BÖLÜM 2

2. BASİT DOĞRUSAL REGRESYON: BİR DOĞRUNUN UYARLANMASI

Pek çok durumda, bir değişkenin bir diğeri üzerindeki etkisini ifade etmek üzere bir doğrusal ilişkinin kullanılabileceği daha önce vurgulanmıştı. En basit doğrusal model sadece tek bir bağımsız değişken içerir. Bu model, bağımsız değişkenin değerinin artması ya da azalması durumunda bağımlı değişkenin gerçek ortalamasının sabit bir oranda değiştiğini ifade eder. Bu kısımda verilerin mevcut olduğu durumlarda en küçük kareler metodu ile böyle bir doğru denkleminin nasıl oluşturulabileceği gösterilecektir. Konunun bir örnek üzerinde ele alınması uygun olacaktır. Güneş enerjisinden elde edilen aylık güç (pound) Y ile ortalama atmosfer ısısı (Fahrenheit) X arasındaki ilişkinin araştırıldığı kabul edilsin, (Draper, N. R.; Simith, H.; 1981). Bu iki değişkenin gözlenmiş yirmibeş değeri Tablo 2.1'de verilmiştir. Gözlem çiftlerinin nokta grafiği ise Şekil 2.1'de gösterilmiştir. Oluşturulan regresyon doğrusunun

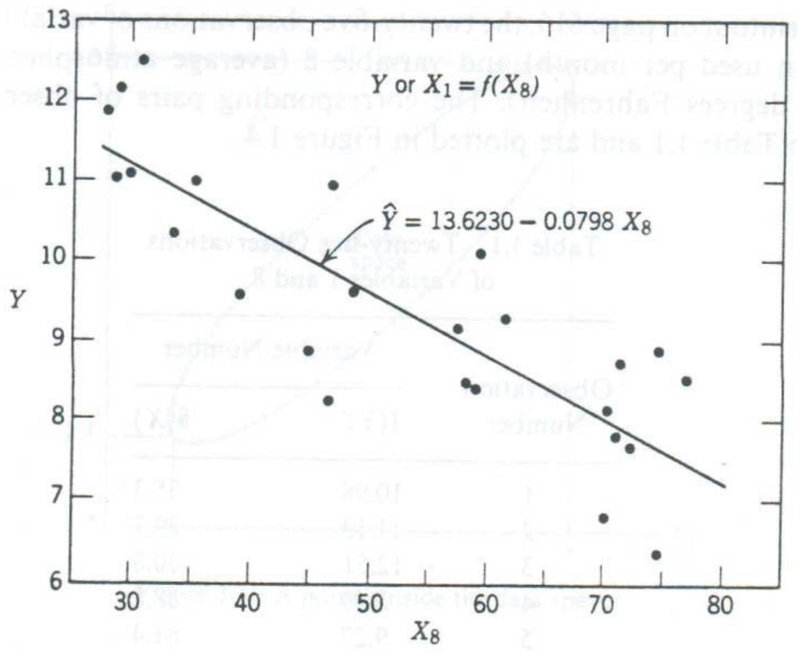
$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \tag{2.1}$$

şeklinde olduğu kabul edilsin. Bu model birinci dereceden olup, parametrelere göre de doğrusaldır.

Tablo 2.1 *Isı ve güç arasındaki ilişkiyi belirlemek için alınan 25 gözlem*

Gözlem Sayısı	Değişkenler	
n	Y	X
1	10,98	35,3
2	11,13	29,7
3	12,51	30,8
4	8,4	58,8
5	9,27	61,4
6	8,73	71,3
7	6,36	74,4
8	8,5	76,7
9	7,82	70,7
10	9,14	57,5
11	8,24	46,4
12	12,19	28,9
13	11,88	28,1
14	9,57	39,1
15	10,94	46,8
16	9,58	48,5
17	10,09	59,3
18	8,11	70
19	6,83	70
20	8,88	74,5
21	7,68	72,1
22	8,47	58,1
23	8,86	44,6
24	10,36	33,4
25	11,08	28,6

Bu modele göre, verilen bir X değerine karşılık gelen Y gözlemi, $\beta_0 + \beta_1 X$ değerine bir ε değerinin ilave edilmesi ile elde edilir. Modeldeki ε değeri nedeni ile herhangi bir Y gözlemi regresyon doğrusunun dışında bulunabilir. Burada; ε, modelin etkisi tamamen ortadan kaldırıldığında Y stokastik değişkeninde, kendi doğal varyasyonundan kaynaklanan bireysel sapmanın simgesidir. Pek çok istatistiksel durumda, ilerleme sağlanabilmesi için bir matematiksel modelin kabulü gereklidir. Seçilen modelin, gerçek modele göre farklılıkları, hata terimi ε' nun bir bileşeni olacağından, seçilen modelin iyileştirme çalışmaları hata terimleri incelenerek yapılmaktadır.



Şekil 2.1 *Veri ve uyumu yapılmış doğru*

Eşitlik (2.1) ile tanımlanan denklemde β_0, β_1 ve ε_i bilinmemektedir ve her bir Y_i gözlemi için bir hata değerinin bulunması gereklidir. Bununla birlikte β_0 ve β_1 sabittir ve X ile Y 'nin tüm olabilir değerleri incelenmeden bu parametre değerleri bulunamazlar. Bu nedenle β_0 ve β_1'i tahminlemek için Tablo 2.1'de verilen yirmibeş gözlemden sağlanan bilgi kullanılacaktır. β_0 ve β_1'in tahminleri b_0 ve b_1 notasyonu ile verilecektir. Tahminlenmek istenen ve eşitlik (2.1) ile verilen modelin, örnekten tahminlenen regresyon fonksiyonu ise,

$$Y_i = b_0 + b_1 X_i + e_i \tag{2.2}$$

ifadesi ile verilebilir. Burada e_i kullanılan modelin doğru olduğu varsayımı altında, ε_i' nin bir tahminidir. Gözlenmiş Y_i değerleri ile regresyon denkleminden hesaplanan $\hat{Y}_i$ değerleri karşılaştırıldığında, model ile veri seti arasındaki uyum için bir ölçüt elde edilir. Bu ölçüt değeri *artık* olarak adlandırılır. Artıklar uyumu yapılan model ile veriler arasındaki farkı tanımlar:

$$e_i = Y_i - \hat{Y}_i \tag{2.3}$$

Modelde sabit terim mevcut olduğunda artıkların toplamı daima sıfırdır. b_0 ve b_1 tahminlendikten sonra, verilen bir X_i değeri için hesaplanan Y_i değeri $\hat{Y}_i$ ile belirtilir ve aşağıdaki formül ile elde edilebilir:

$$\hat{Y}_i = b_0 + b_1 X_i \tag{2.4}$$

Denklem (2.4) bir kestirim ya da uyum denklemi olarak kullanılabilir. Eşitlik (2.4) ile belirli bir X değerine karşılık gelen bağımlı değişken değeri elde edilir. Bu değer daima regresyon doğrusu üzerindedir ve *kestirilmiş değer* ya da *uyumu yapılmış değer* olarak adlandırılır. Uyumu yapılan regresyon doğrusundan hesaplanan her bir değer iki anlama sahiptir:

a) X in belirli bir değeri için, Y nin anakütle ortalamasının tahmini, $E(Y_i)$.

b) X in belirli bir değeri için, Y nin kestirilmiş (uyumu yapılmış) değeri, $\hat{Y}_i$.

Modelin doğru olduğu durumlar için, ileride açıklanacak olan *Gauss-Markov teoremine* göre $\hat{Y}_i$ değeri $E(Y_i)$ nin sapmasız bir tahminleyicisidir.

Gözlenmiş veri iki temel bileşene ayrıştırılabilir: Modelin açıkladığı kısım ve modelin açıklayamadığı kısım.

$$Y_i = \hat{Y}_i + e_i \tag{2.5}$$

$\hat{Y}_i$ bileşeni Y_i gözleminin model tarafından açıklanabilen kısmıdır. e_i ise modelin açıklayamadığı kısımdır.

Regresyon analizinde artıklar ile hata arasındaki farkı kavramak oldukça önemlidir. Artıklar eşitlik (2.3) ile tanımlanmıştır. İleride açıklanacak olan belirli varsayımlar sağlandığında artıklar gözlenmiş (tahminlenmiş) hatalar olarak kabul edilebilir. Regresyon modelindeki bilinmeyen gerçek hata:

$$\varepsilon_i = Y_i - E(Y_i) \tag{2.6}$$

eşitliğinden elde edilir. $E(Y_i)$ ile X_i arasındaki fonksiyonel ilişki,

$$E(Y_i) = \beta_0 + \beta_1 X_i \tag{2.7}$$

olup, X deki birim değişimin Y_i de oluşturduğu değişim oranını tanımlar.

2.1 BASİT DOĞRUSAL REGRESYON İÇİN EKK YÖNTEMİ

Denklem (2.1) kullanılarak hata kareler toplamı;

$$S = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} \left(Y_i - \beta_0 - \beta_1 X_i\right)^2 \tag{2.8}$$

eşitliği ile hesaplanır. Eşitlik (2.8)' den görüldüğü gibi hata terimi kareler toplamı β_0 ve β_1 parametrelerinin bir fonksiyonudur. Diğer bir deyişle,

$$\sum \varepsilon_i^2 = f(\beta_0, \beta_1) \tag{2.9}$$

olduğundan, seçilen her farklı β_0 ve β_1 değerleri $\sum \varepsilon_i^2$ için farklı değerler elde edilmesine neden olacaktır. Bu durum göz önüne alınarak β_0 ve β_1'in tahminleri b_0 ve b_1, bu eşitlikteki S değerlerinin mümkün olan en küçük değerini oluşturacak şekilde seçilir. X_i, Y_i değerleri gözlenmiş sayılardır. β_0 ve β_1 değerleri S'yi minimum yapan değerler olacağından (2.8) eşitliğinin β_0 ve β_1 'e göre türevleri alınır ve elde edilen ifadeler sıfıra eşitlenerek *normal denklemler* elde edilir, bkz. Alıştırma 2.1.

$$b_0 n + b_1 \sum^n X_i = \sum^n Y_i$$

$$b_0 \sum^n X_i + b_1 \sum^n X_i^2 = \sum^n Y_i X_i \tag{2.10}$$

Eşitlik (2.10) ile tanımlanan denklemler b_0 ve b_1 için çözüldüğünde,

$$b_0 = \overline{Y} - b_1 \overline{X} \tag{2.11}$$

$$b_1 = \frac{\sum \left(X_i - \overline{X} \right)\left(Y_i - \overline{Y} \right)}{\sum \left(X_i - \overline{X} \right)^2} \tag{2.12}$$

eşitliği elde edilir, bkz. Alıştırma 2.2. Bu ifadenin payı:

$$\begin{aligned} S_{XY} &= \sum \left(X_i - \overline{X} \right)\left(Y_i - \overline{Y} \right) \\ &= \sum \left(X_i - \overline{X} \right) Y_i = \sum \left(Y_i - \overline{Y} \right) X_i \\ &= \sum X_i Y_i - \frac{\left(\sum X_i \right)\left(\sum Y_i \right)}{n} \\ &= \sum X_i Y_i - n\overline{X}\overline{Y} \end{aligned} \tag{2.13a}$$

Yukarıdaki ifadelerin hepsi birbirine eşittir. Bu eşitlikte $\sum X_i Y_i$ orijine göre (düzeltilmemiş) çarpanlar toplamıdır. $\left(\sum X_i \right)\left(\sum Y_i \right)/n$ terimi ise ortalamaya göre düzeltme terimidir. Aralarındaki fark da X ve Y çarpımlarının ortalamaya göre düzeltilmiş toplamıdır. Eşitlik (2.13*a*)'de Y_i yerine X_i yazılarak payda içinde benzer bir ifade elde edilir.

$$\begin{aligned} S_{XX} &= \sum \left(X_i - \overline{X} \right)^2 \\ &= \sum \left(X_i - \overline{X} \right) X_i \\ &= \sum X_i^2 - \frac{\left(\sum X_i \right)^2}{n} \\ &= \sum X_i^2 - n\overline{X}^2 \end{aligned} \tag{2.13b}$$

$\sum X_i^2$ değeri, X'in orijine göre (düzeltilmemiş) kareler toplamıdır. $\left(\sum X_i \right)^2 / n$ ise ortalamaya göre düzeltilme terimidir. Bu iki terim arasındaki fark ise ortalamaya göre (düzeltilmiş) kareler toplamıdır. Benzer olarak Y_i için ortalamaya göre (düzeltilmiş) kareler toplamı:

$$\begin{aligned} S_{YY} &= \sum \left(Y_i - \overline{Y} \right)^2 \\ &= \sum \left(Y_i - \overline{Y} \right) Y_i \\ &= \sum Y_i^2 - \frac{\left(\sum Y_i \right)^2}{n} \\ &= \sum Y_i^2 - n\overline{Y}^2 \end{aligned} \tag{2.13c}$$

Bu eşitlikler dikkate alınarak b_1 için formül,

$$b_1 = \frac{S_{XY}}{S_{XX}} \tag{2.14}$$

şeklinde yazılabilir.

Modeldeki β_1 parametresi regresyon doğrusunun eğimini verir. Burada dikkat edilmesi gereken nokta eğer tüm X_i değerleri eşit ise $X_i = \bar{X}$ olacak ve (2.12) eşitliğinin paydası sıfır olacaktır. Bunun sonucu olarak da β_0 ve β_1 tahminlerinin elde edilmesi imkansız hale gelecektir. Kısacası bu parametrelerin tahminlenebilmesi için X'in bir değişkenliğe (en az farklı iki değere) sahip olması zorunludur. Bu değişkenliği belirten $\sum(X_i - \bar{X})^2$ değerinin sıfırdan farklı bir değere sahip olması gereklidir. Eğer X'in aralığı X=0 değerini kapsıyor ise, β_0'ın tahmini X=0 noktasında Y'nin ortalaması olarak yorumlanır. Aksi durumda ise β_0 sadece bir regresyon terimi olup dikkatli yorumlanmalıdır.

Tablo 2.1'de verilen verilere yukarıda verilen hesaplamalar uygulanarak,

$$\hat{Y} = 13.623005 - 0.079829X$$

regresyon denklemi elde edilmiştir. Elde edilen regresyon doğrusu Şekil 2.1'de gösterilmiştir. Her bir (X_i Y_i) gözlemi için kestirilmiş değerler ve artıklar Tablo 2.2'de verilmiştir.

Tablo 2.2 *Örnek veri seti için gözlemler, uyumu yapılmış değerler, artıklar*

n	Y_i	$\hat{Y}_i$	e_i
1	10,98	10,81	0,17
2	11,13	11,25	-0,12
3	12,51	11,17	1,34
4	8,4	8,93	-0,53
5	9,27	8,72	0,55
6	8,73	7,93	0,8
7	6,36	7,68	-1,32
8	8,5	7,5	1
9	7,82	7,98	-0,16
10	9,14	9,03	0,11
11	8,24	9,92	-1,68
12	12,19	11,32	0,87
13	11,88	11,38	0,5
14	9,57	10,5	-0,93
15	10,94	9,89	1,05
16	9,58	9,75	-0,17
17	10,09	8,89	1,2
18	8,11	8,03	0,08
19	6,83	8,03	-1,2
20	8,88	7,68	1,2
21	7,68	7,87	-0,19
22	8,47	8,98	-0,51
23	8,86	10,06	-1,2
24	10,36	10,96	-0,6
25	11,08	11,34	-0,26

2.2 SABİT TERİMSİZ MODEL

Eşitlik (2.11) ile tanımlanan b_0 eşitlik (2.4)'de yerine konduğunda,

$$\hat{Y}_i = \bar{Y} + b_1\left(X_i - \bar{X}\right) \tag{2.15}$$

ifadesi elde edilir. Bu eşitlik düzenlenip,

$$Y_i - \hat{Y}_i = \left(Y_i - \overline{Y}\right) - b_1\left(X_i - \overline{X}\right)$$

tüm gözlemler üzerinden toplandığında,

$$\sum\left(Y_i - \hat{Y}_i\right) = \sum\left(Y_i - \overline{Y}\right) - b_1\sum\left(X_i - \overline{X}\right) = 0$$

bulunur. Bu ifadeden de görüleceği üzere artık terimlerinin toplamı sıfırdır. Pratikte, yuvarlama hataları nedeni ile bu toplam tam olarak sıfır değerini vermeyebilir. Modelde β_0 terimi mevcut olduğunda, herhangi bir regresyon problemindeki artıkların toplamı daima sıfırdır. Bir modelde β_0'ın dışlanması, tüm bağımsız değişkenlerin sıfır olması durumunda çıktının da sıfır olacağını belirtir. Bu da genellikle gereksiz olan çok kuvvetli bir kabuldur. Doğrusal regresyon modelinde β_0' ın dışlanması doğrunun $(X,Y) = (0,0)$ noktasından (orjinden) geçtiğini belirtir. Modelden β_0' ın fiziksel olarak çıkarılması, verilerin merkezlenmesi ile her zaman mümkün olabilmektedir. Fakat bu durum β_0' ın sıfıra set edilmesinden tamamen farklıdır. Örneğin, eğer (2.1) denklemi,

$$Y_i - \overline{Y} = \beta_0 + \beta_1\overline{X} - \overline{Y} + \beta_1\left(X_i - \overline{X}\right) + \varepsilon$$

şeklinde veya, $y_i = Y_i - \overline{Y}$, $\beta_0' = \beta_0 + \beta_1\overline{X} - \overline{Y}$ ve $x_i = X_i - \overline{X}$ alınarak,

$$y_i = \beta_0' + \beta_1 x_i + \varepsilon_i$$

olarak yazılırsa, β_1' in EKK tahmini,

$$b_1 = \frac{\sum\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sum\left(X_i - \overline{X}\right)^2} = \frac{\sum\left(x_i - \overline{x}\right)\left(y_i - \overline{y}\right)}{\sum\left(x_i - \overline{x}\right)^2}$$

şeklinde ya da $\overline{x} = 0$ ve $\overline{y} = 0$ olduğu için,

$$b_1 = \frac{\sum x_i y_i}{\sum x_i^2} \tag{2.16}$$

şeklinde olup eşitlik (2.12) ile özdeştir. β_0' ın EKK tahmini ise b_1'in değeri ne olursa olsun, β_0=0 olduğu durumlarda,

$$b_0' = \overline{y} - b_1\overline{x} = 0$$

olarak elde edilir. Bu sonuç merkezleme işleminin her uygulanışında ortaya çıkar. Merkezlenmiş model,

$$y_i = \beta_1 x_i + \varepsilon_i \tag{2.17a}$$

şeklinde β'_0 (kesişim) terimi tamamen ihmal edilerek yazılabilir. Tahminlenen regresyon fonksiyonu ise,

$$y_i = b_1 x_i + e_i \tag{2.17b}$$

şeklinde olup, verilen bir x için y'nin kestirilmiş değeri,

$$\hat{y}_i = b_1 x_i \tag{2.18}$$

ifadesi ile elde edilebilir. Modeldeki tahminlenecek parametre sayısının bir adet azalmasına karşılık $Y_i - \overline{Y}$ değeri gerçekte toplamları sıfır olduğu için sadece (n–1) adet ayrı bilgiyi içermektedir. Buna karşılık merkezlenmemiş modelde $Y_1,\ldots,Y_n$ şeklinde n adet ayrı bilgi içerilmektedir. Bu bilgi kaybı modelde uygun bir düzeltmenin yapılmasında yani kesişim teriminin dışlanmasında kullanılmıştır. Bazı araştırmalarda ise β_0 parametresi bilinmektedir. Bu tip araştırmalarda tahminlenmesi gereken parametre sayısı sadece bir tanedir ve bu nedenle eşitlik (2.8) in sadece β_1 parametresine göre türevi yeterli olacaktır, değişkenler ortalamadan farklara göre yazıldığında:

$$\frac{\partial\left[\sum\left(y_i - \beta_0 - \beta_1 x_i\right)^2\right]}{\partial\beta_1} = \sum\left(y_i - \beta_0 - b_1 x_i\right)x_i = 0$$

b_1 tahminleyicisi genel durum için,

$$b_1 = \frac{\sum x_i y_i - \beta_0 \sum x_i}{\sum x_i^2} \tag{2.19}$$

olarak elde edilir. Orijinden geçen regresyon özel durumu için ise β_0=0 olduğundan eşitlik (2.19), eşitlik (2.16a) ya dönüşecektir.

2.3. TAHMİNLENMİŞ REGRESYONUN HASSASİYETİ

Bu kısımda tahminlenen regresyon doğrusuna ilişkin hassasiyet ölçümünün ne olabileceği sorusu üzerinde durulacaktır. Diğer bir deyişle bağımsız değişkenin bağımlı değişkeni açıklayabilme yeteneği ölçümlenmeye çalışılacaktır. Bu amaç için kullanılabilecek iki temel kriter: Belirlilik katsayısı ve F-testidir. Bu kısımda bu kriter açıklanacaktır. Aşağıda verilen özdeşlik ele alınsın,

$$Y_i = \overline{Y} + \left(\hat{Y}_i - \overline{Y}\right) + \left(Y_i - \overline{Y}\right) \tag{2.20}$$

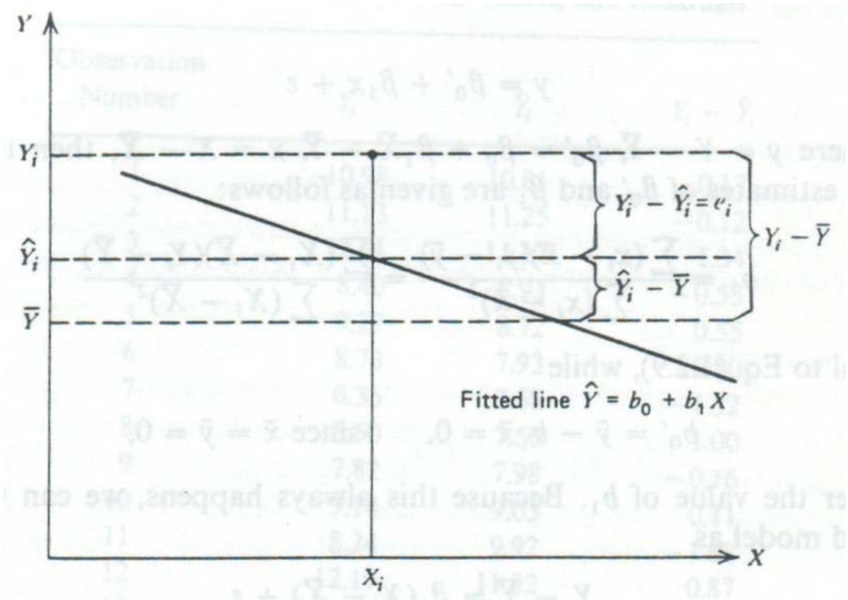
regresyon doğrusu için (2.20) özdeşliğinin geometrik anlamı Şekil 2.2’de verilmiştir. Görüldüğü üzere Y_i şans değişkeni üç bileşene ayrıştırılmıştır: Şans değişkeninin ortalamasının (regresyondaki sabit terimin) etkisi, açıklayıcı değişkenin (regresyonunun) etkisi, artığın etkisi. İlk iki bileşen regresyon modelinin açıklayabildiği, son bileşen ise regresyon modelinin açıklayamadığı kısımdır. Gözlemleri orijine göre ele alan (2.20) eşitliği, ortalamadan farklara göre,

$$Y_i - \overline{Y} = \left(\hat{Y}_i - \overline{Y}\right) + \left(Y_i - \hat{Y}\right) \tag{2.21}$$

şeklinde yeniden yazılabilir. Burada amaç regresyon modelinin etkisini oluşturan iki bileşenden sadece açıklayıcı değişkene diğer bir deyişle regresyona ait olan etkinin araştırılmak istenmesidir. Eğer her iki tarafın karesi alınır ve i=1,…,n gözlem için toplanırsa,

$$\sum\left(Y_i - \overline{Y}\right)^2 = \sum\left(\hat{Y}_i - \overline{Y}\right)^2 + \sum\left(Y_i - \hat{Y}_i\right)^2 \tag{2.22}$$

ifadesi elde edilir. Çapraz çarpan terimi $\sum\left(\hat{Y}_i - \overline{Y}\right)\left(Y_i - \hat{Y}_i\right)$ sıfıra eşittir, bkz Alıştırma 2.3.



Şekil 2.2 *Eşitlik* (2.20)’*in geometrik anlamı*

Eşitlik (2.22) tekrar ele alınsın. $\left(Y_i - \bar{Y}\right)$ değeri *i*-inci gözlemin ortalamadan sapmasıdır ve (2.22)’ nin sol tarafı gözlemlerin ortalamadan farklarının kareler toplamıdır. Başka bir deyişle ortalama etrafındaki kareler toplamıdır ve *Y’* nin düzeltilmiş kareler toplamı olarak da isimlendirilir. Eşitliğin sağındaki ilk terim $\left(\hat{Y}_i - \bar{Y}\right)$, *i*-inci kestirilmiş terimin ortalamadan sapmasını ve $\left(Y_i - \hat{Y}_i\right)$ değeri de *i*-inci gözlemin, kendi kestirilmiş değerinden olan sapmasını belirtmesi nedeni ile (2.22) eşitliği,

Ortalamaya Göre Düzeltilmiş Kareler Toplamı		Regresyon Kareler Toplamı		Artık Kareler Toplamı
$KT(Td)$	=	$KT(R)$	+	$KT(e)$

olarak yazılabilir. Buradan da görüleceği gibi Y ’nin kendi ortalaması etrafındaki değişkenliğinin bir kısmı modeldeki regresyon terimine bir kısmı ise artık terimine bağlı olmak üzere iki temel bileşene ayrılabilmektedir. Bunlardan birincisi Y_i gözlemleri üzerinde açıklayıcı değişkenlerin etkisini belirten regresyonun değişkenliği, ikincisi ise tüm etkili faktörler sabit bir noktada tutulduğunda Y_i değişkeninin modelden bağımsız olan değişkenliğidir. Sonuç olarak gözlemlerin tümünün regresyon doğrusu üzerinde olmadığı söylenebilir. Eğer bu durum gerçekleşseydi artık kareler toplamı sıfır olacaktı.

Sabit terimsiz model için de yukarı açıklanan duruma benzer bir ayrışım gerçekleştirilebilir. Şekil 2.3 bu ayrışımı göstermektedir. Orijinden geçen regresyonda başlangıç noktası (0, 0) yerine $\left(\bar{X}, \bar{Y}\right)$ noktasına taşınmaktadır. Bunun sonucunda her hangi bir $y_i = Y_i - \bar{Y}$ noktası,

$$y_i = \hat{y}_i - e_i \tag{2.23a}$$

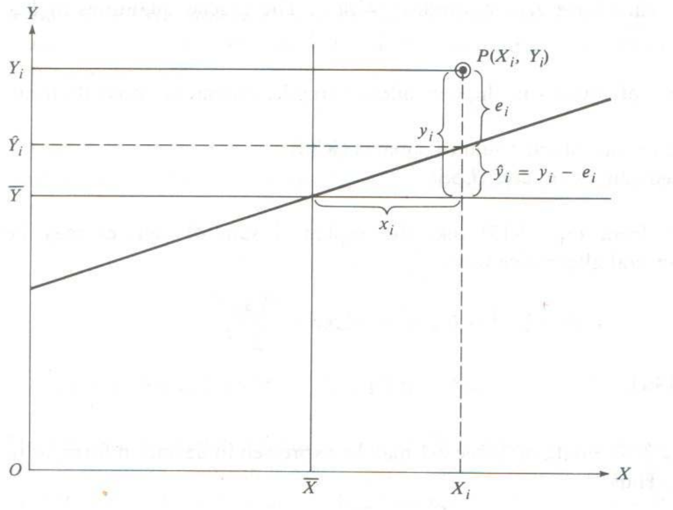
şeklinde tanımlanabilir. Burada uyumu yapılan değer,

$$\hat{y}_i = \hat{Y}_i - \bar{Y} \tag{2.23b}$$

ile hesaplanabilir. Eşitlik (2.23a) nın, eşitlik (2.21) den elde edilebileceği görülmektedir. Eğer her iki tarafın karesi alınır ve i=1,…,n gözlem için toplanırsa,

$$\sum y_i^2 = \sum \hat{y}_i^2 + \sum e_i^2 \tag{2.24}$$

ifadesi elde edilir. Çapraz çarpan terimi $\sum \hat{y}_i e_i$ sıfıra eşittir,



Şekil 2.3 *Orijinin* (0,0) *noktasından* $(\overline{X}, \overline{Y})$ *noktasına kaydırılması.*

2.3.1 Belirlilik Katsayısı

Bu kısımda verilen bir veri seti için uyumu yapılan regresyon doğrusunun uyum iyiliği (goodness of fit) ile ilgilenilecektir. Bu nedenle de veriler ile uyumu yapılan örnek regresyon doğrusu arasındaki ilişki incelenecektir. Eğer tüm gözlemler regresyon doğrusu üzerinde ise mükemmel bir uyum sağlanmış olur. Fakat bu durumla nadiren karşılaşılır. Genelde pozitif ve negatif e_i'ler mevcuttur. Bu nedenle regresyon doğrusu çevresindeki bu e_i'lerin mümkün olduğunca küçük değerli olması istenir. Belirlilik katsayısı r^2 (iki değişkenli regresyon) veya R^2 (çoklu regresyon) veriler ile regresyon doğrusu arasındaki uyumun bir ölçümünü verir.

Bu kriterden yararlanılarak elde edilen regresyon doğrusunun faydası (önemi) değerlendirilebilir. Bunun içinde ortalamaya göre kareler toplamının değeri ve diğer iki bileşenin bu toplamdan aldıkları pay araştırılmalıdır. (2.22) eşitliğinin her iki tarafı $\Sigma(Y_i - \overline{Y})^2$ ile bölündüğünde,

$$1 = \frac{\Sigma(\hat{Y}_i - \overline{Y})^2}{\Sigma(Y_i - \overline{Y})^2} + \frac{\Sigma(Y_i - \hat{Y}_i)^2}{\Sigma(Y_i - \overline{Y})^2}$$

elde edilir. Buna uygun olarak, belirlilik katsayısı,

$$r^2 = \frac{\sum(\hat{Y}_i - \overline{Y})^2}{\sum(Y_i - \overline{Y})^2} = \frac{KT(R)}{KT(Td)} \tag{2.25a}$$

elde edilir. Regresyon kareler toplamının artık kareler toplamından çok daha büyük olması, diğer bir deyişle r^2 değerinin birden çok küçük olmaması arzu edilir. r^2 kriteri $\overline{Y}$ etrafındaki toplam değişkenliğin regresyon tarafından açıklanan kısmını ölçer. Belirlilik katsayısı olarak adlandırılır. Belirlilik katsayısının kare kökü r ise Y ile $\hat{Y}_i$ arasındaki korelasyondur ve çoklu korelasyon katsayısı olarak adlandırılır. r^2'nin hesaplanabileceği diğer formüller aşağıda verilmiştir.

$$r^2 = \frac{\sum \hat{y}_i^2}{\sum y_i^2} \tag{2.25b}$$

$$r^2 = \frac{b_1^2 \sum x_i^2}{\sum y_i^2} \tag{2.25c}$$

$$r^2 = 1 - \frac{KT(e)}{KT(Td)} \tag{2.25d}$$

$$r = \frac{\left(\sum x_i y_i\right)}{\sqrt{\sum x_i^2 \sum y_i^2}} \tag{2.25e}$$

Belirlilik katsayısının Özellikleri:

a) r^2, negatif olmayan bir değerdir.

b)Bu değerin limitleri $0 \leq r^2 \leq 1$ şeklindedir. $r^2=1$ ise mükemmel bir uyum olduğu, $r^2=0$ ise bağımlı değişken ile bağımsız değişken arasında ilişki olmadığı anlamına gelir.

c) R^2 modelde β_0 hariç diğer terimlerin katkısının bir ölçümünü verir.

d) Eğer saf hata (tekrarlı gözlemler) mevcut ise r^2 kesinlikle 1 değerini alamaz.

e)Eğer saf hata yok ise r^2 değeri, β_0 parametresini içeren bir modeldeki parametre sayısına eşit uygun seçilmiş gözlemlere tam bir uyum sağlanarak 1 olacak şekilde belirlenebilir.

r^2 değeri verilerdeki değişkenliğin açıklanmasında regresyon denkleminin başarısının bir ölçüsü olarak kullanılır. Bu nedenle modele yeni bir terim eklenmesine bağlı olarak r^2 de oluşan iyileşmenin sadece modele eklenen parametre sayısındaki artıştan kaynaklanmadığı gerçek bir anlama sahip olduğundan emin olunmalıdır. Belirlilik katsayısı r^2'deki artış yapay olabilir. Bu değerin yüksek çıkmasının iki temel sebebi vardır: regresyonun eğim değeri büyük olabilir ya da bağımsız değişkenin değerlerinin yayılımı $X_1,\ldots,X_n$ fazla olabilir, bkz Alıştırma 2.4.

2.3.2 Varyans Analizi ve F-Testi

Varyans analizi üç ve daha fazla ana kütle ortalamasının eşitliğinin test edilmesi amacıyla kullanılan bir istatistiksel analiz yöntemidir. Analizi gerçekleştirmek amacıyla bir tablo oluşturulur. Bu tablonun bileşenleri (sütunları); *Değişkenlik Kaynağı*, *Serbestlik Derecesi*, *Kareler Toplamları*, *Kareler Ortalaması* ve *F-testidir*. *F*-testinin uygulanabilmesi için şans değişkeninin normal dağılışa uygun bir dağılıma sahip olması gerekir.

Regresyon analizinde incelenecek olan değişkenlik kaynakları ve bunlara ait kareler toplamları Kısım 2.3 te açıklanmıştır. Aşağıda her kareler toplamına ait olan serbestlik dereceleri elde edilecektir.

İstatistikte *serbestlik derecesi*, eldeki veri setindeki gözlemlerin taşıdığı birbirinden bağımsız bilgi sayısı olarak tanımlanır. Herhangi bir kareler toplamı, kendisine ait serbestlik derecesi ile birlikte ele alınır. Serbestlik derecesi, kareler toplamını derlemek için gerekli olan $Y_1,\ldots,Y_n$ şeklinde n adet bağımsız sayının, ne kadar bağımsız bilgi parçasını içerdiğini belirtir. Örneğin ortalamaya göre düzeltilmiş kareler toplamı için (n-1) adet bağımsız bilgi parçası gereklidir. $\left(Y_1-\bar{Y},\ldots Y_n-\bar{Y}\right)$ sayılarının sadece (n-1) tanesi bağımsızdır. Çünkü bu n adet sayının toplamı sıfırdır. Regresyon kareler toplamı $Y_1,\ldots,Y_n$'nin bir tek fonksiyonu $\sum\left(\hat{Y}_i-\bar{Y}\right)^2=b_1^2\sum\left(X_i-\bar{X}\right)^2$ şeklinde hesaplanabilir, bu nedenle basit regresyon için bir serbestlik derecesine sahiptir. Çok değişkenli durum için gerek duyulan serbestlik derecesi ise regresyona ait (β_0 hariç) parametre sayısı kadar olacaktır. Elde edilen bu iki serbestlik derecesini birbirinden çıkararak artık kareler toplamı için (n-2) serbestlik derecesi elde edilir. Genelde, artık kareler toplamının serbestlik derecesi (gözlem sayısı-tahminlenmiş parametre sayısı) şeklinde elde edilir. Eşitlik (2.22) için serbestlik derecesi paylaşımı,

$$n-1=1+(n-2) \tag{2.26}$$

şeklinde gösterilebilir.

Regresyon analizinde artık serbestlik derecesi, örnek hacminin artıklar üzerine oluşturulan kısıtlama sayısından farkı olarak da düşünülebilir. Basit doğrusal regresyon için, artıklar üzerine oluşturulan iki kısıt vardır:

$$\sum\left(Y_i-\hat{Y}_i\right)=0$$

$$\sum\left(Y_i-\hat{Y}_i\right)X_i=0$$

Bu kısıtlar normal denklemlerin bir sonucudur.

Eşitlik (2.22) ve (2.26) dikkate alınarak bir Varyans Analiz Tablosu oluşturulabilir. Oluşturulan varyans analiz tablosu, Tablo 2.3'de verilmiştir. Kareler ortalaması sütunu, kareler toplamlarının serbestlik derecesine bölümü ile elde edilir.

Tablo 2.3 *Basit regresyon için Varyans Analiz Tablosu* I

Değişkenlik Kaynağı	*Kareler Toplamları*	*Serbestlik Derecesi*	*Kareler Ortalaması*	*F-Testi*
Regresyona Bağlı Değişkenlik	$\sum\left(\hat{Y}_i-\bar{Y}\right)^2$	1	$KO(R)=\dfrac{\sum\left(\hat{Y}_i-\bar{Y}\right)^2}{1}$	$F=\dfrac{KO(R)}{KO(e)}$
Artığa Bağlı Değişkenlik	$\sum\left(Y_i-\hat{Y}_i\right)^2$	n-2	$KO(e)=\dfrac{\sum\left(Y_i-\hat{Y}_i\right)^2}{n-2}$	
Toplam Düzeltilmiş Değişkenlik	$\sum\left(Y_i-\bar{Y}\right)^2$	n-1		

Tablo 2.3'ü oluşturmanın alternatif bir yolu da kareler toplamlarının eşitlik (2.20) ye göre elde edilerek ortalamaya bağlı değişkenliğin bir bileşen olarak varyans analiz tablosuna eklenmesidir. Eşitlik (2.20) nin karesi alınıp, n adet gözlem için toplanarak:

$$\sum Y_i^2=n\bar{Y}^2+\sum\left(\hat{Y}_i-\bar{Y}\right)^2+\sum\left(Y_i-\hat{Y}_i\right)^2 \tag{2.27}$$

TK(T) = KT(b_0)+ KT(R) + KT(e)

Eşitlik (2.27) elde edilirken ortaya çıkan tüm çapraz çarpım terimleri sıfırdır. Elde edilen yeni varyans analiz tablosu Tablo 2.4 de gösterilmiştir.

Tablo 2.4 *Varyans Analiz Tablosu* II

Değişkenlik Kaynağı	*Kareler Toplamları*	*Serbestlik Derecesi*	*Kareler Ortalaması*	*F-Testi*
Ortalamaya Bağlı Değişkenlik	$n\overline{Y}^2$	1	$KO(b_0) = \dfrac{n\overline{Y}^2}{1}$	
Regresyona Bağlı Değişkenlik	$\sum(\hat{Y}_i - \overline{Y})^2$	1	$KO(R) = \dfrac{\sum(\hat{Y}_i - \overline{Y})^2}{1}$	$F = \dfrac{KO(R)}{KO(e)}$
Artığa Bağlı Değişkenlik	$\sum(Y_i - \hat{Y}_i)^2$	*n*-2	$KO(e) = \dfrac{\sum(Y_i - \hat{Y}_i)^2}{n-2}$	
Toplam Değişkenlik	$\sum Y_i^2$	*n*		

Bir diğer alternatif yaklaşım ise eşitlik (2.22) ün solundaki toplam düzeltilmiş kareler toplamının iki bileşene ayrıştırılması ile,

$$\sum(Y_i - \overline{Y})^2 = \sum Y_i^2 - n\overline{Y}^2 \qquad (2.28a)$$

elde edilir. Eşitliğin sağındaki ikinci bileşen *düzeltme faktörü* olarak adlandırılır ve β_0 ın yaptığı katkıyı temsil eder. İlk bileşen ise *Toplam Kareler Toplamıdır*. Sonuç olarak eşitlik (2.28*a*) iki bileşenin toplamı olarak,

$$\sum Y_i^2 = \sum(Y_i - \overline{Y})^2 + n\overline{Y}^2 \qquad (2.28b)$$

elde edilir, bkz. Alıştırma 2.5. Kareler toplamlarının bu şekilde elde edilmesine uygun bir varyans analiz tablosu da oluşturulabilir, bkz Alıştırma 2.6.

Artık kareler toplamı nadiren doğrudan hesaplanır. Genelde düzeltilmiş kareler toplamından, regresyon kareler toplamının çıkartılması ile elde edilir. Regresyon kareler toplamının elde edilmesi için aşağıdaki eşitliklerin her hangi biri kullanılabilir:

$$KT(R) = \sum(\hat{Y}_i - \overline{Y})^2 = b_1\left\{\sum(X_i - \overline{X})(Y_i - \overline{Y})\right\} = b_1 S_{XY} \qquad (2.29a)$$

$$= \frac{\left\{\sum(X_i - \overline{X})(Y_i - \overline{Y})\right\}^2}{\sum(X_i - \overline{X})^2} = \frac{S_{XY}^2}{S_{XX}} \qquad (2.29b)$$

$$= \frac{\left\{\sum X_i Y_i - \left(\sum X_i\right)\left(\sum Y_i\right)/n\right\}^2}{\sum X_i^2 - \left(\sum X_i\right)^2/n} = \frac{S_{XY}^2}{S_{XX}} \qquad (2.29c)$$

$$= \frac{\left\{\sum(X_i - \overline{X})Y_i\right\}^2}{\sum(X_i - \overline{X})^2} \qquad (2.29d)$$

KT(*R*) notasyonu, b_0 mevcut iken b_1 için kareler toplamı anlamına da gelmektedir.

Artık kareler ortalaması modelin *tahminlenmiş varyansı* olarak da adlandırılır. İstatistikte örnek varyansını elde etmek için ilk olarak,

$$\sum\left(Y_i-\overline{Y}\right)^2$$

hesaplanır. Bu değer bir kareler toplamıdır. Kareler toplamı kendi serbestlik derecesine (n-1) bölünür. Kaybedilen bir serbestlik derecesi bilinmeyen anakütle ortalamasının tahminlenmesinde kullanılmıştır. Diğer bir ifade ile kaybedilen bir serbestlik derecesinin nedeni artıklar $\sum\left(Y_i-\overline{Y}\right)$ üzerinde oluşturulan kısıtlamadır.

$$\sum\left(Y_i-\overline{Y}\right)=0$$

Sonuç olarak örnek varyansı,

$$s^2=\frac{\sum_{i=1}^{n}\left(Y_i-\overline{Y}\right)^2}{n-1}$$

tanımlanır. Örnek varyansı, serbestlik derecesine bölünmüş bir kareler toplamı olduğundan bir kareler ortalaması olarak da adlandırılır.

Regresyon modeli için her bir Y_i gözleminin varyansı her bir hata teriminin ε_i varyansı ile eşittir.

$$\sigma^2\left(Y_i\right)=\sigma^2\left(\beta_0+\beta_1X_i+\varepsilon_i\right)=\sigma^2\left(\varepsilon_i\right)=\sigma^2$$

Y_i gözlemleri X_i seviyelerine bağlı olarak farklı ortalamalı farklı olasılık dağılımlarından gelir. Bu dağılımların ortalamaları $\hat{Y}_i$ ile tahminlendiği için örnek varyansının formülünde $\overline{Y}$ yerine $\hat{Y}_i$ yazılır. Model parametrelerinin β_0 ve β_1 tahminlenmesi için iki serbestlik derecesi kaybedilir. Eğer model doğru ise, artık kareler ortalaması rassal hatanın (ε) varyansının (σ^2) sapmasız bir tahminleyicisidir.

$$s^2=\frac{KT\left(e\right)}{n-2}=\frac{\sum e_i^2}{n-2} \tag{2.30}$$

Hata varyansı σ^2, regresyon varyansı σ_{YX}^2'e eşit olabilir veya olmayabilir. Eğer kabul edilen model doğru ise $\sigma^2=\sigma_{YX}^2$'dir. Kabul edilen model doğru değil ise $\sigma^2\langle\sigma_{YX}^2$ eşitsizliği ortaya çıkar. Regresyon varyansı σ_{YX}^2'in bir tahmini olan artık kareler ortalaması s^2, eğer model doğru ise σ^2'nin bir tahminini verir, modelin yanlış olduğu durumlarda ise σ^2'nin bir tahmini olarak kabul edilemez. Eğer $\sigma^2\langle\sigma_{YX}^2$ ise model yanlıştır veya uyum yetersizliği mevcuttur. Bu durumların hangisinin mevcut olduğunun araştırılması ileride açıklanacaktır.

Regresyonun kapsamlı olarak anlamlılığı varyans analizi kullanılarak F-testi ile gerçekleştirilir. Bu amaca uygun olarak test edilecek hipotez:

$$H_0:\beta_1=0$$

$$H_1:\beta_1\neq 0$$

Varyans analiz tablosunda elde edilen kareler toplamları birbirinden bağımsız ki-kare şans değişkenleridir. Bu nedenle regresyon kareler ortalamasının artık kareler ortalamasına oranı bir F-değişkeni tanımlayacaktır, bkz. Bölüm 2 Ekler.

$KO(e)$ nin beklenen değeri σ^2 olup, bkz. Bölüm 2 Ekler, bu sonuç X ile Y nin ilişkili olup olmamasına diğer bir deyişle $\beta_1=0$ olup olmamasına bağımlı değildir. $\beta_1=0$ olduğunda $KO(R)$ nin beklenen değeri de σ^2 dir. Ayrıca $\beta_1=0$ ise tüm Y_i ler aynı ortalama $\mu=\beta_0$ ve aynı varyansa σ^2 sahiptir. Bununla birlikte $\beta_1\neq0$ ise $E[KO(R)]$ değeri σ^2 den büyüktür, bkz. Bölüm 2 Ekler. Sonuç olarak $\beta_1=0$ olup olmadığının testi $KO(R)$ ve $KO(e)$ değerleri kullanılarak gerçekleştirilir. Sonuç olarak F-testi

$$F_{1,n-2}=\frac{\chi_1^2/1}{\chi_{n-2}^2/n-2}=\frac{KT(R)/1}{KT(e)/(n-2)}=\frac{KO(R)}{KO(e)} \tag{2.31}$$

olarak tanımlanır.

Belirlilik katsayısı ve F-istatistiği arasındaki ilişki; eşitlik (2.25a) ve (2.25d) kullanılarak, $KT(R)=r^2KT(Td)$ ve $KT(e)=(1-r^2)KT(Td)$ elde edilir. Sonuçlar eşitlik (2.31) da yerine konarak,

$$F_{1,n-2}=\frac{r^2/1}{\left(1-r^2\right)/n-2} \tag{2.32a}$$

sonucu bulunur. F-dağılımı ile t-dağılımı arasındaki ilişki, bkz Bölüm 2 Ekler, kullanılarak örnek belirlilik katsayısının sıfırdan farklı olmadığını tanımlayan boş hipotez, eşitlik (2.32a) den elde edilen,

$$t=\frac{r\sqrt{(n-2)}}{\sqrt{\left(1-r^2\right)}} \tag{2.32b}$$

test istatistiği ile test edilir.

Daha önce verilen örneğin verileri kullanılarak, regresyon kareler toplamı eşitlik (2.29c)'den,

$$KT(R)=\frac{\left\{\sum X_iY_i-\left(\sum X_i\right)\left(\sum Y_i\right)\ /n\right\}}{\sum X_i^2-\left(\sum X_i\right)^2\ \ /n}=45.5924$$

olarak elde edilir. Düzeltilmiş kareler toplamı,

$$\sum Y_i^2-\left(\sum Y_i\right)^2\ \ /n=63.8158$$

şeklindedir. σ_{YX}^2'in 23 serbestlik dereceli bir tahmini s^2 = 0.7925 olarak elde edilmiştir. Sonuçlar Tablo 2.5'de özetlenmiştir.

Tablo 2.5 *Örnek verileri için varyans analiz tablosu*

DK	*sd*	*KT*	*KO*	*F-Değeri*
Model	1	45.5924	45.5924	57.54
Hata	23	18.2234	s^2 =0.7923	
Toplam; düzeltilmiş	24	63.8158		

2.3.3 Sabit Terimsiz Model İçinVaryans Analizi ve F-Testi

Gerçekte orijinden geçen regresyon modeli *kesişim parametresi bilen modellerin* özel bir durumudur. Eğer β_0 biliniyorsa ya da β_0=0 ise eşitlik (2.22) ile verilen kareler toplamlarının bileşenleri,

$$\sum(y_i-\beta_0)^2=\sum(\hat{y}_i-\beta_0)^2+\sum(y_i-\hat{y}_i)^2 \tag{2.34}$$

şeklinde elde edilir. Bu eşitlikteki çapraz çarpan sıfır olduğundan,

$$\sum(\hat{y}_i-\beta_0)(y_i-\hat{y}_i)=b_1\sum x_i(y_i-\hat{y}_i)=0$$

yazılabilir. Bu yaklaşımda $\overline{y}$ yerine β_0 etrafındaki değişkenliğin ayrıştırılması ile ilgilenilmektedir. Eşitlik (2.34) ün sağındaki ilk bileşen,

$$\sum(\hat{y}_i-\beta_0)^2=b_1^2\sum x_i^2 \tag{2.35}$$

regresyon kareler toplamıdır. Normal dağılış varsayımı altında ve β_1=0 koşulu ile bir serbestlik dereceli ki-kare dağılışı gösterir. Orijinden geçen regresyon için tahnimlenmiş hata varyansı,

$$s^2=\frac{\sum(y_i-\hat{y}_i)^2}{n-1} \tag{2.36}$$

formülünden elde edilir. Serbestlik derecesindeki fark tahminlenmesi gerekli parametre sayısının bir azalmasıdır. Bu bileşen s^2 den bağımsız olduğundan H_0: β_1=0 boş hipotezi için bir F-testi,

$$F=\frac{b_1^2\sum x_i^2}{s^2} \tag{2.37}$$

tanımlar. Kritik değer ise $F_{(1,n\text{-}1)}$ olup F-tablosundan bulunur.

2.3.4 Sabit Terimsiz Model İçin Belirlilik Katsayısı

Orijinden geçen regresyonda eşitlik (2.24) kullanılarak orijinden geçen regresyon için,

$$R_0^2=\frac{\sum\hat{y}_i^2}{\sum y_i^2}=\frac{KT(R_0)}{KT(T_0)} \tag{2.38a}$$

bulunur. Orijinden geçen regresyon için elde edilen bu istatistiğn zayıf noktası kesişimli modeller için elde edilen R^2 ile bir kıyaslama olanağı sağlamamasıdır. Uyum hassasiyeti çok iyi olmasa dahi R_0^2 istatistiğinin R^2 den büyük olma yönünde bir eğilimi vardır. Bu özelliğin nedeni ise düzeltilmemiş kareler toplamından elde edilmesidir. Sonuç olarak artık kareler toplamları yaklaşık olarak birbirine denk olan durumlar için R_0^2 istatistiği R^2 değerinden önemli ölçüde büyük olabilecektir.

Rakip modelleri karşılaştırma imkanı verecek şekilde R_0^2 için alternatif hesaplama yöntemleri geliştirilmiştir. Tercih edilebilecek bir istatisitik,

$$R_{0*}^2=1-\frac{\sum(y_i-\hat{y}_i)^2}{\sum(y_i-\overline{y})^2} \tag{2.38b}$$

burada,

$$\sum\left(y_i-\bar{y}\right)^2=\sum y_i^2-\frac{\left(\sum x_i y_i\right)^2}{\sum x_i^2}$$

olup bu istatistik karşılaştırma açısından daha uygundur. Bununla birlikte $\sum\left(y_i-\hat{y}_i\right)^2$ bileşeninin göreli olarak büyük olduğu durumlarda R_{0*}^2 istatistiği negatif değerler alabilir.

2.4 REGRESYON DOĞRUSUNUN ÖZELLİKLERİ

Buraya kadar anlatılanlar dikkate alınarak bu özellikler maddeler halinde özetlenecektir.

1) Regresyon doğrusu X ve Y 'nin örnek ortalamalarından geçer, bkz. Alıştırma 2.7.

2) β_0 ın mevcut olduğu modellerde tahminlenmiş $\hat{Y}_i$ değerlerinin ortalaması, gözlenmiş Y_i değerlerinin ortalamasına eşittir, bkz. Alıştırma 2.8.

$$\bar{\hat{Y}}=\bar{Y} \tag{2.39}$$

3) β_0 ın mevcut olduğu modellerde Alıştırma 2.1 den görüleceği gibi hata terimlerinin toplamı $\sum e_i=0$ 'dır. Bunun sonucu olarak hata terimlerinin ortalaması da $\bar{e}=0$ olacaktır.

4) Artıklar, kestirilmiş değerler ile ilişkili değildir, bkz. Alıştırma 2.9.

5) Alıştırma 2.1 den görüleceği gibi hata terimleri ile açıklayıcı değişkenler ilişkisizdir.

2.5 EKK TAHMİNLEME YÖNTEMİNİN VARSAYIMLARI

Daha önce belirtildiği üzere bir regresyon modelinin parametrelerinin tahminlenebilmesi için hata terimlerinin ortaya çıkışları ile ilgili kesin varsayımlara gerek duyulmaktadır. Bunun nedenini görmek için regresyon denkleminin incelenmesi yeterli olacaktır. Modelden görülebileceği gibi Y_i hem X_i hem de ε_i'ye bağımlıdır. Öyleyse X_i ve ε_i 'nin nasıl oluştuğu veya ortaya çıktığı belirlenmedikçe Y_i ile ilgili herhangi bir istatistiksel yorum yapabilmenin olanağı yoktur. X_i ve ε_i ile ilgili varsayımlar regresyon tahminlerinin geçerli yorumlarının yapılabilmesine olanak verir. Temel regresyon yaklaşımında bağımsız değişkenlerin gözlenen sabit değerler olduğu, her hangi bir istatistiksel dağılıma sahip olmadıkları varsayılır. Bununla birlikte hatalar bir şans değişkenidirler ve ait oldukları dağılımla ilgili kesin varsayımlara gereksinim vardır. Bu varsayımlar aşağıda verilmiştir. Hatalar;

1. Ortalaması sıfır,

$$E(\varepsilon_i)=0 \tag{2.40a}$$

2. Sabit varyanslı,

$$V(\varepsilon_i)=\sigma^2 \tag{2.40b}$$

3. Birbiri ile ilişkisiz,

$$Cov(\varepsilon_i,\varepsilon_j)=0 \tag{2.40c}$$

şans değişkenleridir. Bu varsayımlar EKK metodu kullanılarak parametrelerin sapmasız tahminlerinin elde edilmesi için yeterlidir. Fakat bilindiği gibi regresyon analizinin amacı sadece β_0 ve β_1 'i elde etmek değil aynı zamanda β_0 ve β_1 ile ilgili istatistiksel yorumlamaları oluşturmaktır. Örneğin b_0 ve

b_1'in anakütle parametrelerine olan yakınlığı bilinmek veya $\hat{Y}_i$ ile parametre değeri $E(Y_i)$ arasındaki ilişki belirlenmek istenir. Daha genel olarak *hipotez testleri* ve *güven aralıkları* için ek bir varsayıma gereksinim vardır:

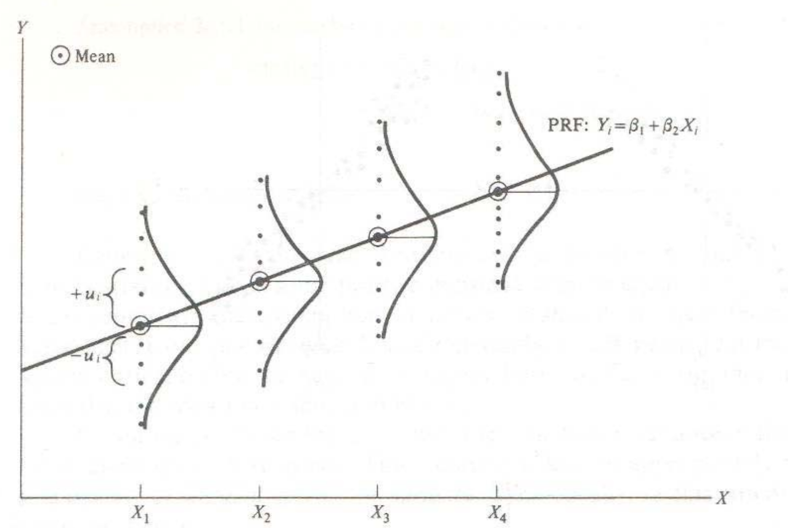
4. Hatalar normal dağılım gösterir. (2.40*d*)

Bu varsayımlara ait açıklamalar aşağıda verilmiştir.

Varsayım 1. Hataların ortalaması sıfırdır. Eşitlik (2.40*a*) için daha uygun bir gösterim,

$$E\left(\varepsilon_i / X_i\right)=0$$

olup ε_i 'nin şartlı beklenen değerinin sıfır olduğunu belirtir. Geometrik olarak bu varsayım Şekil 2.4'de verilmiştir. Verilen bir X değerine karşılık gelen her bir Y değeri kendi anakütle ortalama değeri çevresinde dağılır. Bir diğer deyişle belirli X değerleri için Y_i'ler kendi anakütle dağılımlarına uygun değerler almaktadır. Ancak, Y_i değerlerinin anakütle ortalama ve varyansları bağımsız değişkenin bir fonksiyonudurlar. Bu fonksiyona regresyon modeli adı da verilmektedir. Bu Y_i değerlerinin bazıları ortalamanın altında bazıları da ortalamanın üstündedir. Ortalamanın üstünde ve altında oluşan bu farkların toplamı sıfırdır. Verilen herhangi bir X'e karşılık gelen bu farkların ortalama değeri eşitlik (2.40*a*) a göre sıfır olacaktır. Bu varsayım hatanın, Y'nin ortalama değerini sistematik (düzenli) olarak etkilemediğini belirtir. Bu rassal etki sistematik olmayan hata teriminin ortaya çıkmasına neden olur. Pozitif ε_i değerlerinin negatif ε_i değerleri ile toplamı sıfırdır. Bunun sonucu olarak Y üzerindeki ortalama etkileri sıfırdır. $E\left(\varepsilon_i / X_i\right)=0$ varsayımı, $E\left(Y_i / X_i\right)=\beta_0+\beta_1 X_i$ olduğunu belirtir. Sonuç olarak iki eşitlik birbirine denktir.

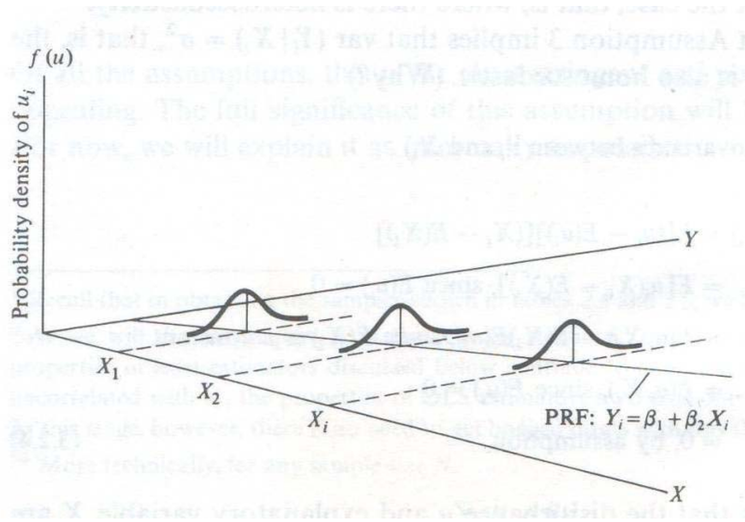


Şekil 2.4 *Hata terimlerinin şartlı dağılışı*

Varsayım 2: Hatalar eşit varyanslıdır. Eşitlik (2.40*b*) in şartlı dağılıma göre ifadesi,

$$\begin{aligned} V\left(\varepsilon_i / X_i\right) &= E\left[\varepsilon_i-E\left(\varepsilon_i\right)\right]^2 \\ &= E\left(\varepsilon_i^2\right) \\ &= \sigma^2 \end{aligned}$$

Eşitlik (2.40*b*), verilen her bir X_i için ε_i'nin varyansının pozitif bir sabit sayıya (σ^2'ye) eşit olduğunu belirtir. Teknik olarak (2.40*b*) eşit varyansı ifade eder. Başka bir deyişle (2.40*b*) çeşitli *X* değerlerine karşılık gelen *Y* anakütlelerinin aynı sabit varyansa sahip olduğunu belirtir. Bu durum geometrik olarak Şekil 2.5'de verilmiştir.

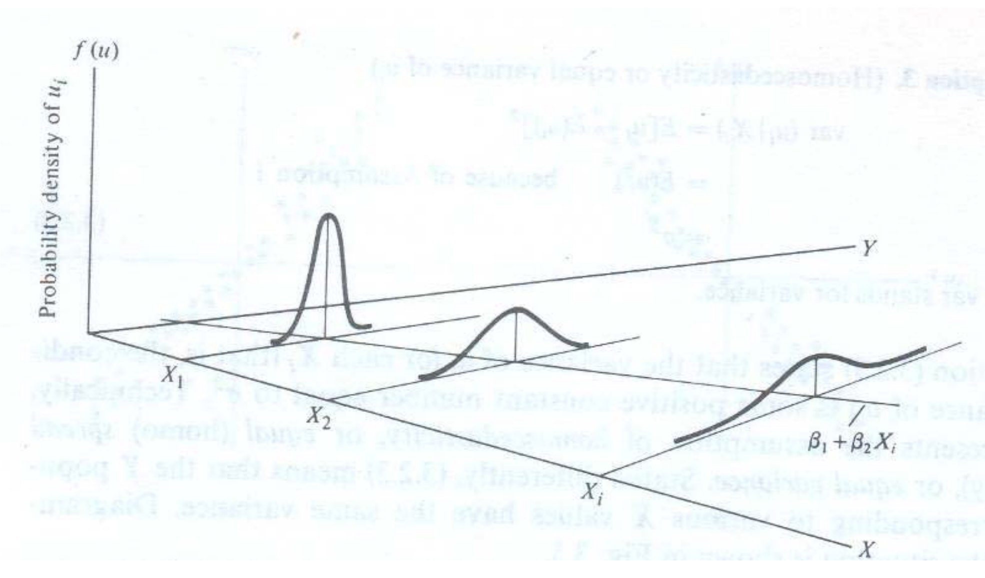


Şekil 2.5 *Hata terimlerinin* (*şartlı*) *eşit varyanslı dağılışı*

Yukarıda verilenlerin tam tersi bir durum olarak Şekil 2.6 incelenebilir. Burada *Y* anakütlesinin şartlı varyansı *X* değerleri arttıkça artmaktadır. Bu durum sabit olmayan varyanslılık veya farklı varyanslılık olarak bilinir. Bu durum sembolik olarak,

$$V\left(\varepsilon_i / X_i\right) = \sigma_i^2$$

şeklinde belirtilebilir. σ^2'nin alt indisi *i*, *Y* anakütlesinin varyansının sabit olmadığını belirtir. Kısaca çeşitli *X* değerlerine karşılık gelen tüm *Y* değerleri eşit güvenirlilikte olmayabilecektir. Burada sözü edilen güvenilirlik, ortalamaları etrafında dağılmış olan *Y* değerlerinin ne kadar uzak veya yakın olduklarını değerlendirmek amacı ile kullanılmıştır.



Şekil 2.6 *Farklı varyanslılık*

Varsayım 3. Hatalar arasında otokorelasyon yoktur.

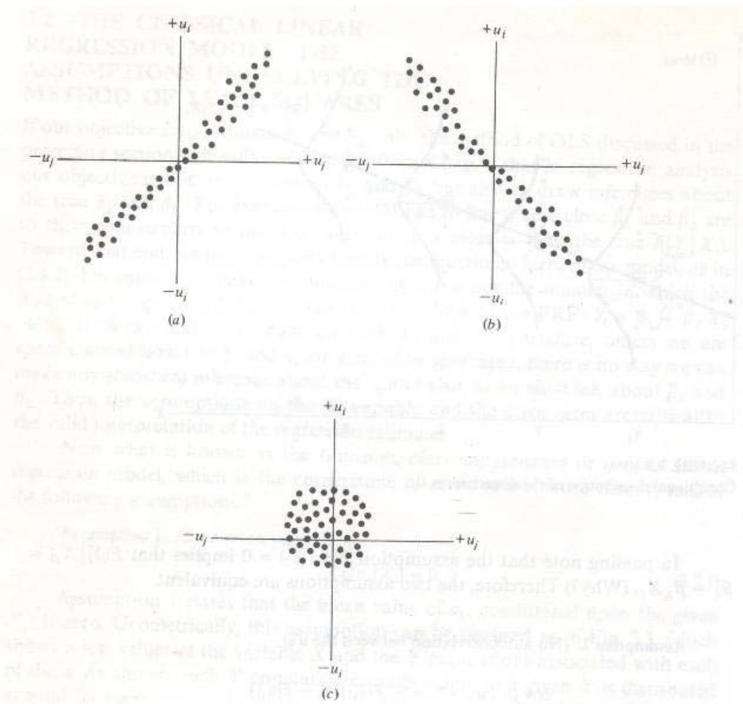
$$Cov(\varepsilon_i;\varepsilon_j) = E\Big[\big(\varepsilon_i - E(\varepsilon_i)\big)\big(\varepsilon_j - E(\varepsilon_j)\big)\Big]$$

$$= E\big[\varepsilon_i \varepsilon_j\big]$$

$$= 0 \qquad \text{her } i \neq j \text{ için.}$$

Burada i ve j iki farklı hatayı ifade eder. Bu varsayım ε_i ve ε_j hatalarının ilişkisiz olduğunu belirtir. Başka bir deyişle dizisel korelasyon veya otokorelasyonun olmadığını belirten varsayımdır. Bunun anlamı, verilen bir X_i için, herhangi iki Y değerinin ortalamadan sapmalarının oluşturduğu çiftlerin Şekil 2.7*a* ve 2.7*b*'de gösterildiği şekilde olmaması gerektiğidir. Şekil 2.7*a* hata terimlerinin pozitif doğrusal ilişkili olduğunu, başka bir deyişle bir pozitif ε'yi yine bir pozitif ε veya negatif ε'yi yine bir negatif ε'nin izlediğini belirtir. Şekil 2.7*b* ise ε'lerin negatif doğrusal ilişkili olduğunu, pozitif bir ε'yi negatif ε'nin izlediğini yada bunun tam tersinin geçerli olduğunu göstermektedir. Bu varsayım, hata terimlerinin sistematik bir davranış içinde olmadığını ifade eder.

Eğer hata çiftleri Şekil 2.7*a* ve 2.7*b*'de gösterildiği şekilde sistematik bir şekilde dağılıyorlarsa otokorelasyon (dizisel korelasyon) mevcuttur. Varsayım 3'nin sağlanabilmesi için bu tip korelasyonların ortadan kaldırılması gerekmektedir. Şekil 2.7*c*, hata çiftleri arasında sistematik bir doğrusal ilişkinin olmadığı, sıfır doğrusal ilişkili bir durum göstermektedir. Varsayım 3 aynı zamanda $Cov(Y_i;Y_j)=0$, $i \neq j$ için, olduğunu da belirtir. Bu varsayım kısaca aşağıdaki şekilde açıklanabilir. Regresyon fonksiyonunun $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ şeklinde olduğu ve ε_t ile ε_{t-1}'in pozitif ilişkili olduğu kabul edilsin. Bu durumda Y_i sadece X_i'ye değil aynı zamanda ε_{t-1}'e de bağımlıdır. Çünkü ε_t'nin bir kısmı ε_{t-1} tarafından belirlenmektedir.



Şekil 2.7 *Hata terimleri arasındaki korelasyon a*) *Pozitif korelasyon*, *b*) *Negatif korelasyon*, c) *Sıfır korelasyon*

Bu varsayımların sağlandığı durumlarda EKK tahmincileri (b_0, b_1, s^2) sapmasızlık, minimum varyanslılık gibi bazı istatistiksel özelliklere sahiptir. Aşağıdaki kısımda bu konu ile ilgili bir teorem ve ispatları verilecek, normallik varsayımı daha sonra incelenecektir.

2.6 GAUSS MARKOV TEOREMİ

Bu teorem, doğrusal regresyon modellerinde en küçük kareler yönteminin yaygın olarak kullanılmasına olanak sağlayan önemli bir teoremdir.

Gauss-Markov Teoremi: Parametrelerin en küçük kareler tahminleri, parametrelerin; doğrusal ve sapmasız tahmincileri içerisinde minimum varyanslı olanıdır.

Bu teorem, kabuller zayıf veya eksik olduğu durumlarda da izlenebildiği için önemli bir teoremdir. Başka bir deyişle hata teriminin dağılışı ile ilgili kabuller yapılmasına gerek yoktur.

Bu önemli teoremi yorumlamak amacı ile β_1'in en küçük kareler tahmincisi b_1 ele alınsın. Teoremin ispatında görüleceği gibi bu tahminci, doğrusal bir tahmincidir. Bu sınıfa giren tahmincilerin anlaşılması ve analizinin kolay olması nedeni ile çalışma doğrusal tahminciler ile sınırlı tutulur. Bir başka sınırlamada doğrusal tahmincilerin sapmasız olmasıdır. En küçük kareler tahmincileri bu tahminciler içerisinde minimum varyanslı olanıdır. Bu nedenle *en iyi doğrusal sapmasız tahminci* olarak adlandırılır.

Gauss-Markov teoremi ilginç bir önermeye sahiptir. Regresyonun özel bir durumu olan, Y bağımlı değişkenin β_1=0 olacak şekilde açıklanması durumunda β_0, Y'nin anakütle ortalamasına eşit olacaktır ve eşitlik (2.11) den görülebileceği gibi aynı zamanda en küçük kareler tahmincisidir. Buna göre bir anakütle ortalamasının en küçük kareler tahmincisi örnek ortalamasıdır. Gauss-Markov teoremine göre bu ifade, aritmetik ortalama bir anakütle ortalamasının en iyi doğrusal sapmasız tahmincisidir, şeklinde açıklanabilir.

Gauss-Markov teoreminin sadece, hem doğrusal hem de sapmasız tahmincilere uygulandığı vurgulanmalıdır. En küçük kareler tahmincilerinden daha iyi (daha küçük varyansa sahip) sapmalı ve doğrusal olmayan bir tahminci mevcut olabilir. Örneğin, örnek medyanı anakütle ortalamasını tahminleyen, doğrusal olmayan bir tahmincidir. Normal olmayan anakütle tipleri için örnek ortalamasından daha iyi bir tahmincidir. Örnek medyanı, parametrik olmayan istatistikler olarak bilinen doğrusal olmayan tahminciler için iyi bir örnektir.

Gauss-Markov teoreminin ispatı: Bu teorem basit doğrusal regresyon durumundaki b_0 ve b_1 tahmincileri için ispatlanacaktır. Çoklu regresyon durumu Bölüm 4'te ele alınacaktır.

1) İlk olarak b_1 tahmincisinin doğrusallığı ele alınsın. Burada doğrusallık, tahmincinin Y 'nin doğrusal bir fonksiyonu olduğu anlamında kullanılmıştır. Bu durumda,

$$b_1 = \sum_{i=1}^{n} c_i \cdot Y_i$$

yazılabilmesi gereklidir. c_i değerleri bir sabiti belirtmektedir. Eşitlik (2.12)'den,

$$b_1 = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2}$$

$$b_1 = \frac{\sum(X_i - \bar{X})Y_i \text{ -} \bar{Y}\sum(X_i - \bar{X})}{\sum(X_i - \bar{X})^2} = \frac{\sum(X_i - \bar{X})Y_i}{\sum(X_i - \bar{X})^2}$$

elde edilir. Bu ifade,

$$b_1 = \sum\left(\frac{X_i - \bar{X}}{\sum(X_i - \bar{X})^2}\right) \cdot Y_i$$

şeklinde düzenlenir ve bu eşitlikte

$$c_i = \frac{X_i - \bar{X}}{\sum(X_i - \bar{X})^2} \quad (2.41)$$

olarak alınır,

$$b_1 = \sum c_i Y_i \quad (2.42)$$

şeklinde ispat tamamlanır. Şimdi b_0'ın doğrusallığı ele alınsın. b_0 için (2.11) eşitliği kullanılarak,

$$b_0 = \bar{Y} - b_1\bar{X}$$

$$b_0 = \frac{\sum Y_i}{n} - b_1\bar{X}$$

elde edilir. b_1'in doğrusallık özelliği kullanılarak,

$$b_0 = \sum\frac{1}{n}Y_i - \sum c_i Y_i \bar{X}$$

$$b_0 = \sum\left(\frac{1}{n}Y_i - c_i\bar{X}Y_i\right)$$

$$b_0 = \sum\left(\frac{1}{n} - c_i\bar{X}\right)Y_i$$

ifadesi bulunur.

$$k_i = \frac{1}{n} - c_i\bar{X} \quad (2.43)$$

alınarak

$$b_0 = \sum k_i Y_i \quad (2.44)$$

elde edilip ispat tamamlanır.

2) İspatın ikinci aşamasında b_0 ve b_1 tahmincilerinin β_0 ve β_1 parametrelerinin sapmasız tahminleyicileri oldukları ispatlanacaktır. İlk olarak b_1 ele alınsın. Bu tahminleyici,

$$b_1 = \frac{\sum(X_i - \bar{X})Y_i}{\sum(X_i - \bar{X})^2}$$

şeklindeydi. Eşitliğin her iki tarafının beklenen değeri alınarak,

$$E(b_1)=\frac{\sum\left(X_i-\bar{X}\right)\,E(Y_i)}{\sum\left(X_i-\bar{X}\right)^2}$$

$$E(b_1)=\frac{\sum\left(X_i-\bar{X}\right)(\beta_0+\beta_1X_i)}{\sum\left(X_i-\bar{X}\right)^2}=\frac{\beta_0\sum\left(X_i-\bar{X}\right)+\beta_1\sum\left(X_i-\bar{X}\right)X_i}{\sum\left(X_i-\bar{X}\right)^2}$$

$$=\beta_1\frac{\sum\left(X_i-\bar{X}\right)X_i}{\sum\left(X_i-\bar{X}\right)^2}$$

$$E(b_1)=\beta_1 \tag{2.45}$$

elde edilip b_1'in sapmasız tahminleyici olduğu ispatlanmış olur. Şimdi de b_0 tahmincisi ele alınsın, bu tahminleyici

$$b_0=\frac{1}{n}\sum Y_i-b_1\bar{X}$$

olarak verilmiştir. Bu ifadenin beklenen değeri alınarak,

$$E(b_0)=\frac{1}{n}\sum E\left(Y_i\right)-E\left(b_1\cdot\bar{X}\right)$$

$$E\left(b_0\right)=\frac{1}{n}\sum\left(\beta_0+\beta_1X_i\right)-\beta_1\bar{X}$$

$$E(b_0)=\frac{1}{n}\left(n\beta_0+\beta_1\sum X_i\right)-\beta_1\bar{X}$$

$$E(b_0)=\beta_0 \tag{2.46}$$

elde edilerek, b_0 tahminleyicisinin β_0 parametresinin sapmasız bir tahmincisi olduğu ispatlanmış olur. EKK tahminleyicilerinin sapmasızlığı için X_i ler sabit (rassal olmayan) değişkenler olmalı ve $E\left(\varepsilon_i\right)=0$ olmalıdır.

3) Gauss-Markov teoremi ile belirtilen minimum varyanslılık özelliği, doğrusallık ve sapmasızlık özelliğini ön koşul olarak almaktadır. Bu aşamada sadece b_1 tahmincisinin doğrusal sapmasız tahminciler sınıfında minimum varyanslı tahminci olduğu ispatlanıp b_0'ın ispatı alıştırma olarak bırakılacaktır.

b_1^* tahmincisi β_1 parametresinin b_1'den farklı başka bir doğrusal sapmasız tahmincisi olsun. Doğrusallık özelliği nedeni ile,

$$b_1^*=\sum\left(c_i+f_i\right)Y_i=b_1+\sum f_iY_i \tag{2.47}$$

yazılabilir. Burada c_i eşitlik (2.41) ile tanımlanmıştır. Sapmasızlık özelliği ile,

$$E(b_1^*)=\beta_1$$

olmalıdır. Eşitlik (2.47)'un beklenen değeri alınarak,

$$E(b_1^*)=E(b_1)+E\left[\sum f_iY_i\right]$$

$$=\beta_1+E\left[\sum f_i Y_i\right]$$

elde edilir. Eşitlik (2.47)

$$b_1^*=\sum\left(c_i+f_i\right)\left(\beta_0+\beta_1 X_i+\varepsilon_i\right)$$

şeklinde yazılıp açılımı yapılarak,

$$b_1^*=\beta_0\sum\left(c_i+f_i\right)+\beta_1\sum\left(c_i+f_i\right)X_i+\sum\left(c_i+f_i\right)\varepsilon_i$$

elde edilir. Burada b_1^*'*in* sapmasız olabilmesi için $\sum\left(c_i+f_i\right)=0$ olması gereklidir, bu da $\sum f_i=0$ olması ile mümkündür. Ayrıca $\sum\left(c_i+f_i\right)X_i=1$ olmalıdır. Bu durumun gerçekleşebilmesi için $\sum f_i X_i=0$ olması gereklidir, bkz Alıştırma 2.15. Yukarıda verilenler dikkate alınarak,

$$b_1^*=\beta_1+\sum\left(c_i+f_i\right)\varepsilon_i \tag{2.48}$$

yazılabilir. b_1 tahmincisinin varyansı,

$$\begin{aligned}V(b_1)&=V\left(\sum c_i Y_i\right)\\&=\sum V\left(c_i Y_i\right)\\&=\sigma^2\sum c_i^2\end{aligned} \tag{2.49}$$

olarak elde edilebilir. b_1^* tahmincisinin varyansı,

$$V\left(b_1^*\right)=E\left[b_1^*-\beta_1\right]^2$$

olup eşitlik (2.48) kullanılarak,

$$\begin{aligned}E\left[\left(b_1^*-\beta_1\right)\right]^2&=E\left[\sum\left(c_i+f_i\right)\varepsilon_i\right]^2\\&=E\left[\sum_{i=j}\left(c_i+f_i\right)^2\varepsilon_i^2+2\sum_{i\neq j}\left(c_i+f_i\right)\left(c_j+f_j\right)\varepsilon_i\varepsilon_j\right]\end{aligned}$$

elde edilir. Varsayımlar nedeniyle ikinci terim sıfıra eşit olduğundan,

$$\begin{aligned}E\left[b_1^*-\beta_1\right]^2&=E\left[\sum\left(c_i+f_i\right)^2\varepsilon_i^2\right]\\&=\sigma^2\sum\left(c_i+f_i\right)^2\\&=\sigma^2\sum\left(c_i^2+f_i^2+2c_i f_i\right)\end{aligned}$$

ve $\sum c_i f_i=0$ olduğundan,

$$\begin{aligned}E\left[b_1^*-\beta_1\right]^2&=\sigma^2\left[\sum\left(c_i^2+f_i^2\right)\right]\\&=\sigma^2\sum c_i^2+\sigma^2\sum f_i^2\end{aligned}$$

elde edilir. Eşitlik (2.49)'de, $Var\left(b_1\right)=\sigma^2\sum c_i^2$ olarak elde edilmişti. Bunun sonucu olarak,

$V\left(b_1^*\right)=V\left(b_1\right)+\sigma^2\sum f_i^2$ elde edilir. $\sum f_i^2 \geq 0$ olduğundan $V\left(b_1^*\right) \geq V\left(b_1\right)$ olup b_1 tahmincisi minimum varyanslıdır. EKK tahminleyicilerinin minimum varyanslılığı için X_i ler sabit (rassal olmayan) değişkenler olmalı ve $V\left(\varepsilon_i\right)=\sigma^2$ ve $Cov\left(\varepsilon_i;\varepsilon_j\right)=0$ olmalıdır.

2.7 KLASİK REGRESYON MODELİ: NORMALLİK VARSAYIMI

Hatırlanacağı gibi EKK yönteminin uygulamaları için klasik doğrusal regresyon modelinde ε_i hata teriminin olasılık dağılımı ile ilgili herhangi bir varsayım yapılmamıştı. Yapılan varsayımlar sadece, ε_i'nin beklenen değerinin sıfır, varyansının sabit ve birbirleri ile ilişkisiz oldukları şeklinde idi. Bu varsayımlar ile EKK tahmincilerinin (b_0, b_1, s^2) sapmasızlıkları, minimum varyanslılık gibi bazı istatistiksel özelliklere sahip oldukları ispatlanmıştır. Eğer amaç sadece nokta tahminlenmesi ise EKK yönteminin ilk üç varsayımı yeterli olacaktır. Fakat nokta tahminlemesi istatistiksel yorumlamanın sadece bir konusunu oluşturmaktadır, diğer önemli konular ise hipotez testleri ve güven aralıklarıdır.

Genellikle sadece parametrelerin tahminlenmesi ile değil aynı zamanda bu parametrelerle ilgili yorumların yapılması ile de ilgilenilmektedir. Amaç, tahminleme olduğu kadar hipotez testlerinin de oluşturulması olduğundan, ε_i hata terimlerinin olasılık dağılımının belirlenmesine ihtiyaç vardır. Buna niçin gerek olduğu sorusunun cevabı ise zor değildir. Daha önce belirtildiği gibi EKK tahmincileri b_0 ve b_1'in her ikisi de ε_i'nin doğrusal fonksiyonudur. Bu tahminciler gerçekte bağımlı değişken Y'nin doğrusal bir fonksiyonudur. Fakat Y şans değişkeninin kendisi (2.1) denkleminden görüleceği gibi ε_i'nin doğrusal bir fonksiyonudur. Eşitlik (2.42)'den,

$$\begin{aligned} b_1 &= \sum c_i\left(\beta_0+\beta_1 X_i+\varepsilon_i\right) \\ &= \beta_0\sum c_i+\beta_1\sum c_i X_i+\sum \varepsilon_i c_i \\ &= \beta_1+\sum \varepsilon_i\ c_i \end{aligned} \tag{2.50}$$

bu tahmincilerin de ε_i'nin doğrusal bir fonksiyonu olduğu görülmektedir. Benzer bir eşitlik b_0 için eşitlik (2.44) kullanılarak elde edilebilir. Bu durumda EKK tahmincilerinin olasılık veya örnekleme dağılımları, ε_i'nin olasılık dağılımı ile ilgili olarak yapılan varsayımlara bağımlı olacaktır. Bu tahmincilerin olasılık dağılımı, onların anakütle değerleri (parametreler) ile ilgili yorumların yapılabilmesi açısından gerekli olduğu için, ε_i'nin olasılık dağılımının hipotez testlerinde ve güven aralıklarında çok önemli bir rolü olduğu kabul edilir.

EKK yöntemi, ε_i'nin olasılık dağılımı ile ilgili herhangi bir varsayım yapmamaktadır. Yapılacak böyle bir varsayım örnek istatistiklerinden, anakütle parametreleri ile ilgili yorumlar yapılmasına yardımcı olabilecektir. Bu eksiklik ε_i'lerin bir olasılık dağılımı izlediği varsayımı ile giderilebilir. Regresyon analizinde genellikle ε_i'lerin *normal dağılış* izlediği varsayılır. Sonuç olarak bu varsayımların tümü kısaca,

$$\varepsilon_i \sim N\left(0;\sigma^2\right) \tag{2.51a}$$

olarak verilebilir.

Normal dağılmış iki şans değişkenin kovaryans veya korelasyonunun sıfır olması durumunda bu iki şans değişkeni birbirinden bağımsızdır. Buna göre normallik kabullerinden (2.51*a*), ε_i ile ε_j'nin sadece ilişkisiz olduğunu değil aynı zamanda da birbirinden bağımsız dağıldığı anlamına gelmektedir.

Normallik varsayımı; b_0 (normal), b_1 (normal) ve s^2 (ki-kare)'nin olasılık dağılımlarını elde etmek için yeterlidir. Ayrıca ε_i'nin sıfır ortalama ve σ^2 varyanslı normal dağıldığı varsayımı sonucunda Y_i değerleride ε_i'nin doğrusal fonksiyonu olduğundan,

$$Y_i \sim N(\beta_0 + \beta_1 X_i,\ \sigma^2) \tag{2.51b}$$

şeklinde normal dağılış gösterecektir. Normallik varsayımının birkaç nedeni aşağıda verilmiştir.

1) Daha önce belirtildiği gibi ε_i regresyon modelinde içerilmeyen bir çok bağımsız değişkenin bağımlı değişken üzerindeki ortak etkisini göstermektedir. Bu ihmal edilen değişkenlerin etkisinin küçük olması ümit edilir. Merkezi Limit Teoremi yardımı ile, çok sayıda bağımsız ve özdeş dağılmış şans değişkeninin toplamlarının, birkaç istisna hariç, normal dağılış göstereceği açıklanabilir. Bu durumda Merkezi Limit Teoremi ε_i'lerin normallik varsayımının teorik bir ispatını sağlar.

Eğer bir hata terimi ε, birkaç değişkenlik kaynağından oluşan hataların toplamı ise, hataların olasılık dağılımlarının ne olabileceği konusunda bir sorun yoktur. Onların toplamı olan ε, bileşen sayısı arttıkça daha çok normal dağılıma yaklaşan bir dağılıma sahip olacaktır (Merkezi Limit Teoremine göre). Pratikte bir deneysel hata, bir ölçüm hatası nedeniyle oluşabilir. Bunun sonucu olarak hatalarda sistematik bir yapı ortaya çıkabilir. Bu gibi durumlarda normallik kabulü uygun değildir. Çalışmalarda hata terimleri incelenerek bu kabulün doğruluğunun araştırılması gerekir.

2) Merkezi Limit Teoremine göre, değişken sayısı çok büyük olmasa da veya bu değişkenler tamamen bağımsız olmasalar bile, onların dağılımı yine de yaklaşık normal olabilecektir.

3) Normallik varsayımı ile EKK tahmincilerinin olasılık dağılımı kolayca elde edilebilir. Çünkü normal dağılışın bir özelliği, normal dağılış gösteren bir değişkenin doğrusal bir fonksiyonunun da normal dağılış göstereceğini belirtir. Hataların normal dağıldığı varsayımı altında b_0 ve b_1 tahmincileri de normal dağılış gösterecektir.

4) Son olarak, normal dağılım diğer dağılımlara göre basit bir dağılım olup sadece iki parametre içermektedir (ortalama ve varyans) ve iyi bilinen bir dağılımdır.

2.8 PARAMETRE TAHMİNLERİ İLE KESTİRİM DEĞERLERİNİN VARYANSLARI VE ÖRNEKLEME DAĞILIMLARI

Herhangi bir şans değişkeni yardımı ile hesaplanan değerin kendisi de şans değişkenidir. Buna göre $\overline{Y}$, $\hat{Y}$, e, b_0 ve b_1 tahminleri Y_i şans değişkenlerinden hesaplandığı için birer şans değişkenidirler ve hepsi Y_i'nin doğrusal fonksiyonudurlar. Bu nedenle tahmincilerin varyansları, bir doğrusal fonksiyonun varyansının temel tanımı kullanılarak belirlenebilir. Y_i'ler şans değişkeni, a_i'ler sabitler olmak üzere $U = a_1Y_1 + a_2Y_2 + ... + a_nY_n = \sum a_iY_i$ ifadesi verilmiş olsun. U şans değişkeninin varyansı için genel formül,

$$V(U) = \sum a_i^2 \ V(Y_i) + \sum_{i \neq j}\sum a_i a_j Cov(Y_i, Y_j) \tag{2.52}$$

şeklindedir. Genel regresyon modelinde kabul edildiği gibi şans değişkenlerinin birbirinden bağımsız olması halinde kovaryansların hepsi sıfır olacak ve ikinci terim dikkate alınmayacaktır. Ayrıca şans değişkenleri eşit varyanslı ise, $V(Y_i)=\sigma^2$ gibi, doğrusal fonksiyonun varyansı,

$$V(U) = a_1^2V(Y_1) + a_2^2V(Y_2) + ... + a_n^2V(Y_n) = \left(\sum a_i^2\right) \sigma^2 \tag{2.53}$$

şeklindedir.

Bilindiği gibi β_1 eşitlik (2.12) ile tahminlenmekteydi. Bu eşitlik,

$$b_1 = \left\{\left(X_1 - \bar{X}\right)Y_1 + ... + \left(X_n - \bar{X}\right)Y_n\right\} \Big/ \sum\left(X_i - \bar{X}\right)^2$$

şeklinde de yazılabilir. b_1 için verilen ifadede X_i değerleri sabit oldukları için,

$$c_i = \left(X_i - \bar{X}\right) \Big/ \sum\left(X_i - \bar{X}\right)^2$$

olup tahmincinin varyansı,

$$\sigma^2(b_1) = \frac{\sigma^2}{\sum\left(X_i - \bar{X}\right)^2} = \frac{\sigma^2}{S_{XX}} \tag{2.54}$$

olarak elde edilebilir. b_1'in standart sapması, varyansın karekökü alınarak bulunur.

$$\sigma\ (b_1) = \frac{\sigma}{\left\{\sum\left(X_i - \bar{X}\right)^2\right\}^{1/2}}$$

Eğer σ bilinmiyor ise modelin doğru olduğu kabul edilerek onun yerine tahmini olan s kullanılır. b_1'in standart sapması,

$$s(b_1) = \frac{s}{\left\{\sum\left(X_i - \bar{X}\right)^2\right\}^{1/2}} \tag{2.55}$$

şeklindedir. Tahminlenmiş standart sapma için alternatif bir terminolojide standart hata ifadesidir. Şans değişkenleri için standart sapma, örnek istatistikleri ve tahminciler için ise standart hata ifadeleri kullanılır. Sonuç olarak b_1 tahmincisi:

$$b_1 \sim N\left[\beta_1; \frac{\sigma^2}{\sum\left(X_i - \bar{X}\right)^2}\right] \tag{2.56}$$

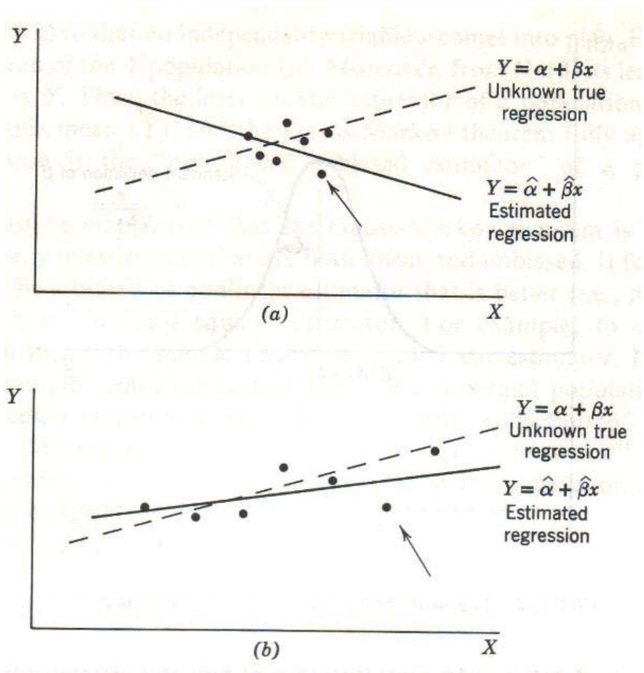
b_1 normal dağıldığı için standardize istatistik $\left(b_1 - \beta_1\right)/\sigma\left(b_1\right)$ bir standart normal değişkendir. Bununla birlikte $\sigma(b_1)$ parametresi bilinmediği için $s(b_1)$ ile tahminlenir. Sonuç olarak belirlenmesi gereken $\left(b_1 - \beta_1\right)/s\left(b_1\right)$ istatistiğinin dağılımıdır. Bu dağılım n-2 serbestlik dereceli bir,

$$\frac{b_1 - \beta_1}{s(b_1)} = t_{n-2} \tag{2.57}$$

t-dağılımıdır, bkz Alıştırma 2.10.

Eşitlik (2.54) ile verilen b_1'in varyansının değerlendirilmesi ele alınacaktır. X_i değerinin birbirine yakın seçilmesi nedeni ile deneyin kötü tasarlandığı kabul edilsin. Bunun sonucunda $\left(X_i - \overline{X}\right)$ sapmaları ve dolayısıyla $\sum\left(X_i - \overline{X}\right)^2$ değeri küçük olacaktır. Böylece eşitlik (2.54) de verilen b_1'in varyansı büyük değer alacak ve b_1 göreceli olarak güvenilir olmayan bir tahminci olacaktır. X gözlemlerinin bir arada toplanması, araştırılan doğrunun hatalar nedeni ile olduğundan farklı tahminlenmesine ve doğrunun eğimini veren b_1'in güvenilir olmamasına neden olur. Şekil 2.8*a*'da elde edilen doğru hatalar ve X gözlemlerinin bir arada olması nedeni ile (özellikle okla belirtilen hata) gerçek durumu yansıtmamaktadır. Şekil 2.8*b*' de ise X gözlemleri tanımlı olduğu aralığa daha iyi dağılmışlardır. Bu nedenle hatalar aynı olduğu halde b_1 çok daha güvenilirdir. Çünkü hatalar artık aynı ağırlık noktasını kullanmamaktadır.

Herhangi bir veri biriktirilmeden önce, X_i değerlerinin Y_i gözlemlerinin alınabileceği noktalarda seçilmesi ve bu seçiminde $V(b_1)$'i minimize edecek bir şekilde olması istenir. Daha sonra seçilen bu X_i'ler $\sum\left(X_i - \overline{X}\right)^2$ değerini maksimize edecektir. Bu problemin teorik cevabı bazı X_i değerlerinin hem artı hem de eksi sonsuzda gözlemlenebileceğidir. Bunun pratik yorumu X_i değerinin, X bölgesinde deney tasarımının mümkün olduğu uç noktalarda yer alabileceğidir.



Şekil 2.8 (*a*) X_i *değerlerinin* (*uzayının*) *birbirine yakın olması nedeni ile güvenilir olmayan tahmin*, (*b*) *Birbirinden uzaklaşan* X_i'ler *nedeni ile daha güvenilir tahmin*.

Sabit terim, eşitlik (2.11) ile tahminlenmişti. Bu eşitlikteki şans değişkenleri $\bar{Y}$ ve b_1 olup katsayıları 1 ve $(-\bar{X}\ )$ 'dır. Denklem (2.52) kullanılarak b_0'ın varyansı,

$$V(b_0) = V(\bar{Y}) + (-\bar{X})^2\ V(b_1) + 2\ (-\bar{X})\ Cov(\bar{Y}, b_1) \tag{2.58}$$

şeklinde elde edilebilir. Bilindiği gibi $V(\bar{Y}) = \sigma^2/n$ olup $V(b_1)$ eşitlik (2.54) ile verilmiştir. Böylece eşitlik (2.58) için belirlenmesi gereken terim sadece $Cov(\bar{Y}, b_1)$ 'dir.

İki doğrusal fonksiyon arasındaki kovaryans, bir tek doğrusal fonksiyonun varyansından biraz daha değişiktir. U katsayıları a_i ile belirtilen ilk doğrusal fonksiyon ve W ise katsayıları d_i ile belirlenen aynı şans değişkeninin ikinci bir doğrusal fonksiyonu olsun:

$$U = \sum a_i Y_i \qquad \text{ve} \qquad W = \sum d_i Y_i$$

U ve W 'nin kovaryansı,

$$Cov(U, W) = \sum_i a_i d_i\ V(Y_i) + \sum_{i \neq j}\sum a_i d_j\ Cov(Y_i, Y_j) \tag{2.59}$$

şeklinde olup Y_i değerleri bağımsız ise kovaryanslar sıfır olup eşitlik (2.59),

$$Cov(U, W) = \sum a_i d_i\ V(Y_i) \tag{2.60}$$

ifadesine indirgenebilir. Bu ifadedeki U ve W sırası ile $\bar{Y}$ ve b_1 tahminlerine karşılık gelmektedir.

$$\bar{Y} = \frac{1}{n}(Y_1 + Y_2 + \ldots + Y_n) = \frac{1}{n}\sum Y_i$$

$$b_1 = \frac{\sum\left(X_i - \bar{X}\right)Y_i}{\sum\left(X_i - \bar{X}\right)^2}$$

olduğundan a_i ve d_i katsayıları;

$$a_i = 1/n \qquad \text{ve} \quad d_i = \frac{\left(X_i - \bar{X}\right)}{\sum\left(X_i - \bar{X}\right)^2}$$

olarak belirlenebilir. Bu durumda $\bar{Y}$ ve b_1 arasındaki kovaryans;

$$\begin{aligned} Cov(\bar{Y}, b_1) &= \sum\left(\frac{1}{n}\right)\cdot\frac{\left(X_i - \bar{X}\right)}{\sum\left(X_i - \bar{X}\right)^2}\ \ V(Y_i) \\ &= \frac{1}{n}\cdot\frac{\sum\left(X_i - \bar{X}\right)}{\sum\left(X_i - \bar{X}\right)^2}\cdot\sigma^2 \\ &= 0 \end{aligned}$$

olarak elde edilir. Çünkü $\sum\left(X_i - \bar{X}\right) = 0$ 'dır. Bu durumda b_0'ın varyansı için;

$$V(b_0) = V(\bar{Y}) + (\bar{X})^2 V(b_1)$$

$$= \frac{\sigma^2}{n} + \bar{X}^2 \frac{\sigma^2}{\sum\left(X_i - \bar{X}\right)^2}$$

$$= \left(\frac{1}{n} + \frac{\bar{X}^2}{\sum\left(X_i - \bar{X}\right)^2}\right) \cdot \sigma^2 \qquad (2.61)$$

ifadesi elde edilir. Sonuç olarak b_0 tahmincisi:

$$b_0 \sim N\left[\beta_0; \left(\frac{1}{n} + \frac{\bar{X}^2}{\sum\left(X_i - \bar{X}\right)^2}\right)\sigma^2\right] \qquad (2.62)$$

b_0 normal dağıldığı için standardize istatistik $\left(b_0 - \beta_0\right)/\sigma\left(b_0\right)$ bir standart normal değişkendir. Bununla birlikte $\sigma(b_0)$ parametresi bilinmediği için $s(b_0)$ ile tahminlenir. Sonuç olarak belirlenmesi gereken $\left(b_0 - \beta_0\right)/s\left(b_0\right)$ istatistiğinin dağılımıdır. Bu dağılım n-2 serbestlik dereceli bir,

$$\frac{b_0 - \beta_0}{s\left(b_0\right)} = t_{n-2} \qquad (2.63)$$

t-dağılımıdır. b_0 ve b_1 tahminleyicileri sadece örnekten örneğe değişkenlik göstermez aynı zamanda verilen bir örnek için birbirine bağımlıdırlar, bkz Alıştırma 2.11.

$$Cov\left(b_0; b_1\right) = \left[-\frac{\bar{X}}{\sum\left(X_i - \bar{X}\right)^2}\right]\sigma^2$$

Şimdi bir şans değişkeni olan $\hat{Y}_0$ tahmincisinin dağılımı araştırılacaktır. Burada $\hat{Y}_0$ verilen bir X_0 değeri için

$$\hat{Y}_0 = b_0 + b_1 X_0 \qquad (2.64)$$

kestirilmiş değerdir. $\hat{Y}_0$ değeri b_0 ve b_1 tahminleyicilerinin doğrusal bir fonksiyonu olduğu için normal dağılıma sahiptir. Dağılımın ortalaması,

$$E\left(\hat{Y}_0\right) = \beta_0 + \beta_1 X_0 \qquad (2.65)$$

şeklindedir. Bu dağılımın varyansı ise,

$$\hat{Y}_0 = \bar{Y} + b_1 (X_0 - \bar{X})$$

eşitliği dikkate alınarak,

$$V(\hat{Y}_0) = V(\bar{Y}) + (X_0 - \bar{X})^2 \; V(b_1) + 2 \; Cov(\bar{Y}, b_1) \; (X_0 - \bar{X})$$

burada $Cov(\bar{Y}, b_1) = 0$ sıfıra eşit olduğu yukarıda ispatlanmıştı. Bu durumda dağılımın varyansı,

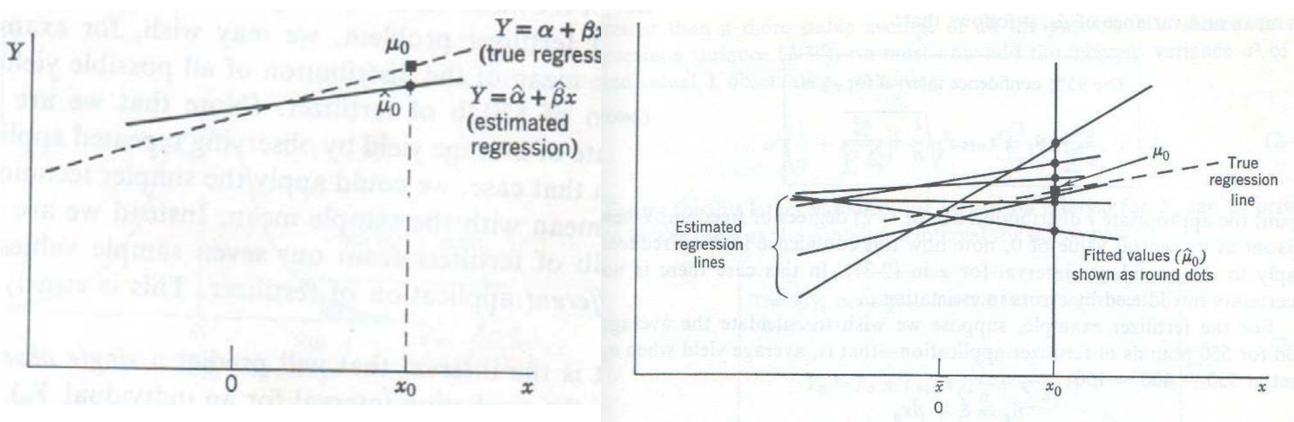
$$V(\hat{Y}_0) = V(\bar{Y}) + (X_0 - \bar{X})^2 \; V(b_1)$$

$$= \frac{\sigma^2}{n} + \frac{(X_0 - \bar{X})^2 \cdot \sigma^2}{\Sigma(X_i - \bar{X})^2} \qquad (2.66)$$

ve $\hat{Y}_0$'ın tahminlenmiş standart hatası,

$$s\left(\hat{Y}_0\right) = s \cdot \left(\frac{1}{n} + \frac{\left(X_0 - \bar{X}\right)^2}{\sum\left(X_i - \bar{X}\right)^2} \right)^{1/2} \tag{2.67}$$

şeklindedir, bkz. Alıştırma 2.12. Eşitlik (2.66) dan görüldüğü gibi $\hat{Y}_0$'nın varyansı tüm X değerleri için sabit değildir ve iki bileşene sahiptir. Bileşenler b_0 ve b_1'in (kesin olmayan) varyanslarıdır. Kesin olmamasının nedeni de b_1'in varyansının X_0 değerine bağlı olmasıdır. X_0 değeri $\bar{X}$'dan uzaklaştıkça $V(b_1)$ bileşeni büyüyecektir. $X_0 = \bar{X}$ noktasında $V(\hat{Y}_0)$ bir minimuma sahiptir. X_0 değeri $\bar{X}$'dan her iki yönde de uzaklaştıkça bu değerin büyümesi X_0 noktasında Y'nin ortalama değeri kestirilirken yapılan hatanın büyüyebileceğini ifade eder. Bu doğal bir sonuçtur. Çünkü en iyi kestirimin X değerlerinin oluşturduğu sınırın orta noktasında olması beklenir ve bu noktadan uzaklaşıldıkça kestirimin hassasiyeti azalır. Ayrıca gözlenmiş X değerlerinin dışında alınan bir X_0 için yapılan kestirimin daha az hassas olması beklenir ve sınırlardan uzaklaşıldıkça bu hassasiyette gittikçe azalacaktır. Bu durum Şekil 2.9 da gösterilmiştir. Şekil 2.9*a* da gerçek regresyon doğrusu ile birlikte tahminlenmiş doğru gösterilerek bu hataların etkisi açıklanmaya çalışılmıştır. Burada $\hat{Y}_0$ olduğundan daha az bir değerde tahminlenmiştir. Şekil 2.9*b* de ise birkaç örnek verisinden uyumu yapılmış tahminlenmiş regresyon doğruları ile gerçek regresyon doğrusu birlikte verilmiştir. Uyumu yapılmış değerlerin bazıları çok düşük bazıları ise çok yüksektir. Fakat ortalama değer tam gerçek regresyon doğrusu üzerindedir.



Şekil 2.9 a) $E(Y_0)$ *ile* $\hat{Y}_0$ *arasındaki ilişki* b) $E(Y_0)$ *değeri* Y_0 *ın sapmasız bir tahmincisidir.*

Sonuç olarak $\hat{Y}_0$ tahmincisinin dağılımı:

$$\hat{Y}_0 \sim N\left[\beta_0 + \beta_1 X_0 ; \frac{\sigma^2}{n} + \frac{(X_0 - \bar{X})^2 \cdot \sigma^2}{\sum(X_i - \bar{X})^2} \right] \tag{2.68a}$$

Yukarıda verilen varyans ve standart sapma formülleri verilen bir X_0 için Y_0'in ortalama değerinin $E(Y_0)$ kestirilmesinde kullanılır ve

$$\frac{\hat{Y}_0 - E(Y_0)}{\sqrt{\frac{s^2}{n} + \frac{(X_0 - \bar{X})^2 s^2}{\sum (X_0 - \bar{X})^2}}} = t_{n-2} \qquad (2.68b)$$

tanımlanmıştır.

Gözlenmiş gerçek Y_0 değeri,

$$Y_0 = \beta_0 + \beta_1 X_0 + \varepsilon_0$$

ise ortalaması çevresinde varyansı ile değişkenlik göstereceğinden bir bireysel gözlemin kestirilmiş değeri $\hat{Y}_0$ olmakla birlikte varyansı, bireysel gözlemin varyansının σ^2, ilave edilmesi ile

$$Var(Y_0) = \sigma^2 \left(\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\Sigma (X_i - \bar{X})^2} \right) + \sigma^2 = \sigma^2 \left(1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\Sigma (X_i - \bar{X})^2} \right) \qquad (2.69a)$$

şeklinde olacaktır, bkz. Alıştırma 2.13. Tahminlenmiş değeri ise σ^2 yerine s^2yazılarak elde edilebilir. Tahminlenmiş standart hatası ise,

$$s(Y_0) = s \cdot \sqrt{\left(1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\Sigma (X_i - \bar{X})^2} \right)} \qquad (2.69b)$$

olarak elde edilebilir. Sonuç olarak,

$$\frac{\hat{Y}_0 - Y_0}{s\sqrt{1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum (X_i - \bar{X})^2}}} = t_{n-2} \qquad (2.69c)$$

Yukarıda elde edilen çeşitli varyans tahmini formülleri incelendiğinde, modelin kestirim yeteneğini arttıran faktörlerin;

a. Örnek hacminin arttırılması

b. $\sum (X_i - \bar{X})^2$ değerinin arttırılması (x-uzayı büyüdüğünde modelin matematiksel yapısı da değişebileceğinden dikkatli olunmalıdır.)

olduğu görülebilir. Bu kısımda elde edilen sonuçlar parametrelerin aralık tahminlerinin elde edilmesinde kullanılacaktır.

2.9 PARAMETRELER İÇİN ARALIK TAHMİNİ VE HİPOTEZ TESTLERİ

Tahmin teorisi iki bölümden oluşur: Nokta tahmini ve aralık tahmini. Bu kısma kadar parametrelerin nokta tahminleri EKK yöntemi ile elde edilmiştir. Bu kısımda parametrelerin aralık tahmini ve hipotez testleri ile ilgilenilecektir.

Parametrelerin nokta tahminleri ile ilgili soru, bu tahminlerin güvenilirliği ne kadardır? şeklindedir. Tekrarlanmış örneklemede parametre tahmininin ortalama değerinin, gerçek değere eşit olması beklense bile $[E(b_1) = \beta_1]$,örneklemdeki düzensizlikler nedeni ile parametrenin bir tek tahmini gerçek değerinden farklılık gösterebilir. İstatistikte bir nokta tahmincisinin güvenirliliği, tahmincinin varyans veya standart sapması ile ölçümlenir. Öyleyse sadece bu nokta tahmincisine güvenmek yerine, gerçek parametrenin kendi nokta tahmincisi çevresinde belirli bir sınır veya aralıkta bulunması olasılığı verilebilir. Bu aralığın veya sınırın genişliği, tahmincinin nokta tahmincisi çevresinde ±2 veya ±3 standart sapma uzaklığını kapsayabilir.

Konuyu daha iyi belirleyebilmek için b_1'in β_1'e olan yakınlığının araştırılması gerekir. Bu amaçla δ ve α şeklindeki iki pozitif sayının belirlenmesi ile uğraşılabilir. α sayısı 0 ile 1 arasında belirlenir. Bunun nedeni ise, gerçek parametre β_1'i içeren $(b_1-\delta\,,b_1+\delta\,)$ şans aralığının olasılığının 1-α ile belirlenmesidir. Bu durum

$$\Pr\left(b_1-\delta\leq\beta_1\leq b_1+\delta\right)=1-\alpha \tag{2.70}$$

şeklinde ifade edilebilir. Eğer mevcut ise böyle bir aralığa *güven aralığı* adı verilir, 1-α güven katsayısı ve α önem seviyesi olarak adlandırılır. Güven aralığının uç noktaları, güven limitleri başka bir deyişle *kritik değerler* olarak bilinir.

Eşitlik (2.70)'den, bir aralık tahminleyicisinin, parametrenin gerçek değerini sınırları içerisinde bulundurma olasılığı 1-α olarak belirlenerek oluşturulmuş bir aralık olduğu görülmektedir. Aralık tahminlemesi ile ilgili önemli noktalar aşağıda verilmiştir.

1) Eşitlik (2.70), β_1'in verilen sınırlar arasında olma olasılığının 1-α olduğunu söylemez. β_1 her ne kadar bilinmesede sabit bir sayı olarak kabul edildiğinden, bu parametre aralığın içindedir veya değildir. Eşitlik (2.70), bu kısımda tanımlanan yöntem tekrarlı olarak kullanılarak, β_1'i içeren bir aralığı oluşturmanın olasılığının 1-α olduğunu ifade eder.

2) Eşitlik (2.70)'de verilen aralık, b_1'e bağımlı olduğundan ve b_1 tahmini de örnekten örneğe değişen bir şans değişkeni olduğu için bu aralık bir şans aralığıdır.

3) Güven aralığı bir şans aralığı olduğu için, olasılık ifadesini tekrarlı örnekleme olarak düşünmek gerekir. Başka bir deyişle, tekrarlı örnekleme güven aralığı eğer (1-α) baz alınarak pek çok defa tekrarlanarak oluşturulur ise bu aralıklar ortalama olarak parametrenin gerçek değerini (*tekrar sayısı*)*(1-α) adet durumu için içerecektir.

4) İkinci adımda belirtildiği gibi, (2.70) aralığı b_1 bilinmediği sürece bir şans aralığıdır. Fakat bir örnek belirlenip bu örnekten b_1'in belirli bir değeri elde edildikten sonra (2.70) aralığı artık bir şans aralığı olmaktan çıkar. Artık o bir sabittir. Bu durumda, verilen bir sabit aralığın gerçek β_1'i içerme olasılığının (1-α) olduğu söylenemez. Çünkü aralık belirlendikten sonra β_1 kesinlikle ya bu aralığın içindedir ya da dışında buna göre olasılık ya 1'dir ya da 0.

Bir güven aralığı nasıl oluşturulur? Yukarıdaki verilenler dikkate alınarak, eğer bir tahmincinin olasılık veya örnekleme dağılımı biliniyorsa eşitlik (2.70)'de verilen şekilde bir güven aralığının oluşturulabileceği görülebilir. Daha önce belirtildiği gibi hataların normal dağıldığı varsayımı altında EKK tahmincileri olan b_0 ve b_1 normal dağılış gösterirler. Ayrıca EKK tahmincisi s^2, χ^2

dağılışı ile ilişkili olan bir istatistiktir. Bu açıklamalar dikkate alınarak parametreler için güven aralıkları oluşturulabilir.

Hipotez testleri klasik istatistiksel yorumlamanın ikinci önemli konusunu oluşturur. Hipotez testi problemi aşağıda açıklandığı gibi tanımlanabilir. Olasılık yoğunluk fonksiyonu $f(X;\theta)$ olan (θ dağılımın parametresidir ve fonksiyon bu parametre dışında tamamen bilinmektedir) bir X şans değişkenine sahip olunduğu kabul edilsin. Bu şans değişkeninin oluşturduğu anakütleden n hacimli bir şans örneği alınır ve $\hat{\theta}$ nokta tahmincisi bu şans örneği yardımı ile tahminlenir. Gerçek parametre değeri θ nadiren bilindiği için, tahminlenen bu $\hat{\theta}$ değerinin, θ'nın hipotez edilen θ^* değeri ile uyumlu olup olmadığı sorusu ortaya çıkar. Burada θ^*, gerçek parametre θ'nın, $\theta=\theta^*$ şeklinde belirlenmiş sayısal bir değeridir. Başka bir deyişle, alınan örnek $f(X;\theta=\theta^*)$ yoğunluğundan mı gelmektedir? Hipotez testlerinde $\theta=\theta^*$ boş hipotez olarak adlandırılır ve genellikle H_o ile ifade edilir. Boş hipotez, H_1 şeklinde belirtilen bir alternatif hipoteze karşı test edilir. Örneğin bu alternatif hipotez $\theta\neq\theta^*$ şeklinde olabilir. Geçerli olduğunu kanıtlama sorumluluğu daima alternatif hipoteze aittir. Alternatif hipotez gibi boş hipotezde basit veya karmaşık olabilir. Bir hipotez dağılımın parametre (veya parametrelerini) değerlerini belirliyorsa basittir. Aksi takdirde karmaşık hipotezdir. Örneğin, eğer $X\sim N(\mu;\,\sigma^2)$ ise, $H_o{:}\mu=15$ ve $\sigma=2$ hipotezi basit, $H_o{:}\mu=15$ ve $\sigma\geq 2$ hipotezi ise karmaşıktır. Bunun nedeni ise σ değerinin belirlenmemiş olmasıdır. Boş hipotezi test etmek için, örnek bilgisi kullanılarak test istatistiği elde edilir. Genelde bu test istatistiği bilinmeyen parametrelerin nokta tahmincisi şeklinde oluşturulur. Daha sonra bu test istatistiğinin olasılık veya örnekleme dağılımı araştırılır ve bu boş hipotezin test edilebilmesi için güven aralığı veya anlamlılık testi yaklaşımları kullanılabilir.

2.9.1 β_1 İçin Güven Aralığı ve Hipotez Testi

Eşitlik (2.56) ve (2.57) ile tanımlanan b_1 in örnekleme dağılımı bilgileri kullanılarak β_1 parametresinin güven aralığı elde edilebilir. Bu amaçla eşitlik (2.57) da verilen t-istatistiği için (1-α) güven katsayılı bir aralık

$$\Pr\ (-t_{\alpha/2}\langle t\ \langle\ t_{\alpha/2}) = 1-\alpha \tag{2.71}$$

oluşturulabilir. Eşitlik (2.57) deki t değeri (2.71)'de yerine yazıldığında,

$$\Pr\left(-t_{\alpha/2}\langle\frac{b_1-\beta_1}{\sqrt{s^2\big/\sum\left(X_i-\bar{X}\right)^2}}\langle t_{\alpha/2}\right)=1-\alpha \tag{2.72}$$

elde edilir. Bu eşitlikte β_1 yalnız bırakılarak,

$$\Pr\left(b_1-t_{\alpha/2}\frac{s}{\sqrt{\sum\left(X_i-\bar{X}\right)^2}}\langle\beta_1\langle b_1+t_{\alpha/2}\frac{s}{\sqrt{\sum\left(X_i-\bar{X}\right)^2}}\right)=1-\alpha \tag{2.73}$$

eşitliği oluşturulabilir. Sonuç olarak β_1 için (1-α) güven aralığı,

$$\beta_1 = b_1 \pm t_{\alpha/2} \frac{s}{\sqrt{\sum\left(X_i - \bar{X}\right)^2}} \tag{2.74}$$

şeklindedir. Daha önce verilen örnek verileri üzerinde gerekli hesaplamalar aşağıdaki gibi oluşturulabilir.

$$s\left(b_1\right) = 0.0105$$

ve α=0,05 kabul edilerek $t_{(23;0,975)} = 2,069$ bulunur. Daha sonra β_1 için %95 güven katsayılı aralık,

$$-0,1015 \leq \beta_1 \leq -0,0581$$

olarak elde edilir. Bu güven aralığının yorumu şu şekilde yapılabilir. Verilen %95 güven katsayısı ile oluşturulan 100 aralıktan 95'i gerçek parametre β_1'i içerecektir. Daha önce belirtildiği gibi, belirlenmiş olan bu aralığın gerçek β_1'i içerme olasılığının %95 olduğu söylenemez, çünkü bu aralık artık şans aralığı değildir. Buna göre β_1 ya bu aralıktadır ya da değildir. Başka bir deyişle, belirlenmiş olan bu sabit aralığın gerçek β_1'i içermesi olasılığı 1 veya 0'dır.

β_1 parametresini test etmek için boş hipotez genel olarak,

$$H_0\text{: } \beta_1 = \beta_{10}$$

şeklinde kurulur. Alternatif hipotez amaca bağlı olarak üç farklı yapıda olabilir. Test istatistiği ise,

$$t = \frac{b_1 - \beta_{10}}{s_{b_1}}$$

olup, basit regresyon için özel bir durum hipotezler H_0:β_1=0 ve H_1: $\beta_1 \neq 0$ olarak tanımlandığında ortaya çıkar. H_0: β_1=0 hipotezinin testi basit regresyon için anlamlılık testidir. Hipotezin testi t ya da F-testi ile gerçekleştirilebilir. Daha esnek olduğundan t-testi tercih edilir. Nedeni ise t-testinin tek yönlü olarak da uygulanabilmesidir. F-testi bu esnekliğe sahip değildir. Basit regresyon özel durumu için test istatistiklerinde t^2=F ilişkisi geçerlidir, bkz. Alıştırma 2.14.

2.9.2 β_0 İçin Güven Aralığı ve Hipotez Testi

β_0 için bir güven aralığı β_1 için tanımlanmış olan aralığa benzer şekilde oluşturulabilir. Bu amaç için kullanılacak bilgiler eşitlik (2.62) ve (2.63) ile verilmiştir. Eşitliklerde tanımlanan bilgiler (2.71) de yerine konarak,

$$\Pr\left(-t_{\alpha/2} \langle \frac{b_0 - \beta_0}{\sqrt{\left(\frac{1}{n} + \frac{\bar{X}^2}{\sum\left(X_i - \bar{X}\right)^2}\right)s^2}} \langle t_{\alpha/2} \right) = 1 - \alpha \tag{2.75}$$

Bu eşitlikte β_0 yalnız bırakılarak,

$$\Pr\left(b_0 - t_{\alpha/2}\sqrt{\left(\frac{1}{n}+\frac{\bar{X}^2}{\sum\left(X_i-\bar{X}\right)^2}\right)s^2}\langle \beta_0 \langle b_0 + t_{\alpha/2}\sqrt{\left(\frac{1}{n}+\frac{\bar{X}^2}{\sum\left(X_i-\bar{X}\right)^2}\right)s^2}\right)=1-\alpha \qquad (2.76)$$

eşitliği oluşturulabilir. Sonuç olarak β_0 için (1-α) güven aralığı,

$$\beta_0 = b_0 \pm t_{\alpha/2}\sqrt{\left(\frac{1}{n}+\frac{\bar{X}^2}{\sum\left(X_i-\bar{X}\right)^2}\right)s^2} \qquad (2.77)$$

olarak elde edilebilir.

β_0 parametresini test etmek için boş hipotez genel olarak,

H_0: $\beta_0=\beta_{00}$

şeklinde kurulur. Alternatif hipotez amaca bağlı olarak üç farklı yapıda olabilir. Test istatistiği ise,

$$t=\frac{b_0-\beta_{00}}{s_{b_0}}$$

olup, özel bir durum hipotezler H_0:β_0=0 ve H_1:$\beta_0 \neq 0$ olarak tanımlandığında ortaya çıkar. H_0:β_0=0 hipotezinin testi regresyon için sabit terimsiz model için anlamlılık testidir.

2.9.3 *E* (Y) İçin Güven Aralığı

Veri setindeki mevcut verilen bir X_i değeri için $\hat{Y}_i$'nin tahminlenmiş ortalama değerinin, $E(Y_i)$ güven aralığı oluşturulabilir. Bunun için ilk olarak $\hat{Y}_0$ nokta tahmini bulunmalıdır. $\hat{Y}_0$ değeri eşitlik (2.64) ile tanımlanmıştır. $\hat{Y}_0$ şeklindeki bu uygun tahminci tahminlenmiş regresyon doğrusunun X_0 noktasındaki tahminlenen değeridir. Daha sonra bu tahmin etrafında bir aralık oluşturulabilir. Fakat bu bir nokta tahmincidir ve b_0 ile b_1 tahminlerindeki hatalar nedeni ile $\hat{Y}_0$'de bazı hataları içerecektir. Eşitlik (2.68*b*) de elde edilen sonuçlar eşitlik(2.71) de yerine konarak, $E(Y_0)$'ın (1-α) güven aralığı,

$$\Pr\left(\hat{Y}_0 - t_{\alpha/2}\cdot s\sqrt{\frac{1}{n}+\frac{\left(X_0-\bar{X}\right)^2}{\Sigma\left(X_i-\bar{X}\right)^2}} \le E\left(Y\right) \le \hat{Y}_0 + t_{\alpha/2}\cdot s\sqrt{\frac{1}{n}+\frac{\left(X_0-\bar{X}\right)^2}{\Sigma\left(X_i-\bar{X}\right)^2}}\right)=1-\alpha \qquad (2.78a)$$

ya da

$$E\left(Y\right)=\hat{Y}_0 \pm t_{\alpha/2}\cdot s\sqrt{\frac{1}{n}+\frac{\left(X_0-\bar{X}\right)^2}{\Sigma\left(X_i-\bar{X}\right)^2}} \qquad (2.78b)$$

şeklinde elde edilebilir. Y_i'nin tekrarlanmış örneklerinin, doğruyu uyarlamakta kullanılan, aynı sabit X değerlerinde ve sabit örnek hacmi kullanılarak oluşturulmuş olduğu kabul edilir. Daha sonra verindeki belirlenen bir X değeri, başka bir deyişle X_0 için Y'nin ortalama değerinin %(1-α) güven katsayılı aralığı eşitlik (2.78) kullanılarak oluşturulur. Bulunan bu aralık X=X_0 noktasındaki yeni bir gözlemin kestirim aralığı ile karıştırılmamalıdır. Bu durum aşağıda açıklanmıştır.

2.9.4 Bireysel Y_0 İçin Kestirim Aralığı

Eğer $X=X_0$ noktasındaki bir Y_0 gözlemi için kestirim aralığı istendiğinde gerçekte araştırılan $Y_0 - E\left[\left(Y_0 / X_0\right)\right] = Y_0 - \hat{Y}_0$ değeri için bir aralıktır. Bireysel Y_0 gözlemi için örnekleme dağılımı ile ilgili bilgi eşitlik (2.69*c*) ile verilmiştir. Eşitlik (2.69*c*) de elde edilen sonuçlar eşitlik (2.71) de yerine konarak, Y_0'ın %(1-α) olasılıklı güven aralığı,

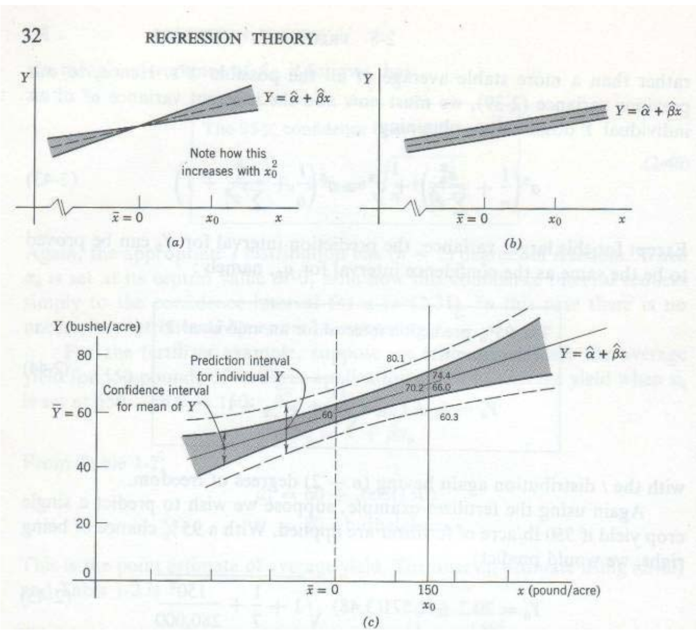
$$\Pr\left(\hat{Y}_0 - t_{\alpha/2}\cdot s\sqrt{1+\frac{1}{n}+\frac{\left(X_0-\bar{X}\right)^2}{\sum\left(X_i-\bar{X}\right)^2}} \le Y_0 \le \hat{Y}_0 + t_{\alpha/2}\cdot s\sqrt{1+\frac{1}{n}+\frac{\left(X_0-\bar{X}\right)^2}{\sum\left(X_i-\bar{X}\right)^2}}\right) = 1-\alpha \qquad (2.79a)$$

ya da

$$Y_0 = \hat{Y}_0 \pm t_{\alpha/2}\cdot s\sqrt{1+\frac{1}{n}+\frac{\left(X_0-\bar{X}\right)^2}{\sum\left(X_i-\bar{X}\right)^2}} \qquad (2.79b)$$

eşitsizliği ile verilebilir.

Ortalamanın güven aralığı ile kestirim aralığı arasındaki ilişki Şekil 2.10'da gösterilmektedir. Ortalamanın güven aralığı için iki potansiyel hata kaynağı mevcuttur ve bunlar Şekil 2.10*a* ve Şekil 2.10*b*'de verilmişlerdir. Bunlar birleşik bant şeklinde Şekil 2.10*c*'de bir araya getirilmişlerdir. Bu şekildeki kesikli bant daha geniş olup bireysel *Y* gözleminin kestirim aralığını verir. Her iki bant da X_0 değeri $\bar{X}$'dan uzaklaştıkça genişlemektedir. Bunun nedeni her iki aralık formülünde de bulunan $\left(X_0 - X\right)^2$ değerinin artmasıdır.



Şekil 2.10 (*a*) *Sadece* β_1*'in tahmininde hata mevcut ise Y'nin ortalamasının aralık tahmini*, (*b*) *Sadece* β_0*'in tahminlenmesinde hata mevcut ise Y'nin ortalamasının aralık tahmini*, (*c*) *Y'nin ortalamasının aralık tahmini ve bireysel Y'nin kestirim aralığı*

2.9.5 σ^2 İçin Güven Aralığı

Normallik varsayımı altında,

$$\chi^2 = (n-2)\frac{s^2}{\sigma^2} \tag{2.80}$$

değişkeni (n-2) serbestlik dereceli bir χ^2 dağılımı gösterir. Bu durumda χ^2 dağılımı σ^2 için bir güven aralığı oluşturmada kullanılabilir.

$$\Pr\ (\chi^2_{1-\alpha/2} \leq \chi^2 \ \leq\ \chi^2_{\alpha/2}) = 1-\alpha \tag{2.81}$$

Burada, çift taraflı eşitsizliğin ortasında bulunan χ^2 değeri eşitlik (2.80)'de verilmiştir. $\chi^2_{1-\alpha/2}$ ve $\chi^2_{\alpha/2}$ değerleri ise (n-2) serbestlik derecesi için ki-kare tablosundan elde edilen iki χ^2 değeridir. Bunlar χ^2 dağılımının uç değerleridir (kritik değerler). Denklem (2.80)'deki χ^2 değeri (2.81) eşitliğinde yerine konarak ve gerekli düzenlemeler yapılarak,

$$\Pr\left[(n-2)\frac{s^2}{\chi^2_{\alpha/2}} \leq \sigma^2 \leq (n-2)\frac{s^2}{\chi^2_{1-\alpha/2}}\right] = 1-\alpha \tag{2.82}$$

σ^2 için bir güven aralığı elde edilir. Burada dikkat edilmesi gereken nokta ki-kare dağılımının simetrik olmadığıdır. Bu nedenle uç noktalardaki kritik değerler mutlak değerce birbirine eşit değildir. Oysa z ve t-dağılımları simetrik olduğundan kritik değerleri mutlak değerce birbirine eşittir.

2.10 EN YÜKSEK OLABİLİRLİK TAHMİNLERİ

En küçük kareler yöntemi kullanıldığında parametrelerin nokta tahmincilerinin elde edilmesinde hata terimlerinin (dolayısı ile Y_i'lerin) normal dağıldığı varsayımına gerek olmadığı Gauss-Markov teoremi yardımı ile doğrulanmıştı. Parametrelerin güven aralığının ve hipotez testinin oluşturulması için yapılacak çalışmalarda ise normallik varsayımı gerekli olacaktır. En yüksek olabilirlik tahmincilerinin (EYOT) elde edilebilmesi için hataların dağılışı ile ilgili varsayıma ihtiyaç vardır.

Bu kısımda β_0, β_1 ve σ^2'nin en yüksek olabilirlik tahminleri elde edilecektir. Bu tahminler β_0, β_1 ve σ^2'nin hipotetik anakütle değerleri olup, gözlenmiş örnek değerleri için en yüksek olasılığı verirler. β_0 ve β_1'in EYOT tahminleri EKK tahminleri ile aynı sonucu verirler. Dolayısı ile en yüksek olabilirlik metodu EKK'nın kullanılmasını haklı çıkaran başka doğrulama sağlamaktadır. Cebirsel işlemler verilmeden önce, konunun geometrik olarak açıklanması faydalı olacaktır. En yüksek olabilirlik metodu niçin verilere uygulanır? Konuyu basitleştirmek amacı ile sadece üç gözlemin (P_1, P_2, P_3) olduğu kabul edilecektir.

İlk olarak Şekil 2.11*a*'daki doğrunun oluşturulmasıyla ilgilenilsin. Bu doğru dikkatli olarak incelenmeden önce, onun gözlenmiş üç nokta için kötü bir uyum sağladığı görülmektedir. Geçici olarak bu doğrunun gerçek regresyon doğrusu olduğu kabul edilsin. Buna göre hataların dağılımı doğru çevresinde merkezlenecektir. Gözlenmiş örnek için bir anakütle ortaya koyan bu olabilirlik, üç ε değerinin belirli bir setinin ortak olasılık yoğunluğudur. Bu üç ε değerinin bireysel olasılık yoğunlukları P_1, P_2 ve P_3 noktalarının üstünde ordinatlar şeklinde gösterilmektedir. Ortak olasılık yoğunluğu bu üç ordinatın çarpımıdır. Çünkü bu üç gözlem istatistiksel olarak bağımsızdır. Bu olabilirlik göreceli olarak küçüktür, çünkü çok küçük olan P_1 ordinatı çarpım değerinin küçük olmasına neden olmaktadır. Bunun sonucu olarak kötü bir tahminde bulunulduğu düşünülebilir. Örnek değerlerini oluşturmak için böyle bir hipotetik anakütle uygun değildir. Daha iyisi elde edilebilir.

Şekil 2.11*b* daha iyi bir hipotetik anakütle oluşturulabileceğini kanıtlamaktadır. Bu anakütle gözlenmiş örneği oluşturmak için daha uygundur. Hata terimleri de müşterek olarak daha küçüktür ve bunun sonucu olarak da onların olasılık yoğunlukları daha büyüktür.

Wonnacott sf 35

Şekil 2.11 *En yüksek olabilirlik tahmini*: *Verilen anakütleler gerçek anakütle olmayıp istatistikçinin dikkate aldığı hipotetik anakütlelerdir*. (*a*) *Gözlenmiş değerleri oluşturmaya uygun olmayan anakütle* , (*b*) *Gözlenmiş değerleri oluşturmaya daha uygun anakütle*.

EYOT tekniğinin çeşitli mümkün anakütleler üzerindeki inceleme ve düşünceleri içerdiği görülmektedir. Bu anakütlelerden her birinin gözlemlenen örneği daha iyi oluşturma özelliği nasıl sağlanacaktır? Geometrik olarak, problem tüm mümkün değerler boyunca anakütleyi hareket ettirmektedir. Bu da uzaydaki tüm mümkün pozisyonlar boyunca regresyon doğrusunu ve onun çevresindeki ε dağılımının hareket ettirilmesiyle gerçekleşir. Her bir pozisyon β_0 ve β_1 için deneme değerlerinin farklı bir setini içerir. Her bir durumda P_1, P_2, P_3'ü gözlemlemenin olabilirliği değerlendirilebilecektir. EYOT için bu olabilirliği maksimize eden hipotetik anakütle seçilir. EYOT değerlerini elde edebilmek için (Şekil 2.11*b*'de gösterilen) küçük bir düzeltme gereklidir. Bu prosedürün iyi bir uyum sağladığı görülmektedir. EYOT sonuçları EKK'ya benzer olduğu için, EKK ve EYOT yöntemleri uygulanarak elde edilen iki sonucun aynı olması sürpriz olmamalıdır.

EYOT 'un iki dezavantajı mevcuttur. Bu tahminciler üç örnek gözleminden elde edilmiştir. Gözlenmiş bir başka örnek seti β_0, β_1 için bir başka EYOT ortaya çıkaracaktır. İkinci dezavantaj ise daha gizli kalmış bir konudur. Herhangi bir anakütlenin olabilirliği sadece örneğin içermiş olduğu ε terimlerinin büyüklüğüne bağlı değildir. Bu olabilirlik aynı zamanda ε dağılımının şekline özelliklede ε'nin varyansı σ^2'ye de bağlıdır. Bunula birlikte en yüksek olabilirlik doğrusu σ^2'ye bağımlı değildir. Başka bir deyişle, σ^2'nin daha büyük olduğu kabul edildiğinde Şekil 2.11'da verilen geometri farklı bir görünüme sahip olacaktır. Çünkü ε'nun dağılımı daha yayvan bir dağılım olacaktır. Fakat sonuç da en yüksek olabilirlik doğrusu değişmeyecektir.

Geometri bu metodu açıklayabildiği halde, en yüksek olabilirlik tahmini elde etmede kesin bir yöntem sağlayamamaktadır. Bu yöntem cebirsel olmalıdır. Durumu genellemek amacı ile eldeki örneğin hacminin *n* olduğu kabul edilecektir $P(Y_1, Y_2, \ldots, Y_n)$ ve β_0, β_1 ile σ^2'nin mümkün anakütle değerlerinin bir fonksiyonu olarak ifade edilen gözlenmiş bu örneğin olabilirliği veya olasılık yoğunluğu bilinmek istenir. İlk olarak *Y*'nin birinci değerinin olasılık yoğunluğu,

$$P(Y_1) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(1/2\sigma^2)\,[Y_1-(\beta_0+\beta_1 X_1)]^2} \tag{2.83}$$

şeklinde verilebilir. Y_1, ortalaması (b_0+b_1X_1) ve varyansı σ^2 olan normal dağılış göstermektedir. Y_2 ve diğer Y değerleri için olasılık yoğunlukları da (2.83)'e benzer şekilde elde edilebilir. Y değerlerinin bağımsız olması nedeni ile ortak olasılık yoğunluğunu bulmak üzere tüm bu olasılık yoğunlukları çarpılır.

$$P(Y_1, Y_2, \ldots, Y_n) = \left[\frac{1}{\sqrt{2\pi\sigma^2}} e^{-(1/2\sigma^2)\,[Y_1-(\beta_0+\beta_1 X_1)]^2}\right]\left[\frac{1}{\sqrt{2\pi\sigma^2}} e^{-(1/2\sigma^2)\,[Y_2-(\beta_0-\beta_1 X_2)]^2}\right] \ldots.$$

$$= \Pi\left[\frac{1}{\sqrt{2\pi\sigma^2}} e^{-(1/2\sigma^2)\,[Y_i-(\beta_0-\beta_1 X)]^2}\right]$$

$$P(Y_1, Y_2, \ldots, Y_n) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n e^{\sum(-1/2\sigma^2)\,[Y_i-(\beta_0-\beta_1 X_i)]^2} \tag{2.84}$$

şeklinde yazılabilir. Gözlenmiş Y değerleri verilmektedir, bu nedenle fonksiyonda bilinmeyen parametrelerin çeşitli değerleri araştırılır. Eşitlik (2.84) olabilirlik fonksiyonu olarak,

$$L(\beta_1, \beta_2, \ldots, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-(1/2\sigma^2)\sum[Y_i-(\beta_0-\beta_1 X)]^2} \tag{2.85}$$

şeklinde verilebilir. Daha sonra yapılacak işlem, hangi β_0 ve β_1 değerlerinin L fonksiyonunu en büyüklediğini araştırmaktadır. β_0 ve β_1 değerleri sadece üslü ifadede de mevcuttur. Negatif işarete sahip üslü bir fonksiyonun maksimizasyonu üst ifadesinin büyüklüğünün minimizasyonunu içerir. Sonuç olarak parametre tahminleri,

$$\sum[Y_i - \beta_0 - \beta_1 X_i]^2 \tag{2.86}$$

ifadesinin minimizasyonu ile elde edilir. Bu işlem ile σ değeri ihmal edilerek β_0 ve β_1 için en yüksek olabilirlik çözümü elde edilir. Daha önce belirtildiği gibi, dağılımın yayılışı ile ilgili herhangi bir varsayım yoktur, bu nedenle en yüksek olabilirlik doğrusu σ'dan etkilenmez.

Eşitlik (2.86) ile (2.8) karşılaştırıldığında önemli bir sonuç elde edilir. En yüksek olabilirlik tahminleri, en küçük kareler tahminleri ile özdeştir (β_0 ve β_1 için).

σ^2 için EYOT'nin elde edilmesi (β_0 ve β_1'den daha zordur. Ayrıca σ^2'nin EYOT'u sapmalıdır. Cebirsel işlemlerde kolaylığı sağlamak amacı ile eşitlik (2.77)'in logaritması alınarak,

$$Q = \text{Log}\, L = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum[Y_i - (\beta_0 + \beta_1 X)]^2 \tag{2.87}$$

elde edilir. Bu fonksiyonun β_0, β_1 ve σ^2'ye göre kısmi türevleri alınıp sıfıra eşitlenerek en yüksek olabilirlik tahminleri elde edilir.

$$\frac{\partial Q}{\partial \beta_0} = \frac{1}{\sigma^2}\cdot\sum(Y_i - \beta_0 - \beta_1 X_i) = 0 \tag{2.88a}$$

$$\frac{\partial Q}{\partial \beta_0} = \frac{1}{\sigma^2} \cdot \Sigma\left(Y_i - \beta_0 - \beta_1 X_i\right)(X_i) = 0 \qquad (2.88b)$$

$$\frac{\partial Q}{\partial \sigma^2} = \frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \cdot \Sigma\left(Y_i - \beta_o - \beta_1 X_i\right)^2 = 0 \qquad (2.88c)$$

Daha önce belirtildiği gibi σ^2'den bağımsız olduğu için (2.88*a*) ve (2.88*b*) denklemleri, (2.88*c*) den bağımsız olarak çözülür ve,

$$\sum Y_i = n\,\tilde{b}_0 + \tilde{b}_1 \sum X_i \qquad (2.89a)$$

$$\sum Y_i X_i = \tilde{b}_0 + \tilde{b}_1 \sum X_i^2 \qquad (2.89b)$$

normal denklemleri elde edilir. Elde edilen bu denklemler EKK normal denklemlerinin aynısıdır. Daha sonra (2.89*c*) denklemi σ^2 için çözülür ve

$$\tilde{s}^2 = \frac{1}{n} \Sigma\left(Y_i - \tilde{b}_0 + \tilde{b}_1 X_i\right)^2 \qquad (2.90c)$$

eşitliği elde edilir. Bu eşitlikte $\tilde{b}_0$ ve $\tilde{b}_1$ yerine konarak $\tilde{s}^2$ bulunabilir.

β_1=0 şeklindeki özel bir durumda (Y 'nin X'e bağımlı olmaması durumu) $b_0 = \overline{Y}$ olacağı daha önce belirtilmişti. Bu durumda $\tilde{s}^2 = s_y^2$ olur. Şüphesiz elde edilen bu sonuçta bir en yüksek olabilirlik tahmincisi vermektedir. Fakat bu sonuç sapmalıdır. Benzer şekilde (2.90*c*)'nin de sapmalı olduğu gösterilebilir. Bu nedenle n serbestlik derecesi yerine (n-2) serbestlik derecesi bölüm olarak kullanıldığında,

$$s^2 = \frac{1}{n-2} \Sigma\left(Y_i - b_0 - b_1 X_i\right)^2$$

sapmasız bir tahminci elde edilir. İki serbestlik derecesinin kaybedilmesinin nedeni, s^2'nin elde edilebilmesinden önce b_0 ve b_1 tahminlerinin elde edilmesinin gerekli olmasıdır.

En yüksek olabilirlik $\tilde{s}^2$ tahmincisinin sapmalı olması bu metodun kullanılmasına bir sınırlama getirmektedir. Küçük hacimli örneklerde bu metod kullanılırken $\tilde{s}^2$ sapmalı olduğu için dikkatli olunmalıdır. Bununla birlikte büyük hacimli örnekler için $\tilde{s}^2$'nin EYOT'u kararlılık özelliğini sağlamaktadır.