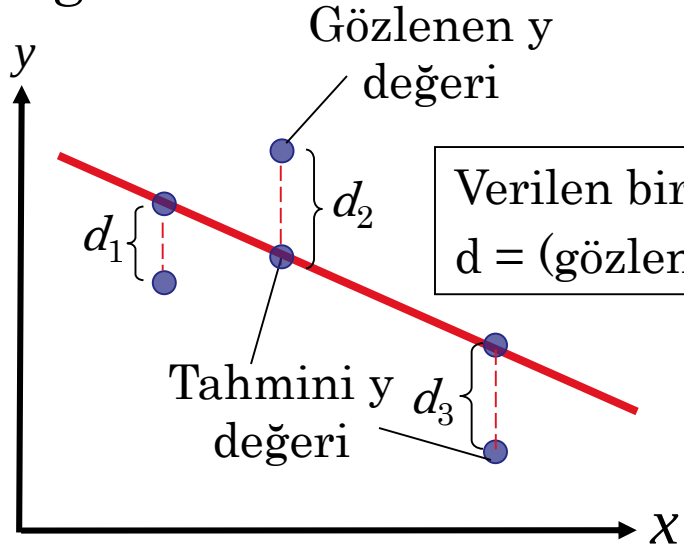
AKT102 İSTATİSTİK

BÖLÜM 12
REGRESYON

Artıklar

İki değişken arasındaki lineer korelasyonun anlamlı olduğunu doğruladıktan sonra, x değeriden y değerini tahmin etmek için kullanılabilecek çizginin denklemini belirleyeceğiz.



Verilen bir x değeri için
d = (gözlenen y değeri) - (tahmin edilen y değeri)

Her veri noktası d_i çizgideki belirli bir x-değeri için gözlenen y-değeri ile tahmini y-değeri arasındaki farkı temsil eder. Bu farklılıklara artıklar denir.

Regresyon Doğrusu

Aynı zamanda en uygun çizgi olarak da adlandırılan bir regresyon doğrusu, artıkların karelerinin toplamının minimum olduğu çizgidir.

Regresyon Doğrusu Denklemi

Bağımsız değişken x ve bağımlı değişken y için olan regresyon doğrusunun denklemi;

$$\hat{y} = mx + b$$

ŷ verilen bir x değeri için öngörülen y değeridir. m eğim ve y-kesim noktası b ise:

$$m = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2} \text{ and } b = \bar{y} - m\bar{x} = \frac{\sum y}{n} - m\frac{\sum x}{n}$$

where $\bar{y}$ is the mean of the y-values and $\bar{x}$ is the mean of the x-values. The regression line always passes through $(\bar{x}, \bar{y})$.

Regresyon Doğrusu

Örnek:

Regresyon doğrusu denklemini bulun.

x	y	xy	x^2	y^2
1	– 3	– 3	1	9
2	– 1	– 2	4	1
3	0	0	9	0
4	1	4	16	1
5	2	10	25	4
$\sum x = 15$	$\sum y = -1$	$\sum xy = 9$	$\sum x^2 = 55$	$\sum y^2 = 15$

$$m = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2} = \frac{5(9) - (15)(-1)}{5(55) - (15)^2} = \frac{60}{50} = 1.2$$

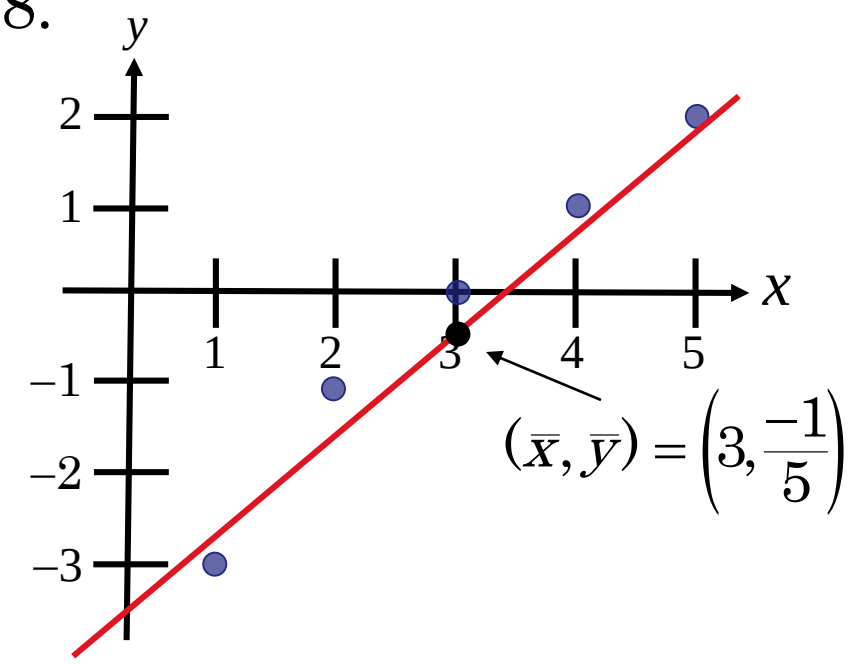
Regresyon Doğrusu

Örneğin devamı:

$$b = \bar{y} - m\bar{x} = \frac{-1}{5} - (1.2)\frac{15}{5} = -3.8$$

Regresyon doğrusu denklemi

$$\hat{y} = 1.2x - 3.8.$$



Regresyon Doğrusu

Örnek:
Aşağıdaki veriler hafta sonu boyunca 12 farklı öğrencinin televizyon izlediği saat sayısını ve ertesi Pazartesi sınava giren her öğrencinin puanlarını göstermektedir.

a.) Regresyon doğrusunun denklemini bulun.

b.) 9 saatlik TV izleyen bir öğrencinin beklenen puanını bulmak için denklemi kullanın.

saat, x	0	1	2	3	3	5	5	5	6	7	7	10
puan, y	96	85	82	74	95	68	76	84	58	65	75	50
xy	0	85	164	222	285	340	380	420	348	455	525	500
x^2	0	1	4	9	9	25	25	25	36	49	49	100
y^2	9216	7225	6724	5476	9025	4624	5776	7056	3364	4225	5625	2500

$\sum x = 54$ $\sum y = 908$ $\sum xy = 3724$ $\sum x^2 = 332$ $\sum y^2 = 70836$

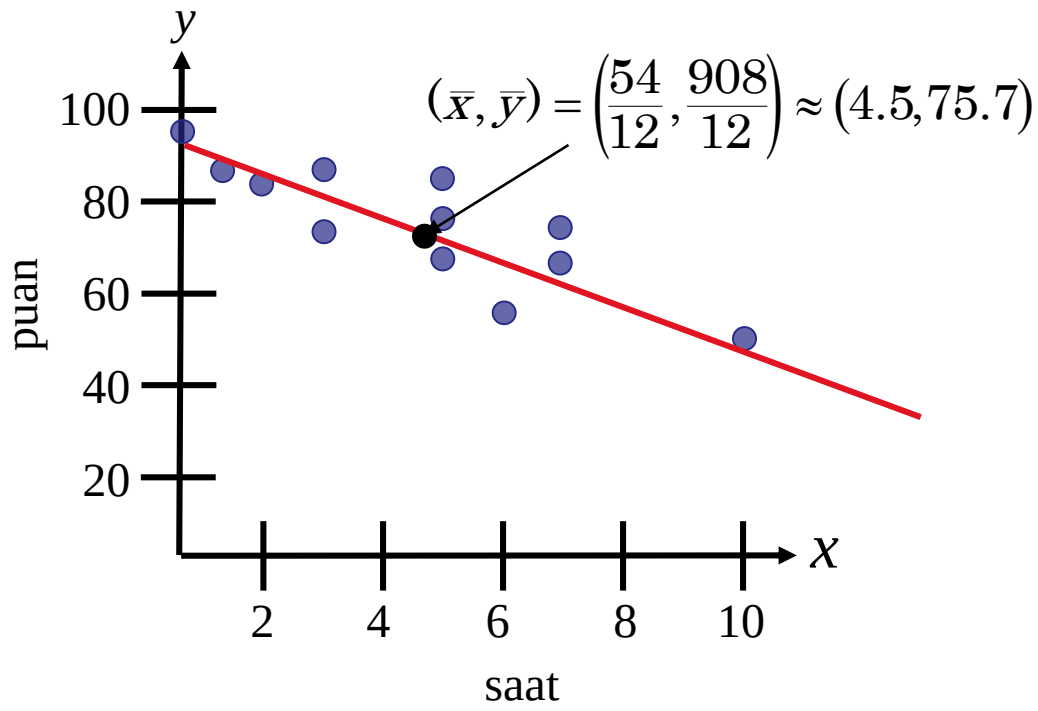
Regresyon Doğrusu

Örneğin devamı:

$$m = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2} = \frac{12(3724) - (54)(908)}{12(332) - (54)^2} \approx -4.067$$

$$b = \bar{y} - m\bar{x}$$
$$= \frac{908}{12} - (-4.067)\frac{54}{12}$$
$$\approx 93.97$$

$$\hat{y} = -4.07x + 93.97$$



Regresyon Doğrusu

Örneğin devamı:

$\hat{Y} = -4.07x + 93.97$ denklemini kullanarak, 9 saatlik TV izleyen bir öğrencinin puanını tahmin edebiliriz.

$$\hat{y} = -4.07x + 93.97$$
$$= -4.07(9) + 93.97$$
$$= 57.34$$

Hafta sonu 9 saat televizyon izleyen bir öğrenci pazartesi sınavından yaklaşık 57.34 puan almıştır.

Regresyon ve Tahmin Aralıklarının Ölçülmesi

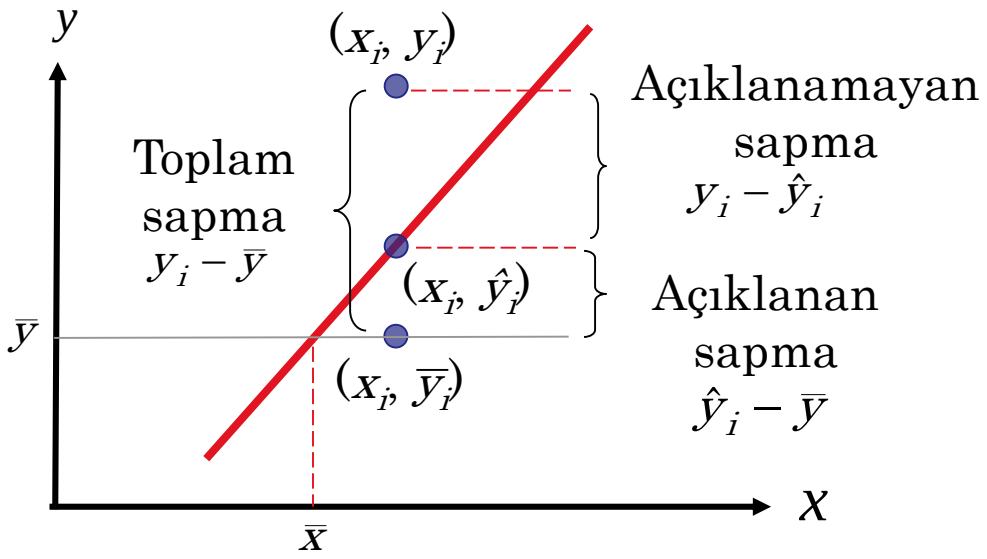
Regresyon Doğrusunun Değişimi

Toplam değişimi bulmak için önce toplam sapmayı, açıklanan sapmayı ve açıklanamayan sapmayı hesaplamanız gerekir

$$\text{Total deviation} = y_i - \bar{y}$$

$$\text{Explained deviation} = \hat{y}_i - \bar{y}$$

$$\text{Unexplained deviation} = y_i - \hat{y}_i$$



Regresyon Doğrusunun Değişimi

Bir regresyon doğrusundaki toplam değişim, her bir sıralı çiftin y değeri ile y'nin ortalaması arasındaki farkların karelerinin toplamıdır.

$$\text{Total variation} = \sum (y_i - \bar{y})^2$$

Açıklanan değişim, tahmin edilen her y-değeri ile y'nin ortalaması arasındaki farkların karelerinin toplamıdır.

$$\text{Explained variation} = \sum (\hat{y}_i - \bar{y})^2$$

Açıklanamayan değişim, her bir sıralı çiftin y değeri ile karşılık gelen tahmin edilen y değeri arasındaki farkların karelerinin toplamıdır.

$$\text{Unexplained variation} = \sum (y_i - \hat{y}_i)^2$$

Total variation = Explained variation + Unexplained variation

Belirtme Katsayısı

Belirtme katsayısı r^2, açıklanan değişimin toplam değişime oranıdır. Yani,

$$r^2 = \frac{\text{Explained variation}}{\text{Total variation}}$$

Örnek:
Öğrencilerin televizyon izlediği saat sayısını ve her öğrencinin puanlarını temsil eden veriler için korelasyon katsayısı $r \approx -0.831$'dir. Belirleme katsayısını bulun.

$$r^2 \approx (-0.831)^2$$
$$\approx 0.691$$

Puanlardaki değişimin yaklaşık% 69,1'i, izlenen TV saatindeki değişimle açıklanabilir. Değişimin yaklaşık% 30.9'u açıklanmamıştır.

Tahminin Standart Hatası

Bir x değerinden bir ŷ değeri tahmin edildiğinde, tahmin bir nokta tahminidir.

Bir aralık da yapılabilir.

s_e tahmininin standart hatası, verilen bir xi-değeri için tahmin edilen $\hat{y}$-değeri ile gözlemlenen yi değerlerinin standart sapmasıdır.

$$s_e = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{n-2}}$$

burada n, veri kümesindeki sıralı çiftlerin sayısıdır.

Gözlenen y değerleri tahmin edilen y değerlerine ne kadar yakınsa, tahminin standart hatası o kadar küçük olacaktır.

Tahminin Standart Hatası

Tahminin Standart Hatasını Bulma

Açıklama	*Gösterim*
1. Gösterilen sütun başlığını içeren bir tablo yapın.	$x_i,\ y_i,\ \hat{y}_i,\ (y_i - \hat{y}_i),\ (y_i - \hat{y}_i)^2$
2. Tahmin edilen y-değerlerini hesaplamak için regresyon denklemini kullanın.	$\hat{y} = mx_i + b$
3. Her gözlemlenen y değeri ve karşılık gelen tahmin edilen y değeri arasındaki farkların karelerinin toplamını hesaplayın.	$\sum(y_i - \hat{y}_i)^2$
4. Tahminin standart hatasını bulun.	$s_e = \sqrt{\dfrac{\sum(y_i - \hat{y}_i)^2}{n-2}}$

Tahminin Standart Hatası

Örnek:
Regresyon denklemi aşağıdaki gibiyse,

$$\hat{y} = 1.2x - 3.8.$$

tahminin standart hatasını bulun.

x_i	y_i	$\hat{y}_i$	$(y_i - \hat{y}_i)^2$
1	– 3	– 2.6	0.16
2	– 1	– 1.4	0.16
3	0	– 0.2	0.04
4	1	1	0
5	2	2.2	0.04
			$\sum = 0.4$

Açıklanamayan değişim

$$s_e = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n-2}} = \sqrt{\frac{0.4}{5-2}} \approx 0.365$$

Belirli bir x değeri için öngörülen y değerinin standart sapması yaklaşık 0,365'tir.

Tahminin Standart Hatası

Örnek:
Hafta sonu boyunca 12 farklı öğrencinin televizyon izlediği saat sayısını ve ertesi Pazartesi sınava giren her öğrencinin puanı temsil eden verinin regresyon denklemi

$$\hat{y} = -4.07x + 93.97.$$

Tahminin standart hatasını bulun

saat, x_i	0	1	2	3	3	5
Puan, y_i	96	85	82	74	95	68
$\hat{y}_i$	93.97	89.9	85.83	81.76	81.76	73.62
$(y_i - \hat{y}_i)^2$	4.12	24.01	14.67	60.22	175.3	31.58

saat, x_i	5	5	6	7	7	10
puan, y_i	76	84	58	65	75	50
$\hat{y}_i$	73.62	73.62	69.55	65.48	65.48	53.27
$(y_i - \hat{y}_i)^2$	5.66	107.74	133.4	0.23	90.63	10.69

Tahminin Standart Hatası

Örneğin devamı:

$$\sum(y_i - \hat{y}_i)^2 = 658.25$$

Açıklanamayan değişim

$$s_e = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{n-2}} = \sqrt{\frac{658.25}{12-2}} \approx 8.11$$

Öğrenci puanlarının, belirli bir saat izlenen TV yayını için standart sapması yaklaşık 8.11'dir.

Aralık Tahmini

İki değişken, eğer x'in herhangi bir sabit değeri için, karşılık gelen y değerleri normal dağılıyor ve y'nin herhangi bir sabit değeri için karşılık gelen x değerleri de normal dağılırsa iki değişkenli normal dağılıma sahiptir.

y'nin gerçek değeri için bir tahmin aralığı oluşturulabilir.

Doğrusal bir regresyon denklemi $\hat{y} = mx + b$ ve x_0, belirli bir x değeri, y için bir c-tahmin aralığı

$$\hat{y} - E < y < \hat{y} + E$$

$$E = t_c s_e \sqrt{1 + \frac{1}{n} + \frac{n(x_0 - \bar{x})^2}{n\sum x^2 - (\sum x)^2}}.$$

Nokta tahmini $\hat{y}$ ve hata payı E'dir. Tahmin aralığının y içerme olasılığı c'dir.

Aralık Tahmini

x'in Belirli Değeri için y için bir Tahmin Aralığı Oluşturun

	Açıklama	*Gösterim*
1.	n veri setindeki sıralı çift sayısını ve serbestlik derecelerini tanımlayın.	$\text{d.f.} = n - 2$
2.	Nokta tahminini $\hat{y}$ bulmak için regresyon denklemini ve verilen x değerini kullanın.	$\hat{y} = mx_i + b$
3.	Verilen c güven seviyesine karşılık gelen kritik t_c değerini bulun.	Ek B deki tabloyu kullanın.

Aralık Tahmini

x'in Belirli Değeri için y için bir Tahmin Aralığı Oluşt

	Açıklama	*Gösterim*
4.	s_e tahminin standart hatasını bulun.	$s_e = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{n-2}}$
5.	Hata payı E yi bulun.	$E = t_c s_e \sqrt{1 + \frac{1}{n} + \frac{n(x_0 - \bar{x})^2}{n\sum x^2 - (\sum x)^2}}$
6.	Sol ve sağ bitiş noktalarını bulun ve tahmin aralığını oluşturun.	Sol bitiş noktası: $\hat{y} - E$ Sağ bitiş noktası: $\hat{y} + E$ Aralık: $\hat{y} - E < y < \hat{y} + E$

Aralık Tahmini

Örnek:
Aşağıdaki veriler hafta sonu boyunca 12 farklı öğrencinin televizyon izlediği saat sayısını ve ertesi Pazartesi sınava giren her öğrencinin puanlarını göstermektedir.

saat, x	0	1	2	3	3	5	5	5	6	7	7	10
puan y	96	85	82	74	95	68	76	84	58	65	75	50

$\hat{y} = -4.07x + 93.97$ $\quad s_e \approx 8.11$

4 saat TV izlenirken puanlar için % 95 bir tahmin aralığı oluşturun.

Aralık Tahmini

Örneğin devamı:
TV izleme saat sayısı 4 olduğunda test puanları için% 95 tahmin aralığı oluşturun.

$n - 2 = 12 - 2 = 10$ serbestlik derecesi

Nokta tahmini

$$\hat{y} = -4.07x + 93.97 = -4.07(4) + 93.97 = 77.69.$$

Kritik değer $t_c = 2.228$, ve $s_e = 8.11$.

$$\hat{y} - E < y < \hat{y} + E$$

$77.69 - 8.11 = 69.58$ $77.69 + 8.11 = 85.8$

Bir öğrencinin hafta sonu boyunca 4 saatlik TV seyretmesi durumunda, öğrencinin sınav notunun 69.58 ile 85.8 arasında olacağından% 95 güven düzeyinde söyleyebiliriz.

Çoklu Regresyon

Çoklu Regresyon Denklemi

Birçok durumda, birden fazla bağımsız (açıklayıcı) değişken kullanarak bağımlı (yanıt) bir değişken için daha iyi bir tahmin bulunabilir.

Örneğin, Pazartesi günkü sınav notunun önceki bölümden daha doğru bir şekilde tahmin edilmesi, öğrencinin alacağı diğer derslerin sayısı ve öğrencinin test materyali hakkındaki önceki bilgileri dikkate alınarak yapılabilir.

Çoklu bir regresyon denkleminin formu;

$$\hat{y} = b + m_1x_1 + m_2x_2 + m_3x_3 + \ldots + m_kx_k$$

x_1, x_2, x_3,..., x_k bağımsız değişkenler, b, y yi kestiği noktadır ve y, bağımlı değişkendir.

* Bu kavramla ilişkili matematik karmaşık olduğundan, genellikle çoklu regresyon denklemini hesaplamak için teknoloji kullanılır.

y- Değerlerinin Tahmini

Çoklu regresyon çizgisinin denklemini bulduktan sonra, denklemi veri aralığında y-değerlerini tahmin etmek için kullanabilirsiniz..

Örnek:
Yıllık ABD pirinç verimini (pound cinsinden) tahmin etmek için aşağıdaki çoklu regresyon denklemi kullanılsın.

$$\hat{y} = 859 + 5.76x_1 + 3.82x_2$$

burada x_1 ekili alanın sayısı (bin cinsinden) ve x_2, hasat edilen dönüm sayısı (bin cinsinden).

a.) $x_1 = 2758$ ve $x_2 = 2714$ olduğunda yıllık pirinç verimini tahmin edin.

b.) Yıllık pirinç verimini, $x_1 = 3581$, ve $x_2 = 3021$ olduğunda tahmin edin.

y- Değerlerinin Tahmini

Örneğin devamı:

a.) $\hat{y} = 859 + 5.76x_1 + 3.82x_2$

$= 859 + 5.76(2758) + 3.82(2714)$

$= 27{,}112.56$

Tahmini yıllık pirinç verimi 27,1125.56 pound.

b.) $\hat{y} = 859 + 5.76x_1 + 3.82x_2$

$= 859 + 5.76(3581) + 3.82(3021)$

$= 33{,}025.78$

Tahmini yıllık pirinç verimi 33,025.78 pound.