**YILDIZ TEKNİK ÜNİVERSİTESİ**
**FEN BİLİMLERİ ENSTİTÜSÜ**

# REGRESYON DİAGNOSTİKLERİ VE SAPAN DEĞERLER

Ekonometrisyen Elif ÖZTÜRK

**FBE İstatistik Anabilim Dalında Hazırlanan**

**YÜKSEK LİSANS TEZİ**

**Tez Danışmanı** : Yrd. Doç.Dr.Atıf EVREN (YTÜ)

Abbas Azimli Atıf Evren Güley Kıroğlu

**İSTANBUL,2005**

**İÇİNDEKİLER**

Sayfa

**SİMGELER**

| | |
|---|---|
| $\mu$ | Ortalama |
| $\sigma^2$ | Anakütle Varyansı |
| $s^2$ | Örneklem Varyansı |
| $df$ | Serbestlik Derecesi |
| $p(Y/X)$ | Y'nin Koşullu Olasılığı |
| $E(Y_i/X_i)$ | $X = X_i$ Değerini Aldığında Y'nin Beklenen Değeri |
| $\beta_0$ | Sabit Parametresi |
| $\beta_1$ | Eğim Parametresi |
| $u_i$ | i.inci Hata Terimi |
| $b_0$ | $\beta_0$'ın Tahmin Edicisi |
| $b_1$ | $\beta_1$'in Tahmin Edicisi |
| $e_i$ | i.inci Kalıntı Değeri |
| $\rho$ | Ardışık Bağımlılık Katsayısı |
| $\rho_{u_t u_{t-1}}$ | Birinci Dereceden Ardışık Bağımlılık Katsayısı |
| H | Dönüşüm Matrisi |
| $R^2$ | Belirginlik Katsayısı |
| $R_a^2$ | Düzeltilmiş Belirginlik Katsayısı |
| $e_i^*$ | i.inci Standart Kalıntı |
| $r_i$ | i.inci Studentize Edilmiş Kalıntı |
| $d_i$ | Öngörü Kalıntısı |
| $t_i$ | Studentize Edilmiş Öngörü Kalıntısı |
| $m_i^2$ | Kare Mahalanobis Uzaklığı |
| $D_i$ | Cook Mesafesi |

## KISALTMA LİSTESİ

| | |
|---|---|
| ARF | Ana Kütle Regresyon Fonksiyonu |
| ÖRF | Örneklem Regresyon Fonksiyonu |
| E.K.K. | En Küçük Kareler Yöntemi |
| K.D.R.M. | Klasik Doğrusal Regresyon Modeli |
| K.D.N.R.M. | Klasik Normal Doğrusal Regresyon Modeli |
| SSE | Kalıntı Kareleri Toplamı |
| MSE | Kalıntı Kareleri Ortalaması |
| SSTO | Toplam Değişkenlik |
| SSR | Regresyon Kareler Toplamı |
| SSPE | Saf Hata Kareleri Toplamı |
| SSLF | Uyum Eksikliği Hata Kareleri Toplamı |

ŞEKİL LİSTESİ

Sayfa

## ÇİZELGE LİSTESİ

Sayfa

## ÖNSÖZ

"Regresyon Diagnostikleri ve Sapan Değerler" isimli yüksek lisans tez çalışmamda öncelikle her zaman bana destek olan anneme, babama ve biricik kardeşim Murat'a,

Tezime başladığım günden bitirdiğim güne kadar beni yönlendiren, derslerinin yoğunluğuna rağmen tezimle en ince ayrıntısına kadar ilgilenen, yardımını ve desteğini benden esirgemeyen, değerli hocam sayın Atıf Evren'e,

En Küçük Kareler Yönteminin geometrisini bana anlatan, sevdiren, sorularıma sabırla cevap veren Gülhayat Şimşek Hocama,

Dar zamanlarımda hep yanımda olan, üzüntümü, sevincimi benimle paylaşan sevgili arkadaşım Fatma Noyan'a,

Manevi desteklerini benden esirgemeyen İstatistik Bölümündeki tüm hocalarıma ve arkadaşlarıma,

Teşekkür ediyorum.

## ÖZET

En Küçük Kareler Metoduyla elde edilen çoklu regresyon modeline ilişkin öngörüde bulunabilmesi modelin varsayımlarının sağlanmasına bağlıdır. Regresyon analizinde varsayımlardan sapmalarla ilgili problemlerin belirlenebilmesi için diagnostik teknikler kullanılır.

Regresyon analizinde bazı veri setleri diğer verilerden ayrılan sapan gözlemler içerir. En Küçük Kareler Yöntemi sapma kareleri toplamını minimize ettiğinden sapan değer, öngörülen regresyon doğrusunu kendisine doğru kaydırabilir. Bu durum sapan değer bir yanlıştan, yanlış veri girilmesinden veya diğer dışsal sebeplerden kaynaklanıyorsa yanlış öngörüye sebep olur. Dolayısıyla sapan değerleri dikkatle incelemek ve modelden dışlanıp dışlanmayacağına karar vermek gerekir. Burada bu sapan gözlemleri ayırmak ve hangi parametre tahminlerinin bu veri setinden daha çok etkilendiğine karar vermek yapıcı bir çözümdür.

Bu çalışmada; 60 ülkenin tarım verileri incelenmiş ve tarımsal katma değeri modellenmiştir. Araştırma içersinde önce tek bir satırın silinmesine dayalı diagnostik ölçüler hesaplanmış ve etkili değerler belirlenmiştir. Ardından birden fazla satırın silinmesine dayalı çoklu diagnostik ölçüler hesaplanmıştır. Bu tekniklerin sonuçları birleştirilmiş ve elde edilen sonuçlar yorumlanmıştır.

Araştırmada, incelenen örnekten yola çıkarak, sapan değerlerin hangi parametre tahminlerini etkilediği belirtilmiştir. Bu etkiler incelenerek sapan değer olarak belirlenen ülkelerin kendine has bazı özelliklerinden dolayı diğer verilerden ayrıldığı sonucuna varılmıştır. Bu özellik ilgilenilen değişken hakkında önemli bilgiler veren özelliktir. Sapan değerlerin dışlanması halinde regresyon modelinde ortalama bir ilişki elde edilebilir. Fakat bu ilişki söz konusu özelliği içinde barındırmaz. Burada varılan sonuç bulunan sapan değerlerin önemli bilgiler içerdiği, dolayısıyla dışlanmaması gerektiğidir.



**Anahtar Kelimeler:** En Küçük Kareler Yöntemi, diagnostik teknikler, sapan ve etkili gözlemler.

**ABSTRACT**

If the regression model is correct and assumptions hold, inferences can be made confidently. Diagnostic techniques are designed to find contradictions between the assumptions of the model and the data at hand.

The data set may contain extreme subsets that differ from the remaining of the data. Under the least squares method, a fitted line may be pulled toward on outlying observation because the sum of squares is minimized. This could cause a misleading fit if indeed the outlying observation is resulted from a mistake or other extraneous cause. It is therefore important to examine the outlying cases carefully and decide whether they should be retained or eliminated. It is reasonable to isolate outliers and to determine which parameter estimates are affected by them.

In this work, agricultural activities of 60 countries were studied and a linear regression model was fitted. First of all single row effects were studied and influential observations were discovered. Then multiple row effects were calculated. At the end the results were interpreted together.

In this thesis, by using the sample examined in the research, it was discovered which parameter estimate(s) was/were affected by the outliers. By investigating these effects, it was found that, the countries which were called outliers had special properties so they differed from the remaining of data. These properties give analysts important information about the interested variable(s). If the outliers are excluded from the model, the regression model can be fitted without causing any problems. But the relationship proposed by this new model doesn't reflect this property. Consequently, outliers give important information. So we shouldn't exclude them from the data set.

**Key Words**: Least Squares Method, diagnostics, outlier and influential observations.

# GİRİŞ

Parametre tahminlerini, tahminlere ilişkin hipotez testlerini ve diğer özet istatistikleri elde etmek için geliştirilen metotlar regresyon analizinin ilk safhasını oluşturmaktadır. Bütün bu metotlar model verilere uygunsa ve varsayımlar gerçekleşmişse uygulanabilir.

$$Y = b_0 + b_1 X_1 + b_2 X_2 + e$$

Modelindeki tahminlerin değişkenler arasındaki gerçek ilişkiyi yansıtıp yansıtmadığını bilmek isteriz. Örneğin $X_1$ değişkenindeki bir birim değişme, $X_2$ değişkeninin etkisinin sabit tutulduğu varsayımı altında Y'de $b_1$ kadar beklenen bir değişim meydana getirir mi? Bu ana kütle tahmininin doğruluğu hakkında güven duymak isteriz. Bu konuda başarılı olmamız, sapan değerler, değişen varyans, otokorelasyon, doğrusal olmama ve normal dağılmama gibi problemlerin üstesinden gelinmesine bağlıdır. Diğer bir deyişle modelden hareketle çıkarsamalar yapılmadan önce modelin verilere uygunluğunun incelenmesi gerekmektedir. Diagnostikler bahsedilen problemleri ortaya çıkaran araçlardır.

Diagnostik tekniklerde birbirini tamamlayan iki soru üzerinde durulur. Birincisi kullanılan modelin verilere ne kadar uyum sağladığıdır. Burada temel istatistik kalıntıların uygun bir dönüşümüdür. Bu kalıntılar modelin varsayımlarının sağlanıp sağlanmadığı hakkında bilgi vermektedir. İkinci soru her bir birimin tahmin, testler ve çıkarsama süreçleri üzerindeki etkisinin ne olduğudur. Bu soru veri setinin incelenmesini gerektirir. Bu amaçla veri setindeki sapan ve etkili değerler belirlenir.

Bu çalışmada ilk bölümde doğrusal regresyon analizi açıklanmıştır. Ardından bu çalışmada kullanılan En Küçük Kareler Yönteminin geometrisi anlatılmıştır. Diagnostikler kısmında öncelikle her bir varsayımdan sapmanın grafiksel araçlarla nasıl tespit edileceği açıklanmıştır. Ardından her bir varsayımdan sapma ayrı bir bölümde incelenmiş ve bu varsayımlarla ilgili biçimsel testlere değinilmiştir. Bu varsayımlardan sapmaların olması halinde gereken dönüşümler ayrı bir bölümde incelenmiştir.

Çalışmanın ana temasını oluşturan sapan ve etkili değerlerin belirlenmesi son bölümü oluşturmaktadır.

Çoklu regresyon modelinin öngörülmesinden sonra sapan ve etkili gözlemlerin belirlenmesi amacıyla kalıntılara, tahmin değerlerine ve açıklayıcı değişkenlere dayalı pek çok süreç tanımlanmıştır. Bu gibi metotlarla önemli bilgiler elde edilebilir. Fakat bu metotlar bir alt veri setinin modelden dışlanması halinde tahmin edilen modelin nasıl değişeceğini gösterememektedir. Bu bölümde etkili veri noktalarını belirleyen ve bu noktaların tahmin edilen katsayılar, standart hatalar ve test istatistikleri üzerindeki etkilerini ortaya çıkaran diagnostik tekniklere değinilmiştir.

# 1. REGRESYON ANALİZİ

## 1.1 Regresyon Teriminin Tarihsel Kökeni

Meşhur bir İngiliz insan bilimci ve meteorolojist olan Sir Francis Galton (1822-1911) "regresyon" terimini takdim eden ilk kişidir. Aslında Galton "eski haline dönme (reversion)" ifadesini "İnsanların Kalıtımlarında Tipik Kurallar "adlı 9 Şubat 1877'de Kraliyet Kurumuna sunduğu bir çalışmasında kullanmıştı. Daha sonra "regresyon" ifadesi 1885'te başkanlığa ait İngiliz derneğinde; 1885'te Nature dergisinde (p.507-510) ve 1885'te "Antropoloji Enstitüsü dergisinin 15.sayısında yayınladığı "Kalıtsal Olarak Boylarda Ortalamaya Doğru Regresyon"(Regression Towards Mediocrity in Hereditary Stature) adlı çalışmasında kullanılmıştır.( Draper ve Smith, 1998)

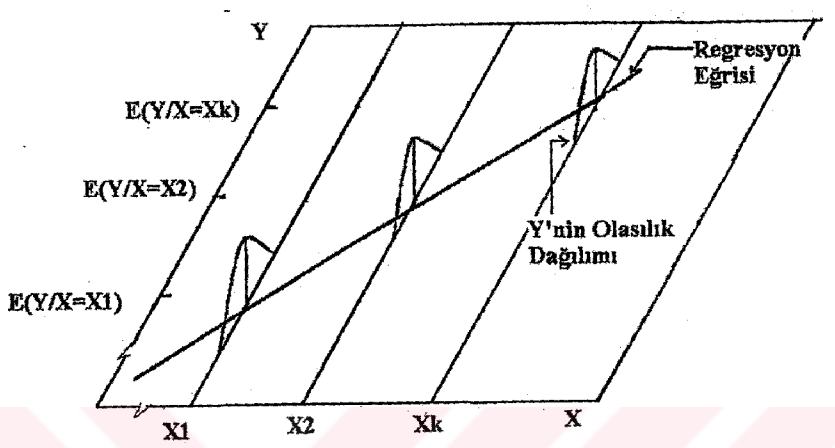
Galton bu yazısında, uzun boylu ana-babaların uzun, kısa boylu ana-babaların kısa çocukları olur eğiliminin geçerliliğine karşın, belli bir boydaki ana-babaların çocuklarının ortalama boyunun genel nüfustaki ortalama boya doğru yaklaşma (İngilizce deyimiyle "regress") eğiliminde olduğunu bulmuştur.

## 1.2 Yalın Doğrusal Regresyon Modeli

İki değişkenli regresyon modeli ile bağımsız X değişkeni ile bağımlı Y değişkeni arasındaki ilişki incelenmektedir. Doğrusal regresyon değişkenler arasındaki ilişkinin bir doğruyla ifade edilebileceğini göstermektedir. Bağımsız değişken $X_i$'nin değerlerinin sabit olması, aynı $X_i$ için, ana kütlede, rastlantısal ilişki gereği en azından iki bağımlı değişkenin bulunmasını zorunlu kılmaktadır. Böylece herhangi bir $X_i$ değişkeni için $Y_{ij}$ değerleri elde edilecek bu da bir dağılım oluşturacaktır.

Y'nin her bir koşullu olsılık dağılımı için koşullu ortalama ya da koşullu beklenen değer diye bilinen, $E(Y/X=X_i)$ ile gösterilen, "X belli bir $X_i$ değerini aldığında Y'nin beklenen değeri" diye okunan ortalama elde edilir.

*Geometrik olarak, bir ana kütle regresyon eğrisi, açıklayıcı değişken(ler)in sabit değerlerine karşılık gelen bağımlı değişkenin koşullu ortalamalarının ya da koşullu beklenen değerlerinin geometrik yeridir.*(-Gujarati,1999)




Şekil 1.1 Regresyon doğrusu

Bu durum şekil 1.1'deki gibi gösterilebilir. Burada her $X_i$ için (daha sonra açıklanacak nedenlerle normal dağıldığı varsayılan) bir Y değerleri alt ana kütlesi ile bir koşullu ortalama yer almaktadır. Regresyon doğrusu da bu koşullu ortalama noktalarından geçer.

Bağımsız değişken $X_i$'nin belirli bir düzeyinde, bağımlı değişkenin tekil bir değerinin aynı $X_i$ düzeyindeki bütün değerlerin beklenen değerinin dolayında dağıldığı görülür. Dolayısıyla tekil bir $Y_i$ değerinin beklenen değerinden sapması aşağıdaki gibi gösterilebilir:

$$u_i = Y_i - E(Y / X_i) \tag{1.1}$$

Burada $u_i$ sapması pozitif ya da negatif değerler alabilen ama gözlemlenemeyen rassal bir değişkendir. Teknik olarak bu değişken olasılıklı hata terimi olarak adlandırılır.

Eğer $E(Y/X_i)$'nin $X_i$'ye göre doğrusal olduğu varsayılırsa yukarıdaki denklem şu şekilde yazılabilir:

$$\begin{aligned} Y_i &= E(Y/X_i) + u_i \\ Y_i &= \beta_0 + \beta_1 X_i + u_i \end{aligned} \qquad (1.2)$$

Uygulamada ilgilenilen ana kütlenin bütününe ulaşmak pek olası olmadığından, ana kütle regresyon fonksiyonu (ARF) idealize edilmiş bir kavramdır. Genellikle elimizde ana kütleden çekilmiş bir örneklemin gözlemleri bulunur. Bu nedenle ana kütle regresyon fonksiyonunu tahmin etmek için olasılıklı örneklem regresyon fonksiyonu (ÖRF) kullanılır. Burada örneklem bilgileriyle ARF'yi tahmin edilmeye çalışılır.ARF'nin örneklemdeki karşılığı şu şekilde yazılabilir:

$$\hat{Y}_i = b_0 + b_1 X_i + e_i \qquad (1.3)$$

$\hat{Y}_i = E(Y/X_i)$'nin tahmin edicisi

$b_0 = \beta_0$ (regresyon sabiti)'ın tahmin edicisi

$b_1 = \beta_1$ (ana kütle parametresi)'in tahmin edicisi

$e_i$= kalıntı terimini ifade eder. $e_i$, $u_i$'nin tahminidir.

*Genelde Regresyon analizi yalnızca parametre tahmini ile değil aynı zamanda verilerin tahminlere dönüştürülmesinde kullanılan yöntem ya da tahmin ediciler üzerinde durur. Bunun nedeni, belirli bir örnekten elde edilen tahminin geçerliliğinin, tahminin elde edilmesinde kullanılan tahmin yönteminin (tahmin edicinin) geçerliliğine bağlı olmasıdır.*(Kennedy,2000)

## 1.3 En Küçük Kareler Yöntemi

Gerçek birimler ile onların tahmin değerleri arasındaki farkların kareleri toplamını minimum

yapma işlemi en küçük kareler yöntemi olarak adlandırılmaktadır.(Genceli,2002)

$$\begin{aligned}\sum e_i^2 &= \sum (Y_i - \hat{Y}_i)^2 \\ &= \sum (Y_i - b_0 - b_1 X_i)^2\end{aligned} \tag{1.5}$$

Bu denklemden görüleceği gibi,

$$\sum e_i^2 = f(b_0, b_1) \tag{1.6}$$

kalıntı karelerinin toplamı, $b_0, b_1$ tahmin edicilerinin bir fonksiyonudur. En küçük kareler ilkesi ya da yöntemi $b_0, b_1$ değerlerini öyle bir biçimde seçer ki, verilmiş bir örneklem y ada veri kümesi için $\sum e_i^2$ en küçük çıkar. Diğer bir deyişle , veri bir örneklem için en küçük kareler yöntemi olanaklar içindeki en küçük $\sum e_i^2$ toplamını veren tek $b_0, b_1$ değerlerini seçer. Bu tahminlerin elde edilmesi için $\sum e_i^2$ 'nin $b_0$ ve $b_1$'e göre kısmi türevleri alınır:

$$\begin{aligned}\frac{\partial(\sum e_i^2)}{\partial b_0} &= -2\sum (Y_i - b_0 - b_1 X_i) = -2\sum e_i \\ \frac{\partial(\sum e_i^2)}{\partial b_1} &= -2\sum (Y_i - b_0 - b_1 X_i) X_i = -2\sum e_i X_i\end{aligned} \tag{1.7}$$

bulunur. Bunlar sıfıra eşitlenip bazı cebirsel kısaltma işlemlerinden sonra şu eşitlikler elde edilir,

$$\begin{aligned}\sum Y_i &= n b_0 + b_1 \sum X_i \\ \sum Y_i X_i &= b_0 \sum X_i + b_1 \sum X_i^2\end{aligned} \tag{1.8}$$

Burada n, örneklem büyüklüğüdür. Bu eşanlı denklemler normal denklemler diye tanınır. Bu eşanlı denklemler çözülürse aşağıdaki eşitlikler elde edilir:

$$b_1 = \frac{\begin{vmatrix} n & \sum Y_i \\ \sum X_i & \sum Y_i X_i \end{vmatrix}}{\begin{vmatrix} n & \sum X_i \\ \sum X_i & \sum X_i^2 \end{vmatrix}} = \frac{n\sum X_i Y_i - \sum X_i \sum Y_i}{n\sum X_i^2 - (\sum X_i)^2}$$

$$= \frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{\sum (X_i - \overline{X})^2} = \frac{\sum x_i y_i}{\sum x_i^2} \qquad (1.9)$$

Burada $\overline{X}$ ile $\overline{Y}$ ,X ile Y'nin örnekleme ortalamalarıdır. $x_i = (X_i - \overline{X})$ ve $y_i = (Y_i - \overline{Y})$ olarak tanımlanmıştır. Küçük harfler *ortalamadan sapma* anlamında kullanılmaktadır.

$$b_0 = \frac{\begin{vmatrix} \sum Y_i & \sum X_i \\ \sum Y_i X_i & \sum X_i^2 \end{vmatrix}}{\begin{vmatrix} n & \sum X_i \\ \sum X_i & \sum X_i^2 \end{vmatrix}} = \frac{\sum Y_i \sum X_i^2 - \sum X_i \sum Y_i X_i}{n\sum X_i^2 - (\sum X_i)^2} \qquad (1.10)$$

$$= \overline{Y} - b_1 \overline{X}$$

Bu yolla elde edilen tahmin ediciler, en küçük kareler ilkesinden türetildikleri için *en küçük kareler tahmin edicileri* adını alırlar.

Bu tahmin edicilerin sağladığı bazı sayısal özellikleri vardır. "Veriler nasıl türetilmiş olursa olsun, en küçük karelerin kullanımının doğurduğu sonuçlarca sağlanan özellikler sayısal özelliklerdir."Bunlar aşağıdaki gibi özetlenebilir:

I. En küçük kareler (EKK) tahmin edicileri, yalnız gözlemlenebilen (örneklem) büyüklükler (yani X ve Y) cinsinden gösterilebilir. Dolayısıyla kolaylıkla hesaplanabilirler.

II. Bunlar nokta tahmin edicileridir, yani örneklem veriyken , her tahmin edici ilgili ana kütle katsayısının tek bir (nokta) tahminini verir.

III. EKK tahmin edicileri örneklem verilerinden bulunduktan sonra, örneklem regresyon doğrusu çizilebilir. Bu yolla elde edilen regresyon doğrusu şu özellikleri taşır:

1. Y ve X'in örnekleme ortalamalarından geçer.

2. Tahmin edilmiş Y'nin ortalama değeri $=\bar{\hat{Y}}$, gözlemlenen Y değerinin ortalamasına eşittir.

3. $e_i$ kalıntılarının ortalaması sıfırdır.

4. $e_i$ kalıntıları tahmin edilmiş $Y_i$'lerle ilişkisizdir. $\sum Y_i e_i = 0$ şeklinde ifade edilebilir.

5. $e_i$ kalıntıları $X_i$'lerle ilişkisizdir, yani $\sum e_i X_i = 0$'dır.

### 1.3.1 En Küçük Kareler Yönteminin Ardındaki Varsayımlar

Eğer regresyon çözümlemesinde amaç sadece $b_0$ ile $b_1$'i tahmin etmek olsaydı, açıklanan EKKY yeterli olabilirdi. Amaç yalnızca $b_0$ ile $b_1$'i tahmin etmek değil, bunun yanı sıra gerçek $b_0$ ile $b_1$'e ilişkin çıkarsamalar da yapmaktır. Örneğin , $b_0$ ile $b_1$'nin ana kütledeki karşılıklarına ya da $\hat{Y}_i$'lerin gerçek $E(Y/X_i)$ değerine ne kadar yakın oldukları bilinmek istenebilir. Bu nedenle, modelin fonksiyon kalıbını belirlemek yeterli olmayıp aynı zamanda $Y_i$'nin türetilme biçimine ilişkin belli bazı varsayımların yapılması gerekmektedir.(Gujarati,1999)

Kullanılan tahmin yöntemlerinin başarısı büyük bir oranda hata teriminin yapısına bağlıdır. *Dolayısıyla hata teriminin yapısı ile ilgili istatistiksel varsayımlar Ekonometri teorisinin en önemli konularını oluşturur.*(Kennedy,2000)

Ekonometri kuramının büyük bir bölümünün temel taşı niteliğindeki **Gausgil, standart ya da klasik doğrusal regresyon modeli** (KDRM), 10 varsayım yapmaktadır. Bu varsayımlar iki değişkenli model çerçevesinde,

1. **Varsayım:** Doğrusal regresyon modeli. Regresyon modeli, katsayılarda doğrusaldır:

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

KDRM'nin ilk varsayımı bağımlı değişkenin bağımsız değişkenin doğrusal bir fonksiyonu artı hata terimi olarak ifade edilebileceğidir. Bağımlı değişken Y ile açıklayıcı değişken X'in kendileri doğrusal olmayabilirler.

**2. Varsayım:** X değerleri yinelenen örneklemelerde değişmez. X açıklayıcı değişkenlerinin yinelenen örneklemlerde aldığı değerlerin aynı kaldığı düşünülür. Başka deyişle X'in olasılıklı olmadığı varsayılır. Bu varsayımın anlamı, regresyon çözümlemesinin, açıklayıcı X değişkenine göre koşullu regresyon çözümlemesi olduğudur.

**3. Varsayım:** Hata teriminin ortalaması sıfırdır. X değerleri veriyken rassal bozucu $u_i$ hata teriminin ortalaması ya da beklenen değeri sıfırdır. Teknik olarak $u_i$'nin koşullu ortalaması sıfırdır. Simgelerle,

$$E(u_i / X_i) = 0$$

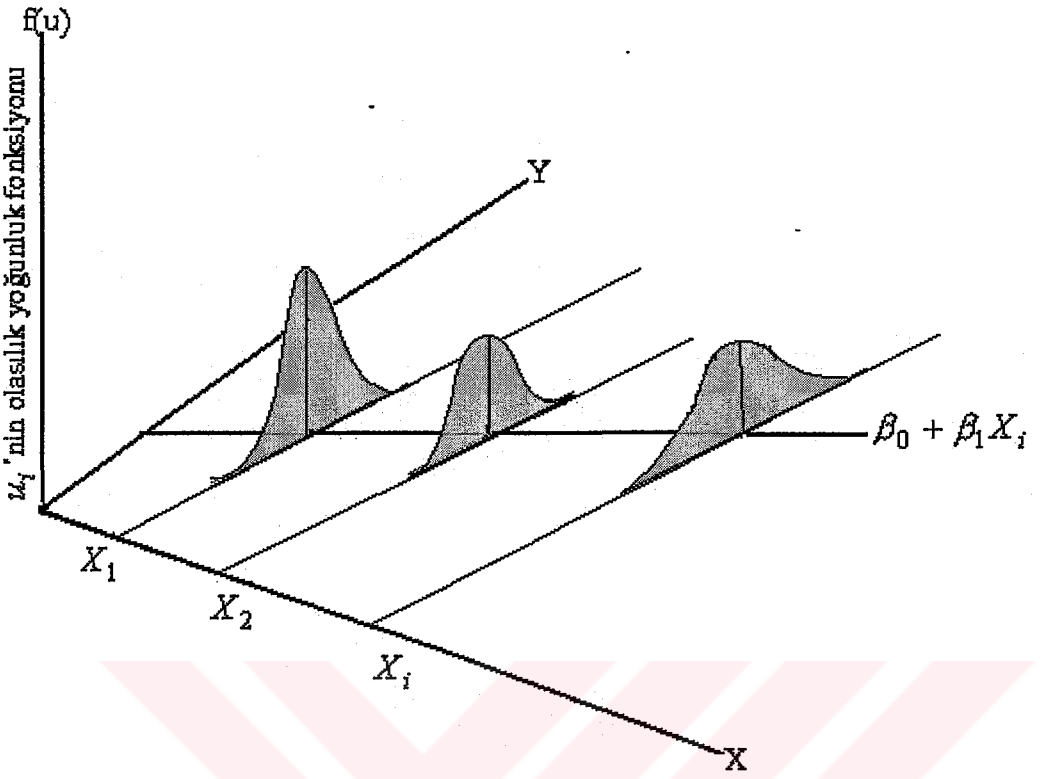
şeklinde ifade edilir.

**4. Varsayım:** X değeri veriyken, $u_i$'nin varyansı bütün gözlemler için aynıdır. Yani, $u_i$'nin koşullu varyansları değişmez. Simgelerle:

$$\begin{aligned} \text{var}(u_i / X_i) &= E[u_i - E(u_i) / X_i]^2 \\ &= E(u_i^2 / X_i) \\ &= \sigma^2 \end{aligned} \qquad (1.11)$$

şeklindedir. Bu eşitliğe göre, her $X_i$ değeri için $u_i$'nin koşullu varyansı $\sigma^2$'ye eşit artı değerli sabit bir sayıdır. Teknik olarak bu eşitlik eş varyans veya homoskedastisite olarak gösterilir.

Bu varsayıma ters olan durum şekil yoluyla incelenirse,




Şekil 1.2 Değişen varyans (Heteroskedastisite)

Burada Y ana kütlesinin koşullu varyansı, X'e bağlı olarak değişir. Bu duruma değişen varyans ya da heteroskedastisite denir. Bu durumda ,

$$\text{var}(u_i / X_i) = \sigma_i^2 \tag{1.12}$$

yazılır. Bu eşitlikteki $\sigma^2$'nin alt imi, hata teriminin varyansının artık sabit olmadığını göstermektedir.

Şekle göre, $\text{var}(u/X_1) < \text{var}(u/X_2) < ... < \text{var}(u/X_i)$ olmaktadır. Bu durumda Y gözlemlerinin, $X = X_1$ alt ana kütlesinden gelen Y gözlemleri, $X = X_2, X = X_3$, v.b. alt ana kütlelerden gelenlere göre ana kütle regresyon doğrusuna daha yakındır. Kısacası, çeşitli X'lere karşılık gelen bütün Y değerlerinin güvenilirlikleri aynı değildir. Burada güvenilirlik, Y değerlerinin kendi ortalamaları, yani ARF üzerindeki noktaları etrafındaki dağılımlarının ne

kadar dar ya da geniş olduğu anlamındadır. Eğer dağılımlar bu şekilde ise, kendi ortalamasına yakın olan Y değerleri içinden örneklem seçme daha yayık dağılımlardan seçmeye göre daha çok tercih edilebilir. Fakat bu yapıldığında, X değerlerinin değişkenliği kısıtlanmış olur.

**5. Varsayım:** Hata terimleri arasında ardışık bağımlılık yoktur. $X_i$ ve $X_j$ gibi $(i \neq j)$ gibi iki X verilmişken, $u_i$ ile $u_j$ $(i \neq j)$ arasındaki korelasyon sıfırdır. Simgelerle:

$$\begin{aligned} \operatorname{cov}(u_i, u_j / X_i, X_j) &= E\left[u_i - E(u_i)/X_i\right]\left[u_j - E(u_j)/X_j\right] \\ &= E(u_i / X_i)(u_j / X_j) \\ &= 0 \end{aligned} \tag{1.13}$$

olmaktadır. burada i ile j iki farklı gözlemi cov ise kovaryansı (ortak varyansı ) göstermektedir.

Sözle ifade edilecek olursa bu varsayım $u_i$ ile $u_j$ hata terimlerinin ilişkisiz olduğunu ileri sürmektedir. Teknik olarak bu **ardışık bağımsızlık** varsayımıdır.

Sezgisel olarak bu varsayım şu şekilde açıklanabilir: $Y_i = \beta_0 + \beta_1 X_i + u_i$ ana kütle regresyon fonksiyonunda $u_i$ ile $u_{i-1}$ aynı yönde ilişkili oldukları varsayılsın. O halde $Y_i$, yalnızca $X_i$'ye değil $u_{i-1}$'e de bağlı olacaktır. Bu varsayımla eğer varsa $X_i$'nin $Y_i$ üzerindeki kurallı etkisi ele alınmakta, Y üzerinde, u'lar arasında olası ilişkilerden doğabilecek başka etkiler ele alınmamaktadır.

**6. Varsayım:** $u_i$ ile $X_i$'nin ortak varyansı sıfırdır, ya da $E(\mathrm{u}_i, X_i) = 0$

$$\begin{aligned} \operatorname{cov}(u_i, X_i) &= E\left[u_i - E(u_i)\right]\left[X_i - E(X_i)\right] \\ &= E\left[u_i(X_i - E(X_i))\right] \\ &= E(u_i X_i) - E(X_i)E(u_i) \\ &= E(u_i, X_i) \\ &= 0 \end{aligned} \tag{1.14}$$

$E(u_i) = 0$ olduğundan ve $E(X_i)$ olasılıklı olmadığından yukarıdaki sonuca varılmıştır.

6. varsayım u hata terimiyle X açıklayıcı değişkeninin ilişkisiz olduğunu ifade etmektedir. Bu varsayımın mantığı şöyledir: ARF $Y_i = \beta_0 + \beta_1 X_i + u_i$ şeklinde yazıldığında , X ve u'nun, Y üzerinde ayrı ayrı (ve toplanabilir) bir etki yarattıkları varsayılmaktadır. Fakat X ile u ilişkiliyse, her birinin Y üzerindeki tekil etkilerini bulmak olanaksızlaşır.

X değişkeni rassal ya da olasılıklı değilse ve üçüncü varsayım olan hata terimlerinin beklenen değeri sıfıra eşitse bu varsayım kendinden gerçekleşir. Dolayısıyla diğer varsayımların sağlanması durumunda bu varsayım önemsenmeyebilir. Burada ifade edilmesinin nedeni, X'ler olasılıklı ya da rassal olsalar dahi, $u_i$'lerden bağımsız ya da en azından ilişkisiz oldukları sürece, regresyon kuramının yine de geçerli olduğunu belirtmesidir.

**7. Varsayım:** Gözlem sayısı n, tahmin edilecek ana kütle katsayılarından fazla olmalıdır. Başka bir deyişle gözlem sayısı n, açıklayıcı değişken sayısından büyük olmalıdır. Bu varsayımın nedeni örneğin tek değişkenli regresyonda iki bilinmeyeni ($\beta_0$ ile $\beta_1$) bulmak için en az iki noktaya gereksinim olmasıdır.

**8. Varsayım:** X değişkenindeki değişkenlik. Belli bir örneklemdeki X değerlerinin hepsi aynı olmamalıdır. Teknik olarak $\text{var}(X)$ artı değerli sonlu bir sayıdır.

Eğer bütün X değerleri aynı olursa $X_i = \overline{X}$ olur. Bu durumda $\beta_0$ ile $\beta_1$'in tahmin edilmesi olanaksız hale gelir ($b_1 = \sum(X_i - \overline{X})(Y_i - \overline{Y}) / \sum(X_i - \overline{X})^2$ ). Sezgisel olarak bağımsız değişkendeki değişme az olursa bağımlı değişkendeki değişmenin çoğu açıklanamaz. Regresyon çözümlemesinde hem X, hem Y'deki değişim zorunludur. Kısacası değişkenler değişmelidir.

**9. Varsayım:** Regresyon Modeli doğru kurulmalıdır. Başka bir deyişle görgül çözümlemede kullanılan modelde kuruluş sapması ya da hatası bulunmamalıdır.

Bir regresyon çözümlemesi model kurmayla başlar. Model sırasında ortaya çıkan bazı önemli sorunlar şunlardır: 1) Modele hangi değişkenler katılmalıdır? 2) Modelin fonksiyon kalıbı nedir? Katsayılarda mı, değişkenlerde mi, yoksa her ikisinde birden mi doğrusaldır? 3) Modele giren $X_i, Y_i, u_i$'ye ilişkin olasılık varsayımları nelerdir? Bu soruların son derece önemli olduğu belirtilmiştir. Nedeni önemli değişkenleri modele koymamak ya da modelin fonksiyonel kalıbını yanlış seçmek , ya da modelin varsayımlarına ilişkin yanlış olasılık varsayımları yapmak, tahmin edilen regresyonun yorumunun geçerliliğini sorgulanır duruma sokmaktadır.

Bütün bu varsayımların örneklem regresyon fonksiyonu ile değil ana kütle regresyon fonksiyonu ile ilgili oldukları belirtilmelidir. Fakat en küçük kareler yönteminin, ana kütle regresyon fonksiyonu üzerine yüklenen varsayımları "kopyaladığı" söylenebilir. Örneğin $\sum e_i = 0$, dolayısıyla $\bar{e} = 0$ bulguları, $E(u_i / X_i) = 0$ varsayımına benzer. Yine, $\sum e_i X_i = 0$ olgusu $\text{cov}(u_i, X_i) = 0$ ile benzeşir. (Gujarati,1999)

### 1.3.2 Hata Terimlerinin Varyansının Tahmini

Y bir rastlantısal değişken ve E(Y)= $\mu$ ise Y değerlerinin beklenen değeri etrafındaki dağılımı ya da yayıklığı olarak tanımlanan *varyansı* aşağıdaki gibi hesaplanır:

***Anakütle Varyansı:*** $Y_1, Y_2, ...., Y_N$ beklenen değeri $\mu$ olan bir ana kütlenin N tane üyesini göstersin. Ana kütle varyansı, $\sigma^2$, bu değerlerin kendi ortalamalarından farklarının karelerinin ortalamasıdır.

$$\sigma^2 = \frac{\sum_{i=1}^{N} (Y_i - \bar{Y})^2}{N} = \frac{\sum_{i=1}^{N} y_i^2}{N} - \mu^2 \qquad (1.15)$$

Ana kütle standart sapması $\sigma$ ise bu varyansın (artı değerli) kareköküdür. Uygulamada ana kütlenin gözlenmesi genelde mümkün olmadığından varyansının elde edilmesi de mümkün

olmaz. Dolayısıyla ana kütle varyansı örneklem varyansından hareketle hesaplanır.

$\sigma^2$ ***'nin Nokta Tahmincisi*** **:** Örneklem varyansı ($s^2$) elde edilirken örneğe ait $Y_i$ gözlemlerinin tahmin edilen ortalamalarından sapmalarının kareleri toplamı elde edilir: $\sum_{i=1}^{n}(Y_i - \bar{Y})^2$. Bu toplam kareler toplamı olarak adlandırılır. Bu kareler toplamı serbestlik derecesiyle bölünür. Burada serbestlik derecesi n-1 çünkü bilinmeyen ana kütle parametresi $\mu$ 'nün tahmini olarak $\bar{Y}$ kullanılmış ve dolayısıyla bir serbestlik derecesi kaybedilmiştir.

$$s^2 = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n-1} \tag{1.16}$$

Bu örneklem varyansı ana kütle varyansının sapmasız bir tahmin edicisidir. Örneklem varyansı kareler ortalaması olarak adlandırılır çünkü kareler toplamı uygun serbestlik derecesine bölünerek elde edilir.

**Regresyon Modeli için $\sigma^2$'nin Tahmini :** Regresyon modeli için $\sigma^2$'nin tahmincisinin elde edilmesi bir ana kütleden elde edilen örneklem yoluyla onun varyansının tahmin edilmesiyle aynı mantığa sahiptir. Regresyon modelinde her bir $Y_i$ gözleminin varyansının $\sigma^2$,her bir hata terimiyle aynı olduğu bilinmelidir. Burada da ortalamadan farkların karelerinin toplamının hesaplanması gerekmektedir fakat $Y_i$'nin her bir $X_i$ düzeyinde farklı ortalamalarla farklı olasılık dağılımlarına sahip olduğu hatırlanmalıdır. Bu durumda $Y_i$ gözleminden sapmalar kendi tahmini ortalamasından, $\hat{Y}_i$, hesaplanmalıdır. O halde bu sapmalar kalıntılardır:

$$Y_i - \hat{Y}_i = e_i \tag{1.17}$$

ve SSE olarak ifade edilen kareler toplamı:

$$SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} e_i^2 \tag{1.18}$$

şeklinde hesaplanır(SSE: residual sum of squares veya error sum of squares olarak açılır). Kareler toplamı SSE n-2 serbestlik derecesine sahiptir. İki serbestlik derecesi kaybedilmiştir. Çünkü $b_0$ ve $b_1$, $\hat{Y}_i$'nin tahmincisini elde etmek için tahmin edilmiştir. İki kısıt söz konusudur. Bu durumda MSE olarak gösterilen kareler ortalaması:

$$MSE = \frac{SSE}{n-2} = \frac{\sum(Y_i - \hat{Y}_i)^2}{n-2} = \frac{\sum e_i^2}{n-2} \tag{1.19}$$

olarak hesaplanır (MSE: Mean Square Error veya Mean Square Residual olarak açılır).

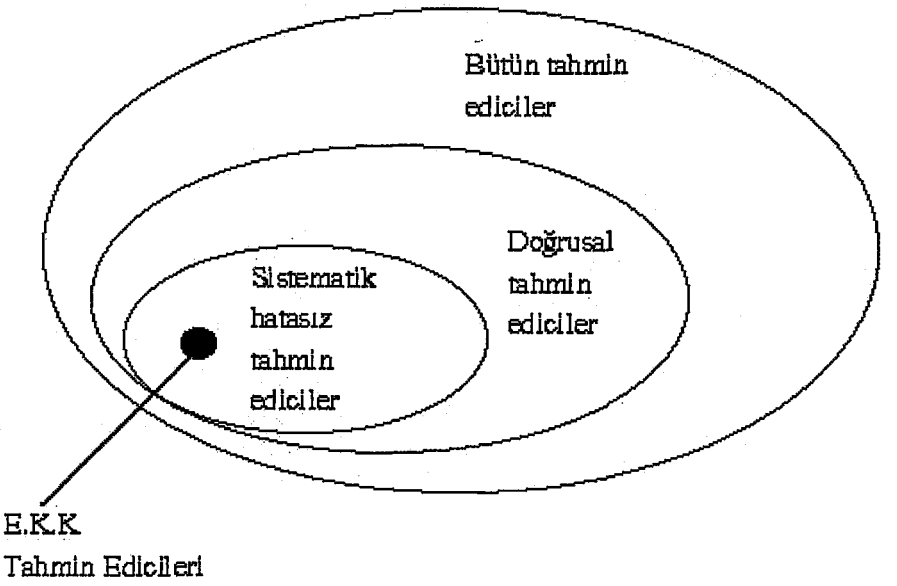
Regresyon modeli için MSE'nin $\sigma^2$'nin yansız bir tahmin edicisi olduğu gösterilebilir:

$$E(MSE) = \sigma^2 \tag{1.20}$$

Standart sapma $\sigma$'nın tahmincisi MSE'nin pozitif karekköüne, $\sqrt{MSE}$, eşittir. (Neter vd.,1996)

### 1.3.3 En Küçük Kareler Tahmin Edicilerinin Özellikleri : Gauss Markov Teoremi

Hata paylarının Klasik doğrusal regresyon modelinde öngörülen varsayımları gerçekleştirmeleri halinde en küçük kareler yöntemi ile elde edilen diyelim $\theta$'nın tahmin edicileri en iyi, doğrusal, sistematik hatasız (BLU) tahmin edicilerdir. Bu sonuç ise *Gauss-Markov Kuramı*'dır. Bu kurama göre: $\theta$'nın bütün doğrusal, sistematik hatasız tahmin edicileri arasında EKK tahmin edicileri minimum varyansa sahiptirler,kısaca en iyi, doğrusal, sistematik hatasız (BLUE) tahmin edicilerdir.




Şekil 1.3 Gauss Markov Teoremi

Fakat burada Gauss Markov Kuramını'nın sadece doğrusal tahmin edicilere uygulandığı vurgulanmalıdır. Doğrusal tahmin edici, $c_i$ sabit olmak üzere, $b = \sum c_i Y_i$'dir. oysa $b^* = \sum c_i \log Y_i$ halinde $b^*$ doğrusal olmayan tahmin edicidir.

***Sistematik Hatasız Tahmin Edici*:** Bir tahmin edicinin sistematik hatası, beklenen değeriyle gerçek parametre arasındaki fark olarak tanımlanır.

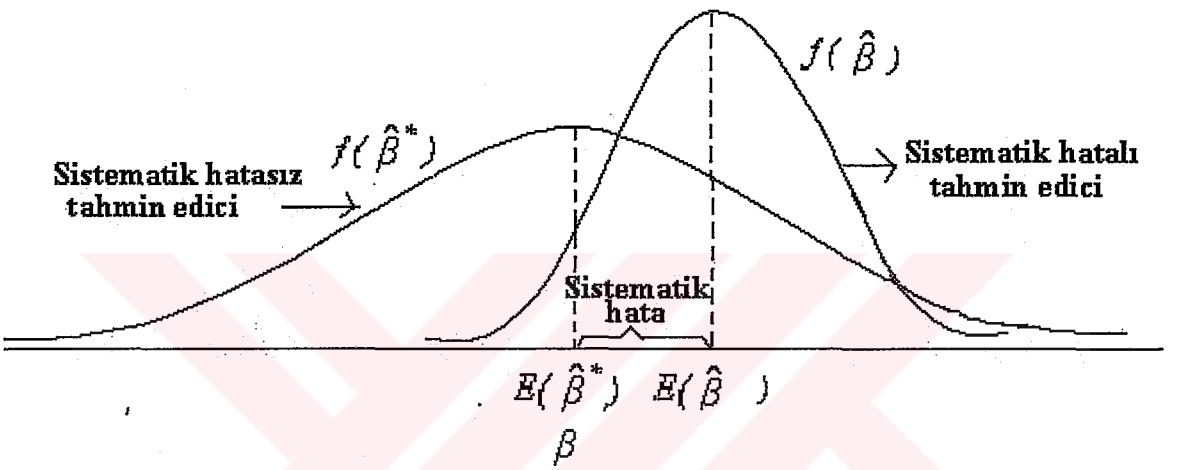
$$sapma = E(\hat{\theta}) - \theta \tag{1.21}$$

Eğer sapma sıfırsa, yani $E(\hat{\theta}) = \theta$ ise, bu tahmin edici standart hatasız olur. Bu da (n sonlu sayısı kadar gözlemi olan) örneklerin sayısı arttıkça, sapmasız tahmin edicinin, parametrenin gerçek değerine yaklaştığı anlamına gelir. Sapmasız bir tahmin edici "ortalama olarak" parametrenin gerçek değerini verir.

$E(\hat{\theta}) > \theta$ halinde yukarıya doğru sistematik hata $E(\hat{\theta}) < \theta$ durumunda ise aşağıya doğru sistematik hata bulunmaktadır. (Genceli,2002)

***Etkinlik*:** Etkin tahmin edici, sapmasız tahmin ediciler arasında en küçük varyanslı tahmin edicidir. Sapmasız bir çok tahmin edici arasından, örnekleme dağılımının varyansı en küçük olan tahmin edici belirlenerek bu tahmin edici diğer sapmasız tahmin ediciler arasında en etkin tahmin edici veya en iyi sistematik hatasız tahmin edici olarak kabul edilecektir.

***Hata Kareleri Ortalaması (MSE)*:** Hata kareleri ortalaması, biri sistematik hatasız fakat büyük varyanslı tahmin edici ile diğeri sistematik hatalı buna karşın daha düşük varyanslı tahmin edici arasındaki tercihe yol gösterici bir ölçüdür.




Şekil 1.4 Hata kareleri ortalaması (MSE) ölçütü

Böyle bir durumda standart hata kabul edilebilir. Küçük varyans ve küçük sistematik hata arasındaki bu ikame edilebilirlik sistematik hatanın ve varyansın ağırlıklı ortalamasının minimize edilmesi şeklinde bir kriterin oluşmasına neden olmuştur. Diğer bir anlatımla bu ağırlıklı ortalamayı minimize eden tahmin edici tercih edilecektir.

$MSE(\hat{\beta}) = E(\hat{\beta} - \beta)^2$ 'dir. Tanımdan da anlaşılacağı gibi MSE $\hat{\beta}$ tahmin edicilerinin kendi parametreleri $\beta$ etrafındaki dağılma ölçüsüdür.

Hata kareler toplamı kriterinin sapmasızlık kriterine tercihi çoğunlukla tahminin hangi alanda kullanılacağına bağlıdır. Örnek olarak at yarışı oynayan bir kişiyi ele alalım. Eğer yalnızca kazanan bir kupon oynamak isteniyorsa kazanacak atın sapmasız bir tahmin edicisi tercih edilecektir. Buna karşılık seçilen atın ilk üç içinde yer alması daha önemli ise bu durumda daha küçük varyansa sahip biraz sapmalı bir tahminci tercih edilebilir.

***Yeterlilik*:** Yeterlilik, R.A.Fisher'in getirdiği bir ölçüttür. Tahmin edicinin yeterlilik ölçütüne sahip olabilmesi için, bu tahmin edicinin tahmin edilecek parametre hakkında örnekte mevcut bulunan bütün bilgiyi kullanması gerekir. Örneğin $\overline{X}$ ana kütle ortalaması $\mu$'nün yeterli bir tahmin edicisidir. Fakat mod ve medyan tüm birimleri kullanmadıklarından $\mu$'nün yeterli bir tahmin edicisi değildirler.(Genceli,2002)

***Büyük Örnek Özellikleri***

Çoğu tahmin edicinin örnekleme dağılımının yapısı örnek büyüklüğüne bağlıdır. Örnek ortalama istatistiğinin örnekleme dağılımı ana kütle ortalamasına odaklanmıştır, ancak varyans büyüklüğü örnek büyüklüğü arttıkça azalır. Çoğu durumda örnek büyüklüğü arttıkça sapmalı tahmin ediciler gittikçe daha az sapmalı hale dönüşür. Diğer bir anlatımla, örnek büyüklüğü arttıkça örnekleme dağılımı değişecek ve örnekleme dağılımının ortalaması tahmin edilmek istenen parametrenin gerçek değerine yakınsayacaktır.

Gittikçe artan örnek çapları kullanılarak $\hat{\beta}$ tahmin edicisine ait bir seri örnekleme dağılımının tanımlandığı varsayılsın. Elde edilen seriyi oluşturan bireysel örnekleme dağılımlarının şekli örnek çapı arttıkça belirli bir dağılıma yakınsıyor ise (normal dağılım gibi) bu dağılım $\hat{\beta}$'nın limitteki ya da asimptotik dağılımı olarak adlandırılır.(Kennedy,2000)

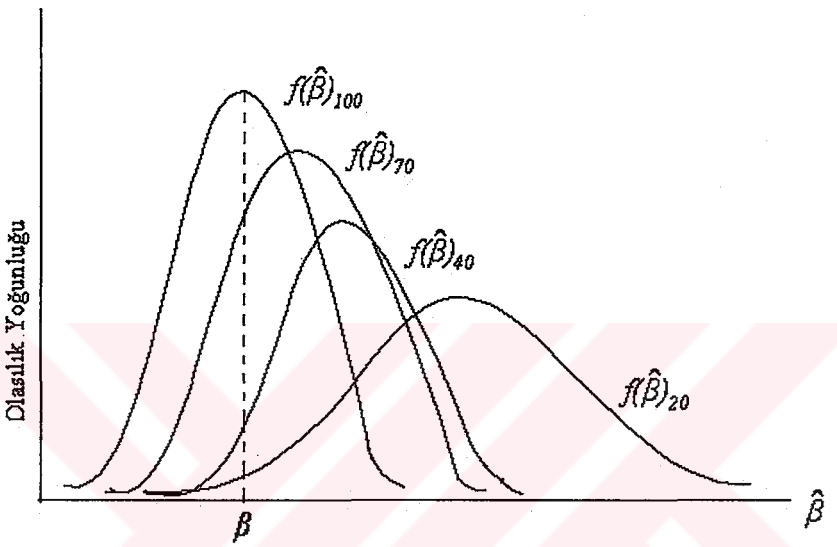
Örneğin örnek birim mevcudu n arttıkça Merkezi Limit Kuramı gereği örnek ortalamalarının dağılımı normal dağılıma yaklaşacaktır. Buna göre normal dağılım örnek ortalamalarının asimptotik dağılımıdır.

Tahmin edicilerin asimptotik dağılımlarının asimptotik ortalama ve asimptotik varyanslarının bulunup bulunmamasına göre büyük örneklerdeki tahmin edicilerin özellikleri iki grup altında toplanabilir: Birinci gruptaki tahmin edicilerin oluşturduğu dağılımın asimptotik ortalaması ve asimptotik varyansı bulunmamaktadır. Bu grup için aranılan özellikler:

1. Asimptotik sistematik hatasızlık
2. Tutarlılıktır.

Momentlere sahip dolayısıyla asimptotik ortalama veya varyansı bulunan ikinci grup için ölçütler ise ;

1. Örnek büyüklüğü sonsuza yaklaştıkça $\hat{\beta}$'nın limit örneklem dağılımı k gibi bir değer etrafında yoğunlaşıyor ise, bu değer $\hat{\beta}$'nın olasılık limiti olarak adlandırılır ve $plim(\hat{\beta}) = k$ şeklinde yazılır. Eğer $plim(\hat{\beta}) = \beta$ ise bu durumda $\hat{\beta}$ **tutarlı** olarak nitelendirilir.




Şekil 1.5 Tutarlılık

Şekilde görüldüğü gibi , örnek birim mevcudu n sonsuza giderken , $n \to \infty$, $\hat{\beta}$ tahmin edicisinin örnekleme dağılımının ortalamasının $\beta$ etrafında yoğunlaşması halinde tutarlılıktan söz edilir. (Genceli,2002)

**2.** $\hat{\beta}$ limit dağılımının varyansı $\hat{\beta}$'nin asimptotik varyansı olarak adlandırılır. $\hat{\beta}$'nın tutarlı olması ve tüm diğer tutarlı tahmincilerin varyansından daha küçük bir varyansa sahip bulunması durumunda **asimptotik etkin** olarak nitelendirilir.

3. Üçüncü daha zayıf bir asimptotik özellik tahmincinin asimptotik beklenen değeri ile ilgilidir. Bu özellik limit dağılımının özelliklerine göre değil, $\hat{\beta}$ beklenen değerlerinin limiti cinsinden tanımlanmıştır. $\hat{\beta}$'nın asimptotik beklenen değeri

$$\lim_{T\to\infty} E(\hat{\beta}) \qquad (1.22)$$

şeklinde yazılabilir. Bu beklenen değerin $\beta$'ya eşit olması durumunda $\hat{\beta}$ **asimptotik sapmasız** olarak nitelendirilir.(Kennedy,2000)

## 1.4 Klasik Normal Doğrusal Regresyon Modeli

Hata terimlerinin dağılımının bilinmemesi en küçük kareler yönteminin $b_0$ ve $b_1$'in bütün sapmasız doğrusal tahmin ediciler arasında en küçük varyanslı nokta tahmin edicileri bulmasını engellemez. Burada dağılımın şeklinin bilinmemesi problem değildir.

Parametre tahmininin yanında tahmin edilen parametreler için güven aralıkları oluşturmak ve parametrelerle ilgili hipotez testleri yapmak regresyon analizinde ikinci adımı oluşturmaktadır. Güven aralığı tahminlerinde bulunmak ve hipotez testleri yapmak için hata terimlerinin dağılımı hakkında bir varsayıma gereksinim duyulur.

Hata payları $u$ regresyon modelinde yer almayan bir çok bağımsız değişkenin etkisini simgelemektedir. Modelin kapsamı dışında bırakılan bu değişkenlerin etkilerinin bireysel düzeyde az fakat çok farklı yönlerde, birbirinden bağımsız ve aynı zamanda rastlantısal olduğu ileri sürülebilir. Bu nedenle hata paylarının küçük değerler etrafında yığılma gösteren tek modlu, simetrik bir dağılımı gerçeklemeleri beklenilir.

Hata paylarının birbirinden bağımsız bu etkilerin toplamını temsil ettiği düşünüldüğünde bunların yaklaşık normal dağıldıkları varsayılabilir çünkü, Lindeberg –Levy Merkezi Limit Teoremine göre rastlantısal değişkenlerin sayısı arttıkça dağılımları ne olursa olsun, toplamları normal dağılıma yaklaşacaktır.

Bu bağlamda daha önce yapılan varsayımlara ek olarak hata paylarının bağımsız ve normal dağıldıkları varsayımı eklenmektedir.

$N(0,\sigma^2)$ ifadesi sıfır ortalama ve $\sigma^2$ varyansla normal dağılımı göstermektedir.

Hata paylarının normal dağıldığı varsayımının bazı önemli sonuçları da bulunmaktadır:

a) Bu varsayım bağımlı değişkenlerin de normal dağıldıklarını söylemektedir.

$$Y_i = \beta_0 + \beta_1 X_i + u_i \text{ (ARF)} \tag{1.23}$$

$$\text{var}(Y_i / X_i) = \text{var}(\beta_0 + \beta_1 X_i + u_i) \tag{1.24}$$

Hata paylarının dağılımı hakkında yapılan varsayımlar, aynı dağılıma sahip olması nedeniyle bağımlı değişken için de geçerlidir; çünkü Y ile $u$ arasındaki ilişkide bunlar dışında, $(\beta_0 + \beta_1 X_i)$ sabit bir değerdir ve sabit sayıların eklenmesi ve çıkartılması dağılımı etkilememektedir. Dolayısıyla,

$$\text{var}(Y_i / X_i) = \text{var}(u_i) = \sigma_u{}^2 \tag{1.25}$$

sonucuna varılacaktır.

b) Dağılımı hakkında varsayımda bulunulmayan regresyon modelinde hata terimlerinin korelasyonsuzluğu varsayımı, $E(u_i u_j) = 0$, hata terimlerinin normal dağılması halinde hata terimlerini istatistik açıdan birbirinden bağımsız yapmaktadır. Kısaca hata terimlerinin normal dağılması halinde korelasyonsuzluk bağımsızlık anlamına gelmektedir. Tersi geçerli değildir.

c) $b_0$ ve $b_1$ tahmincileri hata paylarının doğrusal fonksiyonlarıdır. Normal dağılan bir değişkenin doğrusal fonksiyonları da normal dağılacağından $b_0$ ve $b_1$'de normal dağılacaklardır. Dolayısıyla EKK tahmincilerinin olasılık dağılımları normallik varsayımıyla kolayca türetilebilir.

d) Normal dağılım iki parametre içerdiğinden (ortalama ile varyans) göreli olarak basit bir dağılımdır.(Gujatai,1999)

## 1.5 En Çok Olabilirlik Yöntemi

En Çok Olabilirlik Yöntemi aynı nitelikte ve eşit sayıda birimden oluşan ana kütle dağılımlarının parametre değerlerindeki farklılığın farklı örnek dağılımlarına yol açtığı düşüncesine dayanmaktadır. Buna göre, bir örnek dağılımı mümkün parametre değerleri arasında bazılarına daha yakındır. Böylece, yöntemin esası, ana kütle ve ana kütleden çekilen örnek arasındaki benzerlik ilişkisinden yararlanarak bu örneğin elde edilme ihtimalini maksimum yapan parametre değerinin tahmin edilmesidir. (Genceli,2002)

Burada sonuçtan nedene, geriye doğru bir gidiş vardır. Örnek çekildikten sonra geriye doğru dönülerek bu örneğin çekilme olasılığını maksimum yapan parametre tahmin değeri araştırılmaktadır.

Normal dağılan rastlantısal değişken Y için yoğunluk fonksiyonu:

$$f(Y)=\frac{1}{\sqrt{2\pi\sigma}}\exp\left[-\frac{1}{2}\left(\frac{Y-\mu}{\sigma}\right)^2\right] \qquad -\infty<Y<\infty \tag{1.26}$$

$\mu$ ve $\sigma$ normal dağılımın parametre değerleri ve exp (a) ifadesi $e^a$ anlamına gelmektedir.

En Çok Olabilirlik metodu yoğunlukların çarpımını (bağımsızlık varsayımından dolayı ortak olasılık yoğunluk fonksiyonu tekil yoğunluk fonksiyonlarının çarpımına eşittir) parametre değerlerinin örnek verilerine uygunluğunun bir ölçüsü gibi kullanır. Bu çarpım $\mu$ parametresinin olabilirlik değeri olarak adlandırılır ve $L(\mu)$ olarak gösterilir. Eğer $\mu$ örnek verileriyle uygunsa, yoğunluklar ve dolayısıyla da çarpımlar ($L(\mu)$, olabilirlik değeri) kısmen büyük olacaktır. Eğer $\mu$ örnek verileriyle uygun değilse, yoğunluklar ve çarpımı $L(\mu)$ küçük olacaktır.

### 1.5.1 Regresyonda En Çok Olabilirlik Yöntemi

İki değişkenli $Y_i = \beta_0 + \beta_1 X_i + u_i$ modelinde $Y_i$'nin ortalaması $= \beta_0 + \beta_1 X_i$, varyansı=$\sigma^2$ olmak üzere normal, bağımsız dağılmış olduğunu varsayılsın. Bunun sonucunda yukarıdaki ortalama ve varyans verilmişken, $Y_1, Y_2, ..., Y_n$ 'nin ortak olasılık yoğunluk fonksiyonu şöyle yazılabilir:

$$f(Y_1, Y_2, ..., Y_n / \beta_0 + \beta_1 X_i, \sigma^2) \tag{1.27}$$

Y'lerin birbirinden bağımsız olduğu göz önüne alındığında, ortak olasılık yoğunluk fonksiyonu, n tane tekil yoğunluk fonksiyonunun çarpımı olarak şöyle yazılabilir:

$$f(Y_1, Y_2, ..., Y_n / \beta_0 + \beta_1 X_i, \sigma^2)$$
$$= f(Y_1 / \beta_0 + \beta_1 X_i, \sigma^2) f(Y_2 / \beta_0 + \beta_1 X_i, \sigma^2) ... f(Y_n / \beta_0 + \beta_1 X_i, \sigma^2) \tag{1.28}$$

Burada

$$f(Y_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[ -\frac{1}{2} \frac{(Y_i - \beta_0 - \beta_1 X_i)^2}{\sigma^2} \right] \tag{1.29}$$

olup, bu tekil yoğunluk fonksiyonlarının çarpımı olarak hesaplanan ortak olasılık yoğunluk fonksiyonu,

$$L(\beta_0, \beta_1, \sigma^2) = \prod_{i=1}^{n} \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left[ -\frac{1}{2\sigma^2} (Y_i - \beta_0 - \beta_1 X_i)^2 \right]$$
$$L(\beta_0, \beta_1, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left[ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2 \right] \tag{1.30}$$

şeklindedir.

En Çok Olabilirlik yöntemi, bilinmeyen ana kütle katsayılarına ilişkin tahminleri yaparken, verilmiş $Y_i$'leri gözlemleme olasılığını olabildiğince en yükseğe çıkarma yolunu tutar. Dolayısıyla, bu fonksiyonun en yüksek değerinin bulunması gerekmektedir. Bu da doğrudan

bir türev hesabıdır. Türev almak için En Çok Olabilirlik fonksiyonunu logaritma biçiminde yazmak kolaylık sağlar.

$$\log_e L = -\frac{n}{2}\log_e 2\pi - \frac{n}{2}\log_e \sigma^2 - \frac{1}{2\sigma^2}\Sigma(Y_i - \beta_0 - \beta_1 X_i)^2 \tag{1.31}$$

Bu eşitliğin $\beta_0$, $\beta_1$ ve $\sigma^2$'e göre kısmi türevleri alınırsa aşağıdaki eşitlikler elde edilir:

$$\begin{aligned}
\frac{\partial(\log_e L)}{\partial\beta_0} &= \frac{1}{\sigma^2}\Sigma(Y_i - \beta_0 - \beta_1 X_i) \\
\frac{\partial(\log_e L)}{\partial\beta_1} &= \frac{1}{\sigma^2}\Sigma X_i(Y_i - \beta_0 - \beta_1 X_i) \\
\frac{\partial(\log_e L)}{\partial\sigma^2} &= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\Sigma(Y_i - \beta_0 - \beta_1 X_i)^2
\end{aligned} \tag{1.32}$$

yukarıdaki eşitliklerde $\beta_0$, $\beta_1$ ve $\sigma^2$ yerine sıra ile tahmincileri $\hat{\beta}_0, \hat{\beta}_1$ ve $\hat{\sigma}^2$ konulup kısmi türevleri sıfıra eşitlenirse,

$$\begin{aligned}
\Sigma(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) &= 0 \\
\Sigma X_i(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) &= 0 \\
\frac{\Sigma(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2}{n} &= \hat{\sigma}^2
\end{aligned} \tag{1.33}$$

elde edilir. Birinci ve ikinci denklem En Küçük Kareler Yönteminin normal denklemleri ile aynıdır. Dolayısıyla $\hat{\beta}_0, \hat{\beta}_1$ tahminleri en küçük kareler tahminleri ile aynıdır. Üçüncü denklemden elde edilen $\hat{\sigma}^2$, $\sigma^2$'nin sapmalı bir tahmincisidir.

Çizelge 1.1 En Çok Olabilirlik Tahmin Edicisi

| Parametre | En Çok Olabilirlik Tahmincisi |
|---|---|
| $\beta_0$ | $\hat{\beta}_0 = b_0$ E.K.K.Y ile aynı |
| $\beta_1$ | $\hat{\beta}_1 = b_1$ E.K.K.Y ile aynı |
| $\sigma^2$ | $\hat{\sigma}^2 = \frac{\Sigma(Y_i - \hat{Y}_i)^2}{n}$ |

$\hat{\sigma}^2$ sapmalı bir tahmincidir. Genellikle sapmasız bir tahminci olan MSE kullanılır. n yeterince büyükse MSE En Çok Olabilirlik tahmincisi $\hat{\sigma}^2$'den az bir farklılık gösterir.

$\hat{\beta}_0, \hat{\beta}_1$ En Çok Olabilirlik tahmincileri EKK tahmin edicileriyle aynı özelliklere sahiptirler:

a. sapmasız
b. bütün doğrusal tahmin ediciler arasında minimum varyanslı
c. Tutarlı
d. Etkin
e. Sapmasız tahmin ediciler arasında en küçük varyansa sahiptirler.

Sonuç olarak hata terimleri normal dağılan klasik doğrusal regresyon modelinde $b_0$ ve $b_1$ pek çok arzu edilir özelliğe sahiptirler (Neter vd.,1996).

## 1.6 Çoklu Doğrusal Regresyon Modeli

Genellikle bağımlı değişken birden çok bağımsız değişkenin etkisi altındadır. Dolayısıyla bu aşamaya kadar incelenmiş olan iki değişkenli regresyon modeli uygulamada yetersiz kalmaktadır. Dolayısıyla iki değişkenli regresyon modeli ikiden çok değişkeni içeren modelleri kavrayacak biçimde genişletilmelidir.

Yalın regresyonda olduğu gibi katsayılarında doğrusal regresyon modelleri ele alınacaktır. Çoklu doğrusal regresyon modelinin en basit biçimi bir bağımlı ve iki bağımsız değişkenden oluşan birinci mertebeden bir modeldir:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i \tag{1.34}$$

Çoklu doğrusal regresyon modelinin varsayımları iki değişkenli modelin bir uzantısıdır. Bu bakımdan oradaki varsayımlar yinelenmeyecektir.

Yalın doğrusal regresyon modelinin varsayımlarına ilave olarak, çoklu doğrusal regresyon modelinde bağımsız değişkenler arasında kesin bir doğrusal bağlantı bulunmamaktadır. Bu varsayım "çoklu doğrusal bağlantı" bulunmama varsayımıdır. Varsayımla söylenen bağımsız değişkenlerden hiç birinin diğerlerinin doğrusal bağlantısı olmamasıdır. Örneğin

$X_{2i} = 4X_{3i}$ gibi durumlarda çoklu doğrusal bağlantı söz konusudur. Bu varsayım örneğe ilişkin bir sorundur. Dolayısıyla hipotez testine tabi tutulmaz.

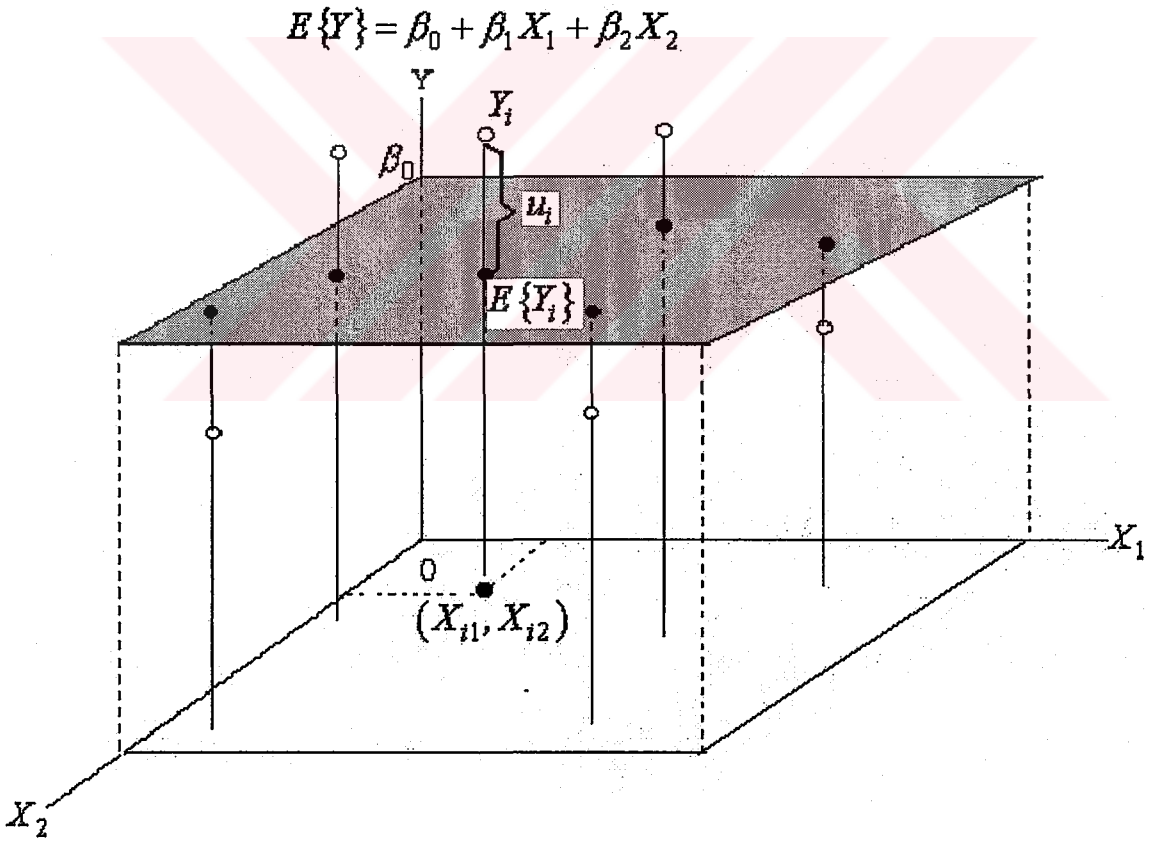
Bu varsayımları gerçekleştiren modele "Klasik Normal Çoklu Regresyon" adı verilmektedir. İki bağımsız değişkenden oluşan en basit çoklu doğrusal regresyonun herhangi bir birimi

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$$

dir. Beklenen değeri alındığı takdirde ana kütle regresyon denklemi elde edilir:

$$E\{Y_i\} = E\{\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i\}$$

$$E\{Y_i\} = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} \qquad E\{u_i\} = 0 \tag{1.36}$$




Şekil 1.6 Regresyon yüzeyi*

* Şekil: Neter,J.,Kutner,M.H.,Nachtsheim,C.J.,Wasserman,W., s.219

$\beta_0$ parametresi regresyon sabiti veya kesişimidir. Regresyon yüzeyi ile Y'nin kesişimini vermektedir, diğer bir deyişle $X_1 = X_2 = 0$ olduğu zaman $E\{Y\}$'nin orjinden yüksekliğini vermektedir. Ekonometrik açıdan ise , $\beta_0$, modele alınmayan diğer değişkenlerin Y'ye yaptıkları ortalama etkidir. $X_1 = 0, X_2 = 0$ modelin veri kapsamında bulunması halinde $E\{Y / X_1 = 0, X_2 = 0\}$ değerini vermektedir. Aksi takdirde özel bir anlamı bulunmamaktadır (Neter vd.,1996).

$\beta_1$ ve $\beta_2$ kısmi regresyon parametreleridir. Kısmi regresyon parametresi $\beta_1$, $X_2$ sabit tutulurken, $X_1$deki bir birimlik değişmeye karşılık Y'nin beklenen değeri olan $E\{Y / X_1, X_2\}$'deki değişmeyi ölçer. Bir başka deyişle, $X_2$ sabitken, $E\{Y / X_1, X_2\}$'ün $X_1$'e göre eğimini verir. Değişik biçimde ifade edilirse, $X_1$'deki bir birimlik değişmenin, Y üzerindeki $X_2$'den ayrı "doğrudan" ya da "net" etkisini gösterir. $\beta_2$ de benzer şekilde yorumlanır.

Genel olarak $\beta_k$ yorumlanırken bu parametreye ilişkin değişken, $X_k$ dışındaki diğer tüm bağımsız değişkenlerin etkilerinin sabit olduğu varsayılmaktadır. Bu varsayım altında $X_k$'daki bir birim değişmenin bağımlı değişkenin ortalamasında ne kadar değişmeye yol açtığını $\beta_k$ göstermektedir.

Matematiksel olarak,

$$\frac{\partial E\{Y / X_1,..., X_k\}}{\partial X_k} = \beta_k \qquad (1.37)$$

$\beta_1, \beta_2,..., \beta_k$ parametrelerinin işaretleri ilişkin oldukları bağımsız değişken ile bağımlı değişken arasındaki bağlantının yönünü vermektedir. Buna karşın hangi bağımsız değişkenin etkisinin daha fazla olduğu hususunda $\beta$'nın büyüklüğü gerekli olmakla birlikte yeterli değildir. Parametrelerin karşılaştırmalarda kullanılabilmesi için tüm bağımsız değişkenlerin aynı ölçü birimine tabi olması gerekmektedir.(Genceli,2002)

## 1.7 Genel Doğrusal Regresyon Modeline Matris Yaklaşımı

Genel doğrusal model,

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_{p-1} X_{i.p-1} + u_i$$

olarak ifade edilirse, matris formunda aşağıdaki matrisler tanımlanmalıdır:

$$\underset{nx1}{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ . \\ . \\ Y_n \end{bmatrix} \qquad \underset{nxp}{X} = \begin{bmatrix} 1 & X_{11} & X_{12} & ... & X_{1.p-1} \\ 1 & X_{21} & X_{22} & ... & X_{2.p-1} \\ . & . & . & . & . \\ . & . & . & . & . \\ 1 & X_{n1} & X_{n2} & ... & X_{n.p-1} \end{bmatrix} \tag{1.38}$$

$$\underset{px1}{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ . \\ . \\ \beta_{p-1} \end{bmatrix} \qquad \underset{nx1}{u} = \begin{bmatrix} u_1 \\ u_2 \\ . \\ . \\ u_n \end{bmatrix} \tag{1.39}$$

Y ve u vektörlerinin yalın doğrusal regresyon modeli için benzer olduklarına dikkat edilmelidir. $\beta$ vektörü ilave regresyon parametrelerinden oluşmaktadır.

Matris terimleriyle genel doğrusal regresyon modeli:

$$\underset{nx1}{Y} = \underset{nxp}{X}\ \underset{px1}{\beta} + \underset{nx1}{u} \tag{1.40}$$

Y :bağımlı değişken vektörü

$\beta$ : parametre vektörü

X :bağımsız değişken matrisi

u :bağımsız normal rastlantısal değişken vektörü

$E\{u\} = 0$ ve varyans- kovaryans matrisi:

$$\underset{nxn}{\sigma^2\{u\}} = \begin{bmatrix} \sigma^2 & 0 & 0 & . & 0 \\ 0 & \sigma^2 & 0 & . & 0 \\ 0 & 0 & \sigma^2 & . & 0 \\ . & . & . & . & . \\ 0 & 0 & 0 & . & \sigma^2 \end{bmatrix} = \sigma^2 I \tag{1.41}$$

$$\underset{nx1}{E\{Y\}} = X\beta \tag{1.42}$$

ve Y'nin varyans kovaryans matrisi u'ya eşittir:

$$\underset{nxn}{\sigma^2\{Y\}} = \sigma^2 I \tag{1.43}$$

**Regresyon Parametrelerinin Tahmini**

Genel doğrusal regresyon modeli için en küçük kareler kriteri :

$$Q = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_{i1} - .... - \beta_{p-1} Xi.p-1)^2 \tag{1.44}$$

$\beta_0, \beta_1, ... \beta_{p-1}$ parametrelerinin en küçük kareler yöntemi tahmincileri Q'yu minimize eden tahmincilerdir. Tahmin edilen regresyon katsayılarını $b_0, b_1, ..., b_{p-1}$ vektörü göstermektedir:

$$\underset{px1}{b} = \begin{bmatrix} b_0 \\ b_1 \\ . \\ . \\ b_{p-1} \end{bmatrix} \tag{1.45}$$

genel doğrusal model için en küçük karelerin normal denklemleri aşağıdaki gibidir:

$$X'Xb = X'Y \tag{1.46}$$

ve en küçük kareler tahmincileri aşağıdaki gibi elde edilir:

$$\underset{px1}{b} = \underset{pxp}{(X'X)^{-1}}\underset{px1}{(X'Y)} \tag{1.47}$$

***Öngörü Değerleri ve Kalıntılar***

$\hat{Y}_i$ vektörü yerine $\hat{Y}$ ve $e_i = Y_i - \hat{Y}_i$ vektörü yerine $e$ notasyonu kullanılacaktır:

$$\underset{nx1}{\hat{Y}} = \begin{bmatrix} \hat{Y}_1 \\ \hat{Y}_2 \\ . \\ . \\ \hat{Y}_n \end{bmatrix} \qquad \underset{nx1}{e} = \begin{bmatrix} e_1 \\ e_2 \\ . \\ . \\ e_n \end{bmatrix} \tag{1.48}$$

öngörü değerleri,

$$\underset{nx1}{\hat{Y}} = Xb = X\left(X'X\right)^{-1}X'Y = HY \qquad (H = X(X'X)^{-1}X') \tag{1.49}$$

sonucuna ulaşılacaktır. H ile gösterilen nx1 boyutlu matris Y gözlem değerlerini $\hat{Y}$, öngörü değerlerine dönüştürmektedir. Bu nedenle de H, dönüşüm matrisi olarak adlandırılmaktadır.

kalıntılar,

$$\underset{nx1}{e} = Y - \hat{Y} = Y - HY = \left(I - H\right)Y \tag{1.50}$$

kalıntıların varyans- kovaryans matrisi,

$$\underset{nxn}{\sigma^2\{e\}} = \sigma^2\left(I - H\right) \tag{1.51}$$

biçimindedir. Tahmincisi,

$$\underset{nxn}{s^2\{e\}} = MSE(I - H) \tag{1.52}$$

şeklinde elde edilir.

## 1.8 Regresyon Analizine Varyans Analizi Yaklaşımı

Özellikle çoklu regresyon açısından önemli olan bu yaklaşım aynı verilere uygulanan çeşitli modellerin seçiminde yol gösterici bir işlev yapmaktadır. İster basit ister çoklu regresyon uygulansın, yaklaşımın ilkeleri aynı olup Y bağımlı değişkenine ilişkin değişkenlik ve serbestlik derecesinin kaynaklarına ayrılması ilkesine dayanmaktadır.

***Toplam Değişkenliğin Ayrıştırılması***

$Y_i - \bar{Y}$ , $Y_i$ gözlemlerinin kendi oratalamalarından sapmalarını ifade etmektedir. ve SSTO ile gösterilen toplam değişkenlik ölçüsü bu sapmaların kareleri toplamına eşittir.

$$SSTO = \sum (Y_i - \bar{Y})^2$$

Bağımsız değişken X 'ten bilgi alındığında $Y_i$ gözlemlerinin tahmin edilen regresyon doğrusundan sapmalarını ifade eden değişkenlik: $Y_i - \hat{Y}_i$'dir. Bu sapmaların kareleri toplamı,

$$SSE = \sum (Y_i - \hat{Y}_i)^2$$

SSE hata kareleri toplamı olarak ifade edilir.
Bu iki kareler toplamı arasındaki fark diğer kareler toplamı:

$$SSR = \sum (\hat{Y}_i - \bar{Y})^2 \tag{1.53}$$

Regresyonla açıklanan değişkenliği ifade eder, dolayısıyla regresyon kareler toplamı olarak gösterilir (Regression sum of squares). Toplam değişkenlik SSTO, regresyon ile açıklanabilen SSR ve açıklanamayan SSE olmak üzere iki kısma ayrılır: $SSTO = SSR + SSE$

$SSTO$'nun bileşenlerine ayrıştırılmasına karşılık ilgili serbestlik derecesi de bileşenlerine ayrıştırılır. $SSTO$ n – 1 serbestlik derecesine sahiptir. $SSE$ p tane parametre tahmin edilmek istendiğinden n – p serbestlik derecesine sahiptir. Son olarak $SSR$, bağımsız değişkenlerin sayısını ifade etmek üzere, $X_1, X_2, ..., X_{p-1}$, p – 1 serbestlik derecesine sahiptir.

Ortalamadan sapmaların kareleri toplamının ilgili serbestlik derecesine bölünmesiyle MS olarak gösterilen ortalama kareler toplamı elde edilir.

Varyans analiz tablosu bütün kareler toplamının ve ilgili serbestlik derecelerinin bileşenlerine ayrıldığı bir tablodur.

Varyans analizinde matris terimleriyle toplam değişkenlik ve bileşenleri tabloda görüldüğü gibi hesaplanır:

Çizelge 1.2 Varyans analiz tablosu (ANOVA)

| **Değişkenlik Kaynağı** | *SS* | *df* | *MS* |
|---|---|---|---|
| Regresyon | $SSR = b'X'Y - \left(\frac{1}{n}\right)Y'JY$ | $p-1$ | $MSR = \frac{SSR}{p-1}$ |
| Hata | $SSE = Y'Y - b'X'Y$ | $n-p$ | $MSE = \frac{SSE}{n-p}$ |
| Toplam | $SSTO = Y'Y - \left(\frac{1}{n}\right)Y'JY$ | $n-1$ | |

$$SSTO = Y'Y - \left(\frac{1}{n}\right)Y'JY = Y'\left[I - \left(\frac{1}{n}\right)J\right]Y \tag{1.54}$$

$$SSE = e'e = (Y - Xb)'(Y - Xb) = Y'Y - b'X'Y = Y'(I - H)Y \tag{1.55}$$

$$SSR = b'X'Y - \left(\frac{1}{n}\right)Y'JY = Y'\left[H - \left(\frac{1}{n}\right)J\right]Y \tag{1.56}$$

biçimindedir. $J$ bütün elemanları 1 olan nxn boyutundaki kare matris ve H dönüşüm matrisidir.

### 1.8.1 Regresyon İlişkisi İçin F Testi

Bağımlı değişken Y ve bağımsız X değişken seti $X_1, X_2, ..., X_{p-1}$ arasında bir regresyon ilişkisinin olup olmadığını test etmek için alternatif hipotezler:

$$H_0 : \beta_1 = \beta_2 = ... = \beta_{p-1} = 0$$

$H_a$ : Bütün $\beta_k$ 'lar (k=1,2,...p-1) sıfıra eşit değildir. (1.57)

(Parametrelerin en az bir tanesi sıfırdan farklıdır.)

şeklinde kurulur. Test istatistiği,

$$F^* = \frac{MSR}{MSE} \tag{1.58}$$

olarak hesaplanır. $\alpha$ anlamlılık seviyesinde karar kuralı:

$F^* \le F(1-\alpha; p-1, n-p)$ ise $H_0$ kabul

$F^* > F(1-\alpha; p-1, n-p)$ ise $H_a$ kabul

şeklindedir.

## 1.9 Çoklu Belirginlik Katsayısı

Çok değişkenli doğrusal regresyon modelinde, modelin bağımsız değişkenlerinin bağımlı

değişkendeki değişkenliği ne derece açıkladığının bir ölçüsü çoklu belirginlik katsayısı $R^2$,

$$R^2 = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO} \tag{1.59}$$

olarak tanımlanmaktadır.

$0 \le R^2 \le 1$ arasında yer alan çoklu belirginlik katsayısı regresyon düzleminin verilere yakınlığını gösteren bir uygunluk ölçüsü olarak kullanılmaktadır. Bu yaklaşım içersinde $R^2$'nin yüksek olması modelin verilere iyi uyduğunun, düşük olması ise uymadığının göstergesidir.

Fakat $R^2$, sadece kısmi bir ölçü olup bazı sınırlamalara tabidir. Öncelikle, $SSTO = \sum\left(Y_i - \bar{Y}\right)^2$ toplam değişkenliğin, SSR, regresyonla açıklanabilen değişkenlik ve SSE, regresyonla açıklanamayan değişkenlik olarak ayrılarak $R^2$'nin yorumlanması üç koşula bağlıdır:

a. $b_0, b_1, .., b_k$ tahmin edicileri EKK tahmin edicileri olmalıdır. EKK hata paylarına ilişkin değişkenliği minimum yaptığından $R^2$'yi de maksimize etmektedir.

b. İlişki doğrusal olmalıdır. $R^2$ istatistiği bağımsız değişkenlerdeki değişkenliğin bağımlı değişkendeki değişkenliği doğrusal olarak açıkladığı bir ölçüdür.

c. Doğrusal model kesinlikle regresyon sabiti içermelidir. Olmadığı durumda $R^2 < 0$ veya $R^2 > 1$ olabilir. Dolayısıyla $R^2$'nin $0 \le R^2 \le 1$ sınırları arasında bulunabilmesi ancak modelde regresyon sabitinin yer alması ile mümkündür.

Bu koşulların geçerli olması halinde yüksek $R^2$'nin kaç olduğu ve yüksek bir $R^2$'nin çözüm olup olmadığı sorularıyla karşılaşılmaktadır.

$R^2$'nin yüksek çıkması her zaman modelin verilere iyi uyduğunun kesin bir ölçüsü değildir. Aynı zamanda $R^2$ büyük olsa dahi MSE, yüksek kesinlik gerektiren çıkarsamalar için oldukça büyük olabilir.

Çok değişkenli doğrusal regresyonda model kapsamına alınan her ilave bağımsız değişkenden sağlanan bilgi hata paylarına ilişkin değişkenlik olan $SSE = \sum e_i{}^2$'yi azaltacak veya en azından aynı kalmasına yol açacaktır. SSE'nin azalması, $R^2$'nin artması demektir. Bu durumda ilave her bağımsız değişkenin $R^2$'nin değerini yükselttiği, yükseltmese dahi düşürmediği söylenebilir. Fakat bağımsız değişken ilave etmenin diğer bir yönü vardır. Modele giren her ilave bağımsız değişken ayrıca bir parametre tahminine yol açacağından $s^2 = SSE/n-p$ 'nin hem payı hem paydası azalacaktır. Çünkü bağımsız değişken sayısı bir artmıştır. Diğer bir deyişle ilave her bağımsız değişken serbestlik derecesini bir azaltmaktadır. Fakat $R^2$'de yer alan toplam değişkenlik $SSTO = \sum\left(Y_i - \overline{Y}\right)^2$ ve serbestlik derecesi (n-1) değişmemektedir. (n-1) serbestlik derecesi sadece birim sayısına bağlıdır. Bu durumu önlemek amacıyla düzeltilmiş $R^2$ kullanılır. Aşağıdaki gibi hesaplanılır:

$$R_a{}^2 = 1 - \frac{\frac{SSE}{n-p}}{\frac{SSTO}{n-1}} = 1 - \left(\frac{n-1}{n-p}\right)\frac{SSE}{SSTO} \qquad (1.60)$$

Serbestlik derecelerini içeren $R_a^2$, düzeltilmiş belirginlik katsayısı olarak adlandırılır. Modele yeni bir bağımsız değişken eklenmesi halinde düzeltilmiş belirginlik katsayısının büyüklüğü $R^2$'den daha az olacaktır.

## 1.10 Matris Notasyonuyla Regresyon Parametrelerine İlişkin Çıkarsamalar

Varyans-kovaryans matrisi, $\sigma^2\{b\}$:

$$\underset{pxp}{\sigma^2\{b\}} = \begin{bmatrix} \sigma^2\{b_0\} & \sigma^2\{b_0,b_1\} & . & . & \sigma^2\{b_0,b_{p-1}\} \\ \sigma^2\{b_1,b_0\} & \sigma^2\{b_1\} & . & . & \sigma^2\{b_1,b_{p-1}\} \\ . & . & . & . & . \\ . & . & . & . & . \\ \sigma^2\{b_{p-1},b_0\} & \sigma^2\{b_{p-1},b_1\} & . & . & \sigma^2\{b_{p-1}\} \end{bmatrix} \quad (1.61)$$

$\underset{pxp}{\sigma^2\{b\}} = \sigma^2(X'X)^{-1}$ şeklinde hesaplanır.

Tahmin edilen varyans- kovaryans matrisi $s^2\{b\}$,

$$\underset{pxp}{s^2\{b\}} = \begin{bmatrix} s^2\{b_0\} & s^2\{b_0,b_1\} & . & . & s^2\{b_0,b_{p-1}\} \\ s^2\{b_1,b_0\} & s^2\{b_1\} & . & . & s^2\{b_1,b_{p-1}\} \\ . & . & . & . & . \\ . & . & . & . & . \\ s^2\{b_{p-1},b_0\} & s^2\{b_{p-1},b_1\} & . & . & s^2\{b_{p-1}\} \end{bmatrix} \quad (1.62)$$

$\underset{pxp}{s^2\{b\}} = MSE(X'X)^{-1}$ ile hesaplanır.

## 1.11 Kısmi Belirlilik Katsayıları

Kısmi belirlilik katsayısı diğer bütün değişkenler modelde iken bir X değişkeninin Y'deki değişkenliği açıklamadaki marjinal katkısını ölçen bir tanımlayıcı ölçüdür.

İki Bağımsız Değişkenli Model İçin, $X_2$ modelde iken $X_1$ ve Y arasındaki kısmi belirlilik katsayısı, $r^2_{Y1.2}$ aşağıdaki gibidir:

$$r^2_{Y1.2} = \frac{SSE(X_2) - SSE(X_1, X_2)}{SSE(X_2)} = \frac{SSR(X_1 / X_2)}{SSE(X_2)} \quad (1.63)$$

$SSE(X_2) - SSE(X_1, X_2)$ farkı $X_2$'nin tek bağımsız değişken olduğu modele göre $X_1$ ve $X_2$'nin bağımsız değişken olduğu modeldeki $SSE$'deki azalış veya $SSR$'deki marjinal artıştır. $SSR(X_1 / X_2) = SSE(X_2) - SSE(X_1, X_2)$, $X_2$ modelde iken $X_1$'in modele dahil edilmesiyle $SSE$'deki marjinal azalış veya $SSR$'deki marjinal artışı ifade etmektedir.

$r^2_{Y1.2}$ iki farklı kalıntı kümesi;

$Y = b_0 + b_1 X_2$ modelinden elde edilen kalıntılar ile

$X_1 = a_0 + a_1 X_2$ modelinden elde edilen kalıntılar arasındaki korelasyon katsayısıdır.

Bu şekilde $X_1$'in marjinal katkısını ölçebilmek için Y'den $X_2$'nin ve $X_1$'den $X_2$'nin katkısı düşülmüş olmaktadır (Neter vd.,1996).

## 1.12 Uyum Eksikliği İçin *F* Testi

Belirli bir regresyon fonksiyonunun veriye uygunluğunu inceleyen bu test doğrusal regresyon fonksiyonunun veriye uygun olup olmadığını araştırmak için kullanılmaktadır.

**Varsayımlar**

Uyum eksikliği testinde verili bir X için Y gözlemlerinin (1) bağımsız, (2) normal dağılan, ve (3) Y'nin dağılımlarının eşit varyansa sahip olduğu varsayımları yapılmaktadır. Uyum eksikliği testinde belirli bir X seviyesinde bir veya daha çok tekrarlı gözlemlerin olması gerekmektedir. Bağımsız değişkenin aynı seviyesinde birden çok Y değeri elde edilmesi halinde tekrarlı gözlemden (replication) söz edilir.

**Notasyon**

c ifadesi X bağımsız değişkeninin seviyelerini, $X_1,...,X_c$, göstermektedir. X'in j.inci seviyesindeki tekrar eden gözlemleri göstermek üzere $n_j$ kullanılır. Toplam gözlem sayısı

$$n = \sum_{j=1}^{c} n_j \tag{1.64}$$

X'in j.inci seviyesi için i.inci tekrar eden Y gözlemleri $Y_{ij}$ ile (i=1,...,$n_j$ ve j=1,...,c) ve $X = X_j$ seviyesinde Y gözlemlerinin ortalaması $\overline{Y}_j$ olarak ifade edilecektir.

**Tam Model**

Genel doğrusal test yaklaşımı tam modeli belirleme ile başlar. Doğrusal resresyon ilişkisinin olup olmadığını test eder. Tam model:

$$Y_{ij} = \mu_j + u_{ij} \tag{1.65}$$

$\mu_j$ parametre değerleri. j=1,...,c

$u_{ij}$ bağımsız normal dağılan hata terimi $N(0,\sigma^2)$

hata teriminin beklenen değeri sıfır olduğundan

$$E\{Y_{ij}\} = \mu_j \tag{1.66}$$

$\mu_j$ parametresi $X = X_j$ iken bağımlı değişkenin regresyon doğrusu üzerindeki değerini ifade etmektedir. Regresyon modeline benzer şekilde tam model iki bileşenden oluşmaktadır: bağımlı değişkenin ortalama değeri ve hata terimi. Tam modelle klasik doğrusal regresyon modeli arasındaki fark tam modelde $\mu_j$ parametresi üzerinde herhangi bir kısıt yokken, klasik regresyon modelinde $\mu_j$ parametresinin X'lerle doğrusal bir ilişkide bulunduğu kısıtı vardır ($E\{Y\} = \beta_0 + \beta_1 X$).

En Küçük Kareler veya En Çok Olabilirlik yöntemleri kullanılarak veri setini tam modelle öngörebilmek için $\mu_j$ parametresinin tahmincileri bulunabilir.

$\mu_j$ parametrelerinin $\overline{Y}_j$ örnek ortalama değerleri oldukları görülebilir.

$$\hat{\mu}_i = \overline{Y}_j \tag{1.67}$$

$Y_{ij}$ gözlemi için tahmin edilen beklenen değer böylelikle $\overline{Y}_j$ olur. Tam model için hata kareleri toplamı:

$$SSE(F) = \sum_j \sum_i (Y_{ij} - \overline{Y}_j)^2 = SSPE \tag{1.68}$$

biçimindedir. Uyum eksikliği için hata kareleri toplamı burada saf hata kareleri toplamı olarak adlandırılır ve SSPE olarak gösterilir. SSPE her bir X seviyesindeki sapma kareleri toplamından oluşmaktadır. $X = X_j$ seviyesinde bu sapma karesi toplamı:

$$\sum_i (Y_{ij} - \overline{Y}_j)^2 \tag{1.69}$$

şeklinde hesaplanılabilir.

Verili bir X seviyesinde n gözleme dayalı sapma kareleri toplamı n-1 serbestlik derecesine sahiptir. Burada $X = X_j$ için $n_j$ gözlem bulunmaktadır dolayısıyla ilgili serbestlik derecesi $n_j$-1'dir. SSPE sapma kareleri toplamının toplamı olduğundan SSPE ile ilgili serbestlik derecesi:

$$df_F = \sum_j (n_j - 1) = \sum_j n_j - c = n - c \tag{1.70}$$

biçimindedir.

**İndirgenmiş Model**

Genel doğrusal test yaklaşımı indirgenmiş modeli $H_0$ hipotezi altında ele almayı gerektirir. Doğrusal regresyon ilişkisinin uygun olup olmadığının testi için alternatif hipotezler:

$$H_0 : E\{Y\} = \beta_0 + \beta_1 X$$
$$H_a : E\{Y\} \neq \beta_0 + \beta_1 X \tag{1.71}$$

biçimindedir. $H_0$, tam modeldeki $\mu_j$'nin $X_j$ ile doğrusal ilişkide olduğunu varsaymaktadır:

$$\mu_j = \beta_0 + \beta_1 X_j \tag{1.72}$$

$H_0$ hipotezi altında indirgenmiş model:

$$Y_{ij} = \beta_0 + \beta_1 X_j + u_{ij} \tag{1.73}$$

biçimindedir ve $Y_{ij}$'nin tahmini aşağıdaki gibidir:

$$\hat{Y}_{ij} = b_0 + b_1 X_j \tag{1.74}$$

indirgenmiş model için hata kareleri toplamı SSE(R):

$$SSE(R) = \sum\sum \left[ Y_{ij} - (b_0 + b_1 X_j)^2 \right] = \sum\sum (Y_{ij} - \hat{Y}_{ij})^2 \tag{1.75}$$

olarak hesaplanır. SSE(R) ile ilgili serbestlik derecesi $df_R = n - 2$ dir.

**Test İstatistiği**

Genel doğrusal test istatistiği:

$$F^* = \frac{SSE(R) - SSE(F)}{df_R - df_F} \div \frac{SSE(F)}{df_F} = \frac{SSE - SSPE}{(n-2) - (n-c)} \div \frac{SSPE}{n-c} \tag{1.76}$$

şeklinde hesaplanır. İki hata kareleri toplamı arasındaki fark uyum eksikliği hata kareleri toplamı olarak adlandırılır ve SSLF olarak gösterilir: SSLF=SSE-SSPE.

Test istatistiği aşağıdaki gibi ifade edilebilir:

$$F^* = \frac{SSLF}{c-2} \div \frac{SSPE}{n-c} = \frac{MSLF}{MSPE} \tag{1.80}$$

$F^*$'ın yüksek değerleri $H_a$'nın kabülüne sebep olur. Karar kuralı:

$$F^* \leq F(1-\alpha;c-2,n-c) \quad H_0 \text{ kabul}$$
$$F^* > F(1-\alpha;c-2,n-c) \quad H_a \text{ kabul} \qquad (1.81)$$

biçimindedir (Neter vd.,1996).

## 2. EN KÜÇÜK KARELERİN GEOMETRİSİ

Öngörü sürecinin açıklanabilmesi; Xβ regresyon doğrusunun Y vektörü üzerine uydurulması için sıradan E.K.K. metodunun geometrik yapısı açıklanacaktır.

Doğrusal modeller için geometrik yaklaşım $Y=(Y_1,...,Y_n)^T$ gözlem vektörünü $R^n$ vektör uzayında bir nokta gibi düşünmeyi gerektirir. Ve benzer olarak Y'nin tahmin vektörünü ifade eden $\mu=(\mu_1,...,\mu_n)^T$ vektörü, $R^n$ uzayında bir vektör olarak düşünülür. Dolayısıyla $\mu$ ve $Y$ nx1 boyutlu vektörlerdir.

$$\mu_i=\beta_1+\beta_2 x_i \quad \text{i=1,...,n} \tag{2.1}$$

Burada $x_1,...,x_n$ bilinen sabitlerdir ve $\beta_1,\beta_2$ bilinmeyen parametrelerdir. Yalın doğrusal regresyon modeli aşağıdaki gibi ifade edilebilir:

$$\mu=\beta_1.1+\beta_2.\mathrm{x} \tag{2.2}$$

$$1=\begin{bmatrix}1\\1\\1\\\vdots\\1\end{bmatrix} \qquad \mathrm{x}=\begin{bmatrix}x_1\\x_2\\\vdots\\\vdots\\x_n\end{bmatrix} \tag{2.3}$$

Böylece yalın doğrusal regresyon modeli $\mu$'nün 1 ve x bilinen vektörlerinin bilinmeyen katsayılarla doğrusal bir kombinasyonunu göstermektedir (Ruud,2000). Doğrusal bir model için genel bazı tanımlamalar aşağıdaki gibidir:

k boyutlu doğrusal bir model, $\mu$ tahmin vektörü için aşağıdaki şekilde tanımlanan bir hipotezdir:

$$H_0 : \mu \in L \tag{2.4}$$

$$H_a : \mu \notin L$$

$L$, $R^n$'in k boyutlu bir alt uzayıdır. Burada bazı lineer cebir kurallarının hatırlanması gerekmektedir:

i) $0 \in L$ (2.5)

ii) $x_1, x_2 \in L \Rightarrow ax_1 + bx_2 \in L \quad \forall a,b \in R$ (2.6)

Yukarıda verilen iki koşulun gerçekleşmesi halinde $L$, $R^n$'in bir alt uzayıdır.

L alt uzayı için bazlar $x_1,...,x_k$ elemanlarından oluşur ve bu elemanlar aşağıdaki koşulları gerçekleştirirler:

(i) L alt uzayında herhangi bir vektör $x_1,...,x_k$'nın doğrusal bir kombinasyonudur.

(ii) $x_1,...,x_k$ doğrusal olarak bağımsızdır ve,

$$a_1x_1 + ... + a_kx_k = 0 \Rightarrow a_1 = ... = a_k = 0 \tag{2.7}$$

dır. (i) ve (ii) koşulları sağlanırsa L alt uzayı k boyutlu kabul edilir. Burada tanımlanan x'ler X değişken matrisinin yalın doğrusal regresyon modeli için sahip olduğu 1 hariç kalan tek değişkenin değerlerini ifade etmektedir.

L doğrusal modeli için $x_1,...,x_k$ baz verilmişse,

$$\mu = \sum_{j=1}^{k} x_j \beta_j \tag{2.8}$$

veya matris notasyonuyla,

$$\mu = X\beta$$

yazılabilir. X, nxk boyutlu, $x_1,...,x_k$ sütunlarından oluşan bir matris ve $\beta = (\beta_1,...,\beta_k)^T$ kx1 boyutlu parametre vektörüdür.

Aşağıda E.K.K.'in geometrisini açıklamak üzere kullanılacak bazı işlemler tanımlanmıştır:

$x$ ve $y$ $; x, y \in R^n$ olmak üzere, bu iki vektörün içsel çarpımı,

$$\langle x.y \rangle = x^T y = y^T x = \sum_{i=1}^{n} x_i y_i \tag{2.9}$$

şeklinde tanımlanır.

$x \in R^n$ olmak üzere bir vektörün uzunluğunun hesabını ifade eden $\|.\|$ işlemi,

$$\|x\| = \sqrt{x^T x} = \sqrt{x.x} = \sqrt{\sum_{i=1}^{n} {x_i}^2} \tag{2.10}$$

şeklinde tanımlanır.

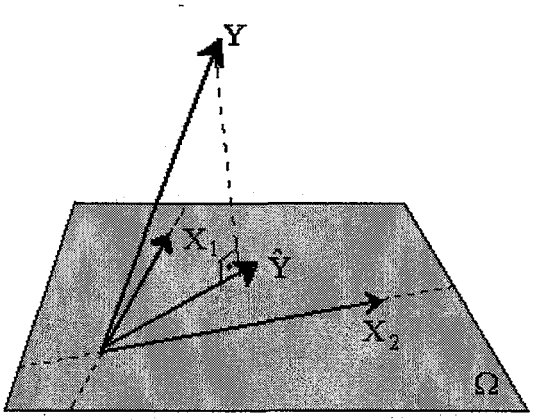
$R^n$'de $x$ ve $y$ için $x.y = 0$ ise $x$ ve $y$ ortogonaldir denir. Bu durum $x \perp y$ ile ifade edilir. $x \perp y$ ise pisagor teoremine dayanarak,

$$\|x + y\|^2 = \|x\|^2 + \|y\|^2 \tag{2.11}$$

yazılabilir (Jorgensen,1992).

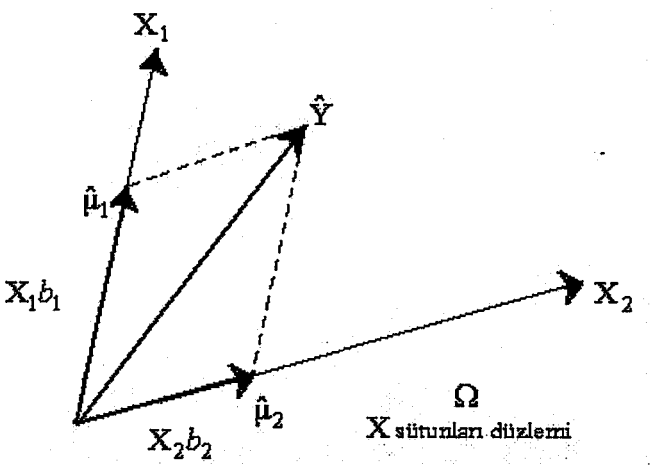
X'in sütunları n boyutlu uzayda vektörleri tanımlamaktadır. Bu şekilde bu vektörler p boyutlu bir alt uzay oluştururlar. X'in p adet (değişken sayısı) sütunlarının doğrusal olarak bağımsız oldukları varsayılmaktadır. Şöyle ki bu sütunlardan herhangi biri diğerlerinin doğrusal bir kombinasyonu olarak elde edilemez. Bu alt uzay $\Omega$ ile gösterilecektir.

Y vektörü , $X_1$ ve $X_2$ ile tanımlanan düzlemde bulunmaz. E.K.K. öngörü süreci, X sütunları düzleminde $\hat{Y}$ olarak gösterilen,Y'ye en yakın olarak elde edilebilecek vektörü bulur. Bu Y'nin gölgesi ya da yansıması gibi düşünülebilir dolayısıyla $\hat{Y}$,Y'nin projeksiyonu olarak adlandırılır:




Şekil 2.1 Veri setinin vektörel gösterimi

Diğer yandan örneğin iki değişkenli (iki sütun) E.K.K. regresyonu için $\hat{Y}$'nin $\hat{\mu}_1$ ve $\hat{\mu}_2$ öngörü bileşenlerine nasıl ayrıştırılabileceği tartışılacaktır. Bu öngörü bileşenleri X'in sütun vektörleriyle E.K.K. katsayılarının çarpılmasından elde edilir: $\hat{\mu}_1 = X_1 b_1, \hat{\mu}_2 = X_2 b_2$ (Ruud,2000). $b_1$ ve $b_2$, $\hat{Y}$'yı elde edebilmek amacıyla $X_1$ ve $X_2$'nin hangi oranlarda biraraya geleceğini gösterir (Evren,1997).




Şekil 2.2 $\hat{Y}$'nin bileşenlerine ayrıştırılması

E.K.K. yöntemi gerçek değerlerden tahmini değerlerin sapmalarının kareleri toplamını minimize eden çözümü bulur. Diğer bir deyişle $b$, Y ve olabilecek $X\beta$ arasındaki kareli uzaklığı minimum yapan $\beta$ değeridir. Y ve $X\beta$ arasındaki sapma kareleri toplamı, Y ve $X\beta$ arasındaki kareli *öklid* uzaklığıdır (Ruud,2000).

E.K.K. yönteminin amaç fonksiyonu $S(\beta)$ ile gösterilirse E.K.K. çözümü bu kareler toplamı fonksiyonunu minimize etmektedir:

$$S(\beta) = (\mathbf{Y} - \mathbf{X\beta})'(\mathbf{Y} - \mathbf{X\beta}) = \|Y - X\beta\|^2 \qquad (2.12)$$

z herhangi bir vektör ise z'z bu vektörün kareli uzunluğudur. Dolayısıyla $S(\beta)$ kareler toplamı fonksiyonu **Y** ve $\mathrm{X}\beta$'yı birleştiren vektörün kareli uzunluğudur (Draper ve Smith, 1998).

Amaç fonksiyonunun çözümü için ilk adım, $\Omega$ alt uzayında bulunmayan bir noktaya, alt uzaydaki en yakın noktayı bulmaktır. Bu alt uzay, $\beta$'daki değişikliğe bağlı olarak değişen olası $\mathrm{X}\beta$ gerçek değerli vektörlerinden oluşan bir uzaydır. Ve tahmin uzayı olarak adlandırılır. Bu tahmin uzayında herhangi bir nokta bu uzayı geren X matrisinin sütunlarının lineer bir kombinasyonu olarak, $\mathrm{X}\beta$ biçiminde ifade edilebilir (Evren,1997). $\theta = \mathrm{X}\beta$ olmak üzere (burada $\theta \in \Omega$) $u'u = \|\mathrm{Y} - \theta\|^2$.'yı minimum yapan çözüm aranır. $\theta$, $\Omega$ uzayında değişebilen bir nokta olarak kabul edilirse, $\theta = \hat{\theta}$ için $(\mathrm{Y} - \hat{\theta}) \perp \Omega$ olduğu zaman bu uzaklık minimum olur.

$\hat{\theta}$, bir simetrik, idompotent matris olan projeksiyon matrisi yoluyla elde edilebilir. **H**, $\Omega$ üzerine dik izdüşümü ifade etmektedir.

$$\mathrm{Y} - \theta = (\mathrm{Y} - \hat{\theta}) + (\hat{\theta} - \theta) \qquad (2.13)$$

$H\theta = \theta, H' = H$ ve $H^2 = H$ olduğundan,

$$\begin{aligned}(Y - \hat{\theta})(\hat{\theta} - \theta) &= (Y - HY)'H(Y - \theta) \\ &= Y'(I_n - H)H(Y - \theta) \\ &= 0\end{aligned} \tag{2.14}$$

elde edilir. Dolayısıyla bu vektörlerin birbirlerine dik olduğu söylenilebilir. Bu vektörler birbirlerine dik olduğundan Pisagor Teoreminden yararlanılabilir.

$$\begin{aligned}\|Y - \theta\|^2 &= \|Y - \hat{\theta}\|^2 + \|\hat{\theta} - \theta\|^2 \\ &\geq \|Y - \hat{\theta}\|^2\end{aligned} \tag{2.15}$$

Dolayısıyla yukarıdaki sonuç elde edilir. Bu eşitlik ancak $\theta = \hat{\theta}$ olması durumunda gerçekleşir. $Y - \hat{\theta}$, $\Omega$'ya dik olduğundan,

$$X'(Y - \hat{\theta}) = 0 \text{ veya } X'\hat{\theta} = X'Y \tag{2.16}$$

sağlanmaktadır. Burada $\hat{\theta}$, Y'nin $\Omega$ üzerine *tek* dik projeksiyonu olarak belirlenir.

X matrisinin sütunları doğrusal olarak birbirinden bağımsız olduğundan $\hat{\theta} = Xb$ sağlayan tek bir $b$ vektörü bulunur. Yukarıdaki eşitlikte yerine konulursa normal denklemler elde edilir:

$$X'Xb = XY \tag{2.17}$$

X matrisinin rankı p'dir ve $X'X$ pozitif tanımlı dolayısıyla tekil olmayan bir matristir. Bu nedenle bu normal denklemler tek bir çözüm verecektir:

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

$$\hat{\theta} = Xb = X(X'X)^{-1}X'Y = HY \tag{2.18}$$

$Xb$ öngörü değerleri, $\hat{Y} = (\hat{Y}_1,..., \hat{Y}_n)$ olarak gösterilecektir. Diğer yandan,

$$\begin{aligned} Y - \hat{Y} &= Y - Xb \\ &= (I_n - H)Y \end{aligned} \tag{2.19}$$

kalıntıları vermektedir ve **e** ile gösterilecektir.

H, n boyutlu öklid uzayı $R_n$'in X'in sütunlarıyla gerilen $\Omega$ alt uzayı üzerine dik projeksiyonunu ifade eden doğrusal bir dönüşümdür. Benzer şekilde $I_n - H$, $R_n$'in, $\Omega$'nin dik tümleyeni olan $\Omega^{\perp}$ üzerine dik izdüşümünü ifade etmektedir. Böylece $Y = HY + (I_n - H)Y$, Y'nin biri $\Omega$ üzerinde, diğeri ona dik olan $\Omega^{\perp}$ üzerinde iki ortogonal alt uzaya bölünmesini ifade etmektedir. Y'nin bu iki alt uzaya projeksiyonundan bu vektör iki bileşene ayrıştırılır. Birinci bileşen regresyonun sistematik bileşenidir. Diğer bileşen bağımlı değişkendeki rasgele değişkenliği ilgilendiren tesadüfi bileşendir. $\Omega^{\perp}$ alt uzayı hata uzayı olarak adlandırılır (Seber ve Lee,2003).

Bu aşamadan sonra p boyutlu $\Omega$ uzayı , $\hat{Y}$ vektörünün, $\vec{1}$ ile giden ve bu vektöre dik olan bileşenlerine ayrılmasıyla iki alt uzaya ayrılır. $\vec{1}$ ile giden alt uzay bir boyutludur ve $\Omega_1$ ile ifade edilecektir. Bu uzaydaki değişkenlik veya değişme regresyon sabitini ilgilendirir. Bu uzaya dik olan uzay p-1 boyutludur ve $\Omega_1^{\perp}$ ile gösterilecektir.

X'ten bilgi alınamaması durumunda gözlemler regresyon modeli terimleriyle aşağıdaki gibi ifade edilebilir:

$$\mathrm{Y} = 1_n \beta + u \tag{2.20}$$

burada $1_n$, sütun değerleri 1'lerden oluşan n boyutlu bir matristir. Bu durumda parametre tahmini:

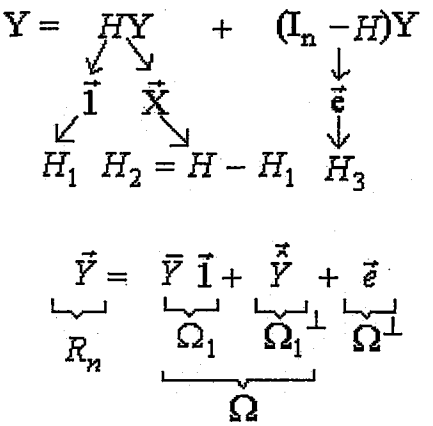
$$\mathrm{b} = (1_n' 1_n)^{-1} 1_n' \mathrm{Y} = \frac{1}{n} 1_n' \mathrm{Y} \tag{2.21}$$

olarak elde edilir. Dolayısıyla,

$$H_1 = 1_n (1_n' 1_n)^{-1} 1_n' = \frac{1}{n} \begin{bmatrix} 1 & 1 & \dots & 1 \\ 1 & 1 & \dots & 1 \\ \dots & \dots & \dots & \dots \\ 1 & 1 & \dots & 1 \end{bmatrix} = \frac{1}{n} \mathrm{J_n} \tag{2.22}$$

şeklinde bulunur. Y'nin $\Omega_1$ alt uzayı üzerine projeksiyonu $H_1$'dir. Bu kısım regresyon sabitiyle açıklanabilen kısmı göstermektedir. $\Omega_1^{\perp}$ alt uzayı üzerine projeksiyonu ise $H_2 = H - H_1$ şeklindedir. Bu kısım regresyona tabi olan kısmı ifade etmektedir.

$H_3 = \mathrm{I_n} - H$, Y'nin hata uzayı üzerindeki projeksiyonunu göstermektedir. Dolayısıyla Y gözlemler vektörü üç bileşene ayrıştırılmış olmaktadır. Sistematik kısım iki bileşenden oluşmaktadır. Üçüncü bileşen hata bileşenidir:




Şekil 2.3 Projeksiyon matrisinin bileşenleri

Bahsedilen ayrıştırma yukarıdaki şekillerle de ifade edilebilir.

Aşağıda en küçük karelerin geometrisinde yararlanılabilecek, projeksiyon matrisine ilişkin teoremler ve bu teoremlerin doğurduğu sonuçlar verilmiştir.

**Teorem:** X'in , nxp boyutlu p ranklı bir matris olduğu varsayılırsa $H = \mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'$ olur. Ardından aşağıdaki özellikler gerçekleşir:

i. $H$ ve $\mathrm{I_n} - H$ simetrik ve idempotent matrislerdir.

ii. $rank(\mathrm{I_n} - H) = iz(\mathrm{I_n} - H) = \mathrm{n} - \mathrm{p}$

iii. $H\mathrm{X} = \mathrm{X}$

**İspat:** (i) H'ın simetrik olduğu kesindir. Ve $(\mathrm{I_n} - H)' = \mathrm{I_n} - H' = \mathrm{I_n} - H$ 'dir. aynı zamanda,

$$\begin{aligned} H^2 &= \mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}' \\ &= \mathrm{X}\mathrm{I_p}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}' = H \end{aligned} \tag{2.23}$$

biçimindedir. Ve $(\mathrm{I_n} - H)^2 = \mathrm{I_n} - 2H + H^2 = \mathrm{I_n} - H$ 'dir. dolayısıyla $\mathrm{I_n} - H$ 'nin

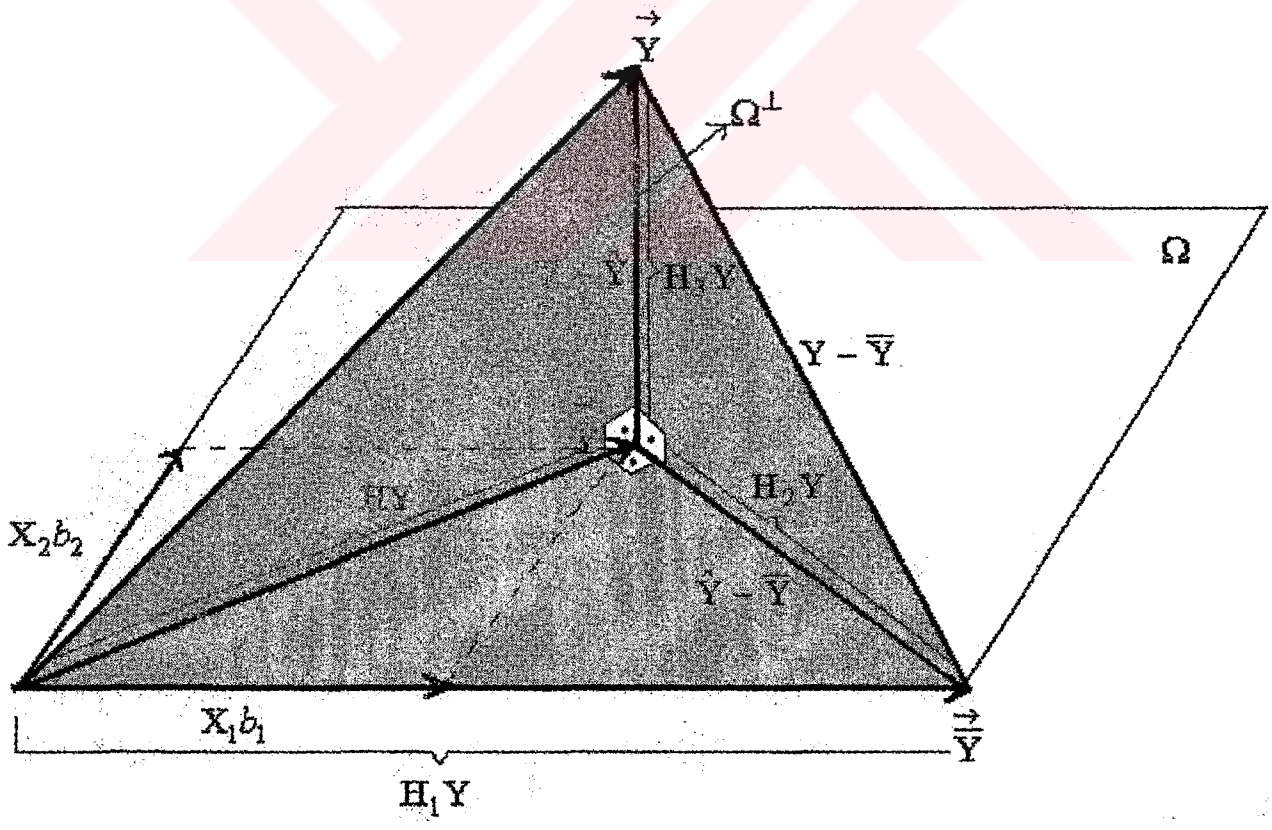
simetrik ve idempotent bir matris olduğu ispatlanmış olmaktadır.

(ii) $\mathrm{I}_n - H$ simetrik ve idempotent bir matris olduğundan,

$$\begin{aligned} rank(\mathrm{I}_n - H) &= iz(\mathrm{I}_n - H) \\ &= n - iz(H) \end{aligned} \tag{2.24}$$

elde edilir. Projeksiyon matrisinin özdeğerleri sıfır ve birlerden oluşur. Rankı $p$ ise bu matrisin $p$ adet 1 değerli özdeğeri ve $n$-$p$ adet 0 olan özdeğeri bulunmaktadır. Bir kare matrisin izi özdeğerlerinin toplamı olduğundan projeksiyon matrisinin izi $p$'dir (Seber ve Lee, 2003).

Yukarıda açıklanan sürecin ardından E.K.K. problemi geometrik olarak şekille aşağıdaki gibi ifade edilebilir:




Şekil 2.4 En Küçük Karelerin Geometrisi

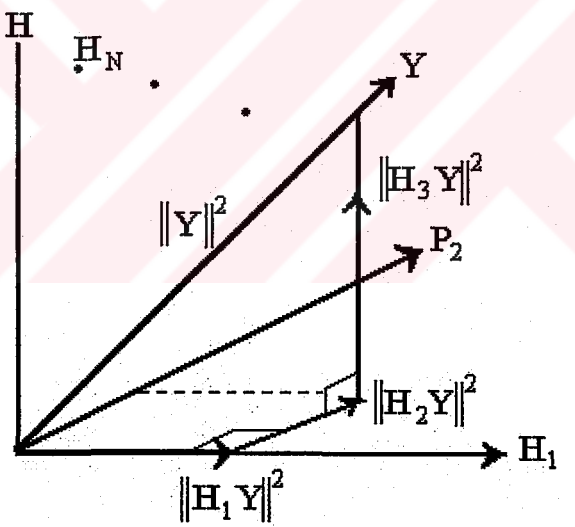
Şekilde görüldüğü gibi n boyutlu uzay birbirine dik üç ortogonal alt uzaya ayrıştırılır. Y

gözlem vektörünün bu alt uzaylara projeksiyonuyla Y üç bileşene ayrıştırılmış olur. Bunlardan ilki yukarıdaki şekilden de görüldüğü gibi sabit olan ortalama bileşenidir, $H_1\mathrm{Y} = \overline{\mathrm{Y}}$. İkinci bileşen $H_2\mathrm{Y} = \hat{\mathrm{Y}} - \overline{\mathrm{Y}}$'dir. Bu Y'nin X ile açıklanan kısmını ifade etmektedir. Üçüncü bileşen $H_3\mathrm{Y} = \mathrm{Y} - \hat{\mathrm{Y}}$, Y'nin regresyon ile açıklanamayan kısmını ifade etmektedir.

N boyutlu öklid uzayında pisagor teoremine göre Y vektörünün uzunluğunun karesi, Y'nin projeksiyonlarının uzunluklarının kareleri toplamına eşittir:

$$\|\mathrm{Y}\|^2 = \|H_1\mathrm{Y}\|^2 + \|H_2\mathrm{Y}\|^2 + \|H_3\mathrm{Y}\|^2 \qquad (2.25)$$

Bu ifade kareler toplamının ayrıştırılmasını ifade etmektedir. Şekil ile gösterilecek olunursa,




Şekil 2.5 Kareler toplamının ayrıştırılması*

Yukarıda belirtildiği gibi N boyutlu uzay ilk olarak birbirine dik iki ortogonal alt uzaya ayrıştırılır. Bu alt uzaylardan biri X sütunlarının oluşturduğu uzay, diğeri bu uzaya dik olan hata uzayıdır. Bu alt uzaylar sırasıyla $\Omega$ ve $\Omega^{\perp}$ ile belirtilmiştir. Bu ayrıştırmanın ardından

*Şekil: Saville,D.J., Wood,G.R., (Ağustos 1986), "A Method For Teaching Statistics Using N Dimensional Geometry",The American Statistician,40:3.'den alınmıştır.

$\Omega$ uzayı biri $\vec{1}$ ile diğeri ona dik merkezileştirilmiş $\vec{x}_1, \vec{x}_2, ...\vec{x}_p$ vektörleri tarafından gerilen iki alt uzaya bölünür (merkezileştirilmiş ifadesi X'in sütun vektörlerinin 1 alt uzayındaki vektörlere dik olacak şekilde ortogonalleştirilmiş halini belirtmek için kullanılmaktadır.). Bu alt uzaylar sırasıyla $\Omega_1, \Omega_1^{\perp}$ ile gösterilmiştir.

Her bir alt uzayın boyut sayısı, o alt uzaydaki bir vektörün ne kadar değişebileceğini belirlemektedir. İstatistiksel bir vektörü içeren uzayın boyutuna, o vektöre karşılık gelen değişkenin *serbestlik derecesi* denilir. Merkezileştirilmemiş toplam uzay ***n*** boyutludur. Regresyon sabitine karşılık gelen serbestlik derecesi ***1***'dir. Regresyona karşılık gelen serbestlik derecesi ***p***'dir. Ve hata uzayı ***n-p-1*** boyutludur. Dolayısıyla hata kısmına ait serbestlik derecesi ***n-p-1***'dir. Bu uzaylar birbirlerine dik olduğundan boyutlar toplanabilmektedir. Dolayısıyla serbestlik derecesi toplamsaldır ve toplanılabilir.(Jorgensen, 1992).

# 3. REGRESYONDA PROBLEMLERİN TEŞHİSİ

Tahminleri, bunlara ilişkin testleri ve diğer özet bilgileri elde etmek için geliştirilen metotlar regresyon analizinin sadece yarısını oluşturmaktadır. Bütün bu metotlar model doğruysa ve varsayımlar gerçekleşmişse uygulanabilir. Fakat herhangi bir uygulama probleminde varsayımlar şüpheli hale gelebilir. Analizin ikinci safhası varsayımları kontrol etmek ve gerekli görülen modeli kurmaktır. Baştaki modelleme safhası verilerden özet istatistikler gibi kombinasyonlar türetir. Sonraki safha bulunan istatistik değerlerinin sorgulanmasını gerektirtir. Regresyon analizinde varsayımlarla ilgili problemleri bulmak için *teşhis araçları* kullanılır.

Bir uygulamada doğrusal regresyon modeli gibi bir regresyon modeli göz önüne alındığında modelin önceden bu uygulama için uygun olduğu kesin olarak söylenemez. Modelin özelliklerinden herhangi biri ya da bir kaçı örneğin regresyon fonksiyonunun doğrusallığı veya hata terimlerinin normal dağılması gibi varsayımlar belirli bir veri seti için sağlanmayabilir. Dolayısıyla modelden hareketle çıkarsamalar yapılmadan önce modelin verilere uygunluğunun incelenmesi önemlidir.

Modelin verilere uygunluğunun incelenmesinde diyagramların faydalı olduğunu gösteren bir örnek Anscombe (1973) tarafından verilmiştir. Bu örnekte suni olarak oluşturulmuş dört farklı veri seti söz konusudur. Her bir veri seti için yalın doğrusal regresyon modeli kurulmuş ve her bir veri seti için aynı sonuçlar elde edilmiştir.

Çizelge 1.1 Diyagramların Kullanımı (Anscombe, 1973)

| Gözlem no | Y | X | Y | X | Y | X | Y | X |
|---|---|---|---|---|---|---|---|---|
| 1 | 8,04 | 10,0 | 9,14 | 10,0 | 7,46 | 10,0 | 6,58 | 8,0 |
| 2 | 6,95 | 8,0 | 8,14 | 8,0 | 6,77 | 8,0 | 5,76 | 8,0 |
| 3 | 7,58 | 13,0 | 8,74 | 13,0 | 12,74 | 13,0 | 7,71 | 8,0 |
| 4 | 8,81 | 9,0 | 8,77 | 9,0 | 7,11 | 9,0 | 8,84 | 8,0 |
| 5 | 8,33 | 11,0 | 9,26 | 11,0 | 7,81 | 11,0 | 8,47 | 8,0 |
| 6 | 9,96 | 14,0 | 8,10 | 14,0 | 8,84 | 14,0 | 7,04 | 8,0 |
| 7 | 7,24 | 6,0 | 6,13 | 6,0 | 6,08 | 6,0 | 5,25 | 8,0 |
| 8 | 4,26 | 4,0 | 3,10 | 4,0 | 5,39 | 4,0 | 12,50 | 19,0 |
| 9 | 10,84 | 12,0 | 9,13 | 12,0 | 8,15 | 12,0 | 5,56 | 8,0 |
| 10 | 4,82 | 7,0 | 7,26 | 7,0 | 6,42 | 7,0 | 7,91 | 8,0 |
| 11 | 5,68 | 5,0 | 4,74 | 5,0 | 5,73 | 5,0 | 6,89 | 8,0 |

Veri analizinin iyi tanımlanması için gerekli suni veri setleri tabloda verildiği gibidir. Her bir Y,X seti 11 gözlemden oluşmaktadır ve basit doğrusal regresyon modeli ,

$$\hat{Y}_i = b_0 + b_1 X_i + e_i$$

şeklinde kurulmuştur. Her bir veri seti bütün analizlerde özdeş neticeler vermektedir :

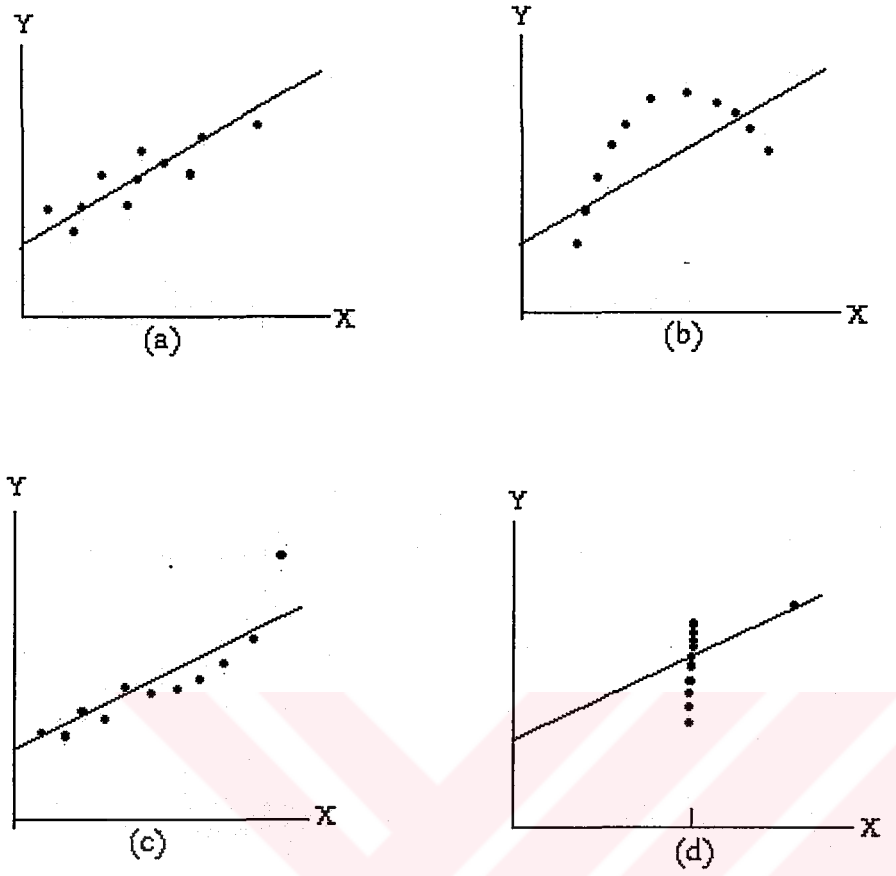
$b_0$=3.0

$b_1$=0.5

$\sigma^2$=13.75

$R^2$=0.667

Her bir veri seti için bütün istatistikler aynı olduğundan doğrusal regresyon modelinin her biri için eş zamanlı olarak uygun olduğu sonucuna varılabilir. Bununla beraber serpilme diyagramları bunun doğru olmadığını aşağıdaki şekillerle görüldüğü gibi ifade eder.




Şekil 3.1 Serpilme Diyagramları

Birinci veri seti doğrusal regresyon modeli uygunsa gözlemlenebilir. (b)'de farklı bir sonuç söz konusudur, basit doğrusal regresyon doğrusu verilere uygun değildir. Polinomiyal veya eğrisel bir model verilere daha uygun olabilir. (c) Verilerin çoğu için basit doğrusal regresyonu öneren bir durum olabilir. Fakat bir değer regresyon doğrusundan uzaktadır. Bu bir sapan değer problemidir. Büyük bir olasılıkla bu sapan değer silindiğinde kalan 10 veriyle regresyon tekrar düzenlendiğinde regresyon denklemi 11 gözlemle yapılandan farklı olacaktır. (d) deki durum diğerlerinden farklıdır. Öngörülen modelle ilgili tahmin ve çıkarsama yapabilmek için yeterli bilgi bulunmamaktadır. Eğim parametresi $\beta$ büyük ölçüde 8 verisi tarafından belirlenmiştir. Eğer bu 8 verisi silinirse tahmin yapılamaz (Weisberg, 1985).

## 3.1 Kalıntılar

Regresyon analizinde bağımlı değişken Y için teşhis diyagramları genellikle çok yararlı değildir. Çünkü bağımlı değişkenin gözlem değerleri bağımsız değişkenin bir fonksiyonudur. Bağımlı değişken için teşhis diyagramları yerine kalıntılar çalışılabilir.

Kalıntı $e_i$, gözlemlenen $Y_i$ değeriyle tahmin edilen $\hat{Y}_i$ değeri arasındaki farka eşittir:

$$e_i = Y_i - \hat{Y}_i$$

Kalıntı gözlenen hata olarak düşünülebilir. Kavram olarak $e_i, u_i$'ye benzer ve $u_i$'nin bir tahminidir.

Klasik normal doğrusal regresyon modeli için hata terimlerinin 0 ortalama ve $\sigma^2$ varyansla bağımsız normal rastlantısal değişken oldukları varsayılır. Model veriye uygunluk sağlıyorsa gözlenen kalıntılar $e_i$, $u_i$ için varsayılan özellikleri yansıtacaktır. Kalıntı analizinin temelinde bu düşünce yer almaktadır.

### 3.1.1 Standart Kalıntılar (Yarı Studentize Edilmiş Kalıntılar)

Kalıntı analizi için bazı durumlarda kalıntıları standardize etmek daha yararlı olabilir. Hata terimlerinin standart sapması $\sigma$, $\sqrt{MSE}$ ile tahmin edilir. Standartlaştırma aşağıdaki şekilde yapılır:

$$e_i^* = \frac{e_i - \bar{e}}{\sqrt{MSE}} = \frac{e_i}{\sqrt{MSE}} \qquad (3.1)$$

$\sqrt{MSE}$, $e_i$ kalıntının standart sapmasının bir tahmini olsaydı $e_i^*$ studentize edilmiş kalıntı olarak adlandırılabilirdi. Bununla birlikte $e_i$'nin standart sapmasının hesaplanması

komplekstir ve farklı $e_i$ kalıntıları için değişkenlik gösterir. $\sqrt{MSE}$ sadece $e_i$'nin standart sapmasının yaklaşık bir tahminidir. Dolayısıyla $e_i^*$ standart kalıntı olarak adlandırılır.

**Kalıntılarla Çalışılabilen Varsayımlardan Sapmalar**

Klasik normal doğrusal regresyon modelinin varsayımlarından altı önemli sapmayı inceleyebilmek için kalıntılar kullanılabilir:

1. Regresyon fonksiyonunun doğrusal olmaması
2. Hata terimlerinin varyansının sabit olmaması
3. Hata terimlerinin bağımsız olmaması
4. Sapan değer problemi
5. Hata terimlerinin normal dağılmaması
6. Bir veya birkaç önemli bağımsız değişkenin modele dahil edilmemesi

## 3.2 Kalıntılar İçin Teşhisler

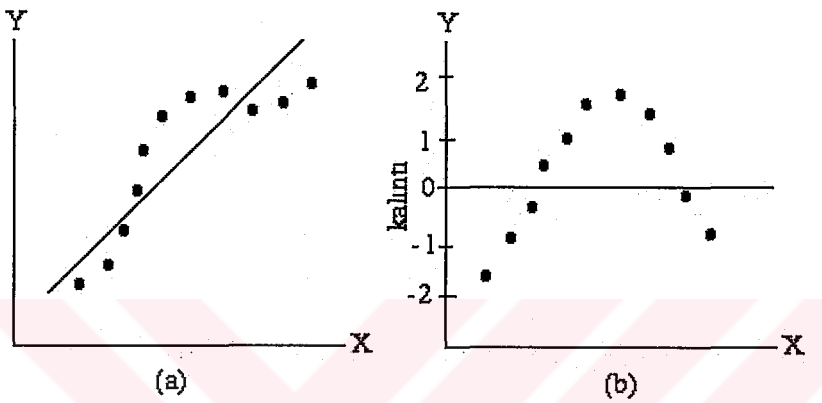
Yalın doğrusal regresyon modelinde bu altı varsayımdan sapmanın olup olmadığı hakkında bilgi edinebilmek için kalıntıların bazı teşhis diyagramları incelemelidir. Bu amaç için kullanılacak kalıntıların (veya semistudentized kalıntıların) grafikleri aşağıdaki şekillerde olabilir:

1. Kalıntıların bağımsız değişkene karşı diyagramı
2. Kalıntıların mutlak veya kareli değerlerinin bağımsız değişkene karşı diyagramı
3. Kalıntıların bağımlı değişkenin öngörü değerlerine karşı diyagramı
4. Zaman serilerinde kalıntıların zamana karşı diyagramı
5. Kalıntıların modelden dışlanan diğer bağımsız değişkenlere karşı diyagramları
6. Kalıntıların kutu diyagramı
7. Kalıntıların normal olasılık diyagramı

Bütün bu diyagramlar ileride görüleceği gibi regresyon modelinin veriye uygunluğunu destekleyici araçlardır.

## 3.3 Regresyon Fonksiyonunun Doğrusal Olmaması

Doğrusal bir regresyonun veri için uygun olup olmadığı kalıntıların bağımsız değişkene karşı diyagramından veya benzer şekilde tahmin değerlerine karşı diyagramından incelenebilir. Regresyon fonksiyonunun doğrusal olmama durumu serpilme diyagramından görülebilir fakat bu diyagram her zaman kalıntı diyagramı kadar etkin değildir.

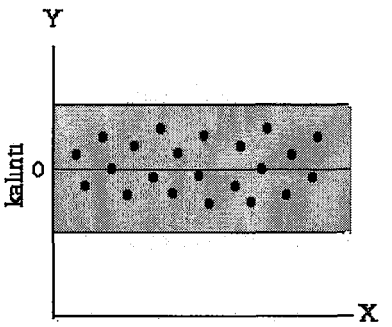



Şekil 3.2 Doğrusallık

(a)'da bir serpilme diyagramı (b)'de ise kalıntıların bağımsız değişkene karşı diyagramı bulunmaktadır. Her iki diyagramdan kuvvetli bir şekilde doğrusal regresyon fonksiyonunun uygun olmadığı görülmektedir. (b)'de doğrusal regresyona uyum eksikliği serpilme diyagramında olduğundan daha kuvvetli bir şekilde görülmektedir. Kalıntıların bu diyagramda sıfırdan sistematik bir biçimde ayrıldığına dikkat edilmelidir: X'in küçük değerlerinde kalıntılar negatif, orta büyüklükteki X değerleri için pozitif ve büyük X değerlerinde negatif değerler almaktadır.

Bu durumda (a) ve (b) diyagramları doğrusal olmayan regresyona işaret etmektedir. bununla beraber kalıntı diyagramı daha çok tercih edilmektedir. Bunun sebebi serpilme diyagramlarına göre bazı önemli avantajlarının bulunmasıdır. İlk olarak kalıntı diyagramı modelin uygunluğunu diğer yönlerden incelemek için kullanılabilir. İkinci olarak serpilme diyagramındaki ölçeklendirme bazı durumlarda $Y_i$ gözlem değerlerini, $\hat{Y}_i$ öngörü değerlerine yakınlaştırır. Bu aşamada serpilme diyagramından doğrusal regresyonun uygun olup olmadığını incelemek zorlaşacaktır. Diğer yandan bir kalıntı diyagramı bu koşullar altında

öngörülen regresyon doğrusu etrafındaki sapmalarda herhangi bir sistematik ilişki olup olmadığını belirgin olarak gösterir.



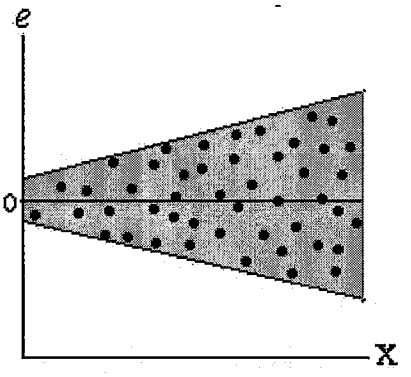

Şekil 3.3 Kalıntı diyagramı

Şekil, kalıntıların X'e karşı diyagramında doğrusal modelin uygun olduğu bir durumu ifade etmektedir. Kalıntılar sıfırın etrafında yatay bir band içinde, pozitif ve negatif herhangi bir sistematik eğilim içinde olmadan dağılmaktadır.

Not: Yalın regresyon modeli için kalıntıların $\hat{Y}$ değerlerine karşı diyagramı kalıntıların X'e karşı diyagramına eşdeğer bilgi sağlar dolayısıyla ayrıca X'e karşı diyagramının çizilmesine gerek kalmaz. İki diyagramın eşdeğer bilgi sağlayabilmesinin nedeni $\hat{Y}_i$ tahmin değerlerinin $X_i$ bağımsız değişkeninin doğrusal bir fonksiyonu olmasıdır. Diğer yandan eğrisel ve çoklu regresyon için kalıntıların öngörü değerlerine ve bağımsız değişkenlere karşı ayrı ayrı diyagramları her zaman yardımcı olur.

## 3.4 Hata Terimlerinin Varyansının Sabit Olmaması

Bağımsız değişkene karşı veya öngörü değerlerine karşı çizilen kalıntı diyagramları sadece doğrusal modelin uygun olup olmadığını göstermez aynı zamanda hata terimlerinin varyansının sabit olup olmadığı da bu diyagramlardan incelenebilir.



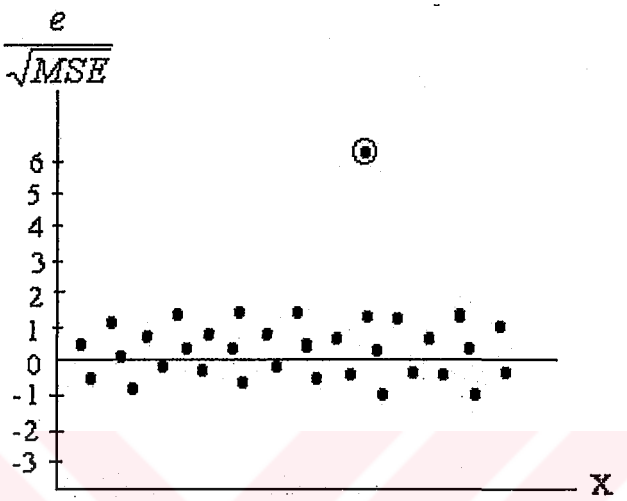

Şekil 3.4 Diyagramlarla değişen varyans

Şekil hata varyansının bağımsız değişken X'le birlikte arttığı kalıntı diyagramını göstermektedir. Pek çok işletme, sosyal bilimler ve biyoloji bilimlerindeki uygulamalarda hata terimlerinin varyansının sabitlikten sapması şekildeki gibi megafon biçiminde olmaktadır. Diğer yandan hata varyansı, bağımsız değişken X'in değerleri düşerken artabilir. Hata terimlerinin varyanslarının sabitliği incelenirken kalıntıların işaretleri bir anlam ifade etmediğinden kalıntıların mutlak değerlerinin veya karelerinin bağımsız değişken X'e karşı veya $\hat{Y}$ öngörü değerlerine karşı diyagramı sabit olmayan hata terimi varyansının teşhisi için kullanılabilir. Mutlak veya kareli kalıntıların diyagramlarının her ikisi de "0" çizgisi etrafındaki kalıntıların büyüklüklerini değiştirerek bütün bir bilgiyi sağlar dolayısıyla (işaretine bakılmaksızın) kalıntıların büyüklüğünün X veya $\hat{Y}$ seviyeleriyle değişip değişmediği kolaylıkla görülebilir.

## 3.5 Sapan Değerlerin Varlığı (Outliers)

Regresyon analizinde modelin bütün veriler için uygun olduğu varsayımı yapılır. Uygunsuzluk, veriler arasında aşırı farklılık gösteren ve "sapan değer" olarak adlandırılan kabul edilemeyen nitelikte verilerin bulunmasından ileri gelebilmektedir. Verilerin kalan kısmıyla oluşturulan modele uymayan veriler sapan değer olarak adlandırılır ve bu değerleri belirlemek kalıntı analizinin önemli bir parçasını oluşturmaktadır. Sapan değerleri belirleme ölçütü ise diğerlerinden çok farklı bir kalıntı sergileyen verilerdir. Buna göre de normal veri ile sapan veriler karşılaştırılmasında ölçüt kalıntılara dayanmaktadır.

Sapan kalıntı değerleri, X veya $\hat{Y}$'e karşı kalıntı diyagramlarından, kalıntıların kutu diyagramlarından, dal ve yaprak diyagramlarından ve nokta diyagramlarından belirlenebilir. Sapan değerleri belirlemek için semistudentized kalıntıların diyagramları özellikle yardımcı olacaktır. Çünkü sıfırdan standart sapmalar şeklindeki kalıntıların aykırı değer olup olmadığını belirlemek daha kolay olacaktır.



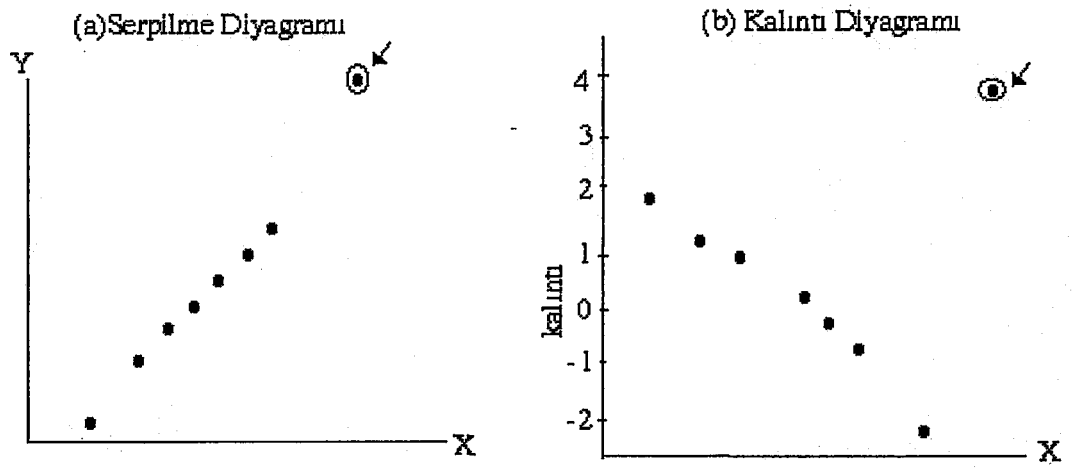

Şekil 3.5 Sapan değerler diyagramı

Şekilde yarı studentize edilmiş kalıntılar gösterilmektedir ve bu diyagram bir sapan değer içermektedir. Bu kalıntı gözlemin tahmin değerinden 6 standart sapma uzağa düştüğünü göstermektedir.

Sapan değerlerin belirlenmesi güçtür. Bir sapan değerle karşılaşıldığında ilk şüphelenilmesi gereken bu gözlemin bir hata veya diğer dışsal etkilerden dolayı sonuçlandığı dolayısıyla diğer gözlemlerden ayrı bir yerde bulunduğudur. En küçük kareler yönteminde sapma kareleri toplamı minimize edildiğinden öngörülen doğru sapan değere doğru kaydırılabilir. Bu durum, sapan değer gerçekten bir yanlıştan veya diğer dışsal sebeplerden kaynaklanıyorsa yanlış öngörüye sebep olur. Diğer yandan sapan değerler anlamlı bir bilgi veriyor da olabilir. Örneğin modelden dışlanan önemli bir bağımsız değişkenle etkileşiminden dolayı sapan değer oluşabilir. Herhangi bir sapan değeri dışlamak için sıklıkla başvurulan güvenli bir yol;sapan değerin bir kayıt hatasından, yanlış hesaplamadan veya benzer tür hatalardan, durumlardan olduğuna dair bir kanıt varsa sapan değerin dışlanmasıdır.

Veri seti için doğrusal regresyon modeli öngörüldüyse, gözlem sayısı az ise ve sapan değer mevcutsa, öngörülen regresyon sapan değer tarafından kaydırılır ve kalıntı diyagramı

doğrusal regresyonun uygun olmadığını gösterir. Kalıntı diyagramı ayrıca sapan değeri gösterir. Bu durum aşağıdaki şekilde ifade edilmiştir:

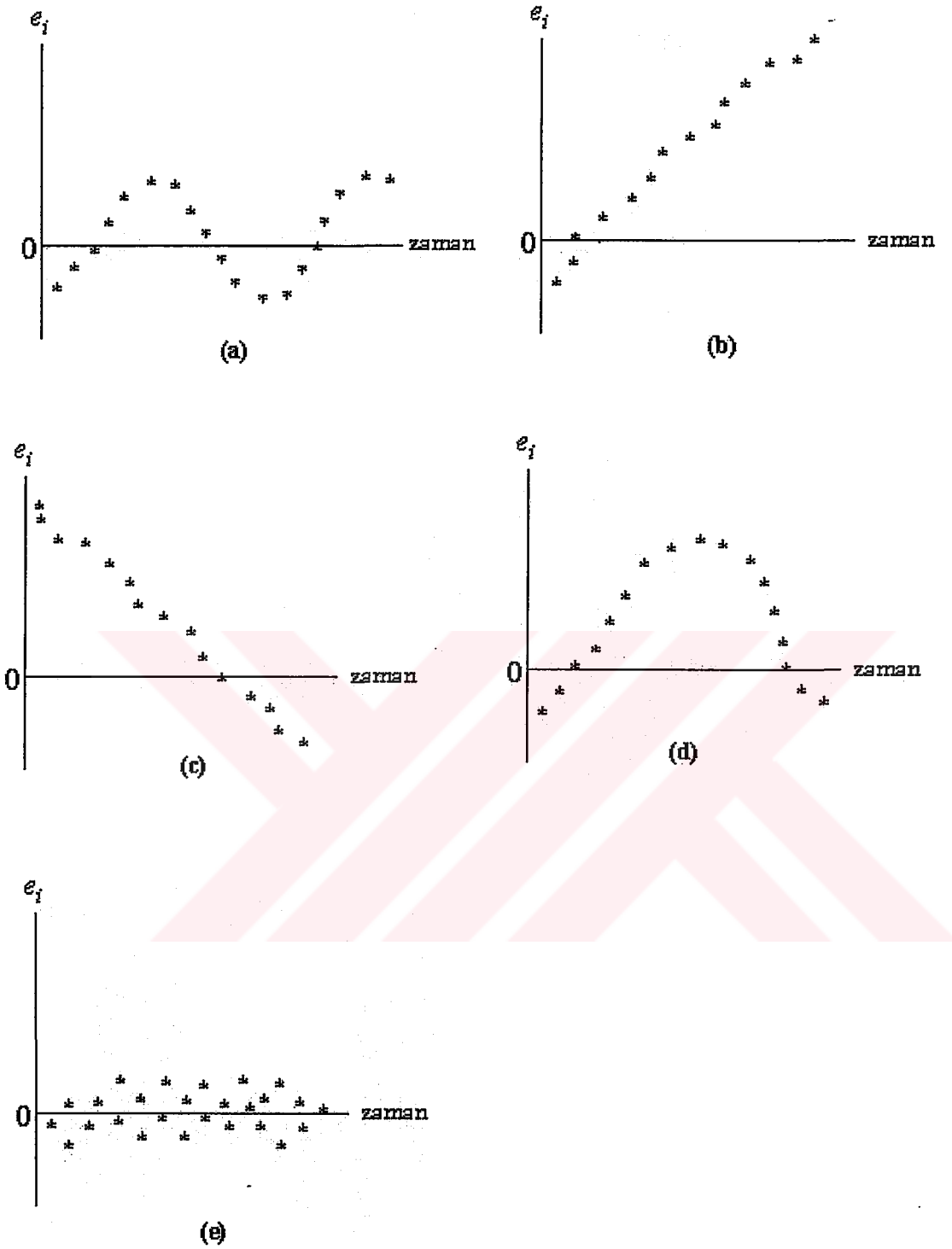



Şekil 3.6 Kalan Veri Doğrusal Regresyonu İzliyorsa Sapan Değerin Kalıntılar Üzerindeki Dışlayıcı Etkisi

Serpilme diyagramı gözlemlerin sapan değer dışında bir doğru etrafında istatistiksel bir ilişki sergilediklerini göstermektedir. Bu veriye doğrusal regresyon öngörüldüğünde sapan değer öngörülen regresyon doğrusunda kalan gözlemlerle tahmin edilen doğrudan sistematik biçimde sapmalara neden olan yer değiştirici bir etki yapar, bu doğrusal regresyona uyum eksikliğini gösterir. (Neter vd.,1996)

## 3.6 Hata Terimlerinin Bağımlılığı

Ardışık bağımlılığın ortaya çıkarılması için yaygın olarak kullanılan bir yöntem regresyon kalıntılarının zamana veya kesit verilerinde seriye karşı diyagramının çizilmesidir. Eğer kalıntılar ardışık dönemlerde düzenli bir yol izliyorlarsa fonksiyonda ardışık bağımlılık olduğu sonucuna varılır.




Şekil 3.7 Otokorelasyon biçimleri

Şekil a periyodik bir ilişki ifadesidir. b ve c, hata terimlerinde aşağıya ve yukarı doğru trendi ima etmektedir. b' de zamanla bağlantılı bazı etkenler dolayısıyla bağımlı değişkenin değeri artmaktadır. Şekil d hata terimlerinde hem doğrusal hem de kuadratik trendin olduğunu

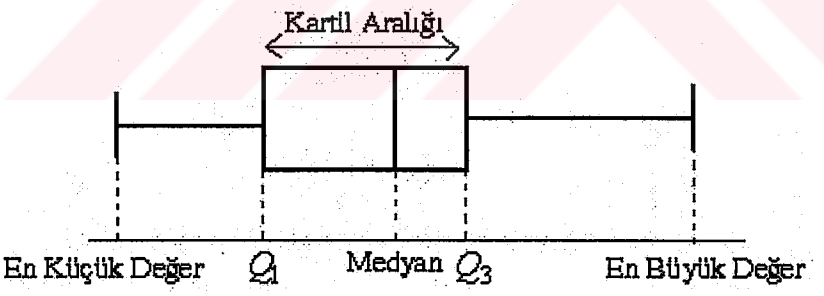
gösterir. Şekil e, klasik doğrusal regresyon modelinin otokorelasyonunun bulunmadığı varsayımını destekleyen sistematik bir kalıbın olmadığını belirtmektedir.

## 3.7 Hata Terimlerinin Normal Dağılmaması

Daha önce de belirtildiği gibi normallikten küçük sapmalar ciddi problemler doğurmaz. Önemli sapmalar dikkate alınmalıdır. Hata terimlerinin normalliği farklı grafiksel yöntemlerle incelenebilir:

### 3.7.1 Dağılım diyagramları

Kalıntıların kutu diyagramları kalıntıların simetrisi ve olası sapan değerler hakkında özet bilgi sağlar. Kutu diyagramları verinin yapısını ortaya çıkarmada kullanılan bir grafik düzenlemedir. Bu diyagramlarda verinin değişim aralığını, medyan ve diğer iki kartili birlikte görmek mümkündür. Bu düzenlemede birinci ve üçüncü kartillerin arası bir kutu olarak gösterilir. Kutunun iki ucundan çıkarılan yatay doğrular verinin en küçük ve en büyük gözlemlerine kadar uzatılır. Kutuyu dikey olarak kesen doğru ise medyanı göstermektedir.




Şekil 3.8 Kutu diyagramı

Diyagramda kutu; kartil aralığını yani merkezdeki %50'ye düşen gözlemleri simgelediğine göre, medyan hangi kartile yakınsa merkezdeki dağılmanın o yöne doğru bir asimetrisi olduğu söylenebilir. Örneğin yukarıdaki grafikte merkezin sol tarafa doğru (büyük değerlere) hafif bir çarpıklığı olduğu söylenebilir. Verinin bütün olarak simetrisi için kutuyu kesen iki yatay çizginin uzunluklarının birbirlerine ve kutunun uzunluğuna kıyaslanması gerekir. Yukarıdaki örnek grafikte sağ taraftaki yatay çizginin soldakine ve kutuya kıyasla uzunluğu dikkate alındığında, sağ tarafta verinin genel karakterinden uzak sapan değerlerin varlığından şüphelenilebilir.

Kutu diyagramlarının yanı sıra kalıntıların histogram, nokta, dal ve yaprak diyagramları normallikten aşırı sapmaları göstermede yardımcı olur. Bununla birlikte bu diyagramlardan herhangi birinin hata terimlerinin dağılımlarının şekli hakkında bilgi verebilmesi için regresyondaki veri sayısının yeterince büyük olması gereklidir.

### 3.7.2 Frekansların Karşılaştırılması

Veri sayısı yeterince büyükse kalıntıların gözlemlenen değerleriyle beklenen değerlerinin karşılaştırılması diğer bir seçenek olarak düşünülebilir. Örneğin kalıntıların, %68'inin $\pm\sqrt{MSE}$ aralığına veya %90'ının $\pm 1{,}645\sqrt{MSE}$ aralığına düşüp düşmediği belirlenebilir. Örnek büyüklüğü yeterince büyükse karşılaştırma için karşılık gelen t değerleri kullanılabilir.

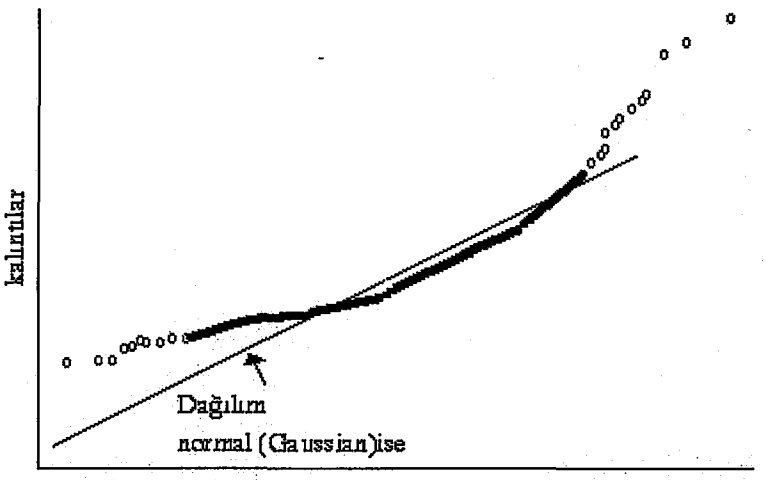
### 3.7.3 Normal Olasılık Diyagramı

Burada her bir kalıntının, normal dağılım varsayımı altında beklenen kalıntı değerine karşı diyagramı çizilir. Bu diyagramda doğruya yakın bir ilişki dağılımın normal dağılıma uyduğunu ifade eder, doğrusallıktan sapmalar hata terimlerinin dağılımının normal olmadığını gösterir.

Normallik varsayımı altında küçükten büyüğe doğru sıralanmış olan kalıntıların teorik değerlerini bulabilmek için, regresyon modelinde hata terimlerinin beklenen değerinin sıfır olduğu ve hata terimlerinin standart sapmasının $\sqrt{MSE}$ ile tahmin edildiği varsayımlarından yola çıkılır. İstatistiksel teori 0 ortalama ve $\sqrt{MSE}$ tahmini standart sapmayla normal dağılan bir rastlantısal değişken için n örnek içinden rassal bir örnekte k.ıncı en küçük gözlemin teorik değeri:

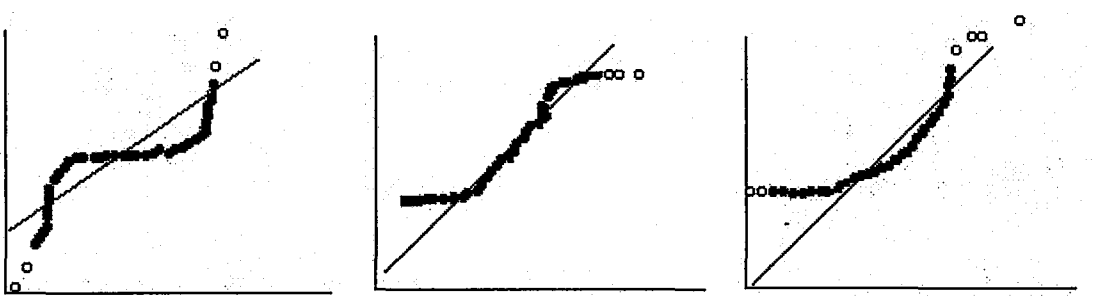
$$\sqrt{MSE}\left[z\left(\frac{k-0{,}375}{n+0{,}25}\right)\right] \tag{3.2}$$

$z(A)$ standart normal dağılımın $(A)100$ yüzdelik dilimini ifade etmektedir. k kalıntının rankını, yani küçükten büyüğe doğru sıralanan kalıntı değerleri arasında teorik değeri bulunmak istenen kalıntının sıra numarasını ifade etmektedir.(Neter vd.,1996)




Şekil 3.9 Normal olasılık diyagramı

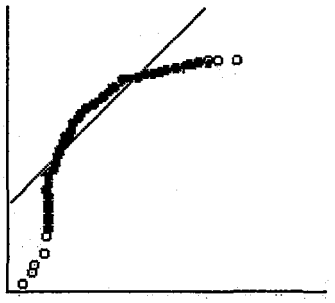
Örneğin şekil, diyagramın orta kısmında kalıntıların normal dağılım sergilediğini göstermektedir fakat kuyruklar eğri veya normal olmayan bir görüntü sergilemektedir. Sağ üst tarafta noktalar hızla yükselmektedir.



Kalın Kuyruklar, Yüksek ve Alçak Sapan Değerler

Hafif Kuyruklar ve Sapan Değer Yok

Pozitif (sağa) Çarpık ve Yüksek Sapan Değerler



Negatif Çarpık ve Düşük
Sapan Değerler

Şekil 3.10 Dağılımın şeklini yansıtan normal olasılık diyagramları

Kalın kuyruklu dağılımlar tepede ve dipte dikleşen dağılımlardır. Diğer bir deyişle dağılımın kuyruklarında normal dağılıma göre daha yüksek olasılıklar vardır. Kuyrukları hafif dağılımlar tepede ve dipte kuyrukları dik olmayan dağılımlardır. Çarpık dağılımlar bir kalın ve bir hafif kuyruğa sahiptir. Sapan değerler üst sağa veya alt sola doğru belirmektedir. Dikey olarak dağılımın kalan kısmından ayrılmıştır (Hamilton, 1992).

### 3.7.4 Normal Dağılımı Belirlemedeki Zorluklar

Pek çok yönden normallikten sapmaların analizi diğer varsayımlardan sapmalardan daha zordur. İlk olarak örnek büyüklüğü yeterince büyük değilse bir olasılık dağılımı çalışıldığında rastlantısal değişkenlik zararlı olabilir. Daha da kötüsü diğer varsayımlardan sapmalar kalıntıların dağılımını etkileyebilir. Örneğin verilere uygun olmayan bir regresyon fonksiyonu öngörüldüğünde veya hata terimlerinin varyansı sabit olmadığında kalıntılar normal dağılmıyor gözükebilir. Dolayısıyla hata terimlerinin normal dağılıp dağılmadığından önce ilk olarak diğer varsayımlardan sapmaların incelenmesi gerekmektedir.

## 3.8 Önemli Bağımsız Değişkenlerin Dışlanması

Kalıntıların modelden dışlandığı halde bağımlı değişken üzerinde önemli etkisi olabilecek bağımsız değişkenlere karşı diyagramları çizilebilir. Bu ilave analizin amacı modelden dışlanan değişkenlerin ilave önemli bilgi sağlayan ve tahmin gücü olan değişkenler olup olmadığını tespit etmek diğer bir deyişle modele eklenebilecek diğer önemli değişkenlerin var olup olmadığını araştırmaktır (Neter vd.,1996)

## 3.9 Kısmi Regresyon Diyagramları

Daha önceki bölümlerde bağımsız değişkene karşı bir kalıntı diyagramının söz konusu değişkenin eğrisel bir etkisinin olup olmadığını belirlemede kullanılabileceğine değinilmiştir. Aynı zamanda modelde olmayan bağımsız değişkenlere karşı diyagramlarının da bu bağımsız değişkenlerin modele ilave edilip edilmemesi konusunda yardımcı olacağı belirtilmiştir.

Bu kalıntı diyagramlarının bir eksikliği bu diyagramların diğer bağımsız değişkenler modelde iken bir bağımsız değişkenin marjinal etkilerini gösterememesidir. *Kısmi regresyon diyagramları* aynı zamanda *ilave değişken diyagramları (added variable plots)* ve *düzeltilmiş değişken diyagramları (adjusted variable plots)* olarak adlandırılırlar. Bu diyagramlar, diğer bağımsız değişkenler modelde iken $X_k$ bağımsız değişkeninin bağımlı değişkenle marjinal ilişkisinin yapısını belirleyen diğer kalıntı diyagramlarıdır. İlave olarak kısmi regresyon diyagramları modele ilave edilmesi gereken bir bağımsız değişkenin marjinal önemi hakkında grafiksel bir bilgi sağlamaktadır (Neter vd., 1996). Regresyon modelindeki bazı veya bütün bağımsız değişkenler için fonksiyonel ilişkiler yanlış belirlenmişse, bu diyagramlar bir bağımsız değişkenin marjinal etkisini uygun olarak göstermez. Kısmi regresyon diyagramları modelin verilere uygun hale gelebilmesi için gerekli dönüşümlere yol gösterir (Mansfield ve Conerly, 1987). Örneğin $X_2$ ve $X_3$ bağımlı değişkenle eğrisel bir ilişki içindeyse fakat regresyon modeli sadece doğrusal terimleri içeriyorsa $X_2$ ve $X_3$ için kısmi regresyon diyagramları uygun ilişkileri göstermeyecektir. Özellikle bağımsız değişkenler korelasyonluysa bu diyagramlar daha da yetersiz olacaktır. Bu bağımsız değişkenlerin kendi aralarında bir etkileşim kısmi regresyon diyagramları tarafından saptanamayacaktır. Sonuç olarak bağımsız değişkenler arasındaki yüksek çoklu doğrusallık bir bağımsız değişkenin marjinal etkisini yanlış bir fonksiyonel ilişki olarak gösterebilir.

Kısmi regresyon diyagramları diğer bağımsız değişkenler modelde iken $X_k$ değişkeninin marjinal rolünü göz önünde tuttuğundan kısmi olarak ifade edilmektedir. Kısmi regresyon diyagramında bağımlı değişken Y ve bağımsız değişken $X_k$'nın her ikisinin de diğer bağımsız değişkenlerle regresyonu kurulur ve bunların her birinden kalıntılar elde edilir. Bu kalıntılar her bir değişkenin diğer bağımsız değişkenlerle doğrusal ilişkide olmayan kısmını yansıtmaktadır. Bu kalıntıların diğer her birine karşı diyagramı (1) bağımsız değişken $X_k$

için marjinal regresyon ilişkisinin yapısı hakkında bilgi verir. (2) kalıntıların değişkenliğini azaltmada bu değişkenin marjnal önemini gösterir.

Bu ifadeleri netleştirmek için basitleştirmek amacıyla birinci dereceden iki bağımsız değişkenli ($X_1$ ve $X_2$) bir çoklu regresyon modeli düşünülecektir. $X_2$ modelde iken $X_1$'in bağımlı değişken üzerindeki etkisiyle ilgilenildiği varsayılsın. Bunun için Y'nin $X_2$'ye regresyonu oluşturulur ve buradan öngörü değerleri ve kalıntılar elde edilir:

$$\hat{Y}_i(X_2) = b_0 + b_1 X_{i2}$$

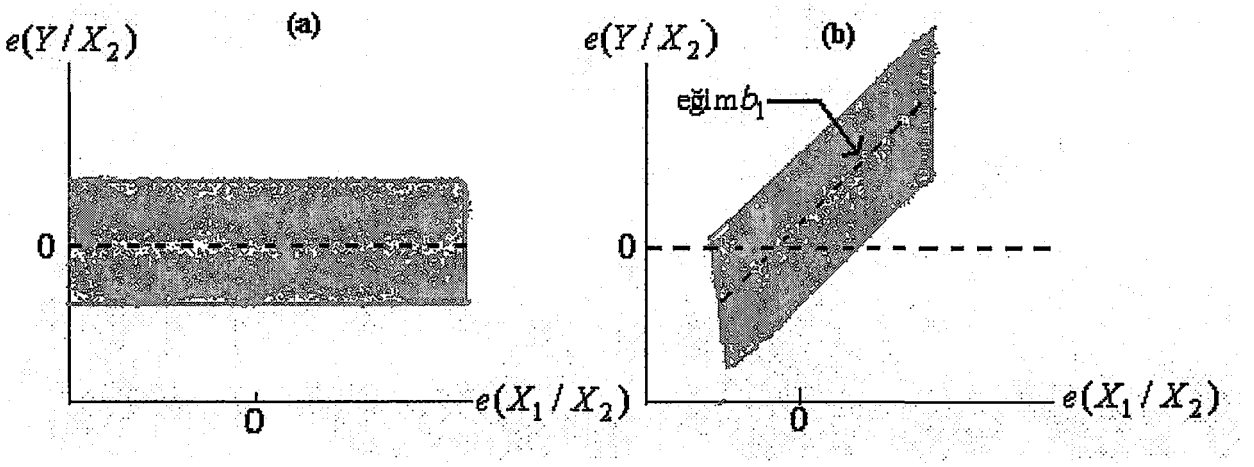
$$e_i(Y/X_2) = Y_i - \hat{Y}_i(X_2) \tag{3.3}$$

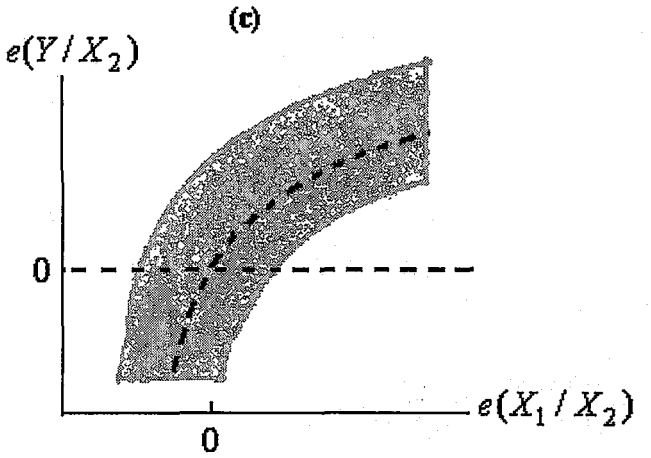
Ardından $X_1$'in $X_2$ üzerine regresyonu oluşturulur:

$$\hat{X}_{i1}(X_2) = b_0^* + b_2^* X_{i2}$$

$$e_i(X_1/X_2) = X_{i1} - \hat{X}_{i1}(X_2) \tag{3.4}$$

$X_1$ değişkeni için kısmi regresyon diyagramı Y kalıntılarının $e(Y/X_2)$, $X_1$ kalıntılarına $e(X_1/X_2)$ karşı diyagramından oluşur.

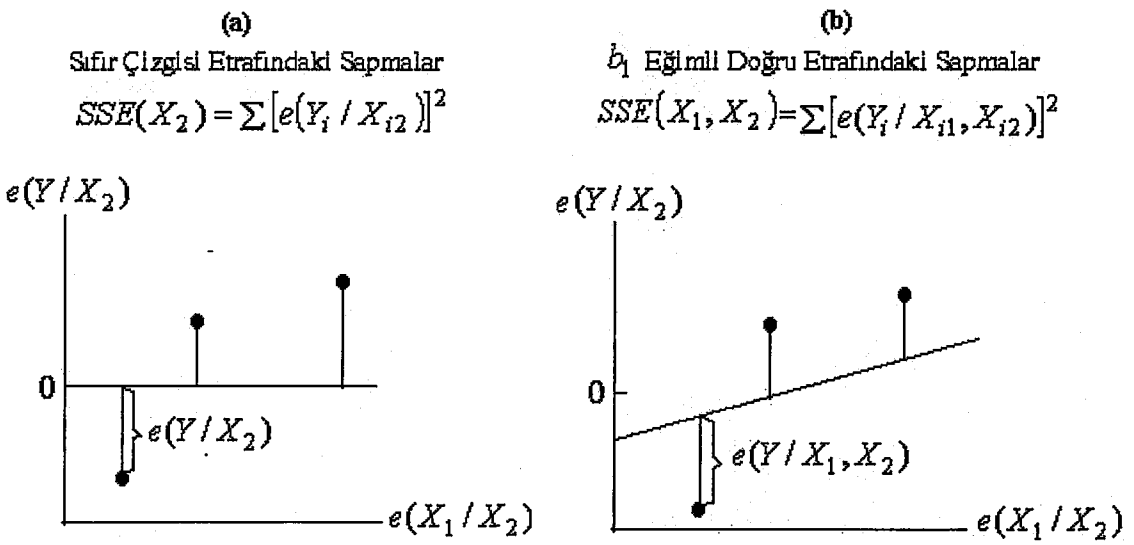





Şekil 3.11 Kısmi regresyon diyagramlarının prototipleri

Yukarıdaki şekil, $X_2$ modelde iken $X_1$ değişkeninin modele dahil edildiği göz önünde tutularak oluşturulan bazı kısmi regresyon diyagramları prototiplerini göstermektedir. (a)'da $X_1$'in Y'yi tahmin etmede ilave yararlı bir bilgi içermediğini ifade eden yatay bir band görülmektedir. Dolayısıyla burada $X_1$'in modele dahil edilmesinin faydalı olmadığına karar verilebilir. (b) eğimi olan bir doğru etrafında doğusal bir bandı göstermektedir. Bu diyagram $X_2$ modelde iken $X_1$'in doğrusal bir teriminin modele dahil edilmesinin faydalı olacağı görülmektedir. (c) eğrisel bir bant göstermektedir. $X_1$'in modele dahil edilmesinin faydalı olacağı ve bu değişkenin eğrisel bir etkisinin olabileceğini göstermektedir.

Kısmi regresyon diyagramları aynı zamanda ilişkinin gücü hakkında bilgi sağlamaktadır. Bu ilave bilgiyi nasıl sağladıklarını görebilmek için aşağıdaki grafikten faydalanılabilir.




Şekil 3.12 Kısmi regresyon diyagramlarıyla gözlenebilen ilişkinin gücü

Şekil (a) $X_2$ modelde iken $X_1$'in kısmi regresyon diyagramını göstermektedir. Bu regresyon üç veri üzerine kurulmuştur. $e(Y/X_2)=0$ yatay doğrusundan noktaların dikey sapmaları Y'nin $X_2$ ile regresyonundan elde edilen kalıntıları göstermektedir. Bu sapmaların kareleri alınıp toplanırsa hata kareleri toplamı $SSE(X_2)$ elde edilir. (b) de aynı noktalar vardır fakat burada, bu noktaların $b_1$ eğimine sahip en küçük kareler doğrusundan dikey sapmaları çizilmiştir. Buradaki sapmalar $X_1$ ve $X_2$' nin her ikisi modelde iken elde edilen kalıntıları $e(Y/X_1,X_2)$ göstermektedir . Dolayısıyla bu sapmaların kareleri toplamı $SSE(X_1,X_2)$ hata kareleri toplamını vermektedir.

(a) ve (b) diyagramlarında iki sapma kareleri toplamı arasındaki fark ekstra kareler toplamına $SSR(X_1/X_2)$ karşılık gelmektedir. Dolayısıyla iki serinin büyüklükleri arasındaki fark $X_2$ modelde iken $X_1$'in doğrusal ilişkisinin marjinal gücünü göstermektedir. Eğer (b)'deki doğrudan sapmalar yatay eksendeki sapmalardan daha az olsaydı $X_1$'i içeren regresyon modeli hata karelerinde önemli bir azalma sağlayabilecekti.

Kısmi regresyon diyagramları aynı zamanda diğer bağımsız değişkenler modelde iken

bağımsız değişkenle bağımlı değişkenin arasındaki ilişkinin tahmininde kuvvetli bir etkisi olan bir sapan değeri ortaya çıkarmada yararlı olabilir (Neter vd., 1996)

# 4 DEĞİŞEN VARYANS

## 4.1 Değişen Varyansın Sebepleri

Değişen varyans zaman serilerinden daha çok kesit verilerinde görülmektedir. Değişen varyans modelin yapısından veya modeli tahmin için kullanılan verilerden kaynaklanmaktadır. Değişen varyansın başlıca nedenleri şu şekilde özetlenebilir:

-Değişen varyans tanımlama hatalarından, ilgili değişkenin modelde yer almamasından kaynaklanabilir.

-Değişen varyansın bir diğer nedeni katsayılardan bir veya bir kaçının; zaman serisi verileri ile çalışılıyorsa zamana, kesit verileriyle çalışılıyorsa birimlere göre değişim göstermesidir. Regresyon modellerinde katsayılar sabit kabul edildiğinden bu durum değişen varyansa sebep olabilir.

-Hata- öğrenme modellerine göre, kişilerin herhangi bir olaydaki deneyimleri arttıkça, davranış hataları da küçülmektedir. Bu durum varyansı azalttığından değişen varyansa sebep olabilir (Güriş ve Çağlayan, 2000).

## 4.2 Değişen Varyansı Belirleyen Biçimsel Testler

### 4.2.1 Modifiye Levene Testi

Modifiye levene testi hata terimlerinin normalliğine dayalı bir test değildir. Doğrusu bu test varsayımlardan sapmalardan etkilenmeyen bir testtir. Tanımlandığı gibi Modifiye Levene testinin basit doğrusal regresyon modeline uygulanabilmesi için hata terimleri varyansının X ile birlikte artması (örneğin megafon şeklinde) veya azalması gerekmektedir. Örnek büyüklüğünün kalıntıların bağımlılığını gözardı edilmesine yetecek kadar büyük olması gerekmektedir.

Bu test kalıntıların değişkenliğine dayalı bir testtir. Büyük hata varyansı kalıntılarda büyük değişkenlik eğilimi demektir. Testi uygulayabilmek için veri seti X'in seviyelerine göre iki guruba ayrılır. Bir grup X'in düşük seviyelerine karşılık gelen gözlemlerden, diğer grup X'in yüksek değerlerine karşılık gelen gözlemlerden oluşturulur. Hata varyansı her iki grupta da azalıyorsa veya artıyorsa bir gruptaki kalıntılar diğer gruptaki kalıntılara göre daha çok değişkenlik göstermektedir. Benzer şekilde her bir grupta kalıntıların kendi grup ortalamalarından mutlak sapmaları bir grupta diğer gruba göre daha fazla olacaktır. Bu testin varsayımlardan etkilenmeyen bir test olarak gerçekleştirilebilmesi amacıyla gruplar için kalıntıların grup medyan değerinden mutlak sapmaları alınarak test gerçekleştirilir.

1.Örnek: $Y_1, Y_2, ...., Y_{n1}$

2. Örnek: $Z_1, Z_2, ...., Z_{n2}$

biçimindedir. Ve ana kütle ortalamasının tahminleri olan örnek ortalamaları

$$\overline{Y} = \frac{\sum_i Y_i}{n_1} \quad \text{ve} \quad \overline{Z} = \frac{\sum_i Z_i}{n_2} \tag{4.1}$$

olmak üzere,

$$t^* = \frac{\overline{Y} - \overline{Z}}{s\{\overline{Y} - \overline{Z}\}} \tag{4.2}$$

istatistiğine dayalı iki örnek t testinden meydana gelmektedir. Testin amacı birinci grup için ortalama mutlak sapmaların ikinci gruptan anlamlı bir şekilde farklılaşıp farklılaşmadığına karar vermektir.

Kalıntıların mutlak sapmalarının dağılımı genellikle normal dağılmasa dahi hata terimlerinin varyansı sabit olduğunda ve iki örnek büyüklüğü aşırı küçük değilse $t^*$ istatistiğinin yaklaşık

olarak t dağılımına uyduğu gösterilebilir.

Birinci grup için i. kalıntı değeri $e_{i1}$ ve 2. grup için i. kalıntı değeri $e_{i2}$ ile ifade edilecektir. $n_1$ ve $n_2$ örnek büyüklüklerini ifade etmektedir: $n = n_1 + n_2$ .

$\tilde{e}_1$ ve $\tilde{e}_2$ sırasıyla birinci ve ikinci gruptaki medyan değerlerini göstermektedir. Modifiye Levene testi kalıntıların medyan değerleri etrafındaki mutlak sapmalarını kullanır. Bu mutlak sapmalar:

$$d_{i1} = |e_{i1} - \tilde{e}_1| \qquad d_{i2} = |e_{i2} - \tilde{e}_2| \tag{4.3}$$

şeklinde hesaplanır. Bu notasyonla yukarıda belirtilen $t^*$ istatistiği aşağıdaki gibi hesaplanır:

$$t_L^* = \frac{\bar{d}_1 - \bar{d}_2}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \tag{4.4}$$

$\bar{d}_1$ ve $\bar{d}_2$, $d_{i1}$ ve $d_{i2}$ 'nin örnek ortalamalarını ifade etmektedir. Ortak varyans $s^2$ aşağıdaki gibi hesaplanır:

$$s^2 = \frac{\sum(d_{i1} - \bar{d}_1)^2 + \sum(d_{i2} - \bar{d}_2)^2}{n-2} \tag{4.5}$$

Modifiye Levene testi için test istatistiği $t_L^*$ ile gösterilir. Hata terimlerinin varyansı sabit ise ve örnek büyüklükleri $n_1$ ve $n_2$ aşırı küçük değillerse $t_L^*$ istatistiği n-2 serbestlik dereceli t dağılımına uyar. $t_L^*$ 'nin yüksek değerleri hata terimlerinin sabit varyanslı olmadıklarını

göstermektedir.

Eğer veri seti pek çok veriden oluşuyorsa hata varyansı için iki örnek *t* testi X'in seviyesine bağlı olarak ve iki ayrık gurubu kullanarak üç veya dört guruba bölünerek de yapılabilir.

Hata terimlerinin varyansı X' seviyesine bağlı olarak basitçe artmıyor ya da azalmıyorsa yani varyans daha kompleks bir görüntü sergiliyorsa veri setinin iki gruptan daha fazla gruba bölünmesi gerekebilir ve bu durumda Modifiye Levene testi F testi ile yapılır.

Hata terimleri varyansı için varsayımlardan etkilenmeyen bir test olması arzulanan bir durumdur çünkü normal olmama ve sabit olmayan varyans genellikle birlikte gerçekleşmektedir.

Yalın doğrusal regresyonda olduğu gibi çoklu regresyonda da Modifiye Levene testi hata varyansının sabitliğini araştırmak için kullanılabilir. Bunun için koşul hata varyansının X'e bağlı olarak artması veya azalmasıdır. Levene testini uygulayabilmek için veri seti bağımsız değişkenin değerinin düşük ve yüksek olduğu iki guruba bölünür. Daha sonra yalın doğrusal regresyonda uygulandığı gibi Levene testi süreci devam eder.

### 4.2.2 Breusch Pagan Testi

Hata terimleri varyansının sabitliği için ikinci bir test Breusch-Pagan testidir. Bu hata terimlerinin bağımsız ve normal dağıldığı bir büyük örnek testidir. Burada hata terimlerinin varyansının X'in seviyeleriyle aşağıdaki gibi bir ilişki içersinde olduğunu göstermek için $\sigma^2$, $\sigma_i^2$ olarak gösterilmektedir. i indisi varyansın değişken olduğunu ifade etmektedir:

$$\log_e \sigma_i^2 = \gamma_0 + \gamma_1 X_i \qquad (4.6)$$

Buradan varyansın X'in farklı seviyelerinde $\gamma_1$'in işaretine bağlı olarak arttığı veya azaldığı görülmektedir. Hata terimlerinin varyansının sabitliği $\gamma_1 = 0$'a karşılık gelmektedir.

$H_0 : \gamma_1 = 0$ hipotezine karşı $H_a : \gamma_1 \neq 0$ hipotezi için kalıntı karelerinin $e_i^2$, $X_i$'ye karşı regresyonu oluşturularak buradan regresyon kareler toplamı, $SSR^*$ elde edilir. Test istatistiği aşağıdaki gibidir:

$$\chi^2_{BP} = \frac{SSR^*}{2} \div \left(\frac{SSE}{n}\right)^2 \tag{4.7}$$

$SSE$, Y'nin X'e karşı regresyonundan elde edilen hata kareleri toplamıdır. $H_0 : \gamma_1 = 0$ hipotezi kabul edilirse ve örnek büyüklüğü mantıklı bir ölçüde büyükse $\chi^2_{BP}$ bir serbestlik dereceli ki-kare dağılımına uyar. $\chi^2_{BP}$'nin büyük değerleri hata terimleri varyansının sabit olmadığı savını, $H_a$, kabule götürür.

Bresusch- Pagan testi basit doğrusal regresyon için olduğu gibi çoklu regresyonda da uygulanır. Bu testin uygulanması için hata varyansının bağımsız değişkenlerden birinin değeriyle birlikte artması veya azalması gerekir. Daha sonra kalıntı karelerinin bağımsız değişkene karşı regresyonu oluşturulur ve regresyona tabi kareler toplamı olan $SSR^*$ elde edilir. Test süreci önceki gibidir fakat burada tam çoklu regresyon modeli için hata kareleri toplamı kullanılarak devam eder.

Hata varyansı bir bağımsız değişkenden daha fazla bağımsız değişkenin bir fonksiyonuysa kareli kalıntıların bu bağımsız değişkenlerle çoklu regresyonu oluşturulur ve regresyona tabi kareler toplamı $SSR^*$ elde edilir. Test istatistiği için yine tam model için hesaplanan $SSE$ kullanılır. Burada basit doğrusal regresyon için yapılan Breusch –Pagan testinden farkı ki kare dağılımı $q$ serbestlik derecesi içermektedir. $q$, kareli kalıntılarla oluşturulan çoklu regresyonda bağımsız değişken sayısıdır.

Not: Bu test istatistiği Cook and Weisberg tarafından geliştirilmiştir. Dolayısıyla bazen Cook-Weisberg testi olarak adlandırılabilir (Neter vd., 1996).

## 5.3 Değişen Varyansın Sonuçları

Değişen varyansın var olup olmadığının belirlenmesi problem çözümünün ilk aşamasıdır. Bunu gerçekleştirmenin çeşitli yolları vardır ve yukarıda bu yöntemlere değinilmiştir.
Sabit varyans varsayımının geçerli olmaması durumunda parametre tahmincileri sapmasızdır. Sapmasız olduklarından asimptotik olarak da sapmasızdırlar. Aynı zamanda doğrusal tahmin edicilerdir.

Sabit varyans varsayımı geçerli değil ise, parametre tahmin edicileri etkin tahmin ediciler değillerdir. Bu nedenle en iyi doğrusal sapmasız tahmin ediciler de değillerdir. Bu sonuç büyük örnekler için de geçerlidir. Yani parametre tahmin edicileri asimptotik olarak da etkin değillerdir.

Parametre tahminlerinin etkin olmaması nedeni ile, varyansları tahmin etmek için kullanılan formüller, varyansları olduğundan küçük ya da büyük tahmin edecektir.

$$\sigma^2\{b_1\} = \frac{\sum(X_i - \bar{X})^2 \sigma^2}{\left[\sum(X_i - \bar{X})^2\right]^2} \tag{4.8}$$

Bu formülden,

$$\sigma^2\{b_1\} = \frac{1}{\sum(X_i - \bar{X})^2}\sigma^2 \tag{4.9}$$

formülüne ulaşılmasının nedeni $E\{u_i^2\} = \sigma^2$ olmasıdır. Değişen varyans durumunda $\sigma^2$'nin yerini $\sigma_i^2$ alacaktır. Bu durumda $\sigma_i^2$ toplam işaretinin dışına çıkamayacağından sabit varyans ve değişen varyans hallerinde tahmin edilen varyans farklıdır (Genceli , 2002).

Parametre tahmincilerinin varyanslarının gerçek değerlerinden büyük veya küçük tahmin edilmeleri, yapılacak aralık tahminlerini, t ve F testlerini etkileyecektir (Neter vd., 1996).

# 5 OTOKORELASYON

Sıradan En Küçük Kareler Yönteminin diğer bir varsayımı, rassal değişken *u*'nun ardışık değerlerinin dönemsel olarak bağımsız olduklarıdır, yani *u*'nun herhangi bir dönemde aldığı değer, daha önceki herhangi bir dönemden bağımsızdır. Bu varsayıma göre $u_i$ ve $u_j$'nin ortak varyansı sıfıra eşittir.

$$\begin{aligned} Kov(u_i,u_j) &= E\left[u_i - E(u_i)\right]\left[u_j - E(u_j)\right] \\ &= E(u_i u_j) = 0 \qquad i \neq j \end{aligned} \tag{5.1}$$

Eğer belli bir dönemdeki *u* değeri, daha önceki kendi değeriyle(değerleriyle) bağlantılıysa, rassal değişkenin ardışık bağımlı olduğu söylenir.

Otokorelasyon haline daha çok zaman serilerinde rastlanılmaktadır. Bu paralelde *u*'nun alt imini değiştirip *t,t-1,t-2* kullanmak daha uygundur, böylelikle *u*'nun dönemsel değerleriyle, zaman içindeki bağımlılığıyla ilgilendiği belirtilmiş olmaktadır. Bu durumda *u*'nun *t* dönemdeki değeri $u_t$ ile *(t-1)* dönemindeki değeri $u_{t-1}$ ile gösterilebilir.

Ardışık Bağımlılık-otokorelasyon korelasyonun özel bir durumudur. Otokorelasyon, iki (ya da daha çok)değişkenin arasındaki ilişkiyle değil, fakat aynı değişkenin ardışık değerleri arasındaki ilişkiyle ilgilidir.

Basit bir durum olan *u*'nun ardışık iki değeri arasındaki doğrusal ilişki aşağıdaki gibi ifade edilebilir:

$$u_t = \rho u_{t-1} + v_t \tag{5.2}$$

Bu birinci dereceden ardışık bağlanımlı dizin olarak bilinir. $\rho$ ardışık bağımlılık (ortak

varyans) katsayısı ($u_t$ ile $u_{t-1}$ arasındaki basit korelasyon katsayısı) olarak adlandırılır. $\rho_{u_t u_{t-1}}$ basit korelasyon katsayısına yöneltilen bütün eleştirilere açıktır. Örneğin $\rho_{u_t u_{t-1}}$, $u_t$ ve $u_{t-1}$ arasındaki doğrusal olmayan ilişkilere uygun değildir. Ayrıca basit $\rho_{u_t u_{t-1}}$, eğer $u$'lar daha karmaşık yapılarla, daha yüksek ardışık dizimlerle bağlıysalar yine uygun değildir.

Ardışık bağımlılığın ortaya çıkarılması için yaygın olarak kullanılan bir yöntem regresyon kalıntılarının zamana karşı diyagramının çizilmesidir. Eğer kalıntılar ardışık dönemlerde düzenli bir yol izliyorlarsa fonksiyonda ardışık bağımlılık olduğu sonucuna varılır.
Birinci dereceden ardışık bağımlılık katsayısı aşağıdaki gibi hesaplanır:

$$r_{e_t e_{t-1}} = \frac{\sum e_t e_{t-1}}{\sqrt{\sum e_t^{\,2}}\sqrt{\sum e_{t-1}^{\,2}}} = \hat{\rho}_{u_t u_{t-1}} \tag{5.3}$$

$r_{e_t e_{t-1}}$, gerçek ardışık bağımlılık katsayısı $\rho_{u_t u_{t-1}}$'nun bir tahminidir.

Eğer $u$'nun belli bir dönemdeki değeri, yalnızca bir dönem önceki kendi değerine bağlıysa, $u$'ların birinci dereceden ardışık bağlanımlı bir dizim olduğunu (ya da birinci dereceden Markov süreci) olduğu söylenir. $u$'lar arasındaki ilişki şu biçimdedir:

$$u_t = f(u_{t-1}) \tag{5.4}$$

Eğer $u$ önceki iki döneminin değerine bağlıysa, yani $u_t = f(u_{t-1}, u_{t-2})$ ise, ardışık bağımlılığın kalıbına ikinci dereceden ardışık bağımlı dizim adı verilir. Bu süreç genelleştirilebilir. Çoğu uygulamalı araştırmada, ardışık bağımlılık varsa, bunun basit birinci dereceden $u_t = f(u_{t-1})$ kalıbı olduğu, özellikle de şu biçimde olduğu varsayılır:

$$u_t = a_1 u_{t-1} + v_t$$

burada $a_1$=ardışık bağımlılık katsayısı, $v_t$=bütün varsayımları sağlayan bir rassal değişkendir:

$$E(v) = 0 \qquad E(v^2) = \sigma_v^{\,2} \qquad E(v_i v_j) = 0 \tag{5.5}$$

$u_t$ ve $u_{t-1}$ arasında doğrusal bir ilişkinin olabilecek en basit ardışık bağımlılık kalıbıdır. Bu modele sıradan E.K.K.Y uygulanırsa,

$$\hat{a}_1 = \frac{\sum_{t=2}^{n} u_t u_{t-1}}{\sum_{t=2}^{n} u^2{}_{t-1}} \tag{5.6}$$

bulunur. Öte yandan ardışık bağımlılık katsayısı $\rho_{u_t u_{t-1}}$ aşağıdaki formülle bulunur:

$$\rho_{u_t u_{t-1}} = \frac{\sum u_t u_{t-1}}{\sqrt{\sum u_t^2}\sqrt{\sum u_{t-1}^2}} \tag{5.7}$$

Büyük örnekler için $\sum u_t^2 \approx \sum u_{t-1}^2$ veriyken,

$$\rho \approx \frac{\sum u_t u_{t-1}}{\sqrt{\left(\sum u_{t-1}^2\right)^2}} = \frac{\sum u_t u_{t-1}}{\sum u_{t-1}^2} \tag{5.8}$$

yazılabilir. Büyük örnekler için $\rho \approx \hat{a}_1$'dir. Dolayısıyla birinci dereceden ardışık bağımlılık kalıbı olarak

$$u_t = \rho . u_{t-1} + v_t$$

verilir. Burada $\rho$ birinci dereceden ardışık bağımlılık katsayısıdır. Eğer $\rho$=0 ise $u_t = v_t$'dir, yani (varsayım gereği $v_t$ ardışık bağımlı değilken) $u_t$ ardışık bağımlı değildir (Koutsoyiannis, 1989).

## 5.1 Otokorelasyonun Kaynakları

Hata terimi *u*'nun ardışık bağımlı değerleri çeşitli nedenlerle ortaya çıkabilir.

*1.Dışlanmış açıklayıcı değişkenler.* Eğer otokorelasyonlu bir değişken, bağımsız değişkenler kümesinden dışlanırsa, bunun etkisi *u* rassal değişkenine yansıyacaktır ve dolayısıyla *u*'lar otokorelasyonlu olurlar.

*2. Modelin matematiksel kalıbının yanlış kurulması.* Eğer ilişkinin gerçek kalıbından farklı bir matematiksel kalıp belirlenmişse hata terimleri ardışık bağımlılık gösterecektir. Örneğin Y ve X arasındaki gerçek ilişki dalgalanma gösterirken doğrusal bir fonksiyon tercih edilmişse, *u* değerleri dönemsel olarak bağımlı olurlar.

*3. Verilerle oynanması.* Görgül çözümlemelerde ham verilerle genellikle "oynanır". Düzeltme yöntemleri, verilere ara değer verme (interpolation) ya da dış değer verme (extrapolasyon) gibi oynamalar hata terimlerinin otokorelasyonlu olmasına neden olabilir.

*4. Gerçek rassal terim u'nun yanlış belirlenmesi.* Pek çok durumda gerçek *u*'nun ardışık değerleri birbirleriyle bağımlı olabilir. Hatta bütünüyle rassal olan etmenler (savaş,fırtına,kuraklık,grev,v.b.) birden çok dönemi etkileyebilirler(Koutsoyiannis, 1989). Örneğin bir grev, üretim süreci üzerinde gelecek birkaç dönemde de sürecek etkiler yaratabilir. Bu ardışık bağımlılığa kesit verilerinde de rastlanılabilir. Örneğin bölgesel yatay kesit verilerinde bir bölgedeki ekonomik etkinlikleri etkileyen tesadüfi bir şok, bölgeler arası yakın ekonomik bağlantılar sonucu, komşu bölgelerdeki ekonomik etkinlikleri de etkileyebilir. Veya hava koşullarından kaynaklanan şoklar da komşu bölgelerdeki hata terimlerinin ilişkili olmasına neden olabilir (Kennedy, 2000).

Bu durumda $E\{u_i, u_j\} = 0$ varsayımı yapıldığında *u* değerlerinin gerçek kalıbı yanlış belirlenmiş olur. Bu duruma *gerçek ardışık bağımlılık* denilebilir, çünkü kökenleri hata teriminin davranış kalıbında yatmaktadır (Koutsoyiannis, 1989).

Otokorelasyonlu hata terimlerinin ortalaması, varyansı ve kovaryansı aşağıdaki gibidir:

$$u_t = \sum_{r=0}^{\infty} \rho^r v_{t-r} \tag{5.9}$$

$$E(u_t) = 0 \tag{5.10}$$

$$\text{var}(u_t) = \frac{\sigma_v^2}{1-\rho^2} = \sigma_u{}^2 \tag{5.11}$$

$$\text{cov}(u_t u_{t-s}) = \rho^s \sigma_u{}^2 \qquad (s \neq t) \tag{5.12}$$

## 5.2 Otokorelasyon Testleri

### 5.2.1 Durbin Watson Testi

Otokorelasyon için şimdiye kadar geliştirilmiş testler arasında en çok rağbet göreni Durbin Watson testidir.

Bundan önce, zaman serilerinde hata paylarının birbirinden bağımsız olduğu savına karşıt olarak hata paylarının birinci mertebeden otoregressif (bağımlı değişkenin kendi gecikmeli değerlerinin açıklayıcı değişken olarak yer aldığı model) modele uyduğu ifade edilmiştir ($u_t = \rho.u_{t-1} + v_t$).

Böylece hata paylarının birbirinden bağımsız olduğu şeklindeki bir $H_0$ hipotezi aslında ρ'yu test etmektedir. Kalıntılardan hareketle $e_t = r.e_{t-1} + \varepsilon_t$ yazıp

$$r^* = \frac{\sum_2^T e_t e_{t-1}}{\sum_2 e_{t-1}{}^2} \tag{5.13}$$

test istatistiği elde edilebilir. Fakat r'nin test istatistiği olarak kullanılması sakıncalıdır, çünkü r tutarlı olmasına rağmen ρ'nun sistematik hatalı bir tahmin edicisidir.

Bununla beraber Anderson tablosu yardımıyla otokorelasyon testi yapılabilir.

Buna göre $r^*$'nin değeri $r_e$ ve $r_u$ arasında yer alırsa $H_0 : \rho = 0$ hipotezini reddetmeye gerek yoktur. Hata paylarının birbirinden bağımsız oldukları kabul edilir. $r^* > r_u$ veya - $r^* < r_e$ ise $H_0$ reddedilerek hata payları için birinci mertebeden otokorelasyon varlığı kabul

edilecektir.

Durbin Watson testi daha emin bir yoldur.

Durbin Watson r'nin sistematik hatalı olmasından hareketle $H_0 : \rho = 0$ için r'yi ikame eden d test istatistiğini geliştirmiş ve kullanmışlardır.

$$d = \frac{\sum_{2}^{T}(e_t - e_{t-1}{}^2)}{\sum_{1} e^2 t} \tag{5.14}$$

olarak tanımlanmaktadır.

İlk kalıntı $e_1$ 'dir. Dolayısıyla ilk fark $(e_2 - e_1)$ olacak, böylece T dönem için payda T-1 fark yer alacaktır. Bunun dışında dönemde veri eksikliği bulunması, ve/veya kabul edilemeyen verilerin çıkartılması halinde her eksik için paydaki fark sayısı 1 inecektir.

*d* örnek otokorelasyon katsayısı *r* ile yakından ilişkilidir. *d*'nin payı,

$$\sum(e_t e_{t-1}) = \sum e_t{}^2 + \sum e_{t-1}^2 - 2\sum e_t e_{t-1} \tag{5.15}$$

biçimindedir. $\sum e_{t-1}^2$ ile $\sum e_t^2$ arasında tek bir kalıntı farkı olduğundan her ikisi yaklaşık eşit kabul edilebilir ($\sum e_{t-1}^2 \cong \sum e_t{}^2$ ). Buna göre,

$$d = \frac{2\sum e_t{}^2 - 2\sum e_t e_{t-1}}{\sum e_t{}^2} = 2\frac{\sum e_t{}^2}{\sum e_t{}^2} - 2\frac{\sum e_t e_{t-1}}{\sum e_t{}^2} \tag{5.16}$$

elde edilir. Buradan da

$$d \cong 2 - 2r = 2(1 - r) \tag{5.17}$$

bulunur. Böylelikle d'nin tanım aralığının (0,4) olduğu ortaya çıkmaktadır. r, ρ'nun bir tahmin edicisi olması nedeniyle $\rho \to +1$ olduğu pozitif otokorelasyon halinde $d \cong 0, \rho \to -1$ olduğu negatif korelasyon için $d \cong 4$ olacaktır. Diğer taraftan ρ=0 için d≅2'dir. O halde d'nin 2 civarında değerler alması arzulanan bir sonuçtur.

d test istatistiğinin uygulanmasının bazı koşullara bağlı olduğuna da değinilmelidir.

1. Modelde regresyon sabiti yer almalıdır, çünkü d istatistiğinin paydasında bulunan $\sum e^2{}_t$ regresyon sabitine göredir.
2. Test sadece sabit veya rassal olmayan bağımsız değişkenler için geçerlidir.
3. Hata payları birinci mertebeden otoregressif modele uymalıdır.
4. Bağımlı değişkenin gecikmeli değerlerinin yer aldığı otoregressif modellerde bu testin geçerliliği yoktur. Böyle bir durumda d yukarı doğru sistematik hatalı sonuçlar vermektedir.

d'nin dağılımı örnekteki bağımsız değişkenlerin aldıkları değerlere, sayısına ve ayrıca birim mevcuduna bağlıdır. Bundan ötürü de d test istatistiğinin değeri bağımsız değişken sayısına ve örnek birim mevcuduna göre değişmektedir. Dolayısıyla da d'nin olasılık dağılımı ancak X'e bağlı olmayan bölgeler için hesaplanabilir. Böylece diğer testlerde olduğu gibi, genel geçerliliği olan kritik değerler yerine d'nin de aralarında yer aldığı ve X'e bağlı olmayan $d_l$ alt ve $d_u$ üst sınırları verilmektedir. d'nin dağılımı daima bu sınırlar arasında yer alacaktır. Bu da 0 ila 4 arasında bulunan d'nin tanım aralığının 5 bölgeye ayrılması demektir.

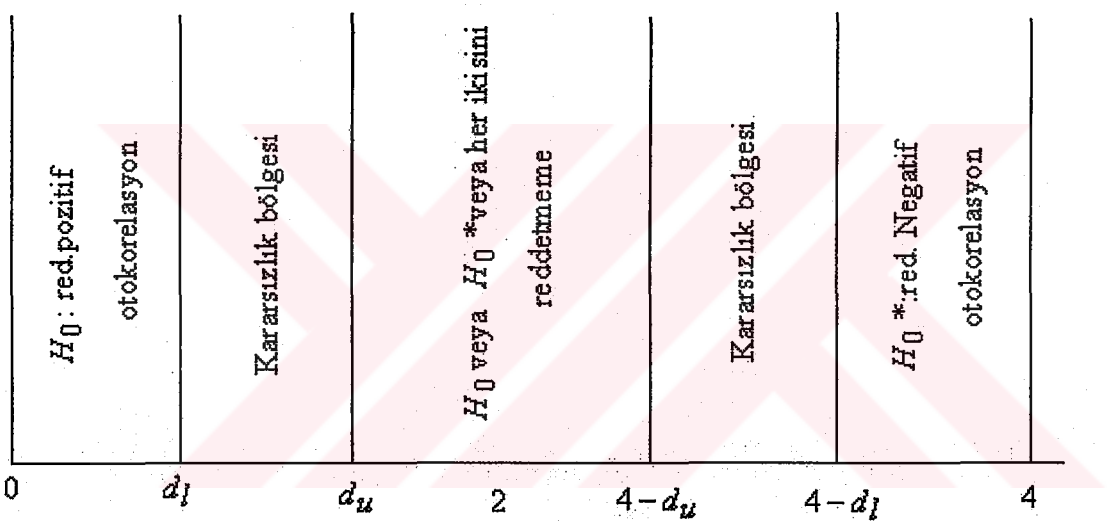
$d_l$ ve $d_u$ 'nun sınırları T ve bağımsız değişken sayısına bağlıdır. Dönem süresi T uzadıkça veya örnek birim mevcudu arttıkça $d_l$ ve $d_u$ birbirine yaklaşmaktadır. Bağımsız değişken sayısı k'nın artması ise aksi yönde etki yapmakta, $d_l$ ve $d_u$ birbirinden uzaklaşmaktadır.

$H_0$ veya $H_0^*$ aşağıdaki gibi kurulur:

$H_0$ :pozitif otokorelasyon yok

$H_0^*$ :negatif otokorelasyon yok

Karşıt hipotezlerin önemi yoktur. Negatif otokorelasyona ilişkin D.W testi çok nadir olarak yapılır. Ayrıca otokorelasyon için Çift taraflı testin sadece kuramsal geçerliliği bulunmaktadır. Uygulamada pek yeri yoktur (Genceli, 2002).




Şekil 5.1 Durbin Watson sınır değerleri

Bu testin gerçekçiliği açısından başka bir sorun da hangi zaman serilerine uygulanacağıdır. Uygulamada aylık, üçer aylık ve yıllık zaman serileri kullanılmaktadır. Örneğin üçer aylık verilerde birinci mertebeden otokorelasyon aranması halinde D.W testinin bu şekilde yapılması ile mevsim etkisi gözönüne alınmamaktadır. Oysa mevsim etkisinin sözkonusu olduğu üçer aylık verilerde birinci mertebeden otokorelasyon saptanmasa bile dördüncü mertebeden otokorelasyon sergileyebilirler. Aylık veriler halinde ise bu kez on ikinci mertebeden otokorelasyon bulunabilir. Kısaca Durbin-Watson testine göre otokorelasyon bulunmadığı şeklinde bir hüküm verilemez, sadece birinci mertebeden otokorelasyon için sonuç alınabilir.

Wallis de bu olgudan hareketle üçer aylık verilerde birinci mertebeden otokorelasyon yerine

dördüncü mertebeden otokorelasyon aranması gerektiğini, çünkü otokorelasyonun aslında büyük bir olasılıkla mevsim etkisini gösteren değişkenin modele dahil edilmemesinden kaynaklandığını savunmuştur. Buna göre Wallis'te,

$$u_t = \theta_4 u_{t-4} + e_t \text{ olup } H_0 : \theta_4 = 0 \text{'dır.} \tag{5.18}$$

Test istatistiği ise

$$d = \frac{\sum_{t=5}^{T} (e_t - e_{t-4})^2}{\sum_{t=1}^{T} e^2{}_t} \tag{5.19}$$

biçimindedir.

$d_l$ ve $d_u$ için Wallis tablolarına bakılmaktadır. Testin mantığı ise D.W ile aynıdır.

Durbin Watson testinin en büyük sorunu kararsızlık bölgesidir. Kararsızlık bölgelerini ortadan kaldırmak için çok daha güçlü testler geliştirilmiştir. Ancak bu testler çok karmaşık yapıya sahip olduklarından ve hesaplama güçlükleri dolayısıyla D-W kadar rağbet görmemektedir. Bununla beraber bu testler arasında Theil- Nagar (T-N) Testi özel bir yere sahiptir (Genceli, 2002).

### 5.2.2 Durbin h Testi

Bağımlı değişkenin gecikmeli halinin (veya hallerinin) açıklayıcı değişken (veya değişkenler) şeklinde bulunduğu otoregressif modellerde otokorelasyon testi için Durbin h testi uygulanabilir. Büyük örneklemler için tam geçerli olan bu test küçük örneklemler için de uygulanabilmektedir. h istatistiğinin formülü şöyledir:

$$h = \hat{\rho}\sqrt{\frac{n}{1-n[V(\gamma)]}} \tag{5.20}$$

burada $\hat{\rho}=\rho$'nun tahmin değeri, n=örneklem hacmi, $V(\gamma)=Y_{t-1}$ değişkeninin tahmin edilen katsayısının varyansıdır. Durbin Watson d testinde belirtildiği gibi $d \cong 2(1-\hat{\rho})$ ifadesinden $\hat{\rho} \cong 1-1/2d$ yazılabilir ve h formülünde yerine konulduğunda aşağıdaki eşitlik elde edilir.

$$h \cong (1-\frac{1}{2}d)\sqrt{\frac{n}{1-n[V(\gamma)]}} \tag{5.21}$$

uygulamada bu ifade kullanılır, $n[V(\gamma)]$<1 olmalıdır. Örneklem hacmi büyük olduğu zaman eğer $\rho$=0 ise h- istatistiği normal dağılım izleyecek ve bu dağılımın ortalaması sıfır ve varyansı bir olacaktır. O zaman örneklemden elde edilecek h- istatistiğinin istatistiksel olarak anlamlı olup olmadığı standart normal dağılım tablosundan yararlanarak belirlenebilir.

Durbin h testinde otoregresif modelde kaç tane X değişkeni ve kaç tane gecikmeli Y değişkeninin olduğu önemli değildir. h istatistiğini bulmak için yanlızca $Y_{t-1}$ değişkeninin tahmin edilen $\gamma$ katsayısının varyansına gerek vardır.

Hipotezler aşağıdaki gibidir:

$$\begin{aligned} H_0 &: \rho = 0 \\ H_a &: \rho \neq 0 \end{aligned} \tag{5.22}$$

$1/2(1-\alpha)$ yoluyla Z değeri standart normal dağılım tablosundan bulunur. $|h|$<Z ise $H_0$ kabul edilir. h>Z ise pozitif otokorelasyon olduğu h<-Z ise negatif otokorelasyon olduğu sonucuna ulaşılır (Akın, 1997).

## 5.3 Otokorelasyonun Doğurduğu Sonuçlar

1. Kalıntılar otokorelasyonlu olsalar dahi E.K.K. tahmin edicileri sapmasızdırlar.
2. Hata terimleri otokorelasyonlu olduklarında parametrelerin E.K.K.Y. ile hesaplanan varyansları diğer tahmin yöntemleriyle bulunanlardan daha yüksek olabilirler.
3. Hata terimleri otokorelasyonlu ise hata terimi u'nun varyansı ciddi biçimde olduğundan küçük tahmin edilecektir. Bunun sonucunda parametreler için güven sınırları olduğundan daha geniş çıkacağı gibi, bağımsız değişkenler arasında negatif otokorelasyon bulunmaması koşuluyla $s\{b\}$ de $\sigma\{b\}$'yi olduğundan daha düşük tahmin ettiğinden $\beta$=0 şeklindeki sıfır hipotezinin olması gerektiğinden daha fazla reddine neden olacaktır. Çünkü hipotez testinde $H_0$'ı kabul bölgesi daralmaktadır. Koşullar altında t ve F testleri de geçerli değildir (Genceli, 2002).
4. Eğer hata terimleri otokorelasyonlu ise E.K.K. yöntemine dayalı kestirimler etkin olmaz. Bu kestirimlerin varyansı diğer yöntemlerle bulunanlara göre daha yüksektir (Koutsoyiannis, 1989).

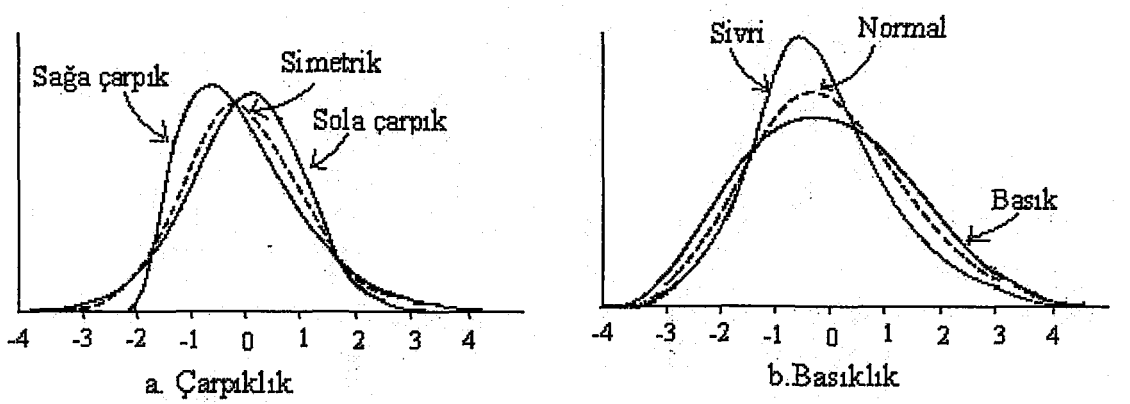
# 6. NORMAL DAĞILIMA UYGUNLUK

E.K.K. yönteminin uygulanması için hata terimlerinin normal dağıldığı varsayımı gerekli değildir. Böyle bir varsayım yapılmasa dahi parametre tahmin edicileri en iyi, doğrusal,sapmasız tahmin edicilerdir. Diğer taraftan böyle bir varsayımın yapılmaması, küçük örneklerde parametreler için hipotez testleri ve güven sınırı uygulamalarını engellemektedir. Bununla beraber hata paylarının normal dağılımdan aşırı farklılık göstermemeleri halinde gerek hipotez testlerinin gerekse güven sınırlarının geçerliliği kabul edilebilir (Genceli, 2002).

## 6.1 Basıklık ve Çarpıklık Ölçüleri

Çarpıklık ve basıklık ölçüleri olarak yalın, ölçü birimlerinin etkisi giderilmiş $\alpha_3$ ve $\alpha_4$ momentleri kullanılmaktadır.

$$\alpha_3 = \frac{E(X-\mu)^3}{\left[E(X-\mu)^3\right]^{1/2}} = \frac{\mu_3}{\sigma^3} \text{ ve } \alpha_4 = \frac{E(X-\mu)^4}{\left[E(X-\mu)^2\right]^{1/2}} = \frac{\mu_4}{(\sigma^2)^2} \tag{6.1}$$




Şekil 6.1 Normal dağılım basıklık ve çarpıklık ölçüleri

Normal dağılımdaki simetriden dolayı çarpıklık ölçüsü $\alpha_3$=0 ve $\alpha_4$=3'tür.

## 6.2 Normallik İçin Korelasyon Testi

Bir normal olasılık diyagramında noktaların doğruya yakın bir ilişki sergileyip sergilemediklerini incelemek için hata terimlerine ilişkin biçimsel bir test, kalıntılarla, normallik varsayımı altında beklenen değerleri arasındaki korelasyon katsayısını hesaplamaktır. Yüksek bir korelasyon katsayısı normal dağıldığını gösterir. Sıralanmış olan kalıntılarla normallik altında bu kalıntıların beklenen değerleri arasındaki korelasyon katsayılarının dağılımı için, farklı örnek büyüklüklerinde kritik değerlerinin (yüzdelik dilimleri) bulundukları tablo Looney ve Gulledge tarafından hazırlanmıştır.

Bu korelasyon katsayısı yaklaşık olarak, sıralanmış kalıntılarla normallik altında beklenen değerleri arasındaki korelasyon katsayısına dayalı olan Shapiro-Wilk testinden daha basit bir testtir (Neter vd., 1996).

## 6.3 Jarque Bera Testi

J.B. sınaması bir kavuşmazlık ya da büyük örnek sınamasıdır. Bu da SEK kalıntılarına dayanır. Bu sınama önce S.E.K.(Sıradan en küçük kareler) kalıntılarının çarpıklık ve basıklık ölçülerini hesaplar, şu sınama istatistiğini kullanır,

$$J.B = n\left[\frac{{\alpha_3}^2}{6} + \frac{(\alpha_4 - 3)^2}{24}\right] \tag{6.2}$$

Burada $\alpha_3$ : çarpıklık (skewness), $\alpha_4$ :basıklık (kurtosis) ölçülerini ifade etmektedir.

Normal dağılımda $\alpha_3$=0, $\alpha_4$=3 olur. hipotez testleri aşağıdaki gibidir:

$H_0 : \alpha_3$=0 ve/veya $\alpha_4$=3

$H_a : \alpha_3 \neq 0$ veya $\alpha_4 \neq 3$ veya hem $\alpha_3 \neq 0$ hem de $\alpha_4 \neq 3$ (6.3)

JB, kalıntıların normal dağıldığı sıfır önsavı altında JB istatistiğinin büyük örneklemde s.d'si 2 olan bir ki-kare dağılımına uyduğunu göstermiştir (Gujarati, 1999). J.B. testine ilişkin olarak, bu testin düzeltici işlevi bulunmamaktadır. Buna göre $H_0$'ın reddedilmesi durumunda nasıl bir yol veya yöntem izleneceği belli değildir (Genceli, 2002).

## 6.4 Ki-Kare Uygunluk Testi

Bağımlı değişken ile bağımsız değişken veya değişkenler için regresyon modeli tahmin edilerek hata terimleri belirlenir ve hata terimlerinin standart hataları hesaplanır. Hata terimleri ortalamadan sapmalara göre büyüklük sırasına göre gruplandırılır. Gruplarda yer alan hata terimi sayısı gözlenen frekansı ($G_i$) verir. Bu yöntemin uygulanması için gözlenen frekansların yanı sıra beklenen frekansların ($B_i$) belirlenmesi gerekir. Beklenen frekanslar normal dağılımdan yararlanılarak belirlenir.

Gözlenen frekanslar belirlenmiş standart sapmalar için sıfırın altında ve üstünde bulunan hata terimlerinin frekans dağılımını verir. Buna sıklık dağılımı da denmektedir. Sıklık dağılımlarını belirlemek için sıklık çiziminden yararlanarak ortalamadan uzaklıkları standart sapmalarla ölçülmüş gözlenen sıklıkların bulunması kolaylaşacaktır.

Beklenen frekanslar ise varsayılan bir olasılık dağılımına göre sıklık dağılımı gösterir. Gözlenen ve beklenen frekanslar arasındaki farkın karesinin beklenen sıklıklara bölünerek toplamlarının alınması ile test istatistiği oluşturulur.

$$\chi^2 = \sum_{i=1}^{k} \frac{(G_i - B_i)^2}{B_i} \quad (6.4)$$

burada,

$G_i$ :i aralığındaki gözlenen frekans

$B_i$ :i aralığında varsayılan dağılımın beklenen sıklığı ifade etmektedir.

Bu şekilde elde edilen ki-kare değeri, rastlantı değişkeninin dağılım fonksiyonundan bağımsızdır.

Gözlenen frekans ile beklenen frekans arasındaki fark küçükse hata terimi normal dağılmaktadır ve $H_a$ alternatif hipotez kabul edilecektir. Tersi durum söz konusu olduğunda ise $H_0$ kabul edilir.

Örnek hacmi büyük olduğunda hesaplanan test istatistiği N-1 serbestlik dereceli ki-kare dağılım tablosu ile karşılaştırılır. Burada N kümelerin sayısıdır (Güriş ve Çağlayan, 2000).

Ki-kare testi, en güçlü olan oranlı ölçekten, en zayıf olan sınıflayıcı ölçekli verilere kadar uygulanabilmektedir. Bu nedenle bazen parametrik test olarak da incelenebilmektedir.

Bu testin en zayıf yönü, hiçbir grupta beklenen ya da gözlenen sıklığın beşten az olmaması koşulunu ileri sürmesidir. Bu kısıtlama da özellikle az sayıda denekten oluşan örneklemler için kullanışsız olmaktadır (Kıroğlu, 1996).

## 6.5 Kolmogorov-Simirnov Tek Örneklem Testi

$\chi^2$ uygunluk testlerinin alternatifi olan Kolmogorov-Smirnov Testi Kolmogorov tarafından 1933'te önerilmiştir. Smirnov tarafından ise kritik değerler tablosu geliştirilmiş ve test iki örnek için uygulanabilir hale getirilmiştir.

Yukarıda belirtildiği gibi $\chi^2$ testinin uygulanabilmesi için her iki özelliğin şıklarına ait frekansların $\geq 5$ olması, bu nedenle de büyük örneklerin gözlemlenmesi gerekmektedir. Kolmogorov- Smirnov Testi böyle bir şarta dayanmadığı için kolayca uygulanabilmektedir. Test en azından sıralı ölçekteki verilere uygulanabilmektedir. Bu testte örnekten elde edilen

birikimli frekans dağılımının $H_0$'da ileri sürülen ana kütle teorik dağılımıyla karşılaştırılmasına dayanmaktadır.

Bu iki dağılımın en farklı oldukları noktadaki farklılığın önemli olup olmadığı araştırılır. Bu en büyük fark önemli değilse dağılımların aynı olduğu söylenebilir.

Örneklem dağılım fonksiyonu $S_n(x)$, teorik dağılım fonksiyonu ise $F_x(x)^*$ olduğunda test için tanımlamalar şu şekilde yapılabilir:

Veriler, $F_x(x)$ dağılımlı ana kütleden alınan n büyüklükteki örneklem, $x_1,...,x_n$ gözlemlerinden oluşur.

Hipotezler:

$H_0$'da ileri sürülen dağılım $F_x(x)^*$ olduğunda kurulabilecek hipotezler aşağıdaki gibi özetlenebilir:

Çift yönlü
$$\begin{aligned} H_0 &: F_x(x) = F_x(x)^* \\ H_a &: F_x(x) \neq F_x(x)^* \end{aligned} \qquad (6.5)$$

Tek Yönlü
$$\begin{aligned} H_0 &: F_x(x) \geq F_x(x)^* \\ H_a &: F_x(x) < F_x(x)^* \end{aligned} \quad \text{veya} \quad \begin{aligned} H_0 &: F_x(x) \leq F_x(x)^* \\ H_a &: F_x(x) > F_x(x)^* \end{aligned} \qquad (6.6)$$

Kritik değer: $S_n(x)$, örneklemde birikimli olasılık fonksiyonu olarak tanımlandığında kullanılan test ölçütü yukarıdaki hipotezler için şu şekildedir:

Çift yönlü hipotez için test ölçütü $D = \max |S_n(x) - F_x(x)^*|$ olup, grafiksel gösterimde bu uzaklık, $S_n(x)$ ile $F_x(x)^*$ arasındaki en büyük dikey uzaklıktır.

Tek yönlü hipotezlerden sırasıyla ilki için test ölçütü $D^+ = \max\left(F_x(x)^* - S_n(x)\right)$ olup grafiksel gösterimde $F_x(x)^*$ fonksiyonu üsttedir ve ölçüt değeri bu iki fonksiyon arasındaki

en büyük dikey uzunluktur. İkinci tek yönlü hipotez için test ölçütü $D^- = \max(F_x(x)^* - S_n(x))$ olup, grafiksel gösterimde $S_n(x)$ fonksiyonu üsttedir ve bu değer iki fonksiyon arasındaki en büyük dikey uzunluğa karşılık gelmektedir.

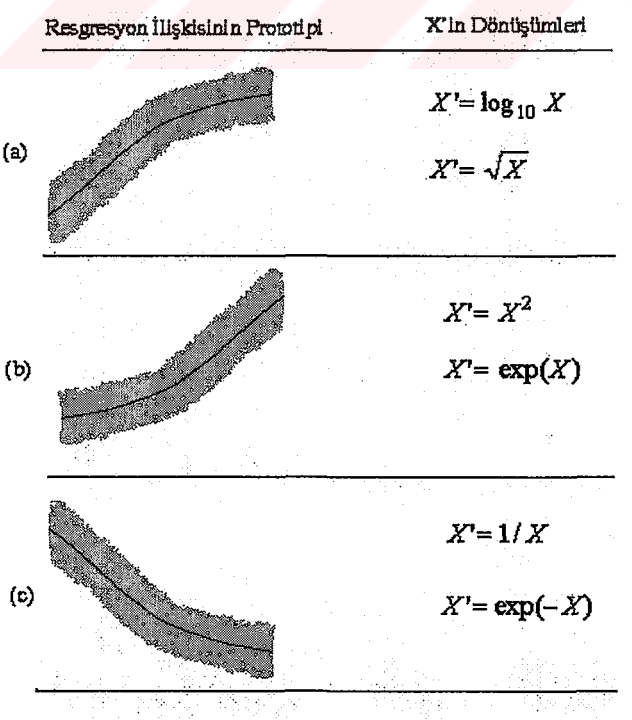
Karar Kuralı: Belirlenen $\alpha$ anlamlılık seviyesinde, hesaplanan $D, D^+$ ve $D^-$ değerleri Kolmogorov- Smirnov tablo değerleriyle karşılaştırılır. $(1-\alpha)$ için bakılan tablo değerleri test değerinden küçükse hipotez erddedilir. Eğer hipotezler doğru ise $S_n(x)$ ile $F_x(x)^*$ çok yakın olmalıdır (Kıroğlu, 1996).

# 7. DÖNÜŞÜMLER

## 7.1 Doğrusal Olmayan İlişki İçin Dönüşümler

Hata terimlerinin dağılımı normal dağılıma yakınsa ve varyansı yaklaşık olarak sabit ise doğrusal olmayan bir regresyon ilişkisini doğrusal hale getirmek için dönüşümler uygulanabilir. Bu durumda X üzerine dönüşüm uygulamak istenebilir. Y üzerinde dönüşüm yapılmak istenmeyişinin nedeni örneğin $Y'=\sqrt{Y}$ gibi bir dönüşüm hata terimlerinin dağılımının şeklini değiştirebilir ve hata terimlerinin varyansının değişiklik göstermesine neden olabilir.

Aşağıdaki şekil bazı doğrusal olmayan, sabit varyanslı regresyon modellerinin prototiplerini ve doğrusallaştırmak için bu örneklere uygun olabilecek X üzerine basit dönüşümleri göstermektedir. Birkaç alternatif dönüşüm denenebilir. Dönüşümler uygulandıktan sonra serpilme diyagramları ve kalıntı diyagramları oluşturulabilir ve hangi dönüşümün daha etkili olduğuna karar verilebilir.




Şekil 7.1 Doğrusal olmayan regresyon ilişkisinin varlığında dönüşümler

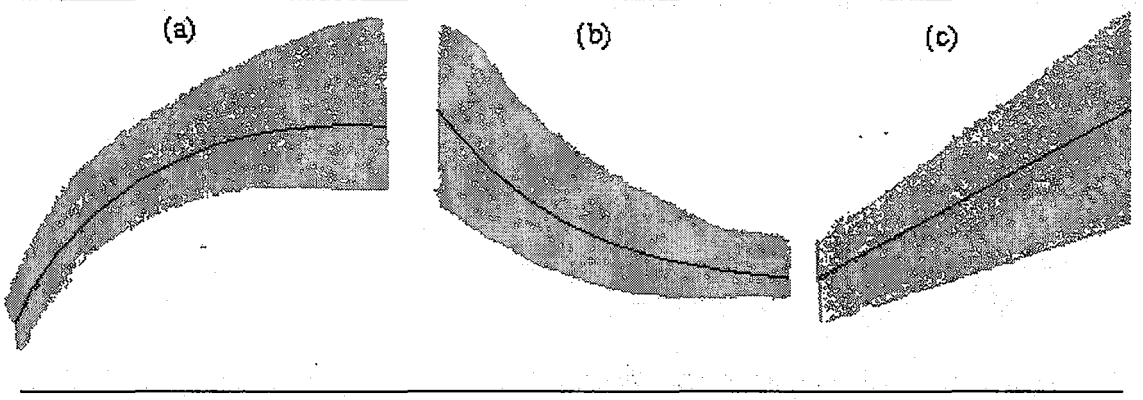
Bazı durumlarda dönüşüme bir sabit eklenmesi yardımcı olabilir. Örneğin bazı X verileri 0'a yakın olabilir dönüşüm yapılmak istenebilir. Dolayısıyla $X'=1/(X+k)$ dönüşümü yapılarak orjinin yeri değiştirilebilir. k seçilen uygun bir sabit değerdir.

## 7.2 Normal Dağılmayan Ve Varyansı Sabit Olmayan Hata Terimlerinin Varlığında Dönüşümler

Hata terimlerinin normal dağılmaması ve varyansının sabit olmaması sıklıkla birlikte gerçekleşir. Basit doğrusal regresyon modelinden bu sapmalara çözüm bulabilmek için Y üzerinde dönüşüm uygulanması gerekmektedir. Dolayısıyla Y'nin dağılımının şekli ve yayılması değişecektir. Y üzerinde bu tür dönüşümler eğrisel regresyon ilişkisinin doğrusallaştırılmasına yardımcı olabilir. Bazı durumlarda eşzamanlı X üzerindeki dönüşümler regresyon ilişkisini korumak veya sürdürmek için yapılabilir.

Doğrusal regresyon modelinden normal dağılmama ve sabit varyanslı olmama gibi sapmalar genellikle bağımlı değişkenin ortalama değerinin arttığı gibi hata terimlerinin çarpıklığının ve değişkenliğinin arttığı bir biçim olmaktadır. Aşağıdaki şekil $E\{Y\}$ ile birlikte artan çarpıklık ve artan hata varyanslarının bulunduğu regresyon ilişkilerinin prototipleri gösterilmektedir. Bu grafikte aynı zamanda her bir durum için uygulanabilir dönüşümler belirtilmiştir. Birkaç alternatif dönüşüm, eş zamanlı olarak X üzerine de uygulanabilir. Dönüşümler uygulandıktan sonra serpilme diyagramları ve kalıntı diyagramları oluşturularak en etkili dönüşümün hangisi olduğuna karar verilebilir.

Prototip Regresyon İlişkileri




Y Üzerindeki Dönüşümler

$$Y' = \sqrt{Y}$$

$$Y' = \log_{10} Y$$

$$Y' = 1/Y$$

Şekil 7.2 Normal dağılmayan ve varyansı sabit olmayan hata terimlerinin varlığı halinde dönüşümler

X üzerinde eşzamanlı dönüşümler yardımcı veya gerekli olabilir.

Bazı durumlarda Y'nin dönüşümüne bir sabit ilave edilmesi gerekebilir. Örneğin Y negatif olduğunda böyle bir dönüşüm yapılabilir. Örneğin Y'nin orjininin yerini değiştirmek için logaritmik bir dönüşüm bütün Y değerlerinin pozitif olması gerektiğinden $Y' = \log_{10}(Y + k)$ şeklinde yapılır. k uygun olarak seçilmiş bir sabit değerdir.

Regresyon ilişkisi doğrusal olmasına rağmen hata terimlerinin varyansı sabit değilse Y üzerinde yapılacak bir dönüşüm başarılı olmayabilir. Bu gibi bir dönüşüm varyansı stabilize ederken doğrusal ilişkinin eğrisel olmasına neden olabilir. Dolayısıyla aynı zamanda X üzerinde de bir dönüşüm gerekebilir.

## 7.3 Box-Cox Dönüşümleri

Dağılımların çarpıklığını, hata terimlerinin dağılımını, sabit olmayan varyanslarını ve doğrusal olmayan regresyon fonksiyonunu düzeltmek için teşhis diyagramlarına bakılarak Y

üzerinde hangi dönüşümlerin yapılacağına karar vermek zor olabilir. Box-Cox prosedürü otomatik olarak Y üzerinde kuvvet (üs) dönüşümleri ailesi tanımlar. Kuvvet dönüşümleri ailesi aşağıdaki biçimdedir:

$$Y' = Y^{\lambda} \tag{7.1}$$

$\lambda$'ya veri setinden karar verilir. Bu aile aşağıdaki basit dönüşümleri içermektedir:

$$\begin{aligned}
&\lambda = 2 && Y' = Y^2 \\
&\lambda = .5 && Y' = \sqrt{Y} \\
&\lambda = 0 && Y' = \log_e Y \\
&\lambda = -.5 && Y' = \frac{1}{\sqrt{Y}} \\
&\lambda = -1.0 && Y' = \frac{1}{Y}
\end{aligned} \tag{7.2}$$

Bu dönüşümlerle klasik normal doğrusal regresyon modeli aşağıdaki biçimi alır:

$$Y_i^{\lambda} = \beta_0 + \beta_1 X_i + u_i \tag{7.3}$$

Bu regresyon modeli farklı olarak $\lambda$ parametresini içermektedir. Bu parametrenin tahmin edilmesi gerekir. Box-Cox prosedürü diğer parametreler $\beta_0, \beta_1$ ve $\sigma^2$ gibi $\lambda$ parametresini En Çok Olabilirlik metodunu kullanarak tahmin eder. Bu regresyon modeli için olabilirlik fonksiyonu şu hali alır:

$$L(\beta_0, \beta_1, \sigma^2, \lambda) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left[ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (Y_i^{\lambda} - \beta_0 - \beta_1 X_i)^2 \right] \tag{7.4}$$

Bu olasılık yoğunluk fonksiyonunun $\beta_0, \beta_1, \sigma^2$ ve $\lambda$'ye göre maksimum değerlerinin elde

edilmesi bu parametrelerin En Çok Olabilirlik tahmincilerini verir. Bu yolla Box-Cox dönüşümleri $\lambda$'nın En Çok Olabilirlik tahmincisi $\hat{\lambda}$'yı kuvvet dönüştürme işleminde kullanmak üzere belirler.

İstatistik bilgisayar paketleri kuvvet dönüşümleri için En Çok Olabilirlik tahmini $\hat{\lambda}$'yı vermez. $\hat{\lambda}$'yı tahmin etmek için basit bir prosedür standart regresyon paketlerini kullanmak olabilir. Bu süreç potansiyel $\lambda$ değerlerinin arasından nümerik bir araştırma yapılmasını gerektirir. Örneğin $\lambda = -2, \lambda = -1.75....., \lambda = 1.75, \lambda = 2.$ Her bir $\lambda$ parametresi için $Y_i^{\lambda}$ gözlemleri standardize edilir dolayısıyla hata kareleri toplamının büyüklüğü $\lambda$ parametresinden bağımsız olur.

$$w_i = \begin{cases} K_1(Y_i^{\lambda} - 1) & \lambda \neq 0 \\ K_2(\log_e Y_i) & \lambda = 0 \end{cases} \quad \text{burada} \quad K_2 = \left(\prod_{i=1}^{n} Y_i\right)^{1/n}, \quad K_1 = \frac{1}{\lambda K_2^{\lambda - 1}} \tag{7.5}$$

şeklinde $w_i$ standardize değerleri elde edilir. $K_2$, $Y_i$ gözlemlerinin geometrik ortalamasıdır.

Verilen $\lambda$ değerleri için standardize edilmiş $w_i$ gözlemleri bir defa elde edildikten sonra, X üzerine regresyonları kurulur ve hata kareleri toplamı $SSE$ elde edilir. En Çok Olabilirlik tahmini $\hat{\lambda}$'nın, $SSE$'yi minimum yapan $\lambda$ değeri olduğu görülebilir.

Mutlaka Box-Cox dönüşümleri yapıldıktan sonra serpilme diyagramları ve kalıntı diyagramları oluşturularak dönüşümlerin uygun olup olmadığı incelenebilir.

Box-Cox prosedürü $\lambda$ değerini 1'e yakın belirliyorsa Y üzerinde herhangi bir dönüştürme işlemine gerek yoktur (Neter vd., 1996).

# 8. ÇOKLU DOĞRUSAL BAĞLANTI

Çoklu doğrusal bağlantı ilk kez 1934 yılında Ragnar Frisch tarafından ortaya atılmış bir kavramdır. Literatürde çoklu doğrusallığın kesin bir tanımı bulunmamaktadır. Eğer iki değişkenin gözlem vektörleri aynı doğru üzerinde ise (bir boyutlu alt uzayda ise) iki değişken arasında çoklu doğrusal bağlantı bulunmaktadır. Daha genel olarak p-1 (bağımsız değişken sayısı) değişkeni temsil eden vektörler p-1'den daha az boyutlu bir uzayda gösterilebiliyorsa, yani vektörlerden biri diğerlerinin doğrusal bir kombinasyonuysa p-1 değişkenler arasında çoklu doğrusal bağlantı bulunmaktadır. Uygulamada bu gibi "tam çoklu doğrusallık" yalnızca verilerin araştırmacı tarafından oluşturulması durumunda mümkündür ve nadiren gerçekleşir. Bu bilgi ikiden fazla bağımsız değişkenli regresyon modelleri için genelleştirilebilir; değişkenlerden birinin diğerleri üzerinde kurulan regresyonunda yüksek çoklu korelasyon bulunursa çoklu doğrusal bağlantı gerçekleşir (Belsley vd., 1980).

Çoklu doğrusal bağlantı bağımsız değişkenler arasındaki teorik ya da gerçek doğrusal ilişkinin varolup olmamasına bağlı değildir; üzerinde çalışılan veri kümesinde yaklaşık doğrusal bir ilişkinin varolup olmadığıyla ilgilidir. Diğer tahmin problemlerinin aksine bu problem üzerinde çalışılan örneğin özelliklerinden kaynaklanır.

Veri setinde çoklu doğrusal bağlantının oluşmasının değişik nedenleri vardır. Örnek olarak bağımsız değişkenler ortak bir trende sahip olabilirler. Bir değişken belirli bir trende sahip diğer bir değişkenin bir dönem gecikmeli değerlerinden oluşabilir. Yeterince geniş bir tabanı kapsaması sonucu bazı bağımsız değişkenler birlikte bir değişim içinde olabilirler ya da bazı bağımsız değişkenler arasında gerçekten de yaklaşık doğrusal bir ilişki bulunabilir.

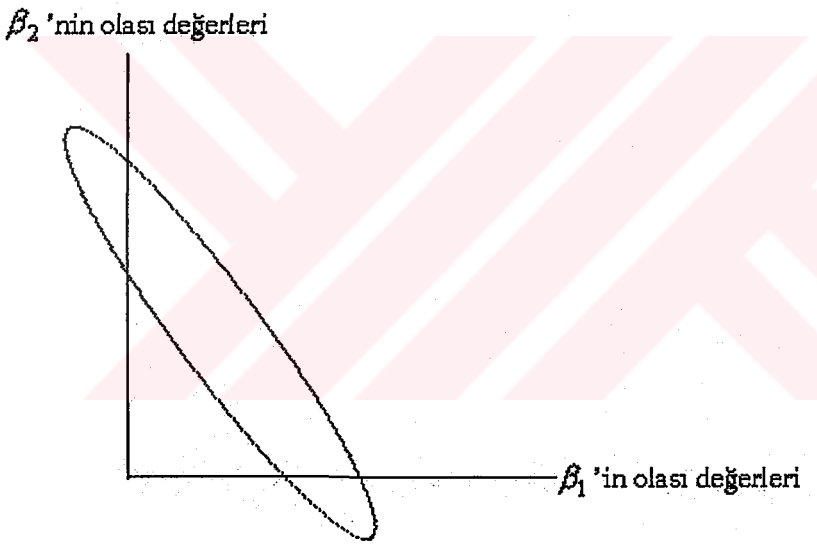
## 8.1 Çoklu Doğrusal Bağlantının Sonuçları:

Çoklu doğrusal bağlantının varlığı halinde elde edilen En Küçük Kareler tahmin edicilerinin BLUE olmalarına karşın, varyansları ve ortak varyansları büyüktür, bu da kesin tahmini güçleştirmektedir. Bu yüksek varyansın oluşmasının temel nedeni ise çoklu doğrusal bağlantının bulunduğu bir ortamda belirli bir değişkene ait yeterli *bağımsız* değişimin bulunmaması sonucu E.K.K. tahmin sürecinin bu değişkenin bağımlı değişken üzerinde sahip

olduğu etkiyi kesin olarak belirleyememesidir.

Bağımsız değişkenler arasında yüksek korelasyonun bulunduğu durumlarda toplam değişimin önemli bir bölümü ortaktır ve geriye değişkenlerin kendilerine özgü çok az değişim kalır. Bu, E.K.K sürecinin parametre tahminlerinde kullanabileceği (çok küçük örnek çaplarında ya da bağımsız değişkenin fazla değişmediği durumlarda olduğu gibi) çok az bilginin bulunduğu anlamına gelir.

Yüksek varyans değeri parametre tahminlerinin kesin olmadığı anlamını taşır (yani araştırmacı parametrelerin itimat edilebilecek bir tahminini elde edemez) ve hipotez testi güçlü değildir (parametrelerle ilgi karşı hipotezler reddedilemez). Bunun bir örneği olarak aşağıda verilen şekildeki durum incelenecektir:



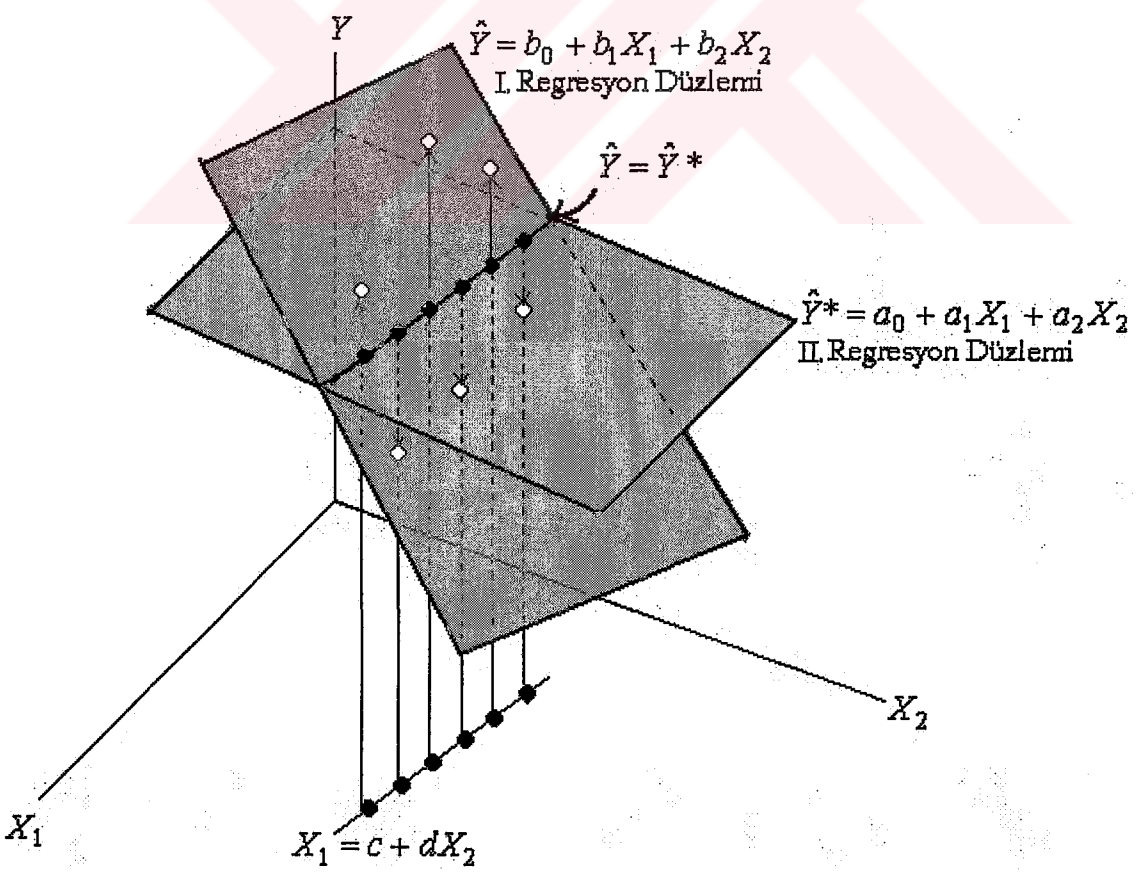

Şekil 8.1 Çoklu doğrusal bağlantı durumunda iki parametre tahminine ilişkin güven elipsi

$\beta_1$ ve $\beta_2$ tahminleri için oluşturulan güven elipsi iki değişken arasındaki ortak doğrusallığı yansıtacak bir şekilde uzun, dar ve devriktir. Ortak değişimin bağımlı değişken üzerindeki etkisinin gerçekte birinci bağımsız değişkene ait olması durumunda $\beta_1$ büyük ve $\beta_2$ küçük olacak ve bu da gerçek parametre kümesinin elipsin sağ alt tarafında olacağını ima edecektir. Aynı bağlamda etkinin ikinci değişkene ait olması durumunda ise $\beta_2$ büyük ve $\beta_1$ küçük olacak ve bu da gerçek parametre kümesinin elipsin sol üst tarafında olacağını ima edecektir. İki tahminci arasında yüksek bir negatif kovaryans vardır. Elips yatay ve dikey eksenlerin bir kısmını içermektedir. Dolayısıyla bireysel $\beta_1$=0 ve $\beta_2$=0 hipotezlerini reddetmek mümkün

değildir. Ancak elips orjin noktasını içermemektedir ve bu da $\beta_1 = \beta_2 = 0$ şeklinde belirlenen birleşik hipotezin reddedilebileceğini ima etmektedir. Bu durum yüksek korelasyonlu iki değişkenin bağımlı değişken üzerindeki tekil etkilerini birbirinden soyutlamanın olanaksız olduğu anlamına gelir (Kennedy, 2000).

Genelde bir regresyon katsayısı bağımsız değişkenin değerindeki bir birimlik değişmeye karşılık bağımlı değişkenin beklenen değerindeki değişimi vermektedir. Bu değişim diğer değişkenlerin etkilerinin sabit olduğu varsayımı altında yapılmaktadır. Değişkenler arasında yüksek korelasyon bulunduğu zaman bu yorumun yapılması mümkün değildir. Çünkü birinci bağımsız değişkendeki değişim aynı zamanda diğer bağımsız değişkendeki değişimi ifade etmektedir.

Geometrik bir yaklaşım içinde tam çoklu doğrusal bağlantı aşağıdaki şekil yardımıyla açıklanabilir:




Şekil 8.2 Tam Çoklu Doğrusallık

Şekilde "o" ile gösterilen verilerin siyah nokta ile gösterilen tahminleri bir doğruyu gerçeklemektedir. Bu doğrunun izdüşümü bağımsız değişkenler arasındaki ilişkiyi vermektedir: $X_1 = c + dX_2$. Şekildeki regresyon düzlemlerinin tamamıyle farklı olduğu görülebilir. $X_1 = c + dX_2$ olduğu durumda iki tahmin uzayı kesişir ve aynı öngörü değerlerini verirler. Şekilde görünümün belirgin olması amacıyla iki farklı regresyon düzlemi gösterilmiştir. Fakat bu doğru etrafında pek çok düzlem oluşturulabilir. Çoklu doğrusal bağlantının bulunmadığı durumda regresyon modelinde paramatrelerin herhangi birinde küçük bir değişiklik kalıntı kareleri toplamı, SSE'de büyük değişikliğe sebep olur fakat burada SSE hata paylarına ilişkin değişkenlikte bir değişme olmadan,yani SSE artıp azalmadan düzlem doğru etrafında dönebilir. Böylece E.K.K. yönteminin SSE'yi minimum yapma özelliği ortadan kalktığından E.K.K. ile elde edilen düzlem tanımlanamamaktadır. Şekilden de görülebileceği gibi her iki regresyon düzlemi de aynı işi görmektedir (Genceli, 2002). $X_1$ ve $X_2$ korelasyonlu ise ve veri seti rassal hata bileşeni içermiyorsa (aralarında kesin bir ilişki söz konusu ise) $X_1 = c + dX_2$ eşitliğini sağlayan $(X_1, X_2)$ kombinasyonları için pek çok regresyon düzlemi aynı tahmini sağlar. Bu tahmin fonksiyonları aynı değildir ve $X_1, X_2$ arasındaki ilişkiyi izlemeyen $(X_1, X_2)$ ikilileri için farklı tahminler üretirler.

Buraya kadar dikkat edilmesi gereken iki anahtar nokta vardır: birincisi $X_1$ ve $X_2$ arasındaki kesin ilişki öngörü yapılmasını engellememektedir , ikincisi pek çok tahmin fonksiyonu aynı iyi tahmini sağladığından regresyon katsayılarından birinin bağımlı değişken üzerindeki etkisinin diğer katsayılardan büyük olduğu söylenemez. Çünkü başka bir regresyon düzleminde söz konusu katsayı diğer katsayılardan düşük olduğu halde aynı tahmin elde edilebilir.

Uygulamada tam çoklu doğrusallık nadir rastlanılan bir durumdur. Güçlü fakat tam çoklu doğrusallığın olmadığı durum hastalıklı durum olarak nitelendirilebilir. Burada E.K.K. düzleminin, aralarında çoklu doğrusallık bulunan $(X_1, X_2)$ ikililerinin (bu bağlantıyı sağlamayan ikili kombinasyonlar haricindeki) oluşturduğu temel eksen etrafında döndürülmesi hata kareleri toplamı $SSE$'de az bir değişime sebep olur. Burada düzlem hastalıklı belirlenmiştir ve istatistiksel olarak E.K.K. tahminleri (sabit ve eğimler) kesin değildir ve yüksek varyanslıdır.

Bu gösterimlerin farklı varyasyonları düzenlenebilir. Örneğin düzlemin Y eksenini kestiği nokta olan sabit kesin fakat tahmin edilen eğim katsayıları $b_1, b_2$ hastalıklı belirlenebilir. Veya $b_0, b_1$ tahminleri kesin değildir, $b_2$ kesin belirlenebilir. Bu durum $X_1$ ve $X_2$ arasında çoklu doğrusallığın olmadığı, $X_1$ ile sabit arasında çoklu doğrusallığın olduğu durumda gerçekleşir (Belsley vd., 1980).

## 8.2 Çoklu Doğrusal Bağlantının Belirlenmesi

Çoklu doğrusallığın belirlenmesi için bir çok süreç tanımlanmıştır. Burada en çok kullanılan teknikler ve bu tekniklerin zayıf yanlarından bahsedilecektir:

*1.Pozitif olması beklenilen bir katsayı negatif çıkabilir. Hipotez testlerinin sonuçları doğru değildir. Önemli açıklayıcı değişkenler düşük t istatistiklerine sahiptir ve dolayısıyla anlamsız kabul edilir. Buna karşın belirginlik katsayısı $R^2$ yüksek çıkar ve dolayısıyla F sınaması anlamlı sonuç verir.*

*2.Açıklayıcı Değişkenlerin Korelasyon matrisi R veya bu korelasyon matrisinin tersi $R^{-1}$ incelenebilir.*

X matrisinin standardize edilmesinden sonra korelasyon matrisi basitçe $X'X$'e eşit olur. İki açıklayıcı değişken arasındaki yüksek korelasyon olası çoklu doğrusal bağlantı probleminin varolduğunu gösterir. İkiden çok değişkenli modellerde ise ikili olarak aralarındaki korelasyon düşük olmasına rağmen üç veya daha fazla değişken arasında yaklaşık olarak doğrusal bir ilişki bulunabilir. Örneğin üç değişkenli bir regresyon modelinde $X_1$, $X_2$ ve $X_3$ bağımsız değişkenleri arasında ikili korelasyonlar düşük olabilir fakat $X_3$ bağımsız değişkeni diğer değişkenlerin doğrusal bir kombinasyonu olarak elde edilebilir. Bu durumda ikili korelasyonların düşük olması çoklu doğrusal bağlantı olmadığı anlamına gelmez. Kısaca ikiden çok açıklayıcı değişken içeren modellerde ikili korelasyonlar çoklu doğrusal bağlantı durumu için yanılmaz bir gösterge değildir.

Çoklu doğrusallığın tespit edilebilmesi için R korelasyon matrisinin bahsedilen eksiklikleri

aynı zamanda $R^{-1}$'in de kullanımını sınırlandırmaktadır. Bu teşhis ölçüsünün yaygın bir kullanımı aşağıdaki gibidir:

X matrisi standartlaştırılmış değerlerden oluşmaktadır. Dolayısıyla korelasyon matrisi bahsedildiği gibi $X'X$'e eşit olacaktır. Aynı zamanda bu $R^{-1}=(X'X)^{-1}$ olduğunu gösterir. $R^{-1}$'in köşegen elemanları $r_{xx}$ , varyans büyütme faktörü $VIF$ olarak adlandırılır. $VIF$ bir parametre tahmininin doğrusal bağlantı nedeniyle kesinlikten ne derece uzaklaştığının bir ölçüsüdür (Genceli, 2002).

$$(VIF)_k=\frac{1}{1-R_k{}^2} \tag{8.1}$$

$(VIF)_k$, $R^{-1}=(X'X)^{-1}$ matrisinin köşegeninde yer alan k değeridir (k=1,2,...,p-1). $X_k$ değişkeni diğer bütün bağımsız değişkenlerin bir fonksiyonu olarak düşünülür ve bu değişkenler üzerine regresyonu oluşturulur. Bu regresyondan $R_k{}^2$ elde edilir . Formülden de görülebileceği gibi çoklu doğrusal bağlantı yüksekse $R_k{}^2$ büyür, dolayısıyla $VIF_k$ artar.

Pratik bir kural maksimum $VIF$ değeri 10'dan büyük bir değer ise söz konusu değişkenle diğer değişkenler arasında ciddi bir çoklu doğrusal bağlantı problemi bulunmaktadır.

Doğrusal bağlantı bulunmaması halinde $R_k{}^2$=0 olduğundan $VIF_k$=1 olur. Kısaca alttan sınırlıdır. $VIF$'e getirilebilecek bir eleştiri ise üst sınırının belli olmamasıdır.

### *3.Farrar ve Glauber Tekniği*

Korelasyon matrisinin köşegeni dışında kalan yüksek değerlerin belirlenmesi tüm olası çoklu bağlantı olasılıklarını içermez; ikili olarak aralarındaki korelasyon düşük olmasına rağmen üç veya daha fazla değişken arasında yaklaşık olarak doğrusal bir ilişki bulunabilir. Bu durumda matrise ait determinant değeri çoklu doğrusal bağlantının bir indeksi olarak kullanılabilir. Veri kümesinin bağımsız (ortogonal) olmaya yakın olması durumunda korelasyon matrisi

birim matrise yakınsayacak ve bu matrise ait determinant değeri yaklaşık bir olacaktır; veri setinin çoklu doğrusal bağlantıya sahip olması durumunda ise korelasyon matrisinin determinant değeri sıfır değerine yakınsar. Farrar ve Glaubert (1967) ve Haitovsky (1969) bu amaca yönelik olarak bağımsız değişkenlerin çok değişkenli normal bir dağılıma sahip oldukları gibi bir kısıtlayıcı varsayım altında determinant değerini kullanan testler geliştirmiştir. Farrar ve Glauber bir örnekteki çoklu doğrusallığı, gözlenen X'lerin ortogonallikten sapması olarak ele almaktadırlar. Yaklaşımlarının kaynağı tam çoklu doğrusallık varsa, katsayılar belirlenememekte ve çeşitli açıklayıcı değişkenler arasındaki ilişkiler, çoklu korelasyon katsayıları ve kısmi korelasyon katsayılarıyla ölçülebilmektedir. Farrar ve Glauber sınaması şöyle özetlenebilir:

*a. İlk olarak, birkaç açıklayıcı değişkeni olan bir fonksiyonda çoklu doğrusallığın varlığını ve ciddiliğini ortaya çıkaracak bir* $\chi^2$ *sınaması.*

Bu aşamada sınanan hipotez, örnekteki X'lerin ortogonal olduklarıdır ($r_{x_i x_i}=1$ ve $r_{x_i x_j}=0$). Bunun için değişkenleri örnek büyüklüğüne ve standart sapmaya göre standartlaştırmak uygun yoldur. Bu yolla elde edilen korelasyon matrisinin determinantı tam çoklu doğrusal bağlantının varlığı halinde sıfıra eşittir. X'lerin ortogonal olduğu durumda ise her X çifti için basit korelasyon katsayıları sıfıra eşittir,dolayısıyla korelasyon matrisinin determinantı bire eşittir. Buradan çıkan sonuç eğer determinantın değeri sıfır ile bir arasındaysa, fonksiyonda bir miktar çoklu doğrusallık bulunmaktadır. Bu olgudan yola çıkan Farar ve Glauber, bütün açıklayıcı değişkenler kümesi için çoklu doğrusallığın gücünü saptamak amacıyla aşağıdaki $\chi^2$ sınamasını önermektedirler:

$H_0$ :X'ler ortogonaldir.

$H_a$ : X'ler ortogonal değildir. (8.2)

Farrar ve Glauber şu büyüklüğün

$$*\chi^2 = -\left[n-1-\frac{1}{6}(2k+5)\right]\log_e .(\text{korelasyon matrisinin determinantı}) \tag{8.3}$$

$v=\frac{1}{2}k(k-1)$ serbestlik derecesiyle bir $\chi^2$ dağılımına sahip olduğunu bulmuşlardır (k açıklayıcı değişken sayısıdır). Örnek verilerinden $*\chi^2$ görgül değeri bulunur ve seçilen

anlamlılık düzeyinde ve $v=\frac{1}{2}k(k-1)$ serbestlik derecesiyle $\chi^2$ tablo değeriyle karşılaştırılır. Gözlemlenen $\chi^2$ değeri $\chi^2$ tablo değerinden büyükse, ortogonallik varsayımı reddedilir, yani .çoklu doğrusallığın varlığı kabul edilir. Gözlemlenen $*\chi^2$ değeri yükseldikçe, çoklu doğrusallığın ciddiliği artar.

*b. İkinci olarak,Çoklu doğrusallığın yerini belirleyecek bir F sınaması*

Çoklu doğrusallığa giren etmenleri belirlemek için Glauber ve Farrar açıklayıcı değişkenler arasındaki çoklu korelasyon katsayısını hesaplarlar ($R^2_{x_1.x_2x_3x_4,...,x_k}$ , $R^2_{x_2.x_1x_3x_4,...,x_k}$ ve genel olarak $R^2_{x_i.x_2x_3x_4,...,x_k}$ ) ve bu çoklu korelasyon katsayılarının istatistik bakımından anlamlılığını bir F sınamasıyla aşağıdaki gibi sınar. Her çoklu korelasyon katsayısı için gözlemlenen $F^*$ değeri hesaplanır:

$$F^*=\frac{(R^2_{x_i.x_1x_2,...,x_k})/(k-1)}{(1-R^2_{x_i.x_1x_2,...,x_k})/(n-k)} \tag{8.4}$$

n:örnek büyüklüğü,k: açıklayıcı değişkenlerin sayısıdır. Bu aşamada sınanan hipotez ve alternatif hipotez şu şekildedir:

$$\begin{aligned} H_0 &: R^2_{x_i.x_1x_2,...,x_k}=0 \\ H_a &: R^2_{x_i.x_1x_2,...,x_k}\neq 0 \end{aligned} \tag{8.5}$$

Gözlemlenen $F^*$ değeri $v_1=(k-1)$ ve $v_2=(n-k)$ serbestlik derecelerinde, seçilmiş anlamlılık düzeyinde F tablo değeriyle karşılaştırılır. $F^*>F$ ise $X_i$ değişkeninin çoklu doğrusal olduğu kabul edilir, yani sıfır hipotezi reddedilir.

*c. Üçüncü olarak, Çoklu doğrusallığın kalıbını belirlemek için bir t sınaması*

Bu da çoklu doğrusallığı doğuran değişkenlerin saptanmasını amaçlayan bir sınamadır. Çoklu doğrusallıktan sorumlu olan değişkenlerin bulunması için açıklayıcı değişkenler arasındaki kısmi korelasyon katsayıları hesaplanır ve t istatistiğiyle bunların anlamlılıkları sınanabilir. İki değişken arasındaki kısmi korelasyon katsayısı bütün öteki değişkenler aynı kalırken, bu iki değişken arasındaki bağlantının derecesini göstermektedir. İki değişkenli model için kısmi

korelasyon katsayısıyla basit korelasyon katsayısı aynıdır. Üç değişkenli model için örneğin 1.inci ve 2.inci değişken arasındaki kısmi korelasyon katsayıları şu formüllerle bulunur:

$$r^2_{x_1x_2.x_3} = \frac{(r_{12} - r_{13}r_{23})^2}{(1-r^2_{23})(1-r^2_{13})} \tag{8.6}$$

Üçten çok açıklayıcı değişken içeren modeller için benzeri formüller çıkarılabilir. Buradaki temel hipotez:

$$\begin{aligned} H_0 &: r_{x_ix_j.x_1x_2,\ldots,x_k} = 0 \\ H_a &: r_{x_ix_j.x_1x_2,\ldots,x_k} \neq 0 \end{aligned} \tag{8.7}$$

şeklindedir. Kısmi korelasyon katsayıları tahmin edildikten sonra, anlamlılıklarının sınanabilmesi için şu istatistik hesaplanır:

$$t^* = \frac{(r_{x_ix_j.x_1x_2,\ldots,x_k})\sqrt{n-k}}{\sqrt{1-r^2_{x_ix_j.x_1x_2,\ldots,x_k}}} \tag{8.8}$$

Burada $r_{x_ix_j.x_1x_2,\ldots,x_k}$, $X_i$ ve $X_j$ arasındaki kısmi korelasyon katsayısını gösterir. Gözlemlenen $t^*$ değeri $v=(n-k)$ serbestlik derecesinde ve seçilmiş anlamlılık düzeyinde t tablo değeriyle karşılaştırılır. Eğer $t^*>t$ ise, kısmi korelasyon katsayılarının istatistik bakımdan anlamlı olduğu, yani fonksiyondaki çoklu doğrusallıktan $X_i$ ve $X_j$ değişkenlerinin sorumlu oldukları kabul edilir (Koutsoyiannis, 1989).

Farrar ve Glaubert önerdikleri test kapsamı içinde her bir bağımsız değişkeni doğrusal ilişkileri belirlemek amacıyla geriye kalan diğer bağımsız değişkenlerle regresyona sokmuşlardır. Verili bir korelasyon farklı çoklu doğrusallık yapılarıyla uyumlu olabileceğinden Farrar ve Glaubert'in kısmi korelasyon sınaması etkin değildir (Gujarati, 1999).

***4.Korelasyon matrisinin özdeğerlerinin ve özvektörlerinin (veya temel bileşenlerinin) İncelenmesi***

$X'X$ çapraz çarpımlar matrisinin veya R korelasyon matrisinin özdeğer ve özvektörleri çoklu doğrusal bağlantıyla ilgili olarak yıllardır kullanılmaktadır. Kloeck ve Mennes(1960) X'in temel bileşenlerini kullanmanın bazı yollarını göstermiştir. Kendall (1957) ve Silvey (1969) $X'X$'in özdeğerlerini çoklu doğrusal bağlantı durumunda anahtar ölçü olarak kabul etmişlerdir ve çoklu doğrusallık "küçük" özdeğerin varlığı durumunda ortaya çıkmaktadır. Fakat bu "küçük" değerin ne kadar küçük olması gerektiği konusunda net bir bilgi bulunmamaktadır. Genel bir eğilim "küçük" değerin sıfır ile karşılaştırılmasıdır. Bazı durumlarda çoklu doğrusallık, bir özdeğer diğerlerine göre küçükse oluşabilir. Burada "küçük" değer sıfıra göre değil diğer büyük özdeğerlere göre belirlenmektedir (Belsley vd.,1980).

*c* koşul sayısı olmak üzere,

$$C = \frac{\lambda_{\max}}{\lambda_{\min}} \tag{8.9}$$

ve koşul indeksi,

$$CE = \frac{\sqrt{\lambda_{\max}}}{\sqrt{\lambda_{\min}}} = \sqrt{C} \tag{8.10}$$

biçiminde hesaplanır. Bu durumda şu parmak hesabı yapılabilir. Eğer C 100 ile 1000 arasında ise çoklu doğrusallık orta ya da güçlü derecededir. 1000'i aşıyorsa ciddidir. Eğer CE ($=\sqrt{C}$) 10 ile 30 arasındaysa çoklu doğrusallık orta ya da güçlü derecededir. 30'u aşıyorsa ciddidir.

***5.Yan Regresyonlar***

Çoklu doğrusallık, bir ya da daha çok açıklayıcı değişkenin, öteki açıklayıcı değişkenlerin tam ya da yaklaşık doğrusal bileşimi olmasından doğduğuna göre, hangi X değişkeninin öteki X değişkenleriyle ilişkili olduğunu bulmanın bir yolu, her bir $X_i$'nin öteki X değişkenlerine göre regresyonunu bulup buna karşılık gelen, $R_i^2$ değerlerinin hesaplanmasıdır. Bu

regresyonlardan her birine yan regresyon denir. F ile $R^2$ arasında kurulan ilişkiden yararlanılırsa

$$F_i = \frac{R^2_{X_1,X_2,...,X_p}/(p-2)}{(1-R^2_{X_1,X_2,...,X_p})/(n-p+1)} \tag{8.11}$$

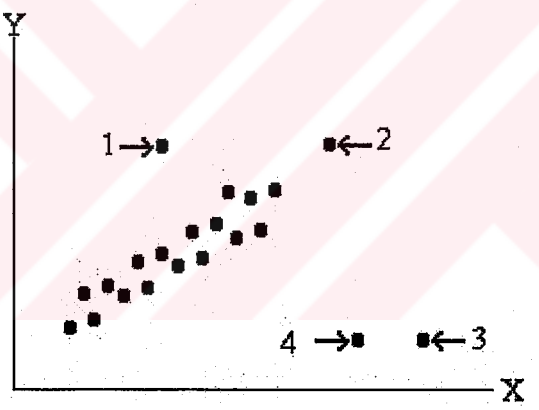
değişkeni *p-2* ve *n-p+1* serbestlik derecesiyle F dağılımına uymaktadır. p sabit terimle birlikte açıklayıcı değişken sayısını, $R^2_{X_1,X_2,...,X_p}, X_i$ değişkeninin kalan X değişkenlerine göre regresyonunun belirginlik katsayısını ifade etmektedir. Hesaplanan $F_i$ seçilmiş anlamlılık düzeyindeki eşik $F_i$ değerini aşıyorsa, bunun anlamı, $X_i$'nin öteki X'lerle ortak doğrusal bağlantı içinde olduğudur.

Bütün $R^2$ değerlerini biçimsel olarak sınamak yerine, *Klein'in parmak hesabı* benimsenebilir. Buna göre, bir yan regresyondan bulunan $R^2$, bütünün yani Y'nin bütün açıklayıcı değişkenlere göre regresyonunun $R^2$'sinden büyükse, çoklu doğrusallık o zaman ciddi bir problem olabilir (Gujarati, 1999).

# 9. SAPAN DEĞERLERİN VE ETKİLİ DEĞERLERİN ANALİZİ

Regresyon analizinde bazı veri setleri diğer verilerden ayrılan sapan veya aykırı gözlemler içerir. Bu sapan gözlemler büyük kalıntılar içerebilir ve öngörülen en küçük kareler regresyon fonksiyonu üzerinde tesirli etkiler oluşturur. Dolayısıyla sapan değerleri dikkatle incelemek ve modelden dışlanıp dışlanmayacağına veya ayrı tutulacağına karar vermek gerekir. Eğer ayrı tutulursa öngörü sürecindeki etkileri azalır ve/veya regresyon modeli tekrar gözden geçirilir.

Bir gözlem Y'ye göre veya X'e göre veya her ikisine göre sapan değer veya aykırı değer olabilir. Sapan değerler (outliers) ifadesi sapan Y gözlemleri için kullanılır. Bağımsız değişkenlerdeki bu değerler aykırı değerler olarak ifade edilir. Bu aykırı gözlemlere ilişkin bağımlı değişken değerleri sapan değer olmayabilir.




Şekil 9.1 Tek açıklayıcı değişkenli regresyon modeli için sapan değerleri gösteren serpilme diyagramı

Şekilde 1 değeri Y değerine göre sapan değer durumundadır. Bu noktanın serpilme diyagramında ayrı görünmesine rağmen bu noktanın X değeri bağımsız değişkenlerin gözlem aralığının ortasında yer almaktadır. 2,3 ve 4 X değerlerine göre aykırı gözlemlerdir. Bu noktalar diğer noktalara göre daha büyük X değerlerine sahiptirler. Bütün sapan değerlerin öngörülen regresyon fonksiyonu üzerinde kuvvetli etkileri yoktur. Şekildeki 1 noktası çok etkili bir gözlem değildir çünkü diğer gözlemlere benzer olarak öngörülen regresyon doğrusunu sapan değere doğru kaydırmayacak X değerlerine sahiptir. Benzer şekilde 2. gözlem değeri de etkili bir sapan değer olmayacaktır çünkü 2. gözleme karşılık gelen Y değeri sapan değer dışındaki gözlemlerle oluşturulan regresyon ilişkisine karşılık gelen Y değeriyle uyumlu olacaktır. 3 ve 4 gözlemlerinin X değerleri diğer X değerlerinden uzağa düşmektedir

ve bu gözlemlere karşılık gelen Y değerleri kalan verilerle oluşturulan regresyon ilişkisiyle uyumlu değildir. Dolayısıyla 3 ve 4 regresyon fonksiyonunun öngörülmesinde etkili gözlemler olarak görülmektedirler.

Göz önünde bulundurulan bir regresyon modelinin veri setindeki bir veya birkaç gözlemden kuvvetli bir şekilde etkilenip etkilenmediğine karar vermek için temel adım, örneğin bir veya iki açıklayıcı değişkenli model için, kutu diyagramları, dal ve yaprak diyagramları, serpilme diyagramları ve kalıntı diyagramları gibi görsel teşhis diyagramlarının incelenmesi ve bu diyagramlardan X veya Y değerlerine göre sapan değerleri belirlenmesidir. Fakat regresyon modelinde iki açıklayıcı değişkenden fazla değişken bulunduğu taktirde basit grafiksel araçlarla sapan değerleri bulmak güçleşecektir. Çünkü bazı tekil sapan değerler çoklu regresyon modelinde sapan değer olmayabilir veya tersine bazı çoklu modellerde sapan değer olarak belirlenen gözlemler tek açıklayıcı değişkenli veya iki açıklayıcı değişkenli modelde sapan değer olarak belirlenmeyebilir.

## 9.1 Kalıntılar ve Standart Kalıntılar

Sapan Y gözlemlerinin belirlenebilmesi için kalıntıların incelenmesi gereklidir.

Kalıntılar : $e_i = Y_i - \hat{Y}_i$

Yarı studentized kalıntılar: $e_i^* = \dfrac{e_i}{\sqrt{MSE}}$ (9.1)

Y sapan gözlemlerini belirlemek amacıyla kalıntı analizini daha etkin yapabilmek için kalıntılar üzerinde iki düzeltme tanımlanacaktır. Bu düzeltmeler dönüşüm matrisinin kullanılmasını gerektirir.

## 9.2 DÖNÜŞÜM MATRİSİ

$$\hat{Y} = Xb$$
$$b = (X'X)^{-1}X'Y \tag{9.2}$$
$$\hat{Y} = X(X'X)^{-1}X'Y$$

Dönüşüm matrisi Y gözlem değerlerini $\hat{Y}$ tahmin değerlerine dönüştürmektedir. Dolayısıyla dönüşüm matrisi olarak adlandırılmaktadır. H ile ifade edilen dönüşüm matrisi:

$$\underset{nxn}{H} = X(X'X)^{-1}X' \tag{9.3}$$

$\hat{Y}_i$ tahmin değerleri $Y_i$ gözlem değerlerinin doğrusal bileşenleri olarak düşünülebilir:

$$\hat{Y} = H.Y \tag{9.4}$$

Ve kalıntılar $e_i$, dönüşüm matrisiyle $Y_i$ gözlem değerlerinin doğrusal bileşenleri olarak ifade edilebilir:

$$e = (I - H)Y \tag{9.5}$$

Kalıntıların varyans kovaryans matrisi:

$$\sigma^2\{e\} = \sigma^2(I - H) \tag{9.6}$$

$e_i$ kalıntının varyansı:

$$\sigma^2\{e_i\} = \sigma^2(1 - h_{ii}) \tag{9.7}$$

Bu ifadede $h_{ii}$ H dönüşüm matrisinin i.inci köşegen elemanıdır ve $e_i$ ve $e_j$ kalıntıları $(i \neq j)$ arasındaki kovaryans

$$\sigma\{e_i, e_j\} = \sigma^2(0 - h_{ij}) = -h_{ij}.\sigma^2 \qquad i \neq j \tag{9.8}$$

$h_{ij}$, H dönüşüm matrisinin i.inci satır ve j.inci sütun elemanıdır.

Bu varyans ve kovaryanslar hata terimleri varyansı $\sigma^2$'nin tahmincisi $MSE$ kullanılarak tahmin edilir:

$$s^2\{e_i\} = MSE(1 - h_{ii}) \tag{9.9}$$

$$s\{e_i, e_j\} = -h_{ij}(MSE) \tag{9.10}$$

H matrisinin diagonal elemanları:

$$h_{ii} = X_i'(X'X)^{-1}X_i \tag{9.11}$$

Bu ifadede yer alan $X_i$ sütun vektörü:

$$\underset{px1}{X_i} = \begin{bmatrix} 1 \\ X_{i.1} \\ \\ : \\ X_{i.p-1} \end{bmatrix} \tag{9.12}$$

$X_{i.p-1}$ elemanı p-1.inci değişkenin i.inci gözlemini ifade etmektedir.

## 9.3 Studentize Edilmiş Kalıntılar

Y sapan değerlerini belirleyebilmek amacıyla kalıntıları daha etkin yapabilmek için yapılacak ilk düzeltme, kalıntıların esasen farklı varyanslara sahip olabileceği gerçeğinin ifade edilmesini gerektirir.

$$s\{e_i\} = \sqrt{MSE(1 - h_{ii})} \tag{9.13}$$

$e_i$'nin $s\{e_i\}$'ye oranı studentize edilmiş kalıntı olarak adlandırılır ve $r_i$ ile gösterilir:

$$r_i = \frac{e_i}{s\{e_i\}} \tag{9.14}$$

$e_i$ kalıntılarının standart sapmaları belirgin bir şekilde farklılaşıyorsa kalıntıların standart hataları (kalıntıların standart sapmalarının örnekleme dağılımının standart sapması) büyük olacaktır. Diğer bir deyişle kalıntılar örnekten örneğe farklı dağılımlara sahip olacaklardır. Studentize edilmiş kalıntılar $r_i$, sabit varyansa sahiptir(model uygun olduğu varsayımı altında). Studentize edilmiş kalıntılar genellikle içsel studentize edilmiş kalıntılar olarak adlandırılır. İçsel ve dışsal studentize edilmiş kalıntılar olarak yapılan ayrımı belirleyen konu bu kalıntıların hesaplanmasında sapan değerlerin veri setinden silinip silinmemesidir. İçsel studentize edilmiş kalıntıların hesaplanmasında sapan değerler silinmemektedir.

## 9.4 Öngörü Kalıntıları (Deleted Residuals)

Kalıntılar için diğer bir düzeltme i.inci gözlem dışındaki gözlemlere dayanarak tahmin edilen regresyon modelinden i.inci kalıntıyı belirlemektir. Bu düzeltmenin sebebi eğer $Y_i$ uzak bir sapan değer ise i.inci gözlem dahil bütün gözlemlere dayalı en küçük kareler regresyon fonksiyonu bu sapan değerden etkilenerek $Y_i$'ye doğru kaydırılabilir. Diğer bir deyişle tahmin edilen $\hat{Y}_i$ tahmin değerini $Y_i$'ye yakın hale getirebilir. Bu durumda $e_i$ olması gerektiğinden küçük olacaktır ve dolayısıyla $Y_i$ gözlemini sapan değer olarak göstermeyecektir. Diğer yandan eğer i.inci gözlem, regresyon fonksiyonu tahmin edilmeden veri setinden dışlanırsa $\hat{Y}_i$ öngörü değeri $Y_i$ sapan değerinden etkilenmeyecektir ve i.inci gözlem için kalıntı değeri daha büyük olacaktır. Dolayısıyla bu düzeltme Y sapan değerini ortaya çıkarmada yardımcı olacaktır.

i.inci gözlem silindikten sonraki süreç kalan n-1 gözlemle regresyon modelini öngörmek ve $Y_i$ gözlemine karşılık gelen X seviyeleri de dışlandıktan sonra $\hat{Y}_i$ tahmin değerini elde etmektir. Bu $\hat{Y}_i$ tahmin değeri i.inci Y gözleminin ve X seviyelerinin dışlandığını belirtmek amacıyla $\hat{Y}_{i(i)}$ olarak gösterilecektir.

Gerçek gözlemlenen $Y_i$ değeriyle tahmin edilen $\hat{Y}_{i(i)}$ değeri arasındaki fark $d_i$ ile ifade edilecektir:

$$d_i = Y_i - \hat{Y}_{i(i)} \tag{9.15}$$

$d_i$ farkı i.inci gözlem için öngörü kalıntısı olarak adlandırılır. $d_i$ için cebirsel olarak eşdeğer bir hesaplama, i.inci gözlemi modelden dışlayarak regresyon fonksiyonunu yeniden tahmin etmeye gerek kalmaksızın aşağıdaki gibi yapılabilir:

$$d_i = \frac{e_i}{1 - h_{ii}} \tag{9.16}$$

$e_i$, i.inci gözlem için kalıntı değeri ve $h_{ii}$ H dönüşüm matrisinin i.inci köşegen elemanıdır. $h_{ii}$'nin büyük değerlerinde öngörü kalıntısı sıradan kalıntıdan büyük olacaktır. Sonuç olarak bazı durumlarda sıradan kalıntılar sapan Y gözlemlerini belirleyemediğinde öngörü kalıntıları bu sapan değerleri belirleyebilir. Bazı durumlarda ise öngörü kalıntılarıyla sıradan kalıntılar aynı sonuca götürebilmektedir. Öngörü kalıntısı aynı zamanda yeni bir gözlem için öngörü hatasına karşılık gelmektedir. Dolayısıyla öngörü kalıntısı olarak ifade edilmiştir. Öngörü yapılırken n gözleme dayalı tahmin edilen regresyon modeline dayanarak n+1. inci yeni gözlemin tahmini yapılmaktaydı. Öngörü kalıntıları için notasyon yeniden bu şekilde düzenlenirse, n-1 gözlem yeni n.inci gözlemin tahmini için kullanılmaktadır.

$s^2\{\text{tahmin}\} = MSE + s^2\{\hat{Y}_h\} = MSE(1 + X_h'(X'X)^{-1}X_h)$ idi. Öngörü kalıntıları için yeniden düzenlenirse,

$$s^2\{d_i\} = MSE_{(i)}(1 + X_i'(X_{(i)}'X_{(i)})^{-1}X_i) \tag{9.17}$$

$X_i$,i.inci gözlem için X gözlem vektörü (i.inci satırın devriği), $MSE_{(i)}$,i.inci gözlemin veri setinden dışlanmasından sonra öngörülen regresyon fonksiyonunun ortalama hata karesi, ve $X_{(i)}$ , i.inci gözlemin silinmesinden sonra kalan X matrisi. $s^2\{d_i\}$ için cebirsel olarak eşdeğer bir ifade:

$$s^2\{d_i\} = \frac{MSE_{(i)}}{1 - h_{ii}} \tag{9.18}$$

biçimindedir. Ve

$$\frac{d_i}{s\{d_i\}} \approx t(n-p-1) \tag{9.19}$$

burada i.inci gözlemin tahmininde n-1 veri kullanıldığından uygun serbestlik derecesi $(n-1)-p=n-p-1$'dir.

## 9.5 Studentize Edilmiş Öngörü Kalıntıları

Yukarıda bahsedilen her iki düzeltmenin birleştirilmesi Y aykırı ya da sapan değerlerini belirlemek için yararlanılan diğer bir düzeltmedir. $d_i$ öngörü kalıntısının kendi standart sapmasına bölünmesiyle studentize edilmiş kalıntılar elde edilir ve $t_i$ ile gösterilir:

$$t_i = \frac{d_i}{s\{d_i\}} \tag{9.20}$$

$t_i$ için cebirsel olarak eşdeğer bir ifade:

$$t_i = \frac{e_i}{\sqrt{MSE_{(i)}(1-h_{ii})}} \tag{9.21}$$

biçimindedir. $t_i$, studentized öngörü kalıntısı aynı zamanda dışsal studentized kalıntı olarak adlandırılmaktadır. Her bir $t_i$, n-p-1 serbestlik derecesiyle t dağılımına uymaktadır. Bununla birlikte $t_i$'ler bağımsız değillerdir.

Studentize edilmiş öngörü kalıntıları, her defasında bir gözlemin veri setinden dışlanıp regresyon modelinin tahmin edilmesine gerek kalmaksızın hesaplanabilir. $MSE$ ve $MSE_{(i)}$ arasında basit bir ilişki bulunmaktadır:

$$(n-p)MSE = (n-p-1)MSE_{(i)} + \frac{e_i^2}{1-h_{ii}} \tag{9.22}$$

Bu ilişkiyi kullanarak $t_i$ için eşdeğer bir hesaplama aşağıdaki gibi yapılabilir:

$$t_i = e_i \left[ \frac{n-p-1}{SSE(1-h_{ii}) - e_i^2} \right]^{1/2} \tag{9.23}$$

Böylece studentize öngörü kalıntıları, n gözleme dayalı tahmin edilen regresyon modeli için elde edilen, $e_i$ kalıntılarından, hata kareleri toplamı SSE'den ve dönüşüm matrisinin $h_{ii}$ elemanlarından yararlanılarak hesaplanabilir.

## 9.6 Weısberg Sapan Değer Testi

Studentize öngörü kalıntıların mutlak değer olarak büyük olanları Y sapan gözlemleri olarak belirlenir. Buna ilave olarak en büyük mutlak studentized öngörü kalıntısına $|t_i|$ sahip gözlemin sapan değer olup olmadığını belirleyebilmek için Bonferroni test sürecinin araçları kullanılarak biçimsel bir test oluşturulabilir. Hangi gözlemin en büyük $|t_i|$ değerine sahip olacağı önceden bilinmediğinden her bir gözlem için n tane ayrı testten oluşan bir aile düşünülebilir. Eğer regresyon modeli uygunsa, diğer bir deyişle sapan değer ve dolayısıyla regresyon modelinde değişim yoksa, her bir studentize öngörü kalıntısı n-p-1 serbestlik dereceli t dağılımına uymaktadır. Bu test Weisberg testi olarak da adlandırılmaktadır. $|t_i|$ ile ilgilenildiğinden test iki yönlüdür.

## 9.7 Sapan X Gözlemlerinin Belirlenmesi

### 9.7.1 Sapan X Gözlemlerini Belirlemede H Dönüşüm Matrisinin Kullanımı

Dönüşüm matrisinin studentize edilmiş öngörü kalıntısının büyüklüğüne karar vermede ve dolayısıyla Y sapan gözlemlerini belirlemede önemli bir role sahip olduğu yukarıda belirtilmiştir. Dönüşüm matrisi aynı zamanda sapan X gözlemlerini belirlemede yardımcı olur. Dönüşüm matrisinin köşegen elemanları X değerlerine göre bir gözlemin sapan değer olup olmadığını belirlemede yararlı bir göstergedir. Dönüşüm matrisinin bazı yararlı özellikleri bulunmaktadır.

H matrisi nxn boyutlu simetrik bir matristir. H matrisinin tanımından elde edilebileceği gibi pek çok özelliği bulunmaktadır. Örneğin H matrisinin X ile soldan çarpılması durumunda yeniden X matrisi elde edilir, $H.X = X$. Benzer şekilde $(I-H).X = 0$'dır. $H.H = H^2 = H$ özelliği (H matrisi idompotent bir matristir) $H(I-H) = 0$ olduğunu gösterir. Dolayısıyla $HY$ öngörü değerleriyle $(I-H)Y$ kalıntılar arasındaki kovaryans $\sigma^2 H(I-H) = 0$'dır (Weisberg, 1947). H, X sütun uzayı üzerinde dik projeksiyon operatörü olarak adlandırılır. X ile aynı ranka sahiptir,p. H matrisinin (i,j).inci elemanı,

$$h_{ij} = X_i^T (X^T.X)^{-1}.X_j = X_j^T (X^T X)^{-1} X_i = h_{ji} \qquad (9.24)$$

köşegen elemanları,

$$h_{ii} = X_i^T (X^T X)^{-1} X_i$$

$h_{ij}$'ler arasında pek çok ilişki bulunabilir. Örneğin

$$0 \le h_{ii} \le 1 \qquad \sum_{i=1}^{n} h_{ii} = rank(X) = p \tag{9.25}$$

dir. Dönüşüm matrisinin özdeğerleri 0 veya 1 değerlerinden oluşur. Özdeğerlerin toplamı matrisin rankını verir. Bu durumda $rank(H) = rank(X) = p$ ve dolayısıyla H matrisinin izi (köşegen elemanlarının toplamı) değişken sayısını verir.

$h_{ii}$ 1 değerini aldığında $\hat{Y}_i = Y_i$, dolayısıyla $e_i$=0 olacaktır. Bu bir parametrenin tamamıyla $Y_i$ tarafından belirlendiğini, ya da diğer bir deyişle bir nokta tarafından atandığı anlamına gelmektedir.

$$\det\left[X'_{(i)} X_{(i)}\right] = (1 - h_{ii}) \det(X'X) \tag{9.26}$$

Buradan açıkça görüldüğü gibi $h_{ii}$=1 olduğunda i.inci satırın silinmesiyle elde edilen yeni $X_{(i)}$ matrisi tekil matris olur ve dolayısıyla en küçük kareler tahminleri elde edilemez (Belsley vd., 1980). Fakat bu durum uygulamada nadir gerçekleşir.

İlave olarak her bir $h_{ii}$'nin alt sınırı (sabiti olan modeller için) $1/n$'dir. Üst sınırı ise $1/r$'dir. r, X'in $X_i$'ye özdeş satır sayısıdır. Aynı zamanda sabiti olan modeller için, $H1 = 1$'dir. veya skaler biçimde,

$$\sum_{i=1}^{n} h_{ij} = \sum_{j=1}^{n} h_{ij} = 1 \tag{9.27}$$

ve

$$\hat{Y}_i = \sum_{j=1}^{n} h_{ij} Y_j = h_{ii} Y_i + \sum_{j \ne i} h_{ij} Y_j \tag{9.28}$$

biçimindedir. 1 ve 2 kombine edilirse $h_{ii}$ 1'e yaklaştıkça $\hat{Y}_i$'nin $Y_i$'ye yaklaşacağı

görülmektedir. Bu sebeple $h_{ii}$ değeri i.inci gözlemin kaldıraç değeri olarak adlandırılır.

$h_{ii}$'nin değerlerinin büyük ya da küçük olması geometrik olarak açıklanabilir.

$\overline{X}$ vektörü X bağımsız değişkenlerinin örnek ortalama vektörleri olmak üzere D, orijinal X gözlemlerinden ortalamalarının çıkarılmasından sonraki nxp boyutlu fark matrisi olarak tanımlanır. X'in düzeltilmiş çapraz çarpımlar matrisi aşağıdaki gibi elde edilir:

$$Düz.KÇT = D'.D \tag{9.29}$$

$X_i^T$ X'in 1 sabiti olmadan oluşturulan satır vektörü olarak ifade edilirse $h_{ii}$ aşağıdaki gibi ifade edilebilir:

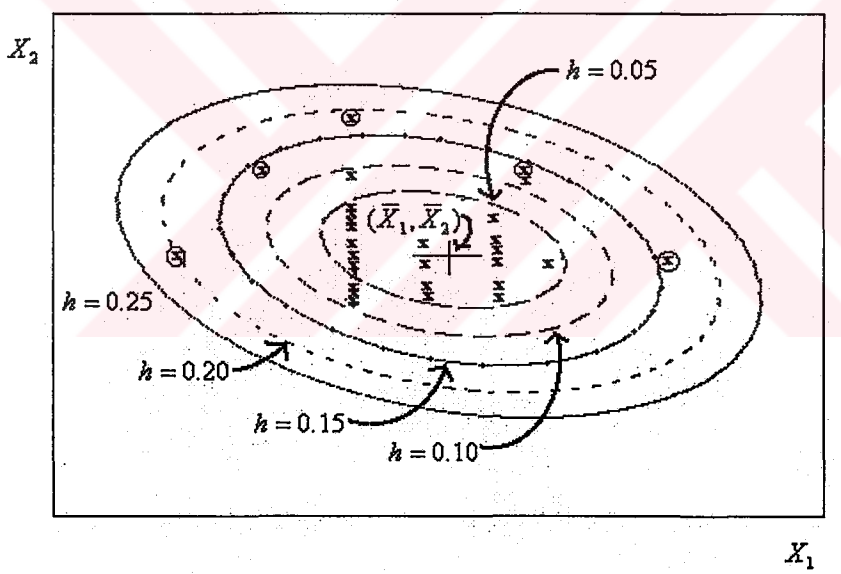
$$h_{ii} = \frac{1}{n} + (X_i - \overline{X})'(D'D)^{-1}(X_i - \overline{X}) \tag{9.30}$$

Denklemde eşitliğin sağ yanındaki $1/n$ terimi modelden çıkarılırsa ve kalan kısım $(n-1)$ ile çarpılırsa elde edilen sonuç $X_i$'nin veri setinin merkezi olan $\overline{X}$'ya olan kare *Mahalanobis uzaklığıdır*. Bu uzaklık $m_i^2$ ile gösterilir. İki değişkenli bir model örnek olarak verilirse $m_i^2$ değerleri matris notasyonuyla aşağıdaki gibi elde edilir ve yorumlanır:

$$m_i^2 = \begin{bmatrix} X_{i1} - \overline{X}_1 \\ X_{i2} - \overline{X}_2 \end{bmatrix}^T [(D'D)/n - 1]^{-1} \begin{bmatrix} X_{i1} - \overline{X}_1 \\ X_{i2} - \overline{X}_2 \end{bmatrix} \tag{9.31}$$

Merkezi $\overline{X}_1$ ve $\overline{X}_2$ olan iki boyutlu uzaydaki bir elipsin geometrik yerini tanımlayan $m_i^2$ değerleri aynı zamanda merkezi $\overline{X}_1$ ve $\overline{X}_2$ olan elipste, $X_{i1}$ ve $X_{i2}$ gözlem

çiftlerinden elips merkezine olan uzaklığın karesi olup kare mahalanobis uzaklığı olarak adlandırılır. Elips üzerindeki bir $X_{i1}$ ve $X_{i2}$ gözlem çiftinin $\overline{X}_1$ ve $\overline{X}_2$'dan uzaklığı aynı $m_i^2$ değerine sahiptir. Çünkü, $m_i^2$,kare mahalanobis uzaklığı $X_1$ ve $X_2$ değişkenlerini, ortalaması sıfır, standart sapması bir ve korelasyon katsayısı sıfir olan $z_1$ ve $z_2$ gibi iki yeni değişkene dönüştürdükten sonra elde edilen dairenin yarıçapının karesidir. Elips merkezinden $m_i^2$ uzaklığında çizilecek elipslere ise kontur denir. Eğer bir gözlem elips merkezinden uzakta ise $m_i^2$'ler büyük değerler olacaktır (Alpar, 1997). Bu uzaklık çok değişkenli analizlerde sıklıkla kullanılır. Bu uzaklık ölçüsünün bir özelliği değişkenler arasındaki korelasyonu dikkate almasıdır. Yukarıdaki formülden de görüldüğü gibi $h_{ii}$ değerleri bu uzaklıklar kullanılarak hesaplanabilmektedir.




Şekil 9.2 İki boyutta $h_{ii}$'nin sabit konturleri*

Yukarıdaki grafikte $h_{ii}$'nin sabit değerleri için ($h_{ii}$=0.25,0.20,0.15,0.10 ve 0.05) elips şeklindeki dış sınırlar görülmektedir. Böylece örneğin en içteki verilerin $h_{ii}$ değeri=0.05. Yüksek $h_{ii}$ değerlerine sahip gözlemler bir modelin öngörülmesinde potansiyel olarak en

* Şekil Weisberg ,S.,(1947), s.112

etkili gözlemlerdir. Potansiyel etki Y'ye değil X'e bağlıdır. Eğer i.inci gözlem X sapan değeri olarak belirlenmişse ve dolayısıyla büyük $h_{ii}$ değerine sahipse bu gözlem $\hat{Y}_i$ öngörü değerini belirlemede önemli bir role sahiptir.

Yukarıda belirtildiği gibi $\hat{Y}_i$ tahmin değerleri $Y_i$ gözlem değerlerinin doğrusal bir kombinasyonudur ve $h_{ii}$, $Y_i$ gözlem değerinin bu tahmin değerini belirlemedeki tartısı olarak düşünülebilir. $h_{ii}$'nin sadece X değerlerinin bir fonksiyonu olduğu hatırlanırsa $h_{ii}$,X değerlerinin $\hat{Y}_i$ tahmin değerlerini belirlemede ne kadar etkili olduğunu göstermektedir (Neter vd., 1996).

$$\overline{h} = \frac{\sum_{i=1}^{n} h_{ii}}{n} = \frac{p}{n} \tag{9.32}$$

Bir deney gerçekleştiriliyorsa, her bir gözlemin p/n ortalama değerine yakın bir $h_{ii}$ değerine sahip olması arzulanır. Fakat X veri seti hazır olarak ele alınmışsa bir $h_{ii}$ değerinin yeteri kadar büyük olup olmadığına (ortalamadan uzaklığının büyüklüğü) karar verecek bazı kriterlere gereksinim duyulur.

$$\frac{n-p}{p-1}\frac{[h_{ii}-(1/n)]}{(1-h_{ii})} \approx F(1-\alpha; p-1, n-p) \tag{9.33}$$

p>10 ve n-p>50 için %95 F değeri 2'den azdır ve dolayısıyla $2p/n$ (ortalamanın iki katı) kabaca iyi bir kritik değerdir. Diğer yandan bu değer $h_{ii}$ değerinin ortalamasının iki katı olarak düşünülebilir ve $h_{ii}$'nin, ortalaması etrafında sınır değeri olarak addedilebilir. $p/n$>0.4 olduğunda parametre başına birkaç serbestlik derecesi düşmektedir ve bütün gözlemler şüpheli hale gelmektedir. Sonuç olarak $h_{ii}$ değeri $2p/n$'i aştığı zaman i.inci gözlem kaldıraç noktası olarak adlandırılır (Belsley vd., 1980). Yukarıda verilen şekil 48

gözlemden oluşan iki açıklayıcı değişkenli ve sabiti olan bir modelden elde edilmiştir. Dolayısıyla kaldıraç noktası için kritik değer $2p/n = 0.12$'dir. Bu değeri aşan gözlemler daire içine alınmıştır. Bu gözlemler kaldıraç noktası olarak belirlenir.

## 9.8 Etkili Gözlemleri Belirleme

Etkili alt veri setlerinin farklı kaynakları olabilir. Öncelikle kaçınılmaz olan veri setinin yanlış kaydedilmesidir. İkinci olarak veri setinin doğasında gözlemsel hatalar bulunabilir. Üçüncü olarak sapan veri noktaları aykırı gözlemler olarak ortaya çıkabilir. Bu gibi veriler genellikle tahminin etkinliğini etkileyebilecek yararlı bilgiler içerirler. Bu faydalı durumuna rağmen bu aykırı noktaları ayırmak ve hangi parametre tahminlerinin bu veri setinden daha çok etkilendiğine karar vermek yapıcı bir çözümdür.

Sapan ve etkili gözlemler arasında bir ayrım vardır. Sapan gözlemler (Y veya bağımsız değişken setindeki) regresyon denklemini etkileyen etkili gözlemler olmayabilir (Barnett ve Lewis, 1994).

Veri setinde küçük bir alt set tahmin edilen parametreler üzerinde gereğinden fazla etkili olabilir. Bu durumda parametre tahminleri veri setinin çoğunluğundan çok bu veri setine dayalı olur.

Çoklu regresyon yapılmadan önce her bir tesadüfi değişkenin dağılımı incelenebilir ve gözle görülen bir aykırılığın olup olmadığına bakılabilir. Aynı zamanda yukarıda grafiksel yöntemlerde bahsedildiği şekilde serpilme diyagramları gözlemlenebilir. Bu yolla farklı gözlemlerin ayıklanması kolay bir yol gibi görünse de, yöntem bu verilerin tahmin edilen modeli nasıl etkilediğini göstermez.

Çoklu regresyonun kurulmasından sonra kalıntılar öngörü değerleri ve açıklayıcı değişkenlere dayalı bir çok test süreci gerçekleştirilir. Bu gibi metotlarla çok şey öğrenilebilmesine rağmen yine de bu metotlar veri setindeki bir alt setin düzeltilmesi veya bir yana bırakılması durumunda tahmin edilen modelin ne olacağı konusunda başarılı olamamaktadır. Farklı

olduğu belirlenen alt veri setlerinin silinmesinin tahmin edilen katsayıları, standart hataları, ve test istatistiklerini nasıl etkilediğini belirleyebilmek için bahsedilen zayıflıkları ortadan kaldıran bir takım teşhis tekniklerine değinilecektir.

Etkili veri noktalarının teşhisi amacıyla öncelikle satır silme tekniği ve ilk önce tek satırın (bir gözlem) bir defeda silinmesi açıklanacaktır. Ardından birden fazla satır silinmesine değinilecektir. Tek bir satırın silinmesinin *tahmin edilen katsayılar, öngörü değerleri, kalıntılar*, ve *katsayıların tahmin edilen varyans kovaryans yapısı* üzerindeki etkileri ele alınacaktır. Tahmin sürecinin bu dört çıktısı çoklu regresyonda en yaygın kullanılan ve teşhis araçları için temel çekirdeği oluşturan çıktılardır.

## 9.9 Tekil Satır Etkileri

Bu teknikler etkili gözlemlerin keşfedilebilmesi için geliştirilmiştir. Tek satır X veri matrisindeki bir satır ve karşılık gelen Y elemanını ifade etmektedir. "Etkili bir gözlem" tek başına veya diğer birkaç gözlemle birlikte farklı tahminler (katsayılar,standart hatalar, t değerleri v.b) üzerinde diğer gözlemlerden daha çok, ispatlanabilir etkiye sahip gözlemdir. Bu gibi bir etkinin incelenebilmesi için bir araç her bir defa bir satırın silinmesi ve ardından elde edilen çeşitli hesaplanan değerlerdeki değişimin elde edilmesidir. Silinmesinin çeşitli hesaplanan değerler üzerinde kısmen büyük etkilere sahip olduğu gözlemler etkili gözlemler olarak adlandırılır. Burada tek bir satırın silinmesinin tahmin edilen katsayılar ve öngörü değerleri üzerinde, kalıntılar üzerinde, tahmin edilen parametrelerin varyans kovaryans matrisi üzerindeki etkileri aşama aşama incelenecektir.

### 9.9.1 Katsayılar Üzerindeki Etki

i.inci satırın silinmesinden sonra elde edilen parametre tahmini için $b_{(i)}$ notasyonu kullanılacaktır. Bu değişim aşağıdaki formülle kolaylıkla hesaplanabilir:

$$DFBETA_i = b - b_{(i)} = \frac{(X'X)^{-1}x_i'e_i}{1-h_{ii}} \tag{9.34}$$

DF harfleri n gözleme dayalı elde edilen katsayı ile i.inci gözlemin dışlanmasından sonra elde edilen katsayı tahmini arasındaki farkı belirtmek için kullanılır. i.inci gözlemin her bir regresyon katsayısı $b_k$ $(k = 0,1,...,p-1)$, üzerindeki etkisinin ölçümü, bütün n gözleme dayalı tahmin edilen regresyon katsayısı $b_k$ ile i.inci gözlemin dışlanmasından sonra tahmin edilen $b_{k(i)}$ arasındaki farka dayanır. Bu fark $b_k$'nın standart sapmasına bölündüğünde $DFBETAS$ ölçüsü elde edilir.

$$(DFBETAS)_{k(i)} = \frac{b_k - b_{k(i)}}{\sqrt{MSE_{(i)}c_{kk}}} \qquad k = 0,1,...,p-1 \tag{9.35}$$

$c_{kk}$, $(X'X)^{-1}$'in k.ıncı köşegen elemanıdır. Regresyon katsayılarının varyans kovaryans matrisinin $\sigma^2\{b\} = \sigma^2(X'X)^{-1}$ olduğu hatırlanırsa $b_k$'nın varyansı:

$$\sigma^2\{b_k\} = \sigma^2 c_{kk} \tag{9.36}$$

şeklinde hesaplanmaktadır. $\sigma^2$ hata varyansı burada $MSE_{(i)}$ ile tahmin edilir. bu ortalama hata karesi i.inci gözlem dışlandıktan sonra öngörülen regresyon modelinden hesaplanır.

$$MSE = \frac{e'e}{n-p} = \frac{(Y-Xb)'(Y-Xb)}{n-p} \tag{9.37}$$

olduğu hatırlanırsa bu durumda $MSE_{(i)}$ aşağıdaki gibi hesaplanır:

$$MSE_{(i)} = \frac{(Y_{(i)} - X_{(i)}b_{(i)})'(Y_{(i)} - X_{(i)}b_{(i)})}{n-p-1} \tag{9.38}$$

$MSE_{(i)}$ için basit bir formül aşağıdaki gibi elde edilir:

$$(n-p-1)MSE_{(i)} = (n-p)MSE - \frac{e_i^{\,2}}{1-h_{ii}} \tag{9.39}$$

buradan

$$DFBETAS_{(i)} = \frac{\sqrt{n}.e_i}{(n-1)MSE_{(i)}} \tag{9.40}$$

elde edilir. Bu formülden görülebileceği gibi $\sqrt{n}$ ile orantılı olarak azalır. $DFBETAS$ ölçüsünün işareti bir gözlemin silinmesinin tahmin edilen regresyon katsayılarında artış veya azalışın bir göstergesidir. Bu ölçünün mutlak değer olarak büyüklüğü $\left|DFBETAS_{k(i)}\right|$ regresyon katsayılarının tahmin edilen standart hatalarına göre değişiminin büyüklüğünü gösterir. $\left|DFBETAS_{k(i)}\right|$'nin büyük olması i.inci gözlemin k.ıncı regresyon katsayısı üzerinde büyük bir etkisi olduğunun göstergesidir. Küçük ve orta büyüklükteki veri setleri için 1 değerini ve büyük veri setleri için $2/\sqrt{n}$ değerini aşan $\left|DFBETAS_{k(i)}\right|$ değeri etkili gözlem olarak belirlenir.

### 9.9.2 Öngörü Değerleri Üzerindeki Etki

i.inci gözlemin silinmesinin $\hat{Y}_i$ öngörü değeri üzerindeki etkisi:

$$(DFFITS)_i = \frac{\hat{Y}_i - \hat{Y}_{i(i)}}{\sqrt{MSE_{(i)} h_{ii}}} \tag{9.41}$$

Bu ölçü n gözleme dayalı i.inci gözlem için tahmin $\hat{Y}_i$ değeri ile i.inci gözlemin regresyon fonksiyonu tahmin edilmeden önce dışlanmasıyla elde edilen i.inci gözlem için $\hat{Y}_{i(i)}$ tahmin değeri arasındaki farka dayanır. Paydada $\hat{Y}_i$'nin tahmin edilen standart sapması bulunmaktadır. Fakat bu tahmin $\sigma^2$'nin tahmincisi olarak i.inci gözlemin regresyon fonksiyonunun tahmininde kullanılmadan hesaplanan ortalama hata karesini ($MSE_{(i)}$)kullanmaktadır. Payda ifadenin standardizasyonunu sağlamaktadır dolayısıyla i.inci gözlem için $(DFFITS)_i, \hat{Y}_i$'in tahmin edilen standart sapmalarının sayısını verir. Bu da i.inci gözlemin dışlanmasıyla $\hat{Y}_i$'nin artış veya azalışını gösterir.

$(DFFITS)_i$, bütün veri seti kullanılarak hesaplanabilir:

$$(DFFITS)_i = e_i \left[ \frac{n-p-1}{SSE(1-h_{ii}) - e_i^2} \right]^{1/2} \left( \frac{h_{ii}}{1-h_{ii}} \right)^{1/2} = t_i \left( \frac{h_{ii}}{1-h_{ii}} \right)^{1/2} \tag{9.42}$$

i.inci gözlem için $(DFFITS)_i$ değeri studentized öngörü kalıntısıdır. Bu değer bu gözlemin kaldıraç değerine bağlı olarak artar veya azalır. i bir X sapan değeriyse ve yüksek kaldıraç değerine sahipse bu gözlem için $(DFFITS)_i$ değeri büyük olacaktır.

Bir gözlemin etkili olup olmadığına karar verebilmek için $|(DFFITS)_i|$ kritik değeri küçük ve orta büyüklükteki veri setleri için 1 ve büyük veri setleri için $2\sqrt{p/n}$'dir. $|(DFFITS)_i|$'nin bu değerleri aşması durumunda söz konusu gözlem etkili bir gözlem olarak belirlenebilir (Belsley vd., 1980).

### 9.9.3 Cook Mesafesi

$(DFFITS)_i$ i.inci gözlemin $\hat{Y}_i$ üzerindeki etkisinin ölçüsüdür. Cook mesafesi i.inci gözlemin bütün n öngörü değerleri üzerindeki etkisini ölçmektedir. Cook mesafesi $D_i$,i.inci gözlemin n öngörü değeri üzerindeki toplu etkisinin ölçüsüdür:

$$D_i = \frac{\sum_{j=1}^{n} (\hat{Y}_j - \hat{Y}_{j(i)})^2}{p.MSE} \tag{9.43}$$

n adet $\hat{Y}_j$ öngörü değerlerinin, karşılık gelen i.inci gözlemin silinmesinden sonra regresyon modelinden elde edilen $\hat{Y}_{j(i)}$ değerleriyle farklarının karelerin toplamı alınır ve ardından toplanılır. Dolayısıyla i.inci gözlemin etkisi etkilerinin işaretine bakılmaksızın ölçülür. Cook mesafesi aşağıdaki gibi ifade edilebilir:

$$D_i = \frac{(\hat{Y} - \hat{Y}_{(i)})'(\hat{Y} - \hat{Y}_{(i)})}{p.MSE} \tag{9.44}$$

Cook mesafesini yorumlamak için $D_i$, $F(p,n-p)$'a karşılık gelen yüzdelik dilimle karşılaştırılır. Eğer yüzdelik dilim yüzde 10 veya 20'den az ise i.inci gözlemin öngörü değerleri üzerinde küçük bir etkisi olduğu söylenilir. Eğer bu yüzdelik dilim yüzde 50 ya da daha fazla ise i.inci gözlemle ve i.inci gözlem olmadan elde edilen öngörü değerleri arasında önemli bir fark vardır. Başka bir ifadeyle %50 ya da daha fazla olması, i.inci gözlemin regresyon fonksiyonunun öngörülmesinde önemli etkiye sahip olduğu anlamına gelmektedir.

Cook mesafesi her bir defa bir gözlemin silinmesine gerek kalmaksızın hesaplanabilir:

$$D_i = \frac{e_i^2}{p.MSE}\left[\frac{h_{ii}}{(1-h_{ii})^2}\right] = \frac{t_i^2 h_{ii}}{p(1-h_{ii})} \tag{9.45}$$

Bu formülden görüleceği gibi $D_i$ mesafesi iki faktöre bağlıdır: a) $e_i$ kalıntıların büyüklüğü b) $h_{ii}$ kaldıraç değeri. $e_i$ ve $h_{ii}$ ne kadar büyük olursa $D_i$ mesafesi de o kadar büyük olur. dolayısıyla i.inci gözlemin etkili bir gözlem olabilmesi için 1)büyük bir $e_i$ ve orta büyüklükte $h_{ii}$ değerine sahip veya 2) orta büyüklükte bir $e_i$ kalıntıyla yüksek $h_{ii}$ değerine sahip veya 3) yüksek $e_i$ ve $h_{ii}$ değerlerine sahip olması gereklidir.

Formülün diğer yanı $D_i$ ve studentized öngörü kalıntıları $t_i$ arasındaki ilişkiyi vurgulamaktadır.

Grafiksel süreçlerde $D_i$'nin karekökünün kullanılmasının bazı avantajları bulunmaktadır. $h_{ii}/(1-h_{ii})$ ile tartılandırılmış standart kalıntıları elde edilir. Atkinson düzenlenmiş Cook istatistiğinin grafiksel kullanımını önermektedir:

$$C_i = \left[\frac{n-p}{p}.\frac{h_{ii}}{1-h_{ii}}\right]^{1/2} |t_i| \tag{9.46}$$

Mutlak değer ve n ve p faktöründen başka düzeltilmiş Cook mesafesi Belsley'in $(DFFITS)_i$ olarak adlandırdığı sonuçla aynıdır. $(DFFITS)_i$'nin kullanımı $C_i$'nin kullanımına eşdeğerdir. Çünkü her ikisinde de hata varyansı $\sigma^2$'nin tahmininde sapan gözlemlere $D_i$'den daha çok tartı vermektedir. $C_i$'nin ya da $D_i$'nin kullanılmasının avantajı Y ve X'in belirleyemediği kaldıraç değerlerine dikkat çekebilmesidir (Atkinson, 1997).

Cook mesafesi bir gözlemin n adet öngörü değeri üzerindeki toplam etkisinin bir ölçüsüdür ve aynı zamanda cebirsel olarak eşdeğer bir biçimde, bir gözlemin p tane regresyon katsayısı üzerindeki toplam etkisinin de bir ölçüsüdür. Cook uzaklık ölçüsü aslında bütün p adet

regresyon katsayısı $\beta_k (k = 0,1,..., p-1)$ için oluşturulan güven bölgesi kavramından elde edilmiştir. Klasik normal doğrusal çoklu regresyon modeli için bu birleşik güven bölgesinin sınırları aşağıdaki gibidir:

$$(b-\beta)'X'X(b-\beta) \leq p.MSE.F(\alpha; p, n-p) \tag{9.47}$$

Bu formül $\beta$ parametre vektörü için $100(1-\alpha)\%$'lık güven bölgesini vermektedir. $MSE$ n-p serbestlik derecesiyle $\sigma^2$'nin bir tahmincisidir ve F, F dağılımının p ve n-p serbestlik derecesiyle $100\alpha\%$ noktasına karşılık gelen değerdir. $\alpha$'nın farklı değerlerinde bu güven bölgeleri $\beta$'nın en küçük kareler tahmincisi $b$ merkezli bir dizi elipsoidlerdir. Bu bölgeler $b$'nin $b_{(i)}$ yakınında olup olmadığına dair bilgi sağlar. İki açıklayıcı değişkenli bir model için n tane $b_{(i)}$'nin yukarıdaki formülle seçilen güven bölgelerine karşı diyagramları oluşturularak grafiksel bir gösterimle etkili değerler belirlenebilir. İki açıklayıcı değişkenden daha fazla değişken varsa bu grafiksel gösterim anlaşılır olmayacaktır. Cook (1977) etkili gözlemleri belirleyebilmek için bütün parametreleri içeren aşağıdaki istatistiği önermiştir:

$$D_i = \frac{(b_{(i)} - b)'X'X(b_{(i)} - b)}{p.MSE} \tag{9.48}$$

$D_i$'nin büyük değerleri modeldeki bütün parametreler ile ilgili birleşik hipotez testleri ve çıkarsamalarda etkili olan gözlemleri gösterir (Atkinson, 1997).

Chatterjee ve Hadi'ye göre Cook mesafesi 4/n-p değerini aşan gözlemler etkili değerlerdir (Fox, 1991). Cook mesafesi >1 olan gözlemler önemli etkili değerlerdir.

### 9.9.4 Bir Gözlemin Katsayıların Kovaryans Matrisine Etkisi

Regresyon teşhisinde diğer bir görüş tahmin edilen katsayıların kovaryans matrisinin incelenmesidir. Burada bu defa bütün veri setini kullanarak elde edilen kovaryans matrisiyle

$\sigma^2(X^TX)^{-1}$,i.inci satır silindikten sonra bulunan kovaryans matrisi $\sigma^2\left[(X_{(i)}{}'X_{(i)})\right]^{-1}$karşılaştırılmaktadır. Bu gibi iki pozitif sonlu simetrik matrisleri karşılaştırmak için çeşitli alternatif araçlardan en basiti determinantların oranıdır: $\det\left[X'_{(i)}X_{(i)}\right]^{-1}/\det\left[X'X\right]^{-1}$.

Bu iki matrisin, sadece kareler toplamında ve çapraz çarpımlarda i.inci satırın silinmesiyle birbirinden farklılaşmasından sonra hesaplanan determinantlarının oranının bire yakın olması iki matrisin birbirine yakın olduğunu, veya kovaryans matrisinin i.inci satırın silinmesinden etkilenmediğini gösterir. Buradaki analiz sadece X matrisinden gelen bilgiye bağlıdır ve $\sigma^2$'nin tahmincisi $MSE$'nin i.inci gözlemin silinmesinden etkilenip etkilenmediğini dikkate almamaktadır.

İki matris $MSE(X^TX)^{-1}$ve' $MSE\left[(X_{(i)}{}'X_{(i)})\right]^{-1}$in determinatlarının oranı karşılaştırılırken Y veri setini gözönünde bulundurabilir.

$$COVRATIO=\frac{\det\left\{MSE_{(i)}\left[X'_{(i)}X_{(i)})\right]^{-1}\right\}}{\det\left[MSE(X^TX)^{-1}\right]} \tag{9.49}$$

$$=\frac{MSE^{2p}{}_{(i)}}{MSE^{2p}}\left\{\frac{\det\left[X'_{(i)}X_{(i)}\right]^{-1}}{\det(X^TX)^{-1}}\right\} \tag{9.50}$$

Bu formülde parantez içindeki ifade $\det\left[X'_{(i)}X_{(i)}\right]=(1-h_{ii})\det(X'X)$ eşitliğinden yararlanılarak yeniden düzenlenilirse,

$$\frac{\det\left[X^T(i)X(i)\right]^{-1}}{\det(X^TX)^{-1}}=\frac{1}{1-h_i} \tag{9.51}$$

olduğu görülür. Böylece,

$$\text{COVRATIO} = \frac{1}{\left[\frac{n-p-1}{n-p} + \frac{t_i^2}{n-p}\right](1-h_{ii})} \tag{9.52}$$

elde edilir.

Bir teşhis aracı olarak bu aşamadan sonra bire yakın olmayan COVRATIO değerleri veren gözlemler muhtemelen etkili gözlemlerdir.

Bu gibi değişme oranının 1'den farklılaşmasının büyüklüğüne ilişkin yol gösterici bir çözüm iki aykırı durumun göz önünde bulundurulmasını gerektirir. Birincisi $|t_i| \geq 2$ ve $h_{ii}$ değeri minimum seviyesi $1/n$'e eşit olduğu durum ve ikincisi $t_i = 0$ ve $h_{ii} \geq 2p/n$. Birinci durumda,

$$COVRATIO \approx \frac{1}{\left(1 + \frac{e_i^{*2} - 1}{n-p}\right)} \leq \frac{1}{\left(1 + \frac{3}{n-p}\right)^p} \tag{9.53}$$

burada,

$$\frac{1}{\left(1 + \frac{3}{n-p}\right)^p} \approx \left(1 + \frac{3p}{n}\right)^{-1} \approx 1 - \frac{3p}{n} \tag{9.54}$$

'e yaklaşır. n-p yerine basitleştirme amacıyla n konulmuştur. Son söylenilen sınırlar n≤ 3p olduğunda kullanılamaz. İkinci durum için,

$$COVRATIO \approx \frac{1}{\left(1-\frac{1}{n-p}\right)^{p}} \frac{1}{(1-h_i)} \geq \frac{1}{\left(1-\frac{1}{n-p}\right)^{p}\left(1-\frac{2p}{n}\right)} \tag{9.55}$$

buradan kabaca fakat daha basit sınırlar aşağıdaki gibi elde edilir:

$$\frac{1}{\left(1-\frac{1}{n-p}\right)^{p}\left(1-\frac{2p}{n}\right)} \approx \frac{1}{\left(1-\frac{p}{n-p}\right)\left(1-\frac{2p}{n}\right)} \approx \left(1-\frac{3p}{n}\right)^{-1} \approx 1+\frac{3p}{n} \tag{9.56}$$

Dolayısıyla 3p/n'e yakın veya büyük $|COVRATIO-1|$ noktaları araştırılmaktadır.

Kabaca söylenen, $COVRATIO$ değerinin $h_{ii}$ büyük olduğunda büyük ve studentized öngörü kalıntıları büyük olduğunda küçük sonuç vereceğidir.

Aynı zamanda bir gözlem silindiğinde $\hat{Y}_i$'nin varyansının nasıl değiştiği araştırılabilir. Bunun için,

$$\begin{aligned} &\text{var}\left\{\hat{Y}_i\right\} = MSE.h_{ii} \\ &\text{var}\left\{\hat{Y}_{i(i)}\right\} = \text{var}\left\{X_{(i)}b_{(i)}\right\} = MSE_{(i)}\left[\frac{h_{ii}}{1-h_{ii}}\right] \end{aligned} \tag{9.57}$$

buradan

$$FVRATIO = \frac{MSE_{(i)}}{MSE(1-h_{ii})} \tag{9.58}$$

elde edilir. Bu teşhis ölçüsü için $COVRATIO$ ölçüsüyle benzer yorumlar yapılabilir (Belsley vd., 1980).

Weisberg testi kullanılırsa bu test y sapan değerlerini tarayacak ve H elemanları veya Mahalanobis uzaklıkları açıklayıcı değişkenler üzerindeki sapan değerleri tanımlayacaktır. Bu gibi sapan değerler ister istemez etkili veri noktaları olmayabilir. Hangi sapan değerin etkili değer olduğuna karar verebilmek için 1'den büyük Cook mesafesine sahip olan verileri bulmamız gerekir. Bu noktalar Cook uzaklığıyla etkili veri olarak adlandırılır ve bunların analizden atılıp atılmayacağına dair karar verebilmesi için dikkatli bir çalışma gerekmektedir. Bunun için o veriler ilişkin hesaplanan DFBETA,DFFIT, COVRATIO gibi teşhis süreçlerinin incelenmesi gerekmektedir (Stevens, 2002).

## 9.10 Çoklu Satır Etkileri

Tekil bir satırın silinmesine dayalı etkili gözlemleri belirleyen diagnostik teknikler çoğu zaman etkili gözlemleri bulmada başarılı olabilir. Fakat bazı durumlarda bu teknikler yetersiz kalabilir. Veri setinde bir sapan değerin etkisi diğer bir sapan değer tarafından maskelenebilir. Bu gibi durumlarda tekil satır silinmesine dayalı teknikler bu gibi sapan değerleri belirleyemez. Dolayısıyla çoklu sapan değerleri belirleyebilmek amacıyla bir takım diagnostik ölçüler geliştirilmiştir.

Öncelikle etkili kabul edilen gözlemlerden oluşan bir küme belirlenmelidir. Bu küme $B*$ ile ifade edilecektir. Tekil satır ölçülerinden elde edilen etkili verilerin dışında da etkili veriler bulunabileceğinden bu kümenin tam anlamıyla belirlenmesinde başarılı olunmayabilir. Bu nedenle tekil satır silinmesinden elde edilen etkili verilerle birlikte kısmi regresyon diyagramlarından belirlenen etkili veriler bu kümenin elemanları olarak seçilebilir.

$B*$ içindeki m büyüklüğündeki alt kümelerden herhangi birinin silinmesinin ardından öngörüdeki değişimi ifade eden çoklu diagnostik ölçü aşağıdaki gibi elde edilir:

$$MDFFIT \equiv [b - b(D_m)]^T X^T(D_m) X(D_m) [b - b(D_m)] \tag{9.59}$$

$b(D_m)$ ifadesi katsayılar matrisinin m satırın silinmesinden sonra hesaplandığını

göstermektedir. Benzer şekilde $X(D_m)$ ifadesi m satırın silinmesinden sonra kalan X matrisini ifade etmektedir (Belsley vd., 1980).

Tekil satır silinmesine dayalı COVRATIO ölçüsü çoklu satırların silinmesini yansıtacak şekilde genelleştirilebilir:

$$COVRATIO(D_m) \equiv \frac{\det s^2(D_m)\left[X^T(D_m)X(D_m)\right]^{-1}}{\det s^2(X^T X)} \tag{9.60}$$

Yukarıda bahsedilen diagnostik ölçüler istatistik paket programları tarafından hesaplanamamaktadır. Dolayısıyla büyük veri setleri için bu ölçülerin hesaplanması oldukça maliyetli ve karmaşık olmaktadır.

## 9.2 Sapan Değerlere Uygulanacak Yaklaşımlar

Literatürde sapan değerlere ne yapılması gerektiği hususunda üç yaklaşım gözükmektedir: gözlemi dışlamak, gözlemi dışlamamak fakat kontrol altında tutmak, bu gözlemlerin ne ifade ettiğine bakmak ve modeli değiştirmek olarak ele alınmaktadır.

### 1. Gözlemi Dışlamak

Modelde değişkenlerde görünen ve diğer gözlemlerden ayrılık gösteren veriler ölçme veya kayıt hatasından oluşmuş ise dışlamak gerekmektedir.

### 2. Gözlemi Dışlamamak Fakat Kontrol Altında Tutmak

İkinci bir yaklaşım ise sapan değeri tahmin edilen model içinde bırakmak fakat etkilerini azaltmaktır. Bunun için varsayımlardan etkilenmeyen (robust) tahmin ediciler (M tahmin ediciler, $L_p$ norm tahmin ediciler, sınırlandırılmış etki (bounded influence) tahmin ediciler) EKKY tahmin edicileri kadar sapan değerlerden etkilenmezler.

### 3. Modeli Değiştirmek

Çoğunlukla sapan değerler modelin spesifikasyonunun bozuk olduğunun göstergesidirler. Sapan değerlerin olduğu modellerde genelde model iki kısma ayrılıp öyle ele alınmalı veya birleştirme (pooling) uygulanmalıdır.

Etkili gözlemler ve sapan değerler bazen önemli bilgi ihtiva ederler. Sapan değerler önemli bir değişkenin eksikliği veya modelin kuruluşunda önemli eksikliklerin olduğunu gösterebilir (Büyüklü, 1997).

# 10. UYGULAMA

## ÜLKELERİN TARIMSAL FAALİYETLERİNİN MODELLENMESİNDE SAPAN VE ETKİLİ GÖZLEMLERİN BELİRLENMESİ

### 10.1 Model: Klasik Doğrusal Regresyon Modeli

$$KTKDGR_i : b_0 + b_1 MKN_i + b_2 İTH_i + b_3 GÜB_i + e_i \qquad (10.1)$$

### 10.2 Değişken Tanımları

**TKDGR:** (cari $), Uluslar arası Standart Endüstriyel Sınıflandırma (International Standard Industrial Classification-ISIC) tanımlamasına göre tarımsal faaliyetler: ormancılık, hububat yetiştirme, hayvancılık,balıkçılık, avcılık faaliyetlerinden oluşmaktadır. Katma değer bir sektörün sahip olduğu bütün üretimden girdilerin çıkartılmasından sonra elde edilir. (Kaynak: Dünya Bankası Ulusal Hesaplar Verileri ve OECD Ulusal Hesaplar Veri Dosyaları)

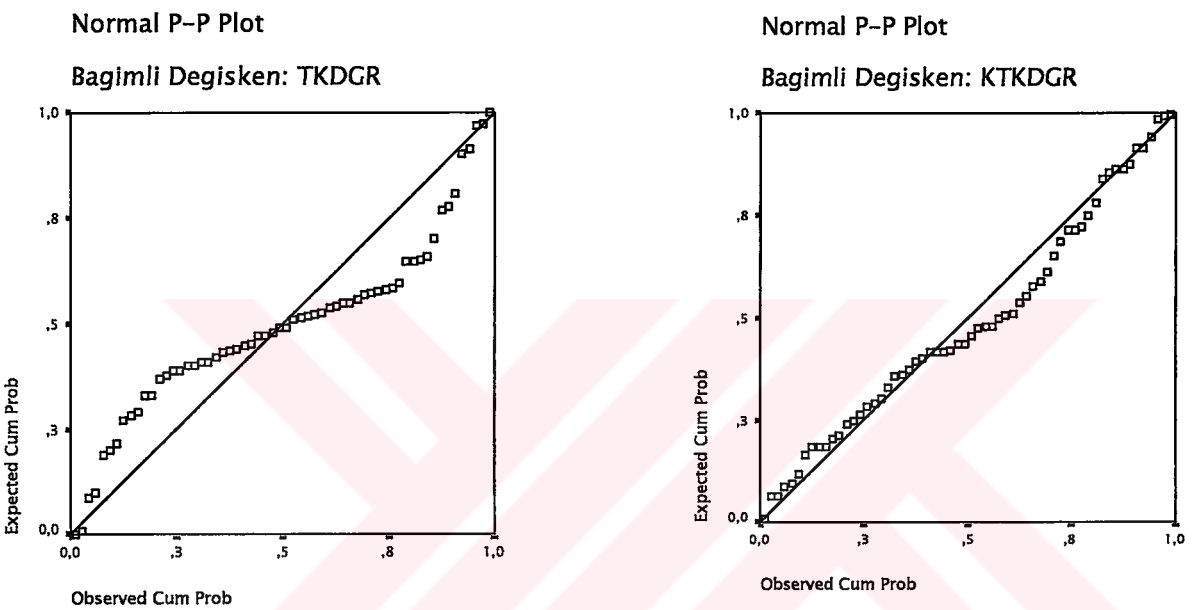
**MKN:** (adet) , Tarımda kullanılan tarımsal makinelerin, paletli ve tekerlekli traktörlerin sayısıdır. Yıl sonundan sonraki yılın ilk çeyreğine kadar tarımda kullanılan tarımsal makine ve traktörlerin sayımı sonucu elde edilmiştir. (Kaynak: Yemek ve Tarım Organizasyonu, Üretim Yıllığı ve Veri Dosyaları)

**İTH:** (Mal İthalatının %'si) : Tarımsal hammadde ithalatının toplam mal ithalatı içersindeki yüzdesini ifade etmektedir. Bu hammaddeler: yakıt dışındaki ham maddeler ; kömür, petrol ve kıymetli taşlar haricindeki ham gübre ve maden cevherleri ve maden kırıntıları. (Kaynak: Dünya Bankası)

**GÜB:** (ton) Kilometre kare başına gübre tüketimi (Kaynak:Dünya Bankası)

## 10.3 Regresyon Modeli:

EKK uygulamamızda bağımlı değişken tkdgr olarak ele alındığında kalıntıların normal dağılmadığı normal olasılık diyagramından görünmektedir. Dolayısıyla bağımlı değişkene uygun olabilecek dönüşümler denenmiş ve karekök dönüşümüne karar verilmiştir. Dönüşümden sonra bu varsayımdan sapma düzeltilmiş ve kalıntıların normal dağılması sağlanmıştır. Bu diyagramlar aşağıdaki gibidir:




Dönüşüm Öncesi Dönüşüm Sonrası

Şekil 10.1 Normallik dönüşümü

Bu dönüşüm uygulandıktan sonra elde edilen regresyon modeli aşağıdaki gibidir:

$$ktkdgr = 26041{,}92 + 0{,}064677mkn + 9083{,}725ith + 0{,}011075güb + e \quad (10.2)$$

$$st.hata: (6277{,}27) \quad (0{,}010) \quad (2742{,}74) \quad (0{,}002) \quad (10.3)$$

$$t \quad : \quad 4{,}153 \quad 6{,}813 \quad 3{,}305 \quad 6{,}130 \quad (10.4)$$

$$R^2 : 0{,}813 \quad (10.5)$$

## 10.4 Kalıntılar Ve Kaldıraç Değerleri

Çizelge 10.1 Kalıntılar ve kaldıraç değerleri

| | | Standart kalıntı | St.öngörü kalıntısı | studnt k. | hii |
|---|---|---|---|---|---|
| 1 | Arnavutluk | -0,11209 | -,11279 | -0,1138 | 0,002902 |
| 2 | Avustralya | 1,1007 | 1,12481 | 1,12215 | 0,015 |
| 3 | Azerbeycan | -0,44501 | -,44672 | -0,44995 | 0,00776 |
| 4 | Bangladeş | 0,12968 | ,14732 | 0,14862 | 0,162 |
| 5 | Barbados | -1,3234 | -1,34923 | -1,33945 | 0,021 |
| 6 | Belarus | -0,90165 | -,90988 | -0,91128 | 0,018 |
| 7 | Bolivya | -0,32392 | -,32482 | -0,32744 | 0,011 |
| 8 | Botsvana | -0,83323 | -,84415 | -0,84633 | 0,002666 |
| 9 | Brezilya | 0,01972 | ,02108 | 0,02127 | 0,123 |
| 10 | Bulgaristan | -0,05292 | -,05305 | -0,05353 | 0,006575 |
| 11 | Şili | 0,98673 | ,99899 | 0,99901 | 0,004293 |
| 12 | Cote d'Ivoire | 0,5659 | ,57026 | 0,57373 | 0,004124 |
| 13 | Hırvatistan | -0,20128 | -,20174 | -0,20349 | 0,009306 |
| 14 | Çek Cumhuriyeti | -0,21758 | -,21772 | -0,2196 | 0,011 |
| 15 | Dominik Cum. | -1,35422 | -1,37992 | -1,36891 | 0,014 |
| 16 | Ekvator | -0,21265 | -,21316 | -0,215 | 0,018 |
| 17 | Mısır Arap Emirliği. | 1,35467 | 1,43148 | 1,41825 | 0,072 |
| 18 | El Salvador Cum. | -0,51361 | -,51715 | -0,52057 | 0,025 |
| 19 | Estonya | -1,52075 | -1,56916 | -1,54907 | 0,036 |
| 20 | Habeşistan | 0,58969 | ,59447 | 0,59793 | 0,00397 |
| 21 | Gürcistan | -0,25014 | -,25280 | -0,25494 | 0,001001 |
| 22 | Almanya | 0,28547 | ,29496 | 0,29739 | 0,074 |
| 23 | Gana | -0,00207 | -,00208 | -0,00209 | 0,015 |
| 24 | Gine | -0,36335 | -,36530 | -0,36816 | 0,004957 |
| 25 | Macaristan | -0,16292 | -,16332 | -0,16476 | 0,004279 |
| 26 | Hindistan | -0,89797 | -2,15905 | -2,09175 | 0,813 |
| 27 | Endonezya | 1,08637 | 1,32707 | 1,31815 | 0,204 |

Çizelge 10.1'in Devamı

| | | Standart kalıntı | St.öngörü kalıntısı | studnt k. | hii |
|---|---|---|---|---|---|
| 28 | İrlanda | 0,19737 | ,19854 | 0,20027 | 0,002679 |
| 29 | İtalya | -0,70471 | -,81520 | -0,81765 | 0,241 |
| 30 | Jamaika | -0,63892 | -,64280 | -0,64619 | 0,008383 |
| 31 | Japonya | 1,57855 | 2,09774 | 2,03681 | 0,394 |
| 32 | Kazakistan | 0,22661 | ,22797 | 0,22993 | 0,002743 |
| 33 | Kenya | -0,21071 | -,21128 | -0,2131 | 0,019 |
| 34 | Kore | 2,34365 | 2,49112 | 2,38283 | 0,032 |
| 35 | Malavi | -0,67958 | -,68366 | -0,68693 | 0,012 |
| 36 | Malezya | 1,14408 | 1,16287 | 1,15923 | 0,008199 |
| 37 | Mauritius | -1,19296 | -1,21350 | -1,20841 | 0,024 |
| 38 | Meksika | 2,72465 | 2,94447 | 2,76143 | 0,01 |
| 39 | Mozambik | -0,35314 | -,35461 | -0,35741 | 0,006467 |
| 40 | Namibia | -0,54923 | -,55594 | -0,5594 | 0,001375 |
| 41 | Nikaragua | -0,16523 | -,16699 | -0,16846 | 0,009144 |
| 42 | Norveç | 0,01956 | ,01957 | 0,01975 | 0,014 |
| 43 | Pakistan | -0,06482 | -,06674 | -0,06734 | 0,062 |
| 44 | Paraguay | 0,09576 | ,09636 | 0,09722 | 0,002682 |
| 45 | Polonya | -2,51553 | -2,89859 | -2,72414 | 0,146 |
| 46 | Portekiz | -0,05289 | -,05297 | -0,05345 | 0,019 |
| 47 | Romanya | 0,76966 | ,77516 | 0,77794 | 0,005743 |
| 48 | Suudi Arabistan | 2,14087 | 2,24919 | 2,17186 | 0,003632 |
| 49 | Seychelles | -0,97684 | -,99616 | -0,99623 | 0,009798 |
| 50 | İspanya | 0,48124 | ,49249 | 0,49585 | 0,054 |
| 51 | St. Lucia | -1,52652 | -1,56518 | -1,5453 | 0,021 |
| 52 | Sudan | 1,35405 | 1,38414 | 1,37296 | 0,003773 |
| 53 | Tanzania | 0,56097 | ,56413 | 0,5676 | 0,02 |
| 54 | Trinidad ve Tobago | -0,80084 | -,81179 | -0,81427 | 0,002119 |
| 55 | Tunisia | -0,27161 | -,27370 | -0,27599 | 0,031 |
| 56 | Türkiye | -0,5795 | -,60664 | -0,6101 | 0,085 |

Çizelge 10.1'in Devamı

| | | Standart kalıntı | St.öngörü kalıntısı | studnt k. | hii |
|---|---|---|---|---|---|
| 57 | Ukrayna | 0,3834 | ,38507 | 0,38803 | 0,01 |
| 58 | Britanya | 0,67378 | ,68028 | 0,68356 | 0,017 |
| 59 | Uruguay | -0,90574 | -,92387 | -0,92508 | 0,04 |
| 60 | Venezuella | 1,06438 | 1,07805 | 1,07649 | 0,005648 |

Normal koşullar altında studentize kalıntıların ($\alpha = 0,05$ için) %5'i $|r_i| \leq 2$ aralığının dışına düşerler. Dolayısıyla studentize kalıntıların $\pm 2$ aralığının dışındaki değerler göze çarpan noktalar olarak ele alınabilir. Bu durumda 26 Hindistan, 31 Japonya, 34 Kore, 38 Meksika, 45 Polonya ve 48 Suudi Arabistan ülkelerinin diğerlerinden farklılık gösterdiği görülmektedir.

Studentized öngörü kalıntılarına bakıldığında 26 Hindistan, 31 Japonya, 34 Kore, 38 Meksika, 45 Polonya ve 48 Suudi Arabistan'a ait studentized öngörü kalıntılarının n-p-1= 60-4-1=55 serbestlik dereceli t dağılımı tablo değeri (2.01)'nden büyük oldukları dolayısıyla sapan değer olarak belirlendikleri görülmektedir. Dikkat edilmesi gereken diğer bir husus yukarıda bahsedilen diğer içsel kalıntıların kullanılmasının sakıncalarına rağmen studentized öngörü kalıntılarının studentized kalıntılarla aynı ülkeleri sapan değer olarak belirlemesidir. Aradaki fark studentized öngörü kalıntılarının daha yüksek olmasıdır. Bu durum bu uygulama için geçerli olabilir. Fakat farklı uygulamalarda bu bahsedilen farklı kalıntı hesaplamaları farklı sonuçlara işaret edebilir.

Standart kalıntılarla studentized kalıntılar iki ülke dışında benzer sonuçlar vermişlerdir.

$$e_i^* = \frac{e_i}{\sqrt{MSE}}$$ standart kalıntılar

$$r_i = \frac{e_i}{\sqrt{MSE(1-h_{ii})}}$$ studentized kalıntılar

Sapan gözlemlerin kalıntı değerleri $e_i$'ler, sapan değerin regresyon doğrusunu kendisine

doğru çekmesinden dolayı olduğundan küçük görülecektir. Sapan değer varyansı büyüttüğünden, $\sqrt{MSE} > \sqrt{MSE(1-h_{ii})}$ olacaktır. Dolayısıyla sapan değerler için $e_i^* < r_i$ olması söz konusudur. Uygulamamızda iki sapan gözlem standart kalıntıları küçük olduğu için standart kalıntıları açısından sapan değer olarak belirlenememiştir. Bu iki ülke 26 Hindistan ve 31 Japonya'dır. Bu ülkeler kaldıraç etkisi yüksek olan ülkelerdir. Yukarıdaki formüllerden de görüleceği gibi $h_{ii}$ değeri yüksek olduğunda $(1-h_{ii})$ değeri küçük olacaktır. Dolayısıyla örneğin standart sapması o ölçüde küçülecektir. Bu durumda studentized kalıntı standart kalıntıdan büyük olacaktır.

Kaldıraç değerlerine bakıldığında 6 ülkenin kesim noktası değeri olan 2p/n=0,1333 değerini aştığını görmekteyiz. Bu ülkeler: 4 Bangladeş, 26 Hindistan, 27 Endonezya, 29 İtalya, 31 Japonya , 45 Polonya. Bu kaldıraç değerleri içerisinde en yüksek değer Hindistan'a aittir. Bu kaldıraç değerlerinin etkili olup olmadıkları diagnostik analizlerin sonucunda belirlenecektir.

## 10.5 Katsayılar Üzerindeki Etkili Değerler

DFBETAS değerleri i.inci gözlemin silinmesinin her bir katsayı üzerindeki etkisini gösteren değerlerdir. Bu değerler aşağıdaki gibi hesaplanmıştır:

Çizelge 10.2 Katsayılar üzerindeki etkili değerler

| | | Sdfbeta Sabit | Sdfbeta mkn | Sdfbeta ith | Sdfbeta güb |
|---|---|---|---|---|---|
| 1 | Arnavutluk | -0,0188 | 0,00408 | 0,01076 | -0,00058 |
| 2 | Avustralya | 0,17358 | -0,0186 | -0,13628 | 0,12101 |
| 3 | Azerbeycan | -0,05345 | 0,01444 | 0,01672 | 0,00575 |
| 4 | Bangladeş | -0,04631 | -0,02168 | 0,0779 | -0,00627 |
| 5 | Barbados | -0,07162 | 0,0665 | -0,06859 | 0,03757 |
| 6 | Belarus | -0,04693 | 0,0491 | -0,03828 | -0,00979 |
| 7 | Bolivya | -0,03408 | 0,0138 | 0,00533 | 0,00419 |
| 8 | Botsvana | -0,14354 | 0,03042 | 0,08393 | -0,00467 |
| 9 | Brezilya | 0,00308 | 0,00001 | -0,00379 | 0,00667 |

Çizelge 10.2'nin Devamı

| | | Sdfbeta Sabit | Sdfbeta mkn | Sdfbeta ith | Sdfbeta güb |
|---|---|---|---|---|---|
| 10 | Bulgaristan | -0,00677 | 0,00206 | 0,00254 | 0,00001 |
| 11 | Şili | 0,14344 | -0,03865 | -0,07286 | 0,02054 |
| 12 | Cote d'Ivoire | 0,08759 | -0,02309 | -0,04486 | 0,00283 |
| 13 | Hırvatistan | -0,0226 | 0,0102 | 0,00533 | -0,00062 |
| 14 | Çek Cumhuriyeti | -0,01816 | 0,00579 | -0,0002 | 0,00215 |
| 15 | Dominik Cum. | -0,12142 | 0,06341 | -0,00815 | 0,02407 |
| 16 | Ekvator | -0,01335 | 0,01099 | -0,00764 | 0,00269 |
| 17 | Mısır Arap Emirliği. | -0,18253 | -0,12233 | **0,38735** | -0,01503 |
| 18 | El Salvador Cum. | -0,01757 | 0,02731 | -0,03872 | 0,01522 |
| 19 | Estonya | 0,0194 | 0,05892 | -0,20478 | 0,08564 |
| 20 | Habeşistan | 0,09212 | -0,02576 | -0,04802 | 0,00609 |
| 21 | Gürcistan | -0,04904 | 0,00789 | 0,03311 | -0,00326 |
| 22 | Almanya | 0,02049 | 0,06934 | -0,02259 | -0,01071 |
| 23 | Gana | -0,00017 | 0,00009 | -0,00003 | 0,00004 |
| 24 | Gine | -0,05365 | 0,01466 | 0,02537 | 0,00011 |
| 25 | Macaristan | -0,02179 | 0,00201 | 0,0105 | 0,00053 |
| 26 | Hindistan | 0,19602 | **0,73831** | **0,34503** | **-4,03847** |
| 27 | Endonezya | **-0,54736** | -0,24866 | **0,85834** | 0,0126 |
| 28 | İrlanda | 0,03248 | 0,00015 | -0,02163 | 0,00246 |
| 29 | İtalya | 0,11852 | **-0,43682** | -0,11768 | 0,24563 |
| 30 | Jamaika | -0,07674 | 0,02708 | 0,02227 | 0,00527 |
| 31 | Japonya | -0,19416 | **1,66208** | 0,06641 | **-0,8963** |
| 32 | Kazakistan | 0,03725 | -0,00518 | -0,02168 | -0,00023 |
| 33 | Kenya | -0,0125 | 0,01047 | -0,00867 | 0,00399 |
| 34 | Kore | -0,05876 | -0,02668 | **0,31814** | -0,09801 |
| 35 | Malavi | -0,0673 | 0,03088 | 0,0053 | 0,0091 |
| 36 | Malezya | 0,15697 | -0,07841 | -0,0753 | 0,0726 |
| 37 | Mauritius | -0,0506 | 0,06286 | -0,07931 | 0,03513 |
| 38 | Meksika | **0,38211** | 0,01975 | -0,25735 | 0,17491 |

Çizelge 10.2'nin Devamı

| | | Sdfbeta Sabit | Sdfbeta mkn | Sdfbeta ith | Sdfbeta güb |
|---|---|---|---|---|---|
| 39 | Mozambik | -0,04713 | 0,01431 | 0,01853 | 0,00137 |
| 40 | Namibia | -0,10537 | 0,01953 | 0,06884 | -0,00632 |
| 41 | Nikaragua | -0,03271 | 0,00594 | 0,02204 | -0,00246 |
| 42 | Norveç | 0,00138 | -0,00013 | 0,0003 | -0,00066 |
| 43 | Pakistan | 0,00746 | 0,00392 | -0,0146 | -0,00403 |
| 44 | Paraguay | 0,01613 | -0,00335 | -0,00938 | 0,00067 |
| 45 | Polonya | -0,09284 | **-1,12893** | 0,13777 | **0,51** |
| 46 | Portekiz | -0,00203 | -0,0001 | -0,00284 | 0,0025 |
| 47 | Romanya | 0,09724 | 0,00392 | -0,04545 | -0,01104 |
| 48 | Suudi Arabistan | **0,3584** | -0,11159 | -0,19895 | 0,05981 |
| 49 | Seychelles | -0,19689 | 0,03539 | 0,13331 | -0,01417 |
| 50 | İspanya | 0,03173 | 0,09581 | -0,02685 | -0,01944 |
| 51 | St. Lucia | -0,07928 | 0,07792 | -0,08452 | 0,04436 |
| 52 | Sudan | 0,21536 | -0,05259 | -0,11354 | 0,00719 |
| 53 | Tanzania | 0,03119 | -0,02665 | 0,02664 | -0,01532 |
| 54 | Trinidad ve Tobago | -0,14433 | 0,02984 | 0,08847 | -0,00711 |
| 55 | Tunisia | -0,00161 | 0,01247 | -0,02962 | 0,01125 |
| 56 | Türkiye | 0,07134 | -0,13605 | -0,1075 | 0,08549 |
| 57 | Ukrayna | 0,0455 | 0,021 | -0,0245 | -0,01312 |
| 58 | Britanya | 0,07374 | 0,04531 | -0,05145 | 0,01678 |
| 59 | Uruguay | 0,02334 | 0,04622 | -0,13758 | 0,04561 |
| 60 | Venezuella | 0,14081 | -0,03778 | -0,05914 | 0,00628 |

Değerler incelendiğinde 26 Hindistan,29 İtalya, 31 Japonya ve 45 Polonya ülkelerine ait verilerin $2/\sqrt{60}=0,258$ kesim noktasını aştıklarını dolayısıyla mkn (tarımsal makine,traktör,teçhizat sayısı) değişkenine ait katsayı üzerinde etkili olduklarını görmekteyiz. Bu ülkelerden İtalya ve Polonya'nın etkileri negatif yöndedir. 17 Mısır, 26 Hindistan, 27 Endonezya ve 34 Kore ülkelerinin ith (tarımsal hammadde ithalatı) değişkenine ilişkin katsayının üzerinde etkili oldukları görülmektedir. Bu ülkeler arasında Endonezya en etkili

ülke konumundadır. 26 Hindistan, 31 Japonya, ve 45 Polonya ülkelerinin güb (kilometre kare başına gübre tüketimi) değişkenine ilişkin katsayı üzerinde etkili oldukları görülmektedir. Hindistan ve Japonya bu katsayı üzerinde negatif yönde etki yapmaktadır.

Sapan değerlerin katsayılar üzerindeki etkileri genel olarak değerlendirildiğinde 26 Hindistan, 27 Endonezya, 31 Japonya, 45 Polonya ülkelerinin etkili veri noktaları oldukları görülmektedir. Bu ülkeler aynı zamanda yüksek kaldıraç etkisine sahip ülkelerdir. Bu verilerden herhangi birinin silinmesinin katsayıları önemli ölçüde değiştireceği söylenebilir.

## 10.6 Kovaryans Matrisi Üzerindeki Etkili Değerler

COVRATIO oranı i.inci satırın silinmesinden sonra kalan verilerle tahmin edilen kovaryans matrisinin determinantının tüm verilere dayanarak elde edilen kovaryans matrisinin determinantına oranıdır. Bu değer dolayısıyla tahmin edilen genelleştirilmiş varyansların oranıdır. O halde bu değer i.inci gözlemin silinmesinin katsayı tahminleri üzerindeki etkisini belirlemede kullanılabilir.

Birden büyük COVRATIO değerleri , ilgili gözlemin eksikliğinin etkinliği bozduğunu gösterir. $1 \pm 3(p/n)$ aralığının dışına düşen COVRATIO değerleri aykırı değerler olarak tanımlanır.

Çizelge 10.3 COVRATIO değerleri

| | | COVRATIO |
|---|---|---|
| 1 | Arnavutluk | 1,1068 |
| 2 | Avustralya | 1,01991 |
| 3 | Azerbeycan | 1,08291 |
| 4 | Bangladeş | 1,4094 |
| 5 | Barbados | 0,96651 |
| 6 | Belarus | 1,03413 |

Çizelge 10.3'ün Devamı

| | | COVRATIO |
|---|---|---|
| 7 | Bolivya | 1,08983 |
| 8 | Botsvana | 1,05314 |
| 9 | Brezilya | 1,25106 |
| 10 | Bulgaristan | 1,09928 |
| 11 | Şili | 1,0252 |
| 12 | Cote d'Ivoire | 1,07894 |
| 13 | Hırvatistan | 1,09528 |
| 14 | Çek Cumhuriyeti | 1,09098 |
| 15 | Dominik Cum. | 0,95841 |
| 16 | Ekvator | 1,09505 |
| 17 | Mısır Arap Emirliği. | 1,01764 |
| 18 | El Salvador Cum. | 1,08282 |
| 19 | Estonya | 0,93593 |
| 20 | Habeşistan | 1,07702 |
| 21 | Gürcistan | 1,11113 |
| 22 | Almanya | 1,15901 |
| 23 | Gana | 1,0982 |
| 24 | Gine | 1,09278 |
| 25 | Macaristan | 1,09696 |
| 26 | Hindistan | 4,21181 |
| 27 | Endonezya | 1,39481 |
| 28 | İrlanda | 1,10331 |
| 29 | İtalya | 1,37894 |
| 30 | Jamaika | 1,06692 |
| 31 | Japonya | 1,31514 |
| 32 | Kazakistan | 1,10232 |
| 33 | Kenya | 1,09572 |
| 34 | Kore | 0,72441 |
| 35 | Malavi | 1,06157 |
| 36 | Malezya | 1,00123 |

Çizelge 10.3'ün Devamı

| | | COVRATIO |
|---|---|---|
| 37 | Mauritius | 0,99215 |
| 38 | Meksika | **0,6147** |
| 39 | Mozambik | 1,09089 |
| 40 | Namibia | 1,09017 |
| 41 | Nikaragua | 1,11487 |
| 42 | Norveç | 1,09509 |
| 43 | Pakistan | 1,15975 |
| 44 | Paraguay | 1,10709 |
| 45 | Polonya | **0,71375** |
| 46 | Portekiz | 1,09761 |
| 47 | Romanya | 1,05129 |
| 48 | Suudi Arabistan | **0,7779** |
| 49 | Seychelles | 1,04066 |
| 50 | İspanya | 1,12108 |
| 51 | St. Lucia | 0,92517 |
| 52 | Sudan | 0,96352 |
| 53 | Tanzania | 1,07519 |
| 54 | Trinidad ve Tobago | 1,05939 |
| 55 | Tunisia | 1,10358 |
| 56 | Türkiye | 1,15986 |
| 57 | Ukrayna | 1,08905 |
| 58 | Britanya | 1,06969 |
| 59 | Uruguay | 1,05415 |
| 60 | Venezuella | 1,01112 |

Uygulamamızda COVRATIO değerleri 1,199 ve 0,8 aralığının dışında kalan ülkeler: 4 Bangladeş, 9 Brezilya, 26 Hindistan, 27 Endonezya, 29 İtalya, 31 Japonya, 34 Kore, 38 Meksika,5 Polonya ve 48 Suudi Arabistan'dır. Burada 9 Brezilya'nın önceden incelenen diagnostiklerde etkili ülke olarak belirlenmediğine dikkat edilmektedir. Diğer yandan çoğu ülke diğer diagnostiklerde etkili değerler olarak belirlenmiştir.

i.inci gözlemin silinmesinin varyansın $s^2$ veya $(X^T.X)^{-1}$'ın elemanları ya da her ikisi üzerinde büyük değişiklik yaratması durumunda COVRATIO için i.inci gözlem sapan değer olarak belirlenebilir. Dolayısıyla 26 Hindistan,31 Japonya,45 Polonya yüksek kaldıraç etkisi ve büyük kalıntı değerinden dolayı sapan değer olarak belirlenmiştir. 27 Endonezya, 29 İtalya ve 4 Bangladeş yüksek kaldıraç etkisine sahiptirler. 34 Kore, 38 Meksika ve 48 Suudi Arabistan yüksek kalıntı değerlerine sahiptirler. Bu ülkelerin COVRATIO değerlerinin sınırların dışında çıkması bu sebeplerden dolayı olabilir. Fakat 9 Brezilya ne yüksek kaldıraç etkisine sahiptir ne de yüksek kalıntı değerine. Bundan dolayı COVRATIO diagnostik ölçüsü kaldıraç etkisi ve kalıntı değerleri bilgilerini birleştirdiği gibi yeni etkili noktaların yerini kesin olarak belirtmektedir. Buradan COVRATIO'nun ayrıntılı bir diagnostik ölçü değerine sahip olduğu sonucuna varılmaktadır.

## 10.7 Öngörü Değerleri Üzerindeki Etkili Veriler ,Cook, Mahalanobıs Uzaklıkları

i.inci gözlemin silinmesinin tekil bir öngörüde yarattığı değişim değerleri (DFFITS) aşağıdaki gibidir.

Çizelge 10.4 DFFITS,Cook,Mahalanobıs değerleri

| | | DFFITS | COOK | MAHALANOBIS |
|---|---|---|---|---|
| 1 | Arnavutluk | -0,01979 | 0,0001 | 0,77884 |
| 2 | Avustralya | 0,22317 | 0,01239 | 1,25136 |
| 3 | Azerbeycan | -0,06673 | 0,00113 | 0,30432 |
| 4 | Bangladeş | 0,08248 | 0,00173 | 13,09729 |
| 5 | Barbados | -0,2108 | 0,01095 | 0,42254 |
| 6 | Belarus | -0,13335 | 0,00446 | 0,25721 |
| 7 | Bolivya | -0,04801 | 0,00059 | 0,27816 |
| 8 | Botsvana | -0,15027 | 0,00567 | 0,82886 |
| 9 | Brezilya | 0,00854 | 0,00002 | 7,33371 |
| 10 | Bulgaristan | -0,00805 | 0,00002 | 0,34561 |
| 11 | Şili | 0,15811 | 0,00625 | 0,45844 |

Çizelge 10.4'ün Devamı

| | | DFFITS | COOK | MAHALANOBIS |
|---|---|---|---|---|
| 12 | Cote d'Ivoire | 0,09519 | 0,00229 | 0,61604 |
| 13 | Hırvatistan | -0,03002 | 0,00023 | 0,29448 |
| 14 | Çek Cumhuriyeti | -0,02971 | 0,00022 | 0,09491 |
| 15 | Dominik Cum. | -0,20384 | 0,01022 | 0,27662 |
| 16 | Ekvator | -0,03181 | 0,00026 | 0,30228 |
| 17 | Mısır Arap Emirliği. | 0,44369 | 0,04831 | 4,188 |
| 18 | El Salvador Cum. | -0,08539 | 0,00185 | 0,58244 |
| 19 | Estonya | -0,3042 | 0,02255 | 1,15374 |
| 20 | Habeşistan | 0,0997 | 0,00251 | 0,63089 |
| 21 | Gürcistan | -0,04972 | 0,00063 | 1,21377 |
| 22 | Almanya | 0,08612 | 0,00188 | 3,65135 |
| 23 | Gana | -0,00031 | 0 | 0,2772 |
| 24 | Gine | -0,05969 | 0,0009 | 0,55084 |
| 25 | Macaristan | -0,02459 | 0,00015 | 0,32412 |
| 26 | Hindistan | **-4,54232** | **4,8416** | **47,14345** |
| 27 | Endonezya | **0,91192** | **0,20511** | **17,94066** |
| 28 | İrlanda | 0,03412 | 0,0003 | 0,70896 |
| 29 | İtalya | -0,47965 | 0,05786 | 14,18941 |
| 30 | Jamaika | -0,09726 | 0,00239 | 0,33723 |
| 31 | Japonya | **1,71049** | **0,68957** | **22,57852** |
| 32 | Kazakistan | 0,03919 | 0,00039 | 0,70992 |
| 33 | Kenya | -0,03193 | 0,00026 | 0,33408 |
| 34 | Kore | 0,45741 | 0,04786 | 0,94094 |
| 35 | Malavi | -0,10085 | 0,00257 | 0,27316 |
| 36 | Malezya | 0,18988 | 0,00896 | 0,54885 |
| 37 | Mauritius | -0,19593 | 0,00952 | 0,51572 |
| 38 | Meksika | 0,48544 | 0,05182 | 0,5779 |
| 39 | Mozambik | -0,05533 | 0,00078 | 0,41872 |
| 40 | Namibia | -0,10745 | 0,00292 | 1,14144 |

Çizelge 10.4'ün Devamı

| | | DFFITS | COOK | MAHALANOBIS |
|---|---|---|---|---|
| 41 | Nikaragua | -0,03317 | 0,00028 | 1,25589 |
| 42 | Norveç | 0,0027 | 0 | 0,1148 |
| 43 | Pakistan | -0,01881 | 0,00009 | 3,35959 |
| 44 | Paraguay | 0,01691 | 0,00007 | 0,77977 |
| 45 | Polonya | **-1,20471** | **0,32047** | 7,70706 |
| 46 | Portekiz | -0,00777 | 0,00002 | 0,25784 |
| 47 | Romanya | 0,11401 | 0,00327 | 0,26601 |
| 48 | Suudi Arabistan | 0,38404 | 0,03438 | 0,68807 |
| 49 | Seychelles | -0,19946 | 0,00995 | 1,291 |
| 50 | İspanya | 0,12228 | 0,00379 | 2,44259 |
| 51 | St. Lucia | -0,24628 | 0,01478 | 0,4421 |
| 52 | Sudan | 0,2321 | 0,01325 | 0,63026 |
| 53 | Tanzania | 0,08699 | 0,00191 | 0,38689 |
| 54 | Trinidad ve Tobago | -0,14928 | 0,00561 | 0,94661 |
| 55 | Tunisia | -0,0493 | 0,00062 | 0,87096 |
| 56 | Türkiye | -0,19971 | 0,01009 | 4,78588 |
| 57 | Ukrayna | 0,06001 | 0,00091 | 0,41582 |
| 58 | Britanya | 0,11631 | 0,00341 | 0,69237 |
| 59 | Uruguay | -0,19194 | 0,00923 | 1,45798 |
| 60 | Venezuella | 0,16308 | 0,00663 | 0,33652 |

DFFITS değerlerine göre 26 Hindistan, 27 Endonezya, 45 Polonya ve 31 Japonya etkili ülkeler olarak belirlenmiştir.

Cook mesafesi i.inci gözlemin bütün öngörü değerleri üzerindeki etkisini ölçen bir diagnostik araçtır. Bu ölçü sadece bağımlı değişken üzerindeki sapan değerleri değil aynı zamanda bağımsız değişken seti üzerindeki sapan değerleri gösterir. Cook ve Weisberg birden büyük Cook mesafesinin büyük olarak dikkate alınabileceğini ifade etmişlerdir.

Chatterjee ve Hadi'ye göre Cook mesafesi 4/n-p değerini aşan gözlemler etkili değerlerdir. Bu kritere göre 4/60-4=0,0714 değeri sınır değeridir. 26,27,31,45 gözlemleri Cook mesafesine göre etkili gözlemlerdir.

Uygulamamızda sadece 26 Hindistan'a ait Cook mesafesi 1'i aşmıştır. Eğer bir gözlem bağımlı değişken üzerinde anlamlı bir sapan değer ise fakat Cook mesafesi <1 ise bu noktayı silmeye gerek yoktur çünkü regresyon analizi üzerinde önemli bir etkisi olmayacaktır.

Mahalanobis uzaklığı i.inci gözlemin bağımsız değişken değerlerinin merkezinden ne kadar uzağa düştüğünü göstermektedir. Büyük bir uzaklık söz konusu gözlemin bağımsız değişkenler üzerinde sapan değer olduğunu gösterir. Mahalanobis uzaklığının büyüklüğü konusunda kritik değer Barnett&Lewis, 1978 tarafından hazırlanan tablo aracılığıyla belirlenebilir.* n=50 ve p-1=3 için tablo değeri 14,18. n=60 için tablo değeri bulunmamaktadır. Fakat n=5'den sonraki değerler artmaktadır. Dolayısıyla 14.18 değeri minimum uzaklık olarak düşünülmüştür.

Uygulamamızda Mahalanobis değerleri küçükten büyüğe doğru sıralanarak uzaklıklar incelenmiştir. En uzak değerin 26 Hindistan'dır ve ardından sırasıyla 27 Endonezya ve 31 Japonya'nın gelmektedir.

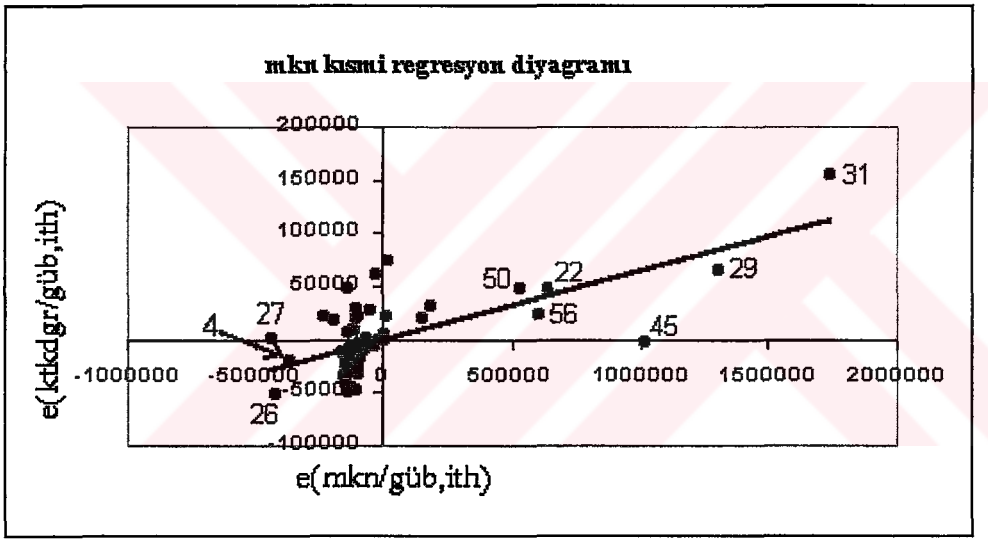
**ARA SONUÇ**: 26 Hindistan yüksek kaldıraç etkisine sahiptir ve bütün katsayıların üzerinde anlamlı bir etkisi bulunmaktadır. Bu ülke potansiyel bir etkili veri olarak düşünülebilir ve ciddi problemlere sebep olacağı söylenilebilir. Sırasıyla 31 Japonya, 27 Endonezya ve 45 Polonya'nın da potansiyel etkili ülkeler oldukları ve regresyon analizinin sonuçlarını değiştirdikleri söylenebilir. Yukarıdaki uygulamadan aynı zamanda bütün yüksek kaldıraç etkisine sahip ülkelerin aynı zamanda tahmin edilen katsayılar veya öngörüler üzerinde etkili olmadıkları görülebilir: 4 Bangladeş.

---

* Tablo için Bkz Ek.2

## 10.8 Kısmi Regresyon Diyagramları

Yukarıda bahsedilen diagnostik teknikler çoğu zaman etkili gözlemleri belirlemede tatmin edici olabilirken her zaman başarılı olmayabilirler. Etkili bir sapan değerin etkisi diğer bir sapan değer tarafından maskelenebilir. Dolayısıyla gözlemlerin alt gruplarının potansiyel etkilerinin incelenmesi için, diğer bir deyişle çoklu sapan değerlerin belirlenebilmesi için teknikler geliştirilmiştir. Bu diagnostik ölçülerde öncelikle bu alt grupların belirlenmesi gerekmektedir. Bu belirlemede kısmi kalıntı diyagramları ve tekil diagnostik ölçülerden elde edilen sonuçlar yardımcı olabilir. Aşağıda uygulamamıza ilişkin kısmi regresyon diyagramları yer almaktadır.

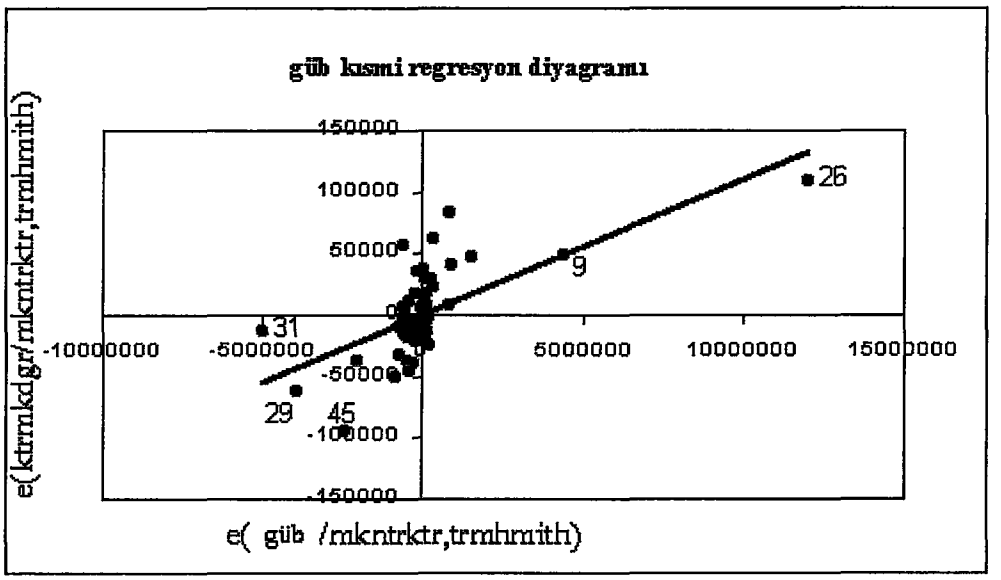



Şekil 10.2 mkn değişkeni kısmi regresyon diyagramı

Yukarıda tarımsal makine, traktör,teçhizat sayısı değişkeninin (mkn) kısmi regresyon diyagramı görülmektedir. Bu diyagramdan bu değişken üzerinde etkili olan gözlemler incelenebilir. 31 Japonya mkn değişkenine ilişkin eğimi arttırmaktadır. Diğer gözlemlerden oldukça uzaktadır. 31 Japonya'nın etkisi 29 İtalya'nın etkisini maskelemiş olabilir. 45 Polonya da bu değişken için sapan değer durumundadır. 26 Hindistan değişkene ait eğimi arttırıcı yönde etki yapmaktadır. 26 regresyon doğrusunun başlangıç noktasını aşağıya çekmekte, 31 ise yukarıya kaldırmaktadır. Dolayısıyla 26 ve 31 birlikte mkn değişkeninin eğimini arttırmaktadır.

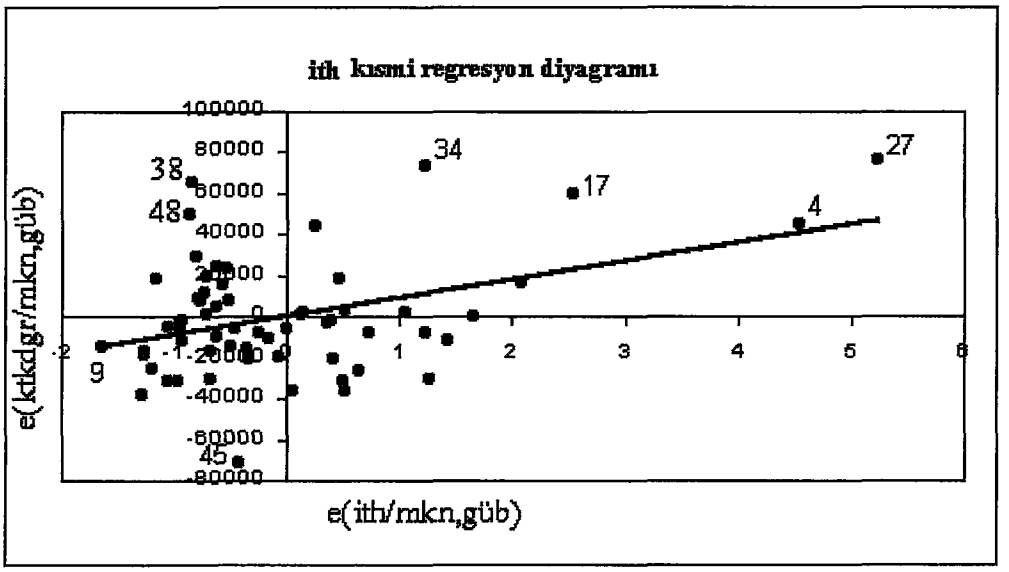
Buradan elde edilen sonuçlar tekil satır diagnostiklerinde elde edilen DFBETAS ölçüsü ile

aynı sonuçlardır.




Şekil 10.3 güb değişkeni kısmi regresyon diyagramı

Tarımsal gübre tüketimi değişkenine (güb) ilişkin kısmi regresyon diyagramı yukarıdaki gibidir. Diyagramda 9,26,29,31 ve 45 gözlemlerinin sapan değerler oldukları görülmektedir. Tekil satır etkilerinde güb değişkenine ait DFBETAS ölçüsüne göre 26,31 ve 45 ülkeleri sapan değerler olarak görülmekteydi. Burada yukarıda bahsedilen bir sapan değerin, diğer bir sapan değerin etkisini maskelemesi durumu söz konusudur. 9'un etkisi 26 tarafından, 29'un etkisi 31 ve 45 tarafından maskelenmiştir. Dolayısıyla tekil analizde bu değerler sapan değerler olarak görünmemektedir.




Şekil 10.4 ith değişkeni kısmi regresyon diyagramı

Tarımsal hammadde ithalatı değişkenine (ith) ilişkin kısmi regresyon diyagramı incelendiğinde 4'ün etkisinin 27 tarafından maskelendiği ; 27,4,17,34,38 ve 48 verilerinin regresyon doğrusunu yukarıya doğru çekerek 45'in etkisini maskelediği görülmektedir. Dolayısıyla tekil satır etkileri incelenirken 45 söz konusu değişken için etkili bir değer olarak görülmemektedir.

## 10.9 Çoklu Satır Etkileri

Çoklu satır tekniklerine geçilmeden önce potansiyel olarak etkili olan gözlemlerin belirlenmesi gerekmektedir. Önceki tekil satır tekniklerinden yararlanılarak belirlenen gözlemler ve kısmi regresyon diyagramlarında belirlenen gözlemlerin oluşturduğu potansiyel etkili noktalar aşağıdaki gibidir:

Çizelge 10.5 Etkili veri kümesi

| Gözlem Sırası | Ülkeler |
|---|---|
| 4 | Bangladeş |
| 9 | Brezilya |
| 17 | Mısır |
| 26 | Hindistan |
| 27 | Endonezya |
| 29 | İtalya |
| 31 | Japonya |
| 34 | Kore |
| 38 | Meksika |
| 45 | Polonya |
| 48 | Suudi Arabistan |

Tekil satır etkilerinde izlenen süreç gibi, bir gözlemin etkili olup olmadığını incelemek amacıyla o satırın silinmesinin kalıntılar, katsayılar, öngörü değerleri, kovaryans matrisi üzerindeki etkileri incelenecektir. Burada m=1,2,... birimden oluşan alt gruplar belirlenecektir. Ardından bu alt gruplara ilişkin geliştirilen çeşitli diagnostik ölçüler hesaplanacak ve söz konusu alt grupların etkili olup olmadığı değerlendirilecektir.

MDFFIT için oluşturulan tablo aşağıdaki gibidir. Bu tabloda aynı zamanda her bir alt grup büyüklüğü için gözlemlerin o grubun en büyük MDFFIT değerine yüzdesi alınmıştır. Bu yüzdelikler diyagramlarda belli gözlemlerin oluşturdukları grupları kalan gruplardan ayıran bir gösterge durumundadır. Örneğin m=1 için 26 ve 31 belirgin bir şekilde kalanlardan ayrı görülmektedir. m=2 için 9,26 ve 26,31 ikililerinin kalan ikili gruplardan ayrıldığı görülebilir. m=3 için 9,26,31 gözlemlerinin oluşturduğu alt grup diğerlerinden belirgin bir şekilde farklıdır. MDFFIT çoklu satır silinmesinin öngörü üzerindeki etkisini ifade ettiğinden bu grup diğer gruplardan daha etkilidir. Diğer oluşturulan alt gruplar bu gruba oranla analizi etkileyen gruplar olarak değerlendirilmemektedir. m=4 ve m=5 için etkili alt gruplar "*" simgesiyle belirtilmiştir.

Çizelge 10.6 MDFFIT değerleri

| m | alt küme | MDFFIT | En Büyük Değere Oranı |
|---|---|---|---|
| 1 | 26* | 25,93 | 1,000* |
| | 29 | 1,227 | 0,047 |
| | 31 | 11,87 | 0,458 |
| | 34 | 1,327 | 0,051 |
| | 45 | 7,93 | 0,306 |
| 2 | 4,27 | 7,522 | 0,110 |
| | 26,31* | 51,52 | 0,751* |
| | 29,31 | 8,499 | 0,124 |
| | 26,9* | 68,57 | 1,000* |
| | 29,45 | 20,08 | 0,293 |
| 3 | 9,26,31* | 114,7 | 1,000* |
| | 4,17,27 | 23,46 | 0,205 |
| | 34,38,45 | 7,609 | 0,066 |
| | 17,27,34 | 24,09 | 0,210 |
| | 29,31,45 | 0,363 | 0,003 |
| 4 | 26,29,31,45 | 31,27 | 0,253 |
| | 9,26,29,31* | 123,8 | 1,000* |
| | 4,17,27,34 | 44,97 | 0,363 |
| | 22,50,29,31 | 19,54 | 0,158 |
| | 9,26,31,45* | 94,16 | 0,761* |
| | 26,27,31,45 | 31,01 | 0,250 |
| 5 | 9,26,29,31,45* | 91,99 | 1,000* |
| | 4,17,27,34,45* | 64,82 | 0,705* |

Bu durumda etkisi maskelenen gözlemlerin belirlenebilmesi amacıyla bir çok gözlemin etkisini maskelediği düşünülen 26 Hindistan veri setinden dışlanarak oluşturulan yeni modelde studentize öngörü kalıntıları, kaldıraç değerleri ve DFFITS değerleri aşağıdaki gibidir:

Çizelge 10.7 26 Hindistan gözleminin dışlanması halinde diagnostik ölçüler

| sıra | Ülkeler | st.öng.kalıntısı | kaldıraç | DFFITS | DFBETAS (mkn) | DFBETAS (ith) | DFBETAS (güb) |
|---|---|---|---|---|---|---|---|
| 4 | Bangladeş | 0,05082 | 0,22341 | 0,02859 | -0,00786 | 0,02619 | 0,0011 |
| 9 | Brezilya | -2,62496 | 0,62083 | -3,4832 | 0,9922 | 1,20271 | -3,31442 |
| 17 | Mısır, Arap Emr. | 1,34637 | 0,074 | 0,42585 | -0,1353 | 0,34761 | 0,06482 |
| 27 | Endonezya | 0,94699 | 0,33097 | 0,69172 | -0,23391 | 0,5868 | 0,1748 |
| 29 | İtalya | -0,66662 | 0,24516 | -0,39731 | -0,35678 | -0,10396 | 0,14315 |
| 31 | Japonya | 2,67831 | 0,40246 | 2,27636 | 2,20335 | 0,1637 | -0,98772 |
| 34 | Kore | 2,60298 | 0,01572 | 0,47838 | -0,01989 | 0,33143 | -0,06598 |
| 38 | Meksika | 2,68863 | 0,0326 | 0,61386 | -0,1183 | -0,30097 | 0,44573 |
| 45 | Polonya | -2,96492 | 0,1306 | -1,23352 | -1,10939 | 0,13109 | 0,29118 |
| 48 | Suudi Arabistan | 2,30649 | 0,01147 | 0,39448 | -0,11551 | -0,205 | 0,04863 |

26 verisinin silinmesinin ardından diagnostik değerler tekrar incelendiğinde belirgin farklılıklar görülmektedir. Öncelikle göze çarpan 9 Brezilya verisidir. 9 verisine ait studentize öngörü kalıntısı 0,02108 iken 26 silindikten sonra bu kalıntı değeri -2,62496'a, $\pm 2$ sınırlarının dışına düşmüştür. Kaldıraç değerleri incelendiğinde de Brezilya'nın kaldıraç değerinin yükseldiği ve sınır değeri olan 2p/n=0,135 değerini aştığı görülmektedir. Aynı durum DFFITS diagnostik ölçüsü için de geçerlidir. 9 verisine ait DFFITS değeri 0,00854 iken 26 verisi silindikten sonra -3,4832 olmuştur. Bu değer mutlak olarak kesim noktası değeri olan $2\sqrt{p/n}$=0,5207 değerini aşmıştır. Yukarıda görüldüğü gibi 26 verisi 9 verisinin etkisini maskelemiştir. Bu durum katsayıların duyarlılıkları incelendiğinde de görülmektedir. 9 verisi hiçbir değişken üzerinde etkili görünmezken bu durumda oldukça etkili konumdadır. Şöyle ki, mkn değişkeni üzerindeki etkisi (DFBETAS) mutlak olarak 0,00001'den 0,9922'a; ith değişkeni üzerindeki etkisi –0,00379'dan 1,20271'e; güb değişkeni üzerindeki etkisi 0,00667'den -3,31442'e yükselmiştir. En büyük etkisi güb değişkeni üzerinde gözükmektedir. Bu durum kısmi kalıntı diyagramından da açıkça görülebilir.

26 verisi 9 verisinin etkisini maskelediği gibi 38 verisinin de etkisini maskelemektedir. Bu maskeleme 9 verisi kadar büyük değildir. 26 verisinin silinmesinden sonra 38 verisi DFFITS sınır değerini aşmıştır. Katsayılar üzerindeki etkileri incelendiğinde 26'nın silinmesinden sonra 38 verisinin ith değişkeni üzerindeki etkisi 0,17491'den 0,44573'e; güb değişkeni üzerindeki etkisinin –0,25735'den –0,30097'ye değiştiği görülmektedir.

38 verisinin etkisinin, daha çok ith değişkeninde 26 tarafından maskelendiği kısmi kalıntı diyagramından da görülmektedir.

Buradan etkili bir gözlemin diğer etkili gözlemlere ilişkin diagnostik ölçüleri nasıl etkilediği görülmektedir.

Benzer şekilde 31 verisi silindiğinde de belirgin olmasa da bazı farklılıklar ortaya çıkmaktadır. COVRATIO değerlerinde 31 silinmeden önce etkili veri durumunda bulunmayan 56 Türkiye, 31 verisinin silinmesinin ardından etkili veri haline gelmiştir. COVRATIO için çoklu satır istatistikleri aşağıdaki gibi hesaplanmıştır:

Çizelge 10.8 Çoklu satırlarla maksimum COVRATIO değerleri

| m | Gözlemler | Büyük COVRATIO | Maksimuma Oranları |
|---|---|---|---|
| 1 | 26 | 4,211 | 1,000 |
| | 27 | 1,3948 | 0,331 |
| | 4 | 1,409 | 0,335 |
| | 29 | 1,3789 | 0,327 |
| | 31 | 1,315 | 0,312 |
| | 9 | 1,251 | 0,297 |
| 2 | 26,9 | 7,744 | 0,977 |
| | 26,31 | 4,752 | 0,600 |
| | 29,31 | 7,924 | 1,000 |
| 3 | 26,29,31 | 8,674 | 0,399 |
| | 26,9,31 | 6,180 | 0,284 |
| | 26,27,4 | 21,767 | 1,000 |

29,31 verilerinin yüksek COVRATIO değerleri bu gözlemlerin kovaryans matrisi üzerinde oldukça büyük etkilerinin olduğunu göstermektedir. 26,9 değerlerinin de kovaryans matrisi üzerinde önemli bir etkiye sahip olduğu söylenilebilir.

Çizelge 10.9 Çoklu satırlarla minimum COVRATIO değerleri

| m | Gözlemler | Küçük COVRATIO | Minimuma Oranları |
|---|---|---|---|
| 1 | 34 | 0,724 | 1,178 |
| | 38 | 0,615 | 1,000 |
| | 45 | 0,714 | 1,161 |
| | 48 | 0,778 | 1,265 |
| 2 | 38,45 | 0,404 | 1,000 |
| | 34,38 | 0,415 | 1,026 |
| | 38,48 | 0,443 | 1,097 |
| 3 | 34,38,45 | 0,224 | 1,000 |
| | 38,45,48 | 0,268 | 1,200 |
| | 34,45,48 | 0,336 | 1,499 |
| | 34,38,48 | 0,273 | 1,220 |

Küçük COVRATIO değerleri bu gözlemlerin herhangi bir yeni bilgi sağlamadığını göstermektedir. Dolayısıyla 38,45 gözlemlerinin yeni bir bilgi sağlamadığı kalan veri setiyle uyum içinde olduğu söylenilebilir.

## 11. SONUÇLAR VE ÖNERİLER

Uygulamamız 271 ülke arasından tabakalı örnekleme yoluyla seçilen 60 ülke üzerine yapılmıştır. Tabakalar Birleşmiş Milletlerin tasnifine göre az gelişmiş, gelişmekte olan ve gelişmiş ülkeler olarak belirlenmiştir.

Ülkelerin tarım verilerine En Küçük Kareler Yöntemi kullanılarak doğrusal regresyon modeli öngörülmüş ve anlamlı sonuçlar elde edilmiştir.

Bu veri seti üzerinde sapan ve etkili veriler belirlenmiştir. Sapan değerlerin belirlenmesi amacıyla kalıntı diyagramları kullanılmıştır. Uygulamada değinilen diagnostik teknikler her bir değişken sütununu, sapan gözlemler çıkarıldıktan sonra tek tek inceleme yolundan daha etkilidir. Örneğin güb değişkeninde 9 verisi sapan bir değer olduğu halde diğer katsayılar üzerinde bu verinin etkili olduğunu belirten bir sonuç yoktur. Üstelik diyagramlarda ortaya çıkmadığı halde tekil satır tekniklerinde örneğin ith değişkeni için 17 verisi sapan değer olarak ortaya çıkmıştır.

Etkili gözlemlerin belirlenmesi amacıyla öncelikle tekil satır silinmesine dayalı teşhis ölçüleri kullanılmış ve sonuçlar yorumlanmıştır. Bu veriler arasında 26. sırada bulunan Hindistan güçlü bir sapan değerdir.

Tekil ve çoklu satır teknikleriyle Hindistan'ın yanı sıra Japonya,Endonezya, Brezilya,Bangladeş, Meksika,Polonya,İtalya, Mısır,Kore ve Suudi Arabistan'ın sapan değerler olarak belirlenmişlerdir. Bu gözlemlerin bazıları farklı değişkenler üzerinde etkili gözlemlerken bazıları ise Hindistan gibi bütün değişkenler üzerinde etkili gözlemlerdir.

Hindistan verisi tahmin ve çıkarsama süreçlerini etkilemektedir. Bunun yanı sıra veri setindeki diğer bazı sapan gözlemlerin etkisini maskelemektedir. Örneğin güb değişkeni için 9 verisi tekil satır teknikleriyle sapan değer olarak belirlenmemiştir. Kısmi regresyon diyagramından 9 verisinin etkisinin 26 tarafından gizlendiği dolayısıyla belirlenemediği görülmektedir. Kısmi regresyon diyagramlarının dışında 26 verisinin modelden dışlanmasından sonra hesaplanan diagnostik ölçüler bu verinin etkisinin gizlendiğini doğrulamaktadır. Uygulamamızda söz konusu verinin dışlanmasının ardından elde edilen

sonuçlara değinilmiştir.

Bu ülkenin sapan ve etkili bir değer olarak ortaya çıkmasındaki sebep bu ülkenin bir tarım ülkesi olmasıdır. Ülkenin son 20 yıl boyunca tarımdaki performansı oldukça başarılıdır. Tarımda gelişmiş teknoloji kullanımında lider olan bu ülke hububat üretiminde kendine yeten ve bazı tarımsal ürünlerin üretiminde dünyada lider ülke konumundadır. Dolayısıyla bağımsız değişkenlerimizden biri olan tarımsal ürünlerin ithalatı, ülkenin toplam ithalatı içinde oldukça küçük bir yüzdeyi kapsamaktadır. Hindistan'ın gübre tüketimi yıllar boyunca sürekli artmaktadır. Gıda ürünlerinde gübre tüketimi 1980-81'den 1999-2000'e yaklaşık üç katından fazla artmıştır. Bu geçen süre içinde Hindistan sulama kanallarını geliştirmiş ve yaymıştır. Dolayısıyla gübre tüketimi kalıbında büyük bir değişim gerçekleşmiştir.

Ülkenin ekilebilir arazi oranı diğer ülkelerin üzerindedir. Dolayısıyla model kurma safhasında bağımsız değişkenlerin seçilmesi aşamasında ekilebilir arazi oranı bağımsız değişken olarak ele alınmamıştır. Ekilebilir arazi oranı bağımsız değişken olarak seçildiğinde bu değişkenin katsayısı negatif olarak hesaplanmaktadır. Benzer bir durum tarımsal nüfusun toplam nüfus içindeki payı için de geçerlidir. Söz konusu değişkenler tarımsal katma değer üzerinde pozitif etkisi olan değişkenlerdir. Sapan değerlerin etkisiyle bu değişkenlerin işareti negatif çıkmaktadır.

Hindistan ve diğer sapan değerlere ilişkin veriler hesaplama veya kayıt hatasından kaynaklanmamaktadır. Veriler Dünya Bankası'ndan alınmıştır.

Sonuç olarak uygulamamızda veri setindeki sapan ve etkili gözlemler belirlenerek bu gözlemlerin tahmin edilen katsayılar, öngörü değerleri, kalıntılar, ve katsayıların tahmin edilen varyans kovaryans yapısı üzerindeki etkileri incelenmiştir.

## ÖNERİ

Bir sapan değerle karşılaşıldığında ilk şüphelenilmesi gereken bu gözlemin bir hata veya diğer dışsal etkilerden dolayı sonuçlandığı dolayısıyla diğer gözlemlerden ayrı bir yerde bulunduğudur. En küçük kareler yönteminde sapma kareleri toplamı minimize edildiğinden öngörülen doğru sapan değere doğru kaydırılabilir. Bu durum, sapan değer gerçekten bir

yanlıştan veya diğer dışsal sebeplerden kaynaklanıyorsa yanlış öngörüye sebep olur. Örneğin bir katsayı pozitif olması gerekirken negatif tahmin edilebilir. Diğer yandan bu gibi veriler genellikle tahminin etkinliğini etkileyebilecek yararlı bilgiler içerirler. Bu faydalı durumuna rağmen bu aykırı noktaları ayırmak ve hangi parametre tahminlerinin bu veri setinden daha çok etkilendiğine karar vermek yapıcı bir çözümdür.

Herhangi bir sapan değeri dışlamak için başvurulması gereken güvenli bir yol; bu sapan değerin bir kayıt hatasından, yanlış hesaplamadan veya benzer tür hatalardan, durumlardan olduğuna dair bir kanıt varsa sapan değerin dışlanmasıdır.

Uygulamamız için Hindistan'ın ve diğer sapan gözlemlerin dışlanmaması gerektiğini savunmaktayım. Bu gözlemlere ilişkin herhangi bir kayıt hatası veya diğer dışsal sebepler bulunmadığından söz konusu gözlemlerin anlamlı bilgiler içerdikleri sonucu doğmaktadır. Bağımsız değişkenler ayrı ayrı incelendiğinde her biri üzerindeki sapan gözlemlerin anlamlı olduğu görülmektedir. Örneğin tarımda kullanılan makine teçhizat değişkenine ilişkin sapan gözlemler incelendiğinde bu gözlemlerin Japonya gibi tarımda gelişmiş teknolojiyi kullanan ülkeler oldukları görülmektedir. Bu ülkelerin veri setinden çıkarılması halinde ortalama bir ilişki belirlenebilir fakat teknolojik gelişmenin bu değişken üzerinde önemli etkisi olduğu sonucu elde edilemez.

**KAYNAKLAR**

Akın,F.,(1997), Ekonometri II Ders notları, M.S.Ü.,İstanbul.

Alpar,R., (1997), "Çok Değişkenli İstatistiksel Yöntemler",Bağıran Yayınevi,Ankara.

Atkinson, A.C.,(1997), Plots, Transformations and Regression, Clarendon Press, Oxford.

Barnett,V., Lewis,T., (1994), Outliers in Statistical Data, John Wiley&Sons, Chichester.

Belsley,D.A., Kuh,E., Welsch,R.E.,(1980), Regression Diagnostics:Identifying Influential

Data and Sources of Collinearity, John Wiley&Sons,U.S.A.

Büyüklü, A.H.,(1997), "Ekonometrik Modellerde Sapan Değerlerin Etkisi ve Sonuçları", İstatistik ve Ekonometri Araştırma ve Uygulama Merkezi Dergisi, I:83-85.

Draper,N.R., Smith,H.,(1998) ,Applied Regression Analysis, John Wiley&Sons,New York.

Evren,A.A., (1997), Doğrusal Olmayan Regresyon Teknikleri ve bir Uygulama, Doktora Tezi,İ.Ü,S.B.E, İstanbul.

Fox,J.,(1991), Regression Diagnostics, Sage Publications,California.

Genceli,M.,(2002) Ekonometri ve İstatistik İlkeleri, Filiz Kitapevi, İstanbul.

Gujarati,D.N.,Çev: Şenesen,Ü.,Şenesen, G.G.,(1999), Temel Ekonometri ,Literatür Yayıncılık.

Güriş,S.,Çağlayan,E.,(2000), Ekonometri, Der Yayınları,İstanbul.

Hamilton,L.C.,(1992), Regression With Graphics, Brooks/Coole Publishing Company,California.

James Stevens,(2002), Applied Multivariate Statistics for the Social Sciences, Lawrance Erlbaum Associates, London.

Jorgensen,B.,(1992), The Theory of Linear Models, Chapman&Hall, London.

Kennedy,P., çev. Sarımeşeli,M.,(2000), Ekonometri El Kitabı, Gazi Yayınları, Ankara.

Kıroğlu,G.B.,(1996), Uygulamalı Parametrik Olmayan İstatistiksel Yöntemler-Ders Notu, M.S.Ü.,İstanbul.

Koutsoyiannis,A.,(1989), Ekonometri Kuramı, Teori Yayınları,İstanbul.

Mansfield,E.R.,Conerly,M.D., (May 1987), "Diagnostic Value of Residual and Partial Residual Plots",American Statistical Association, 41,2.

Neter,J.,Kutner,M.H.,Wasserman,W.,Nachtsheim,C.J.,(1996), Applied Linear Regression Models , Irwin,London.

Ruud,P.A., (2000), An Introduction to Classical Econometric Theory ,Oxford University Press,New York.

Saville, D.J.,Wood,G.R., (Ağustos 1986),"A Method for Teaching Statistics Using N Dimensional Geometry", The American Statistician, 40:3.

Seber, G.A.F., Lee, A.J., (2003), Linear Regression Analysis, John Wiley&Sons,New Jersey.

Weisberg,S.,(1985), Applied Linear Regression, John Wiley&Sons,U.S.A.

**INTERNET KAYNAKLARI**

[1] www.devdata.org

[2]www.undp.org

**EKLER**

EK 1 Ülkelerin Tarım Verileri

| | | tkdgr | ktkdgr | mkn | ith | güb |
|---|---|---|---|---|---|---|
| 1 | Arnavutluk | 9,99E+08 | 31600 | 7947 | 0,863974 | 18700 |
| 2 | Avustralya | 1,28E+10 | 113000 | 315000 | 1,06814 | 2463000 |
| 3 | Azerbeycan | 8,47E+08 | 29100 | 30093 | 1,428353 | 11900 |
| 4 | Bangladeş | 1,09E+10 | 104000 | 5530 | 6,522537 | 1354800 |
| 5 | Barbados | 1,24E+08 | 11100 | 585 | 2,276763 | 3000 |
| 6 | Belarus | 1,28E+09 | 35800 | 67070 | 2,326596 | 780000 |
| 7 | Bolivya | 1,06E+09 | 32600 | 6000 | 1,626176 | 12119 |
| 8 | Botsvana | 1,32E+08 | 11500 | 6000 | 0,822524 | 4600 |
| 9 | Brezilya | 2,76E+10 | 166000 | 806000 | 1,376757 | 6773000 |
| 10 | Bulgaristan | 1,62E+09 | 40200 | 25000 | 1,346979 | 156630 |
| 11 | Şili | 5,18E+09 | 72000 | 54000 | 1,160951 | 481000 |
| 12 | Cote d'Ivoire | 2,66E+09 | 51600 | 3800 | 1,030112 | 62500 |
| 13 | Hırvatistan | 1,39E+09 | 37300 | 2386 | 1,558582 | 215200 |
| 14 | Çek Cumhuriyeti | 2,28E+09 | 47700 | 94778 | 1,87574 | 394717 |
| 15 | Dominik Cum. | 39400000 | 6280 | 90 | 1,840285 | 3000 |
| 16 | Ekvator | 1,89E+09 | 43500 | 14700 | 2,169568 | 230600 |
| 17 | Mısır Arap Emirliği. | 1,55E+10 | 124000 | 89527 | 4,543308 | 1307296 |
| 18 | El Salvador Cum. | 1,30E+09 | 36100 | 3430 | 2,520439 | 73174 |
| 19 | Estonya | 2,84E+08 | 16900 | 55564 | 3,062953 | 42218 |
| 20 | Habeşistan | 2,80E+09 | 52900 | 3000 | 1,020359 | 134913 |
| 21 | Gürcistan | 6,62E+08 | 25700 | 17100 | 0,529996 | 42000 |
| 22 | Almanya | 2,10E+10 | 145000 | 1030800 | 1,72679 | 2612317 |

EK 1'in Devamı

| | | tkdgr | ktkdgr | mkn | ith | güb |
|---|---|---|---|---|---|---|
| 23 | Gana | 1,91E+09 | 43700 | 3600 | 1,912884 | 10200 |
| 24 | Gine | 6,96E+08 | 26400 | 542 | 1,109374 | 3200 |
| 25 | Macaristan | 1,92E+09 | 43800 | 113500 | 1,236301 | 323162 |
| 26 | Hindistan | 1,09E+11 | 330000 | 1525000 | 4,128088 | 1,74E+07 |
| 27 | Endonezya | 2,40E+10 | 155000 | 70000 | 7,420155 | 2523600 |
| 28 | İrlanda | 3,17E+09 | 56300 | 167000 | 0,85742 | 572980 |
| 29 | İtalya | 2,76E+10 | 166000 | 1650000 | 3,697495 | 1681000 |
| 30 | Jamaika | 4,98E+08 | 22300 | 3080 | 1,448186 | 11700 |
| 31 | Japonya | 5,74E+10 | 240000 | 2028000 | 2,767179 | 1354000 |
| 32 | Kazakistan | 1,93E+09 | 43900 | 49000 | 0,88161 | 50400 |
| 33 | Kenya | 1,83E+09 | 42800 | 12568 | 2,206419 | 144642 |
| 34 | Kore | 1,92E+10 | 139000 | 201089 | 3,17837 | 716700 |
| 35 | Malavi | 5,57E+08 | 23600 | 1430 | 1,710833 | 22756 |
| 36 | Malezya | 7,11E+09 | 84300 | 43300 | 1,335611 | 1130717 |
| 37 | Mauritius | 2,66E+08 | 16300 | 370 | 2,422371 | 37200 |
| 38 | Meksika | 2,34E+10 | 153000 | 324890 | 1,302111 | 1869692 |
| 39 | Mozambik | 8,29E+08 | 28800 | 5750 | 1,280128 | 24900 |
| 40 | Namibia | 2,82E+08 | 16800 | 3150 | 0,588156 | 300 |
| 41 | Nikaragua | 7,11E+08 | 26700 | 2700 | 0,514129 | 22715 |
| 42 | Norveç | 3,07E+09 | 55400 | 133000 | 1,994653 | 191000 |
| 43 | Pakistan | 1,37E+10 | 117000 | 320500 | 4,376097 | 2923320 |
| 44 | Paraguay | 1,46E+09 | 38200 | 16500 | 0,854234 | 66800 |
| 45 | Polonya | 6,04E+09 | 77700 | 1308500 | 1,933423 | 1557000 |

EK 1'in Devamı

| | | tkdgr | ktkdgr | mkn | ith | güb |
|---|---|---|---|---|---|---|
| 46 | Portekiz | 3,56E+09 | 59700 | 169000 | 2,382205 | 228000 |
| 47 | Romanya | 5,32E+09 | 72900 | 164221 | 1,305144 | 327469 |
| 48 | Suudi Arabistan | 9,53E+09 | 97600 | 9925 | 0,985617 | 383760 |
| 49 | Seychelles | 17600000 | 4200 | 40 | 0,492531 | 20 |
| 50 | İspanya | 1,88E+10 | 137000 | 885000 | 1,825936 | 2183200 |
| 51 | St. Lucia | 34800000 | 5900 | 146 | 2,3076 | 5300 |
| 52 | Sudan | 5,37E+09 | 73300 | 11856 | 1,0022 | 79000 |
| 53 | Tanzania | 3,88E+09 | 62300 | 7600 | 2,246578 | 22500 |
| 54 | Trinidad ve Tobago | 1,30E+08 | 11400 | 2700 | 0,731133 | 10865 |
| 55 | Tunisia | 2,31E+09 | 48100 | 35100 | 2,854926 | 108700 |
| 56 | Türkiye | 1,60E+10 | 126000 | 948416 | 3,940213 | 1668650 |
| 57 | Ukrayna | 5,50E+09 | 74200 | 318000 | 1,320311 | 474000 |
| 58 | Britanya | 1,24E+10 | 111000 | 500000 | 1,469574 | 1909000 |
| 59 | Uruguay | 1,19E+09 | 34500 | 33000 | 3,242383 | 119586 |
| 60 | Venezuella | 5,35E+09 | 73100 | 49000 | 1,307306 | 300000 |

Ek 2. X Veri Setindeki Sapan Gözlemlerin Belirlenmesi İçin Mahalanobis Kritik Değerleri

Critical Values for an Outlier on the Predictors as Judged by Mahalanobis $D^2$

| | *Number of Predictors* | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | *k = 2* | | *k = 3* | | *k = 4* | | *k = 5* | |
| *n* | *5%* | *1%* | *5%* | *1%* | *5%* | *1%* | *5%* | *1%* |
| 5 | 3.17 | 3.19 | | | | | | |
| 6 | 4.00 | 4.11 | 4.14 | 4.16 | | | | |
| 7 | 4.71 | 4.95 | 5.01 | 5.10 | 5.12 | 5.14 | | |
| 8 | 5.32 | 5.70 | 5.77 | 5.97 | 6.01 | 6.09 | 6.11 | 6.12 |
| 9 | 5.85 | 6.37 | 6.43 | 6.76 | 6.80 | 6.97 | 7.01 | 7.08 |
| 10 | 6.32 | 6.97 | 7.01 | 7.47 | 7.50 | 7.79 | 7.82 | 7.98 |
| 12 | 7.10 | 8.00 | 7.99 | 8.70 | 8.67 | 9.20 | 9.19 | 9.57 |
| 14 | 7.74 | 8.84 | 8.78 | 9.71 | 9.61 | 10.37 | 10.29 | 10.90 |
| 16 | 8.27 | 9.54 | 9.44 | 10.56 | 10.39 | 11.36 | 11.20 | 12.02 |
| 18 | 8.73 | 10.15 | 10.00 | 11.28 | 11.06 | 12.20 | 11.96 | 12.98 |
| 20 | 9.13 | 10.67 | 10.49 | 11.91 | 11.63 | 12.93 | 12.62 | 13.81 |
| 25 | 9.94 | 11.73 | 11.48 | 13.18 | 12.78 | 14.40 | 13.94 | 15.47 |
| 30 | 10.58 | 12.54 | 12.24 | 14.14 | 13.67 | 15.51 | 14.95 | 16.73 |
| 35 | 11.10 | 13.20 | 12.85 | 14.92 | 14.37 | 16.40 | 15.75 | 17.73 |
| 40 | 11.53 | 13.74 | 13.36 | 15.56 | 14.96 | 17.13 | 16.41 | 18.55 |
| 45 | 11.90 | 14.20 | 13.80 | 16.10 | 15.46 | 17.74 | 16.97 | 19.24 |
| 50 | 12.23 | 14.60 | 14.18 | 16.56 | 15.89 | 18.27 | 17.45 | 19.83 |
| 100 | 14.22 | 16.95 | 16.45 | 19.26 | 18.43 | 21.30 | 20.26 | 23.17 |
| 200 | 15.99 | 18.94 | 18.42 | 21.47 | 20.59 | 23.72 | 22.59 | 25.82 |
| 500 | 18.12 | 21.22 | 20.75 | 23.95 | 23.06 | 26.37 | 25.21 | 28.62 |

## ÖZGEÇMİŞ

| | | |
|---|---|---|
| Doğum Tarihi | 14.05.1979 | |
| Doğum Yeri | Çayeli | |
| Orta Öğrenim | 1991-1993 | Otakçılar Lisesi |
| Lise | 1994-1997 | Yabancı Dil Ağırlıklı Çağlayan Lisesi |
| Lisans | 1997-2001 | İstanbul Üniversitesi İktisat Fakültesi Ekonometri Bölümü |

**Çalıştığı Kurum**

| | |
|---|---|
| 2002-Devam ediyor | Y.T.Ü. Fen Edb. Fakültesi Araştırma Görevlisi |