EN KÜÇÜK KARELER YÖNTEMİ

(En küçük kareler doğrusu / Regresyon)

Biri diğerine bağlı olabilen X, Y gibi iki değişken gözönüne alalım. Bunlar, aşağıdaki gibi olayları temsil edebilirler.

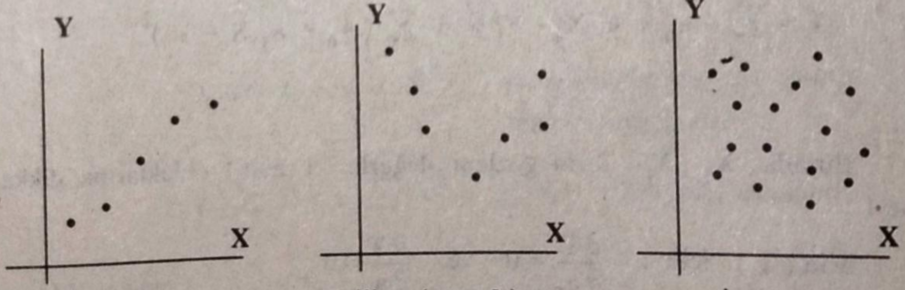
X	Y
Matematik notları	Fizik notları
Öğrencinin boyu	Öğrencinin ağırlığı
Yıllar	Bir firmanın üretimi
Sıcaklık	Bir firmanın meşrubat satışı
Yıllar	Doğum hızı
Birim zaman	Birim hacimdeki bir kültürde, bakteri sayısı
. . . .	. . . .
. . . .	. . . .

Bu örnekler çoğaltılabilir.

Bu örneklerden herhangi birini ele alalım. Meselâ, belirli bir topluluktaki kişilerin boy / ağırlık verilerinin aşağıdaki gibi olduğunu varsayalım.

X	Y
X_1	Y_1
X_2	Y_2
X_3	Y_3
. . .	. .
X_N	Y_N

(X_i, Y_i) $i = 1, 2, 3, \ldots, N$ ikilisi, bir nokta ile gösterilerek saçılım (*scatter*) diyagramı yapılır ve eğilim (*trend*) saptanır.

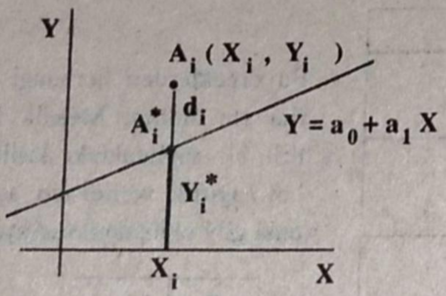


Eğilim (**trend**): doğrusal — Eğilim (**trend**): Eğrisel / Parabol — İlişki yok

Saçılım diyagramında, eğilim'in doğrusal olduğunu varsayalım ve gözlem değerlerini temsil eden A_i noktaları arasından uygun bir doğru geçirelim. Bu doğru üzerinde,

$A_i(X_i, Y_i)$ $(i = 1,2,3,\ldots,N)$ gözlem noktasına karşı gelen nokta

$A^{\bullet}_i(X_i, Y^{\bullet}_i)$ $(i = 1,2,3,\ldots,N)$ olsun.



$d_i = Y^{\bullet}_i - Y_i$ farkı **kalıntı'dır**.

Kalıntıların $\sum\limits_{i=1}^{N} d_i^2$ karelerinin toplamını minimum yapacak şekilde belirtilen a_0, a_1 parametreleriyle bulunan

$$Y = a_0 + a_1 X$$

doğrusu **en uygun doğru**'dur. Bu doğru **En küçük kareler doğrusu** adını alır. En küçük kareler doğrusu, A_i gözlem noktaları arasından geçirilebilecek doğrular içinde kestirim hatasını minimum yapan doğru'dur.

$T = \sum\limits_{i=1}^{N} d_i^2$ toplamını minimum yapan a_0, a_1 parametrelerini belirtmek:

$$T = (Y_1^* - Y_1)^2 + (Y_2^* - Y_2)^2 + \ldots + (Y_N^* - Y_N)^2$$

$$= (a_0 + a_1X_1 - Y_1)^2 + (a_0 + a_1X_2 - Y_2)^2 + \ldots + (a_0 + a_1X_N - Y_N)^2$$

$$T = \sum_{i=1}^{N} (a_0 + a_1X_i - Y_i)^2 = \sum (a_0 + a_1 X - Y)^2$$

dir.

Burada X_i, Y_i lerin gözlem değerleri (sabit) olduklarına dikkat edilirse.

Min (T) için : $\dfrac{\partial T}{\partial a_0} = 0$ ve $\dfrac{\partial T}{\partial a_1} = 0$

olmalıdır.

$$\frac{\partial T}{\partial a_0} = 2.\sum (a_0 + a_1 X - Y) = 0 , \qquad N.a_0 + a_1 \sum X = \sum Y$$

$$\frac{\partial T}{\partial a_1} = 2.\sum [(a_0 + a_1 X - Y).X] = 0 , \qquad a_0 \sum X + a_1 \sum X^2 = \sum XY$$

olmalıdır.

$$\boxed{\begin{aligned} N.a_0 + a_1 \sum X &= \sum Y \\ a_0 \sum X + a_1 \sum X^2 &= \sum X Y \end{aligned}}$$

Normal denklemler adını alırlar.

Normal denklemler çözülerek a_0 ve a_1 parametreleri belirtilir.

Böylece bulunan

$$\boxed{Y = a_0 + a_1 X}$$

doğrusu, **En küçük kareler doğrusu** adını alır.

En küçük kareler doğrusu, saçılım diyagramında geçirilebilecek doğruların en uygunudur. Çünkü, kalıntıların karelerinin toplamının minimum olduğu doğru en küçük kareler doğrusudur.

Uyarı 1) Aşağıdaki yol izlenirse, normal denklemler pratik olarak elde edilebilir.

$Y = a_0 + a_1 X$ denklemine $\sum$ işlemcisi uygulanırsa

$N.a_0 + a_1 \sum X = \sum Y$ normal denklemi;

$Y = a_0 + a_1 X$ denkleminin her iki yanı X ile çarpılarak oluşan

$X Y = a_0 X + a_1 X^2$ denkleminin her iki yanına $\sum$ işlemcisi uygulanırsa

$a_0 \sum X + a_1 \sum X^2 = \sum XY$ normal denklemi

elde edilir.

Uyarı 2) Saçılım diyagramında , olayın eğilim'ine (*trend'ine göre*) gözlem noktaları arasından ikinci dereceden parabol veya daha yüksek dereceden eğriler geçirilebilir. Bu eğriler için de normal denklemler , **Uyarı 1)** deki yol izlenerek oluşturulur.

Örnek 5. 11)

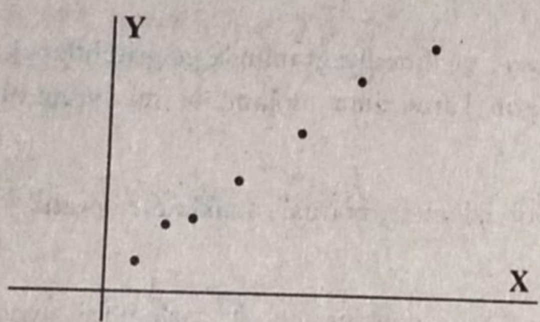
X	2	5	6	10	15	18	23
Y	3	9	12	15	28	40	45

Gözlem değerleri için

1) Saçılım diyagramını çiziniz.
2) En küçük kareler doğrusunun denklemini bulunuz.
3) Kalıntıların karelerinin toplamını bulunuz.

Yanıt :

1)



Eğilim'in (*trend'in*) doğrusal olduğu görülmektedir.

2)

X	Y	$\sum X^2$	$\sum XY$	Y^*	$Y^* - Y$	$(Y^* - Y)^2$
2	3	4	6	0.191	−2.809	7.890
5	9	25	45	7.745	−1.255	1.575
6	12	36	72	10.263	−1.737	3.017
10	15	100	150	20.335	5.335	28.462
15	28	225	420	32.925	4.925	24.256
18	40	324	720	40.779	0.779	0.607
23	58	529	1334	53.069	−4.931	24.315
$\sum X =$ 79	$\sum Y =$ 165	$\sum X^2 =$ 1243	$\sum XY =$ 2747			$\sum (Y^* - Y)^2 =$ 90.122

Normal denklemler

$$\left.\begin{aligned} 7a_0 + 79a_1 &= 165 \\ 79a_0 + 1243a_1 &= 2747 \end{aligned}\right]$$

Normal denklemlerin çözümü

$$a_0 = \frac{\begin{vmatrix} 165 & 79 \\ 2747 & 1243 \end{vmatrix}}{\begin{vmatrix} 7 & 79 \\ 79 & 1243 \end{vmatrix}}, \quad a_1 = \frac{\begin{vmatrix} 7 & 165 \\ 79 & 2747 \end{vmatrix}}{\begin{vmatrix} 7 & 79 \\ 79 & 1243 \end{vmatrix}}$$

$$a_0 = \frac{-11918}{2460} = -4.845, \quad a_1 = \frac{6194}{2460} = 2.518$$ olduğundan,

en küçük kareler doğrusunun denklemi

$$\boxed{Y = -4.845 + 2.518\,X}$$

dir.

Kontrol :

$G(\bar{X}, \bar{Y})$ santroid'dir. En küçük kareler doğrusu santroid'den geçer.

$$\bar{X} = \frac{\sum X}{N} = \frac{79}{7}, \quad \bar{Y} = \frac{\sum Y}{N} = \frac{165}{7} \qquad G\left(\frac{79}{7}, \frac{165}{7}\right)$$ noktasının

koordinatları, en küçük kareler doğrusunun denklemini sağlamalıdır. Gerçekten,

$$-4.845 + 2.518\left(\frac{79}{7}\right) = \frac{165}{7}$$ dir.

Niçin $G(\bar{X}, \bar{Y})$ noktasının kordinatları, en küçük kareler doğrusunun denklemini sağlar? Bu soruyu yanıtlayınız.

3) X gözlem değerleri, en küçük kareler doğrusu'nun denkleminde yerlerine yazılarak bunlara karşı olan Y^* kestirim değerleri hesaplanır

$x=2$ için $Y^* = -4.845 + 2.518(2) = 0.191$,

$x=5$ için $Y^* = -4.845 + 2.518(5) = 7.745$,

$x=6$ için $Y^* = -4.845 + 2.518(6) = 10.263$,

$x=10$ için $Y^* = -4.845 + 2.518(10) = 20.335$,

$x=15$ için $Y^* = -4.845 + 2.518(15) = 32.925$,

$x=18$ için $Y^* = -4.845 + 2.518(18) = 40.479$,

$x=23$ için $Y^* = -4.845 + 2.518(23) = 53.069$

dir. Y^* nin bu değerleri, tablo'nun 5. Sütununa yazılmıştır.

$d = Y^* - Y$ kalıntıları tablo'nun 6. Sütununa ve $d^2 = (Y^* - Y)^2$ kalıntıların karelerinin toplamı ise tablo'nun son sütununa yazılmıştır.

En küçük kareler doğrusunun denkleminden, meselâ $X = 7.4$ için Y^* değerini hesaplayalım :

$$X = 7.4 \text{ için } \quad Y^* = -4.845 + (2.518).(7.4) \approx 13.79$$

dır.

En küçük kareler doğrusu (*regresyon doğrusu*)'nun Zaman serilerinde kullanımı

Belirli bir bölgede, zeytin yağı üretiminin yıllara göre dağılımı aşağıdadır.

Yıllar	1985	1986	1987	1988	1989	1990	1991
Üretim (ton)	400	480	535	560	520	580	625

1) En küçük kareler doğrusu (*regresyon doğrusu*)'nun denklemini bulunuz.

2) 1992 yılındaki üretimi kestiriniz.

Yanıt :

Burada iki yol izlenebilir.

I. Yol : 1985 yılı başlangıç olarak kabul edilir ve yıllar buna göre kod'lanır.

Yıllar	X (kod)	Y (Üretim)	X^2	XY
1985	0	400	0	0
1986	1	480	1	480
1987	2	535	4	1070
1988	3	560	9	1680
1989	4	520	16	2080
1990	5	580	25	2900
1991	6	621	36	3726
	$\sum X = 21$	$\sum Y = 3696$	$\sum X^2 = 91$	$\sum XY = 11936$

Normal denklemler :

$$7.a_0 + 21a_1 = 3696$$
$$21a_0 + 91a_1 = 11936$$

Sistemin çözümü :

$$a_0 = \frac{\begin{vmatrix} 3696 & 21 \\ 11936 & 91 \end{vmatrix}}{\begin{vmatrix} 7 & 21 \\ 21 & 91 \end{vmatrix}} = \frac{85680}{196} = 437.143 ,$$

$$a_1 = \frac{\begin{vmatrix} 7 & 3696 \\ 21 & 11936 \end{vmatrix}}{\begin{vmatrix} 7 & 21 \\ 21 & 91 \end{vmatrix}} = \frac{5936}{196} = 30.286 \quad \text{dir.}$$

Regresyon doğrusu'nun denklemi

$$\boxed{Y = 437.143 + 30.286X}$$

dir.

1992 yılı için $X = 7$ olacağından, 1992 yılı için

$$Y^* = 437.143 + 30.286(7) = 649.145 \text{ ton}$$

olur.

2. Yol : En ortadaki yıl, yani 1988 yılı başlangıç kabul edilerek kotlama yapılır.

	Yıllar	X (kod)	Y (Üretim)	X^2	XY
	1985	-3	400	9	-1200
	1986	-2	480	4	-960
	1987	-1	535	1	-535
Başlangıç	1988	0	560	0	0
	1989	1	520	1	520
	1990	2	580	4	1160
	1991	3	621	9	1863
		$\sum X = 0$	$\sum Y = 3696$	$\sum X^2 = 28$	$\sum XY = 848$

Normal denklemler :

$$7a_0 = 3696$$
$$28\,a_1 = 848$$

Sistemin çözümü :

$$a_0 = 528$$
$$a_1 = 30.286$$

dir.

Regresyon doğrusunun denklemi

$$\boxed{Y = 528 + 30.286\,X}$$

olur. 1992 yılı için $X = 4$ olacağından ,

1992 yılı için : $Y^* = 528 + 30.286\,(4) = 649.144$ ton

olur.

Not : Yılların sayısı , meselâ aşağıdaki tabloda olduğu gibi çift ise bir yıl ikiye bölünerek (6 ay + 6 ay) kod'lama yapılır.

Yıllar	X
1975	−5
1976	−3
1977	−1
1978	1
1979	3
1980	5
	$\sum X = 0$

Burada 1 yıl 2 tane 6 aya bölünmüştürve

$\sum X = 0$ olduğu görülmektedir.

Ayrıca , saçılım diyagramındaki gözlem değerlerinin eğilimine (*trend'ine*) göre , bunların arasından **en küçük kareler parabolü** veya daha yüksek dereceden eğriler geçirilebilir. Aşağıda , bu konu ele alınacaktır.

GENELLEME

X	X_1	X_2	. . .	X_N
Y	Y_1	Y_2	. . .	Y_N

, $A_i(X_i, Y_i)$ $i = 1, 2, \ldots, N$

gözlem değerleri veriliyor. Amacımız ,

$$(E): \quad Y = a_0 + a_1 X + a_2 X^2 + \ldots + a_{D-1} X^{D-1} + a_D X^D .$$

en küçük kareler (*regresyon*) eğrisi için **normal denklemler'i** oluşturmak ve bu sistemi çözerek

$$a_0, a_1, a_2, \ldots, a_{D-1}, a_D$$

katsayılârını belirtmektir. Aşağıdaki işlemler bu amaca yöneliktir.

(E) eşitliğinin her iki yanına $\sum_{i=1}^{N}$ işlemcisi uygulanırsa

$$a_0 N + a_1 \sum X + a_2 \sum X^2 + \ldots + a_{D-1} \sum X^{D-1} + a_D \sum X^D = \sum Y$$

normal denklemi elde edilir.

(E) eşitliğinin her iki yanı X ile çarpıldıktan sonra $\sum_{i=1}^{N}$ işlemcisi uygulanırsa

$$a_0 \sum X + a_1 \sum X^2 + a_2 \sum X^3 + \ldots + a_{D-1} \sum X^D + a_D \sum X^{D+1} = \sum XY$$

normal denklemi elde edilir. Böylece devam edilirse

....

(E) eşitliğinin her iki yanı X^D ile çarpıldıktan sonra $\sum_{i=1}^{N}$ işlemcisi uygulanarak

$$a_0 \sum X^D + a_1 \sum X^{D+1} + a_2 \sum X^{D+2} + \ldots + a_D \sum X^{2D} = \sum X^D Y$$

normal denklemi elde edilir. Böylece, normal denklemler oluşturulmuş olur.

Not : Normal denklemlerin oluşturulması için uygulanan yukardaki yöntem daha kısa bir şekilde ifade edilebilir :

(E) eşitliğinin iki yanı X^{t-1} $(t = 1, 2, \ldots, D+1)$ ile çarpıldıktan sonra $\sum_{i=1}^{N}$ işlemcisi uygulanırsa ,

$$a_0 \sum X^{t-1} + a_1 \sum X^t + a_2 \sum X^{t+1} + \ldots + a_D \sum X^{D+t-1} = \sum X^{t-1} Y$$

normal denklemleri elde edilir .

Şu halde , normal denklemlerden oluşan sistem

$$
\begin{array}{llllll}
a_0 . N & + a_1 \sum X & + a_2 \sum X^2 & + \ldots\ldots + a_D \sum X^D & = \sum Y \\
a_0 \sum X & + a_1 \sum X^2 & + a_2 \sum X^3 & + \ldots\ldots + a_D \sum X^{D+1} & = \sum XY \\
\ldots & \ldots & \ldots & \ldots & \ldots \\
a_0 \sum X^D & + a_1 \sum X^{D+1} & + a_2 \sum X^{D+2} & + \ldots + a_D \sum X^{2D} & = \sum X^D Y
\end{array}
$$

dir. Bu sistemin matrislerle gösterimi

$$
\begin{bmatrix}
N & \sum X & \sum X^2 \ldots & \sum X^D \\
\sum X & \sum X^2 & \sum X^3 \ldots & \sum X^{D+1} \\
\ldots & \ldots & \ldots & \ldots \\
\sum X^D & \sum X^{D+1} & \sum X^{D+2} \ldots & \sum X^{2D}
\end{bmatrix}
\begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_D \end{bmatrix}
=
\begin{bmatrix} \sum Y \\ \sum XY \\ \vdots \\ \sum X^D Y \end{bmatrix}
\qquad (1)
$$

şeklindedir.

(1) sisteminin çözümü :

$$
\mathbf{X} = \begin{bmatrix}
1 & X_1 & X_1^2 & X_1^3 \ldots & X_1^D \\
1 & X_2 & X_2^2 & X_2^3 \ldots & X_2^D \\
1 & X_3 & X_3^2 & X_3^3 \ldots & X_3^D \\
\ldots & \ldots & \ldots & \ldots & \ldots \\
1 & X_N & X_N^2 & X_N^3 \ldots & X_N^D
\end{bmatrix}, \quad
\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ \vdots \\ Y_N \end{bmatrix}, \quad
\mathbf{A} = \begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ \vdots \\ a_D \end{bmatrix}
$$

ile gösterilirse

$$
\mathbf{X^T . X} = \begin{bmatrix}
N & \sum X & \sum X^2 \ldots \sum X^D \\
\sum X & \sum X^2 & \sum X^3 \ldots \sum X^{D+1} \\
\ldots & \ldots & \ldots \\
\sum X^D & \sum X^{D+1} & \sum X^{D+2} \ldots \sum X^{2D}
\end{bmatrix}, \quad
\mathbf{X^T . Y} = \begin{bmatrix} \sum Y \\ \sum XY \\ \sum X^2 Y \\ \\ \sum X^D Y \end{bmatrix}
$$

olduğundan , (1) sistemi

$$(X^T . X) . A = X^T . Y \qquad (1)$$ şeklinde yazılabilir.

Bu eşitliğin her iki yanı soldan $(X^T . X)$ matrisinin inversi ile çarpılırsa

$$\boxed{A = (X^T . X)^{-1} . (X^T . Y)} \qquad (2)$$

elde edilir.

Sonuç :

(2) eşitliğinden yararlanılarak ,

(E) : $$Y = a_0 + a_1 X + a_2 X^2 + \ldots . + a_D X^D$$

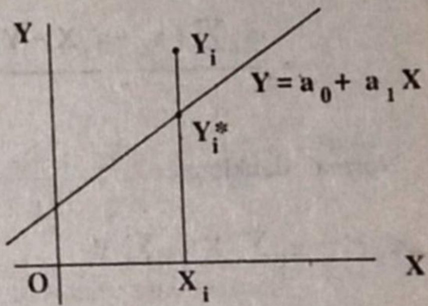
regresyon eğrisinin denklemindeki

$$a_0 , a_1 , a_2 , \ldots . , a_D$$

parametreleri hesaplanabilir.

STANDART HATA

X	X_1 X_2 . . . X_N
Gözlem değeri Y	Y_1 Y_2 . . . Y_N
Teorik değer Y*	Y_1^* Y_2^* . . . Y_N^*



Bileşik dağılımı gözönüne alınıyor. Y' lerin kestirimindeki standart hata

$$\boxed{s = \sqrt{\frac{\sum (Y - Y^*)^2}{N}}} \qquad (3)$$

olarak tanımlanır.

Eğer regresyon doğrusunun denklemi $Y = a_0 + a_1 X$ şeklinde ise bu tanım

$$s = \sqrt{\frac{\sum Y^2 - a_0 \sum Y - a_1 \sum XY}{N}} \qquad (4)$$

şekline dönüştürülebilir.

İspat :

Tanım'dan

$$s^2 = \frac{\sum (Y^* - Y)^2}{N} = \frac{\sum (a_0 + a_1 X - Y)^2}{N}$$

$$= \frac{\sum (a_0 + a_1 X - Y)(a_0 + a_1 X - Y)}{N}$$

$$= \frac{a_0 \sum (a_0 + a_1 X - Y) + a_1 \sum X(a_0 + a_1 X - Y) - \sum Y(a_0 + a_1 X - Y)}{N}$$

Normal denklemler

$$a_0 N + a_1 \sum X = \sum Y \quad \rightarrow \quad a_0 N + a_1 \sum X - \sum Y = 0 ,$$

$$a_0 \sum X + a_1 \sum X^2 = \sum XY \quad \rightarrow \quad a_0 \sum X + a_1 \sum X^2 - \sum XY = 0$$

olduğundan

$$\sum (a_0 + a_1 X - Y) = a_0 N + a_1 \sum X - \sum Y = 0 ,$$

$$\sum X (a_0 + a_1 X - Y) = a_0 \sum X + a_1 \sum X^2 - \sum XY = 0$$

dir. Bu değerler, yerlerine yazılırsa

$$s^2 = \frac{-\sum Y(a_0 + a_1 X - Y)}{N} = \frac{-a_0 \sum Y - a_1 \sum XY + \sum Y^2}{N}$$

$$s = \sqrt{\frac{\sum Y^2 - a_0 \sum Y - a_1 \sum XY}{N}} \qquad (4)$$

elde edilir. Bu sonuç, lineer olmayan regresyon denklemleri için genelleştirilebilir. Meselâ

$Y = a_0 + a_1 X + a_2 X^2$ parabolünden yapılacak olan kestirimlerin standart hatası

$$s = \sqrt{\frac{\sum Y^2 - a_0 \sum Y - a_1 \sum XY - a_2 \sum X^2 Y}{N}}$$

dır.

Örnek 5. 12)

X	2	5	6	10	12	16	21
Y	1	7	11	15	18	22	27

bileşik dağılımı için

1) $Y = a_0 + a_1 X$ regresyon doğrusu'nun denklemini bulunuz.

2) Standart hatayı hesaplayınız.

Yanıt :

X	Y	X^2	Y^2	XY
2	1	4	1	2
5	7	25	49	35
6	11	36	121	66
10	15	100	225	150
12	18	144	324	216
16	22	256	484	352
21	27	441	729	567

$\sum X = 72$, $\sum Y = 101$,

$\sum X^2 = 1006$, $\sum Y^2 = 1933$,

$\sum XY = 1388$.

1) Normal denklemler :

$$7\,a_0 + 72\,a_1 = 101$$
$$72\,a_0 + 1006\,a_1 = 1388$$

$$\Delta = \begin{vmatrix} 7 & 72 \\ 72 & 1006 \end{vmatrix} = 1858 ,$$

$$a_0 = \frac{\begin{vmatrix} 101 & 72 \\ 1388 & 1006 \end{vmatrix}}{1858} = 0.8988 , \quad a_1 = \frac{\begin{vmatrix} 7 & 101 \\ 72 & 1388 \end{vmatrix}}{1858} = 1.3154$$

Regresyon doğrusunun denklemi

$$Y = 0.8988 + 1.3154\,X$$

dir.

2) Bu regresyon doğrusu'ndan yapılacak kestirimlerin standart hatası

$$s = \sqrt{\frac{1933 - 0.8988\,(101) - 1.3154\,(1388)}{N}} = 1.53$$

dir.

Not : En küçük standart hatalı regresyon eğrisini bulan bir bilgisayar programı kitabın sonunda'dır .

KOVARİANS

Tanım 5. 5)

X , Y rasgele değişkenlerinin bileşik dağılımında , **Kovarians**

$$\sigma_{XY} = E\,[\,(X - \mu_X)(Y - \mu_Y)\,] \qquad (1)$$

olarak tanımlanır. Bu tanım

$$\sigma_{XY} = E\,[\,XY\,] - \mu_X\,\mu_Y \qquad (2)$$

şekline dönüştürülebilir.

İspat :

$$\sigma_{XY} = E[(X-\mu_X)(Y-\mu_Y)] = E[XY - \mu_Y X - \mu_X Y + \mu_X \mu_Y]$$

$$= E[XY] - \mu_Y E[X] - \mu_X E[Y] + \mu_X \mu_Y$$

Burada

$$E[X] = \mu_X \quad , \qquad E[Y] = \mu_Y$$

olduğundan , yerlerine yazılırsa

$$\boxed{\sigma_{XY} = E[XY] - \mu_X \mu_Y} \qquad (2)$$

elde edilir .

KORELASYON KATSAYISI

X , Y rasgele değişkenlerinin bileşik dağılımında **korelasyon katsayısı**

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \cdot \sigma_Y} \qquad (3)$$

olarak tanımlanır . ρ_{XY} korelasyon katsayısı , seçilen birime bağlı olmaksızın X , Y rasgele değişkenlerinin birlikte değişiminin bir ölçüsüdür. Korelasyon katsayısı

$$|\rho_{XY}| \leq 1 \quad \text{yani} \quad -1 \leq \rho_{XY} \leq 1 \qquad (4)$$

dir .

İspat :

λ parametresinin her değeri için

$$E[[\lambda(X-\mu_X) - (Y-\mu_Y)]^2] \geq 0$$

dır.

$$E[[\lambda^2(X-\mu_X)^2-2\lambda(X-\mu_X)(Y-\mu_Y)+(Y-\mu_Y)^2]]\geq 0,$$

$$\lambda^2 E[(X-\mu_X)^2]-2\lambda E[(X-\mu_X)(Y-\mu_Y)]+E[(Y-\mu_Y)^2]\geq 0.$$

λ nın $\quad \lambda=\dfrac{E[(X-\mu_X)(Y-\mu_Y)]}{E[(X-\mu_X)^2]}$

değeri için de eşitsizlik geçerlidir. λ'nın bu değeri yerine yazılırsa,

$$\frac{\{E[(X-\mu_X)(Y-\mu_Y)]\}^2}{E[(X-\mu_X)^2]}-2.\frac{\{E[(X-\mu_X)(Y-\mu_Y)]\}^2}{E[(X-\mu_X)^2]}+E[(Y-\mu_Y)^2]\geq 0$$

Burada

$$E[(X-\mu_X)(Y-\mu_Y)]=\sigma_{XY},\ E[(X-\mu_X)^2]=\sigma_X^2,\ E[(Y-\mu_Y)^2]=\sigma_Y^2$$

olduğundan yerlerine yazılırsa

$$-\frac{\sigma_{XY}^2}{\sigma_X^2}+\sigma_Y^2\geq 0,\quad \left(\frac{\sigma_{XY}}{\sigma_X.\sigma_Y}\right)^2\leq 1,\quad \rho_{XY}^2\leq 0,$$

$$\boxed{-1\leq \rho_{XY}\leq 1} \qquad (4)$$

dir.

İrdeleme :

1) X, Y rasgele değişkenleri birbirinden bağımsız ise

$E[XY]=E[X].E[Y]=\mu_X.\mu_Y$ olacağından (2) eşitliğine göre, korelasyon katsayısı $\rho_{XY}=0$ dır.

Fakat bunun karşıtı doğru değildir. Yani $\rho_{XY} = 0$ ise X, Y rasgele değişkenlerinin birbirinden bağımsız olması gerekmez.

2) $\rho_{XY} = 1$ veya $\rho_{XY} = -1$ ise X ile Y arasında doğrusal bir bağıntı vardır. Bu durumda (X , Y) noktaları $Y = mX + n$ doğrusu üzerindedir.

$\rho_{XY} = 1$ ise doğrunun eğimi : $m > 0$,

$\rho_{XY} = -1$ ise doğrunun eğimi : $m < 0$

dır. ρ_{XY} korelasyon katsayısı +1 ya da −1'e yaklaşırken X, Y nin birlite değişimi de doğrusal bir bağıntıya yaklaşır.

3) $\rho_{XY} = 0$ ise X , Y rasgele değişkenleri arasında doğrusal bir korelasyon yoktur.

Kesikli X , Y rasgele değişkenleri arasında doğrusal bir ilişki olabileceği kabul edilirse doğrusal korelasyon katsayısı

$$\rho_{XY} = \frac{\sum (X-\overline{X}).(Y-\overline{Y})}{\sqrt{[\sum (X-\overline{X})^2].[\sum (Y-\overline{Y})^2]}} \qquad (5)$$

dir. Bu formül

$$\rho_{XY} = \frac{N\sum XY - (\sum X).(\sum Y)}{\sqrt{[N\sum X^2 - (\sum X)^2].[N\sum Y^2 - (\sum Y)^2]}} \qquad (6)$$

şekline dönüştürülebilir.

Örnek 5.13)

X	0	1	2	3	4	5	6
Y	16	21	24	25	24	21	16

Yukarıdaki tablo ile verilen X, Y rasgele değişkenleri için lineer korelasyon katsayısını hesaplayınız. Sonucu açıklayınız.

X	Y	X Y	X^2	Y^2
0	16	0	0	256
1	21	21	1	441
2	24	48	4	576
3	25	75	9	625
4	24	96	16	576
5	21	105	25	441
6	16	96	36	256
21	147	441	91	3171

(6) Formülünden yararlanalım. Tablo'dan

$\sum X = 21$, $\sum Y = 147$,

$\sum XY = 441$, $\sum X^2 = 91$,

$\sum Y^2 = 3171$ hesaplanır.

N = 7 olduğundan, bu değerler (6) formülü'nün payında yerlerine yerleştirilirse,

$7(441) - (21)(147) = 0$, $\rho_{XY} = 0$ bulunur.

Bunun anlamı " X , Y rasgele değişkenleri arasında doğrusal bir korelasyonun bulunmadığı" dır. Fakat bu sonuç, X ile Y arasında bir ilişkinin bulunmadığı anlamına gelmez. Gerçekten

(X , Y) noktaları $Y = -X^2 + 6X + 16$ parabolü üzerindedir.

Örnek 5. 14)

Sayfa 233 - (Örnek 5. 12) bileşik dağılımında lineer korelasyon katsayını hesaplayınız.

Yanıt :

Tablo'daki sonuçlardan yararlanılır.

$\sum X = 72$, $\sum Y = 101$, $\sum X^2 = 1006$, $\sum Y^2 = 1933$, $\sum XY = 1388$

Bu sonuçlar (6) eşitliğinde yerlerine yerleştirilirse X , Y rasgele değişkenlerinin bileşik dağılımında lineer korelasyon katsayısı

$$= \frac{7(1388) - (72)(101)}{\sqrt{[7(1006) - (72)^2].[7(1933) - (101)^2]}} = \frac{2444}{2487} = 0.98$$

Örnek 5. 15)

80 öğrencinin **FİZİK (X)** ve **MATEMATİK (Y)** puanlarının bileşik dağılımı aşağıdaki frekanslı dağılım tablosunda verilmiştir. Kod'lama yöntemini kullanarak (6) formülü yardımıyla ρ_{XY} korelasyon katsayısını hesaplayınız.

Yanıt :

Sınıflardan herhangi birinin orta noktası 0 (sıfır) olarak kod'lanırsa (6) formülü

$$\rho_{XY} = \frac{N\sum f k_X k_Y - (\sum f_X k_X)(\sum f_Y k_Y)}{\sqrt{[N\sum f_X k_X^2 - (\sum f_X k_X)^2].[N\sum f_Y k_Y^2 - (\sum f_Y k_Y)^2]}} \qquad (7)$$

şekline dönüşür. Aşağıda, bileşik dağılım tablosu (7) formülü kullanılacak şekilde genişletilmiştir. Buradan hesaplanan değerler, (7) formülünde yerlerine yazılırsa ρ_{XY} korelasyon katsayısı elde edilir .

FİZİK → ; MATEMATİK ; ← Sınıfların orta noktaları

	X	34.5	44.5	54.5	64.5	74.5	84.5	94.5				
	k_X	−3	−2	−1	0	1	2	3	f_Y	$f_Y k_Y$	$f_Y k_Y^2$	$f_Y k_X k_Y$
Y	k_Y											
34.5	−3	1	2		2				5	−15	45	21
44.5	−2			2	5	3			10	−20	40	−2
54.5	−1		1			2			3	−3	3	0
64.5	0			2	12	10			24	0	0	0
74.5	1			1	8		2		11	11	11	3
84.5	2				5	6	9		20	40	80	48
94.5	3					1	3	3	7	21	63	48
f_X		1	3	5	32	22	14	3	$\sum f_X$ $\sum f_Y$ = 80	$\sum f_Y k_Y$ = 34	$\sum f_Y k_Y^2$ = 242	$\sum f_Y k_X k_Y$ = 118
$f_X k_X$		−3	−6	−4	0	15	28	9	$\sum f_X k_X$ = 39			
$f_X k_X^2$		9	12	5	0	15	56	27	$\sum f_X k_X^2$ = 124			
$f_X k_X k_Y$		9	14	3	0	7	58	27	$\sum f_X k_X k_Y$ = 118			