	.060	.059	.126	−.139	.078
33	−.044	−.179	.291	.027	.048	.249
34	.584*	−.752*	−.309*	−.183	.533*	−.406*
35	−.125	.041	.213	.307*	−.341*	.210
36	−.119	.206	.015	−.114	.039	−.002
37	.060	−.023	−.069	−.079	.076	−.119
38	−.026	.035	.020	−.023	.011	−.002
39	−.001	.009	−.001	−.015	.009	−.020
40	−.059	.065	.047	−.043	.018	.033
41	.033	−.081	.058	.017	−.044	.026
42	.010	.001	−.024	−.004	.009	−.020
43	.000	.000	−.001	−.000	.000	−.000
44	−.127	.075	.180	.080	−.105	.159
45	−.056	.013	.109	.025	−.024	.083



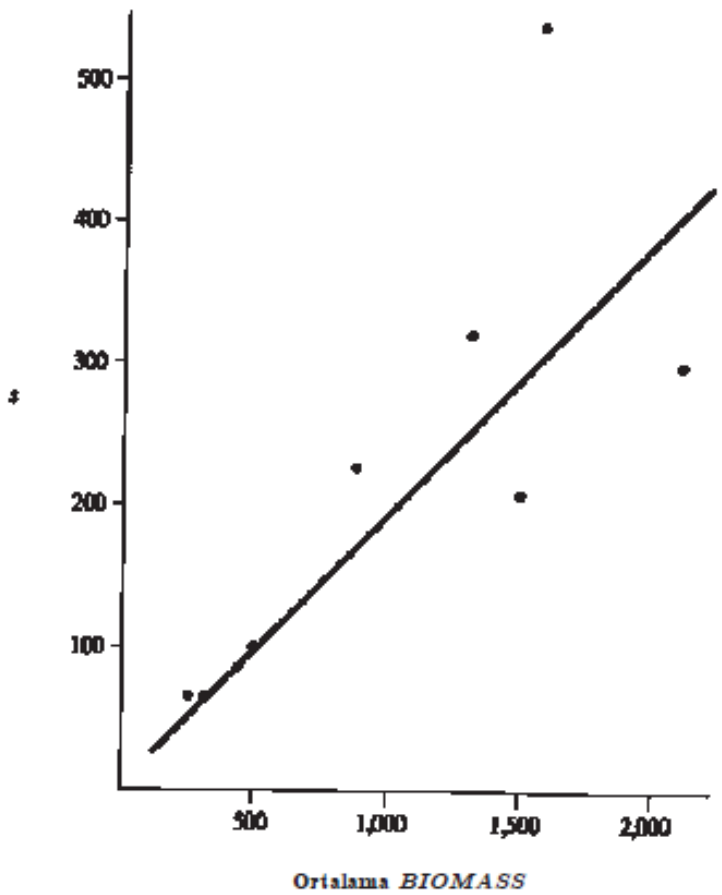
ŞEKİL 11.17. *En küçük kareler artıkları kestirilen değerler (a) ve her bir beş bağımsız değişkene [(b)–(f)] karşı çizilmiştir. Linthurst Eylül verisi kullanılmıştır.*



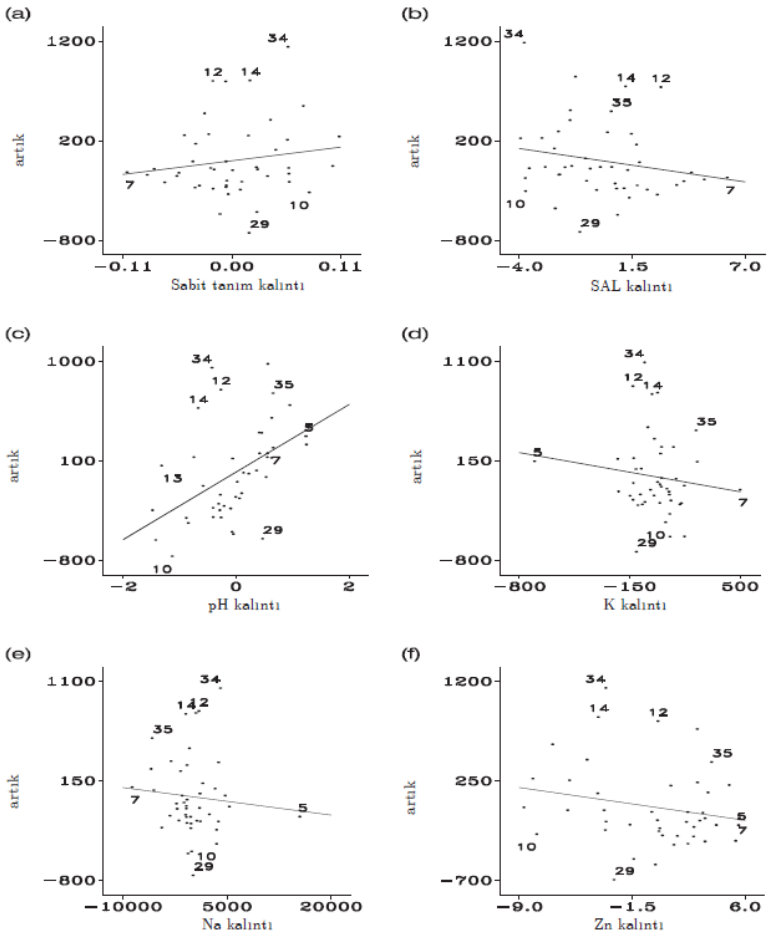
ŞEKİL 11.18. *BIOMASS ın beş bağımsız değişken SALINITY, pH, K, Na ve Zn üzerine regresyonundan elde edilen standartlaştırılmış artıkların frekans poligonu. Linthurst Eylül veri seti.*



ŞEKİL 11.19. *BIOMASS ın beş bağımsız değişken üzerine regresyonundan elde edilen standartlaştırılmış artıkların normal grafiği. Linthurst Eylül veri seti.*



ŞEKİL 11.20. *Her bir bölgedeki beş gözlemden elde edilen ortalama BIOMASS ile gözlemler arasında standart sapmanın grafiği. Linthurst Eylül verisi.*



ŞEKİL 11.21. *BIOMASS ın sabit ve beş bağımsız değişken üzerine regresyonundan elde edilen kısmi regresyon kaldıraç grafikleri. Çizilen grafiğin eğimi o değişken için kısmi regresyon katsayısıdır. Belirli noktalarla ilişkili sayılar gözlem sayısına işaret eder.*

ETKİ İSTATİSTİKLERİ

Tablo 11.4 ve Tablo 11.5'te (Cook'un D'*si*) etki istatitikleri sunulmaktadır. Bu örnek için etki istatistikleri referans değerleri $p' = 6$ ve $n = 45$'dir.

- v_{ii}, $\boldsymbol{P}$ nin elemanları (PROC REG deki HAT DIAG): Ortalama değer $p'/n = 6/45 = .133$. Eğer $v_{ii} \geq 2p'/n = .267$ ise nokta potansiyel olarak etkilidir.
- Cook'un D'si: Cook'un D'si için kesim değeri $4/n = 4/45 = .09$ dur eğer DFFITS e ilişkiler kullanılmışsa.
- DFFITS: Mutlak değerler $2\sqrt{p'/n} = 2\sqrt{6/45} = .73$ den büyükse $\widehat{Y}_i$ üzerine etkiye işaret etmektedir.
- DFBETAS$_j$: Mutlak değerler $2/\sqrt{n} = .298$ den büyükse $\widehat{\beta}_j$ üzerine etkiye işaret etmektedir.
- COVRATIO: $1 \pm 3p'/n = (.6, 1.4)$ aralığı dışındaki değerler genelleştirimiş varyans üzerine büyük etkiye işaret eder.

TABLO 11.7. *Potansiyel etki (vii) yi ya da etkiyi gösteren 9 gözlem. (Linthurst veri seti). Yıldız işareti, (sütunlarda) ölçütün kesme noktasını geçtiğine işaret etmektedir.*

				DFBETAS					
Gözlem	v_{ii}	*Cook's D*	*DFFITS*	*Sabit*	*SAL*	*pH*	*K*	*Na*	*Zn*
5	*								
7	*								
10				*		*			*
12		*	*		*				
13				*		*			*
14		*	*			*			*
29		*	*						
34		*	*	*	*	*		*	*
35							*	*	

TABLO 11.8. *BIOMASS ın beş bağımsız değişken, SAL, pH, K, Na ve Zn üzerine regresyonu için doğrusal bağımlılık tanılama. Linthurst veri seti (SAS PROC REG ve COLLIN seçeneği ile)*

Temel Bileşen Boyutu	*Özgeçler*	*Durum Endeksi*	*Varyans Ayrıştırma Oranı*					
			Sabit	*SAL*	*pH*	*K*	*Na*	*Zn*
1	5.57664	1.000	.0001	.0002	.0006	.0012	.0013	.0011
2	.21210	5.128	.0000	.0007	.0265	.0004	.0000	.1313
3	.15262	6.045	.0015	.0032	.0141	.0727	.1096	.0155
4	.03346	12.910	.0006	.0713	.1213	.2731	.2062	.0462
5	.02358	15.380	.0024	.0425	.1655	.5463	.5120	.0497
6	.00160	58.977	.9954	.8822	.6719	.1062	.1709	.7561

DOĞRUSAL BAĞIMLILIK TANILARI

Doğrusal bağımlılık tanılama (Tablo 11.8) PROC REG de "COLLIN" seçeneğinden elde edilmektedir. Doğrusal bağımlılık ölçütleri standartlaştırılmış $\boldsymbol{X}'\boldsymbol{X}$ nin özanalizi sonucu elde edilmiştir. Her bir sütunun kareler toplamı birdir ve özdeğerlerin toplamı $p' = 6$'dır. $\boldsymbol{X}$ için koşul sayısı 58,98'dir ve ılımlıdan güçlüye doğrusal bağımlılık işaret eder. Dördüncü ve beşinci boyutlar için koşul endeksleri 10'dan büyüktür ve X- uzayının bu iki boyutunun doğrusal bağımlılık problemlerine yol açmasına sebep olur.