TEMEL İSTATİSTİK

Regresyon I

Prof. Dr. Ezel Tavşancıl

Regresyon Nedir?

- Galton, 19. yy'da yaptığı çalışmada;
  - Uzun boylu ebeveynlerin çocuklarının ortalama olarak anne babalarından daha kısa; kısa boylu ebeveynlerin çocuklarının da ortalama olarak anne babalarından daha uzun olduğunu belirliyor.
  - Buna örnekteki, boy uzunluklarının grup ortalamasına çekilmesine *regresyon* deniyor.
- **Regresyon**, bir değişkene ilişkin ölçümlerin grup ortalamasına doğru çekilmesidir.

Regresyon Tarihsel Gelişim

- Öncelikle regresyon yöntemini geliştirir. (Galton tarafından)
- Galton, diagramda verileri Z standart puanlarına dönüştürür, aralarındaki doğrusal bağıntıyı bulur ve regresyon denklemini kurar.
- Daha sonra Pearson, değişkenler arasındaki ilişkiyi sayısal olarak ifade eden yöntemi bulur ve Pearson korelasyon katsayısını geliştirir.
- Galton'un regresyon denklemindeki regresyon katsayısı (regresyon doğrusunun eğimi) de aslında korelasyon katsayısıdır. Zaten, Pearson korelasyonu da X ve Y puanlarının standart puanlara dönüştürüldüğünde regresyon doğrusunun eğimidir.

- Regresyon analizi, aralarında ilişki olan iki ya da daha fazla değişkenden birinin bağımlı değişken, diğerlerinin bağımsız değişkenler olarak ayrımı ile aralarındaki ilişkinin matematiksel bir eşitlik ile açıklanması sürecidir.
  - Öğrencinin zekâ puanı, öğrenme motivasyonu, sorumluluk duygusu değişkenlerinden yararlanarak başarısı yordanabilir.
  - Kişinin suç işlemeye yatkınlığı, cinsiyet, ırk, din, fizikî özellikler ve stres gibi değişkenlere bakılarak kestirilebilir.

- y=f(X) > Değişkenler arası ilişkinin matematiksel fonksiyonu doğrusal, üssel, eğrisel olabilir.

- İlişkinin Yönü:
  - Değişim aynı yöne mi yoksa ayrı yönlerde mi? > Artan fonksiyon ya da azalan fonksiyon

- İlişkinin Gücü:
  - Çok kuvvetli, zayıf ya da ilişki yok.

- Regresyon denkleminin iki değişken arasındaki ilişki 0'dan farklı olduğunda kurulması uygundur.

(Büyüköztürk, Çokluk ve Köklü, 2018)

Regresyon ile İlgili Temel Kavramlar

- Olaylar ve olgular arasındaki ilişkilerin betimlenmesindeki temel amaç çoğu kez ortaya konulan ilişkiye dayanarak ileriye yönelik tahmin yapmaktır.
- Öğrencilerin sınav kaygıları ile depresyon düzeyleri arasındaki ilişkiye dayanarak, depresyonun kaygıya dayalı olarak ne derece kestirilebilir olduğunu araştırabilirsiniz.
- Sadece bilgi ve deneyimlere bağlı olarak iki olay arasında kurulan ilişkiyi temel alan tahminler yapılabilir.
- Bu tahminlerde, yanılma payı bilinmez ve yapılan tahmini formülleştirmek mümkün değil. (sistematik veri yok)

- **Bağımsız Değişken**: Genellikle X ile gösterilir. Başka bir değişken tarafından etkilenmeyen ama y'nin nedeni olan ya da onu etkilediği düşünülen (açıklayıcı) değişkendir.
- **Bağımlı Değişken**: Genellikle Y ile gösterilir. X değişkenine bağlı olarak değişebilen ya da ondan etkilenen (açıklanan) değişkendir.

- **!!!** X neden, Y sonuç demek doğru değil; eşitlik tam tersi de kurulabilirdi. Nedensellik için ancak iyi bir kuramsal dayanak ve deneysel düzenek gereklidir.

BASİT DOĞRUSAL REGRESYON MODELİ (POPULASYON MODELİ)

POPULASYON MODELİ

y = D+ βx+ ε

- $\widehat{y_i}$= a+ bx_i+e_i

  $\hat{y}$= bağımlı değişken

  x= bağımsız değişken

  a= sabit (y-eksenini kestiği nokta)

  b =regresyon doğrusunun eğimi (regresyon katsayısı)

  ε= hata terimi veya artık

ÖRNEKLEM MODELİ

$\hat{y}$= a + bx

$\hat{y}$ = Tahmin edilen y değeri (bağımlı değişken)

a = regresyon sabit değerinin yansız tahmini

b = regresyon eğiminin yansız tahmini

x = bağımsız değişken değeri

Hata terimi minimum olsun istiyoruz.

Bu nedenle, i. gözlem için eşitliği yazarken hata terimini eşitlikten çıkarıyoruz ve sağdaki gibi formulize ediyoruz.

İki değişken arasındaki korelasyon katsayısı, ortalama ve standart sapma değerleri biliniyorsa eşitlik (Tanis, 1987):

$$\widehat{Y} = \overline{Y} + r\frac{S_Y}{S_X}(X - \overline{X})$$

- Regresyon eşitliğindeki katsayı ve sabit değeri inceleyelim:
  - **Eğim (b):** X'deki bir birimlik değişmenin Y'de yol açtığı değişim miktarı.

  İşareti bağımlı ve bağımsız değişken arasındaki ilişkinin yönü hakkında bilgi verir.
  X (bağımsız değişken): Terapi seansı saati
  Y (bağımlı değişken): Özgüven düzeyi

$$\hat{Y} = 50 + 25\,X$$

  1 saatlik terapi seansı alan bireyin özgüven düzeyi 25 puan artmaktadır.

  - **Sabit (a):** Hiç X değişkeni etkisi söz konusu olmadığında Y değişkeninin değeri.

  Hiç terapi almayan bireyin özgüveni 50 puan değerinde olacaktır.

- Regresyon analizinde, bağımlı ve bağımsız değişkenler en az aralık ölçeği düzeyinde olmalıdır.
  - Bağımsız değişken sıralama ölçeği düzeyindeyse logaritmik dönüşüm
  - Bağımsız değişken sınıflama ölçeği düzeyindeyse kukla değişken atama yapılır.

Regresyon Denklemindeki Terimler

β = b = eğim

- Bağımsız değişkendeki değişime dayalı olarak bağımlı değişkende görülen değişimdir.
- Eğimin alacağı katsayının işareti iki değişken arasındaki ilişkiye bağlı olarak pozitif veya negatif olabilir.
- b= $r_{XY}\frac{S_Y}{S_X}$ (Y ve X'in standart sapması eşitse, b=r_{XY})

ya da

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

α = a = sabit

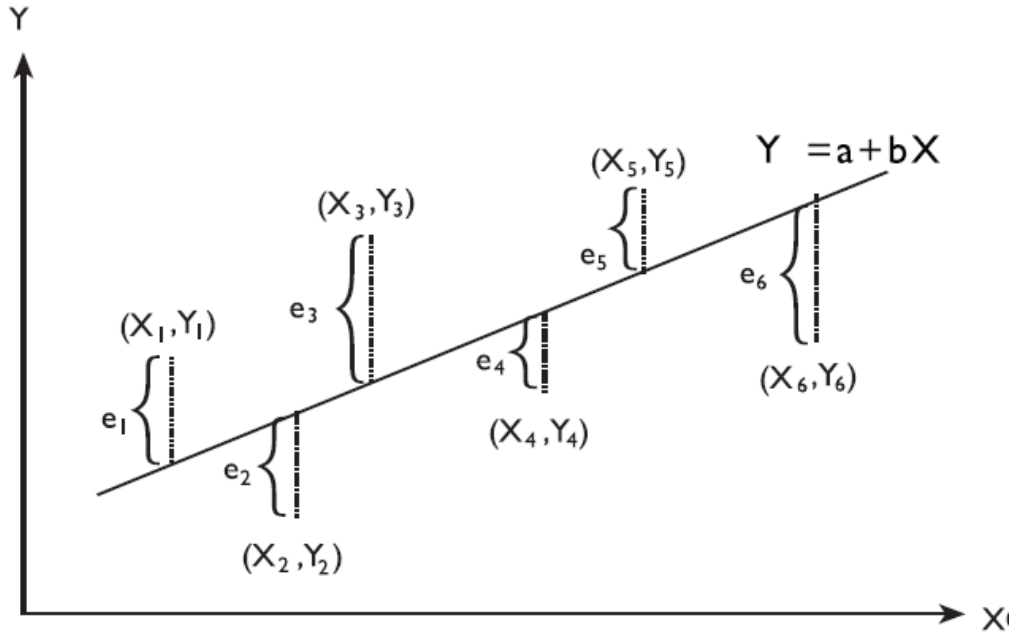
- Doğrunun y eksenini kestiği nokta.
- Bağımsız değişkenin değerinin 0 olduğu durumda bağımlı değişkenin aldığı değerdir.

$$a=\bar{y}-b\bar{x}$$

ε = e = Hata terimi (artık)

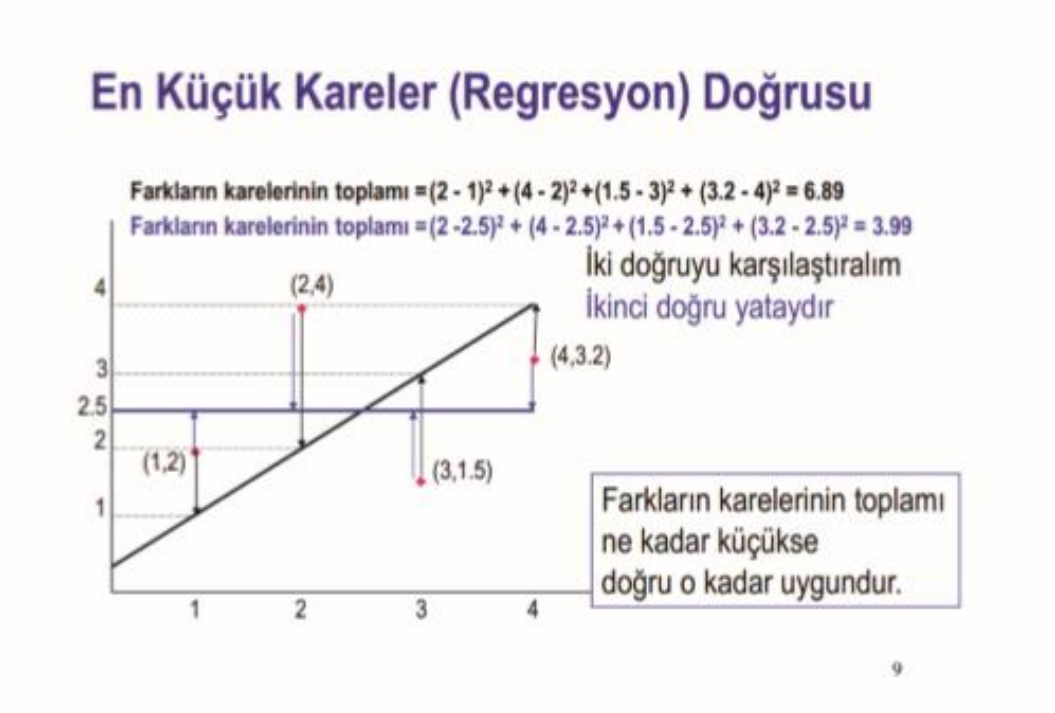
- Regresyon modelleri tam (%100) doğru tahmin yapma özeliğine sahip değillerdir.
- Ana kütleden yapılan gözlem değerleri genellikle bir doğru üzerinde sıralanmayıp rassallığa bağlı olarak doğrudan sapmalar gösterirler.
- Hata terimi (artık), gözlenen değer ile model tarafından tahmin edilen değer arasındaki farktır.
- $\varepsilon = y - \hat{y}$

$$Y = a + bX + e$$



En Küçük Kareler Yöntemi

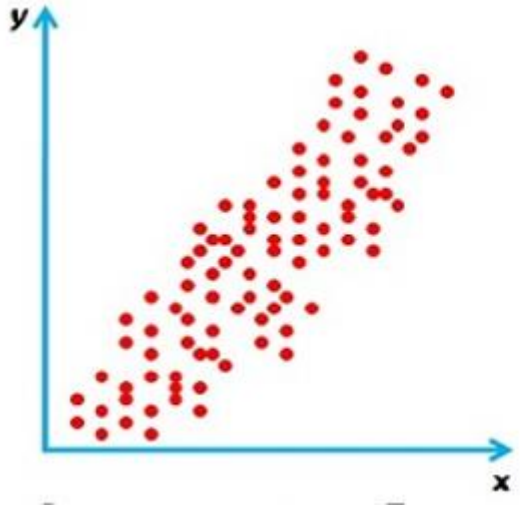
- Regresyon eşitliğinde bilinmeyen a ve b parametrelerinin tahmini, gözlenen veri çiftlerinin (X_i ve Y_i) oluşturduğu noktalar ile regresyon doğrusu arasındaki sapmaların kareleri toplamını en küçük yapacak şekilde gerçekleştirilen yöntemdir.

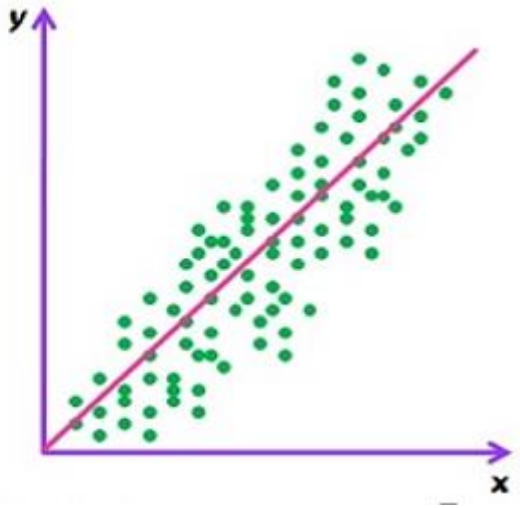


Determinasyon Katsayısı

- r^2(açıklanan varyans) değişkenlerin birindeki değişimin ne kadarının diğer değişkenler tarafından açıklandığını yüzde olarak eden bir değerdir.
- Determinasyon katsayısı olarak da isimlendirilir.
- Regresyon modeli ile Y'deki değişimlerin yüzde ne kadarı açıklanabilir?
- r= $\mp 1\ ise?$

Korelasyon vs Regresyon





Karşılaştırma Yönü	Korelasyon	Regresyon
Anlam	Korelasyon, iki değişken arasındaki ilişkiyi belirleyen istatistiksel bir ölçüdür.	Regresyon, bağımsız bir değişkenin, bağımlı değişkenle sayısal olarak nasıl ilişkili olduğunu açıklar.
Kullanım	İki değişken arasındaki doğrusal ilişkiyi göstermek	En iyi satıra sığdırmak ve bir değişkeni başka bir değişken temelinde tahmin etmek.
Bağımlı ve Bağımsız Değişken	Hangisinin bağımlı hangisinin bağımsız değişken olduğu fark etmez.	İki değişken de farklıdır.
Gösterge	Korelasyon katsayısı, iki değişkenin birlikte hareket etme derecesini gösterir.	Regresyon, yordayıcı değişkendeki (x) bir birim değişikliğinin yordanan değişken (y) üzerindeki etkisini gösterir.
Amaç	Değişkenler arasındaki ilişkiyi ifade eden sayısal bir değer bulmak.	Sabit değişkenli değerleri baz alarak rasgele değişkenin değerlerini tahmin etme

Regresyon Analizinin 4 Temel Amacı

- Bağımlı ve bağımsız değişken arasındaki ilişkiyi regresyon ile açıklamak
- Regresyon modelinin bilinmeyen parametreleri tahmin edildiğinde, bağımsız değişken/lerin bilinen değeri için bağımlı değişkenin değerini tahmin etmek
- Bağımsız değişken/lerin bağımlı değişkende gözlenen değişmelerin ne kadarını açıkladıklarını determinasyon katsayısı ile belirlemek
- Bağımsız değişken ya da değişkenlerin bağımlı değişkeni manidar bir şekilde kestirip kestirmediklerini; birden fazla bağımsız değişken var ise bunların bağımlı değişken üzerindeki göreli önemliliklerini saptamak

- Değişkenler arası ilişki doğrusalsa: **doğrusal regresyon**
- Tek bağımsız değişken varsa**: basit regresyon**

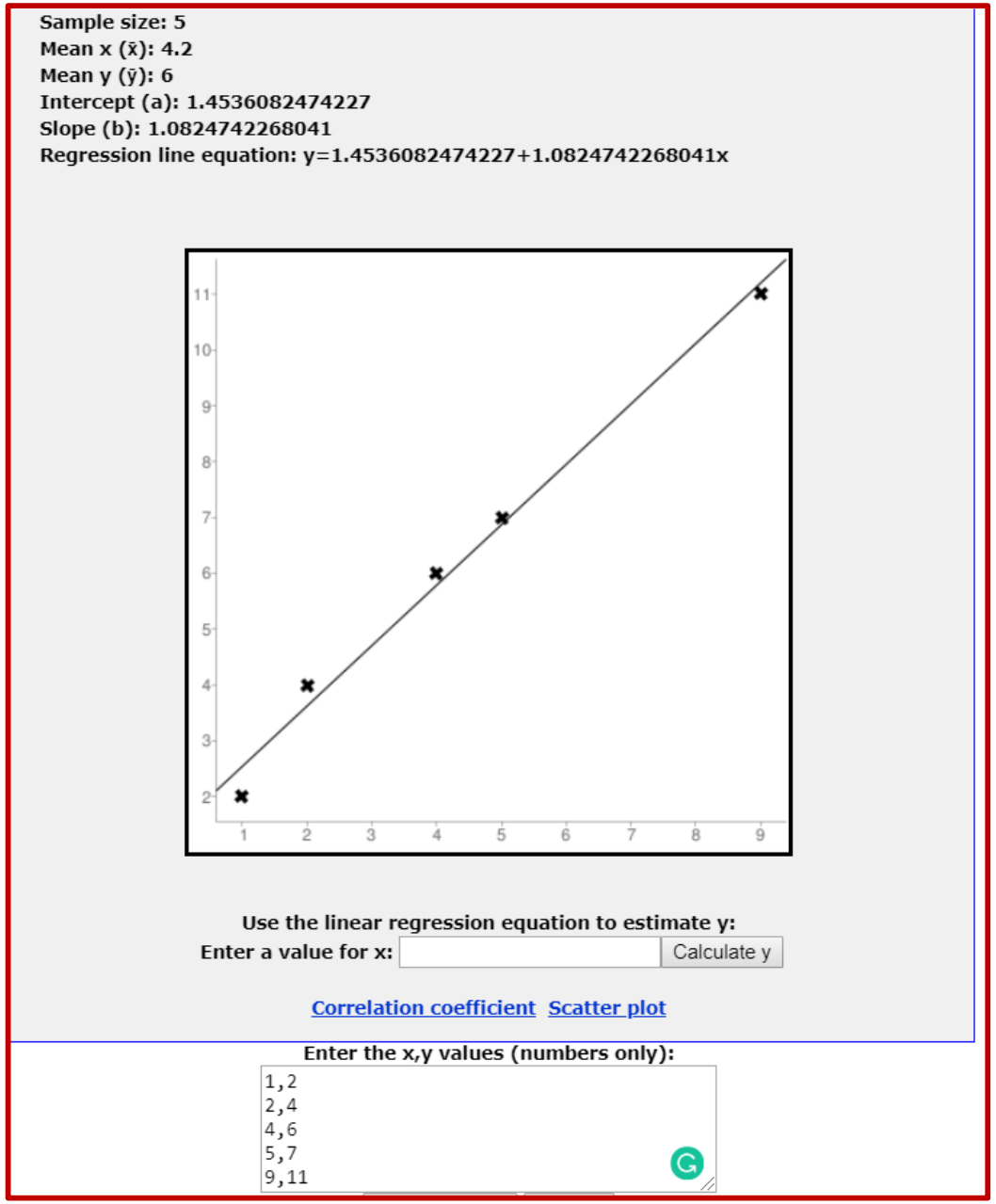
Basit doğrusal regresyondaki basit kelimesi iki değişken arasındaki ilişkiyi açıklamak için kullanılmasından, doğrusal kelimesi ise kurulan modelin parametreleri açısından doğrusal bir model olmasındandır.

Bağımlı değişken mutlaka sürekli olmalıdır.

- Değişkenlerde uç değer varsa:
  - Dönüştürme yapılabilir.
  - Uç değerler veriden çıkarılabilir.
  - Gözlem sayısı artırılabilir.

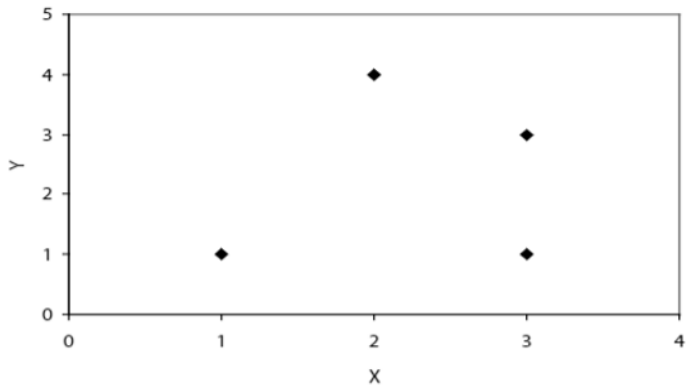
Online Hesaplama Sitesi

- http://www.alcula.com/calculators/statistics/linear-regression/
- Her bir (x, y) ikilisini *enter* ile alt satıra girerek *submit data*'ya basılır.
- r=0 ve r=∓1 olduğu durumlar için kurulacak regresyon modeli hakkında ne söylenebilir?

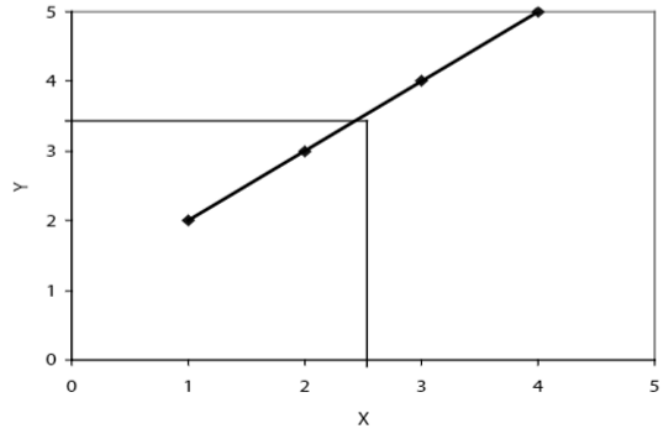


Tablo 6.1: *X ve Y Değişkenlerine Ait Veri Çiftleri*

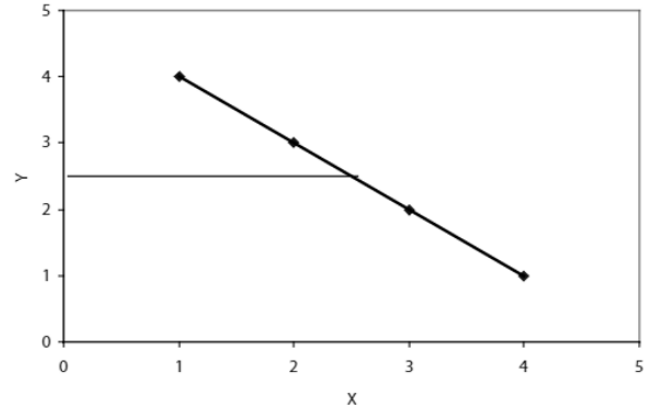
a		b		c	
X	Y	X	Y	X	Y
1	1	1	2	1	4
2	4	2	3	2	3
3	3	3	4	3	2
3	1	4	5	4	1



Şekil 6.1a: *Tablo 6.1a'daki Veriler İçin Saçılma Diyagramı*



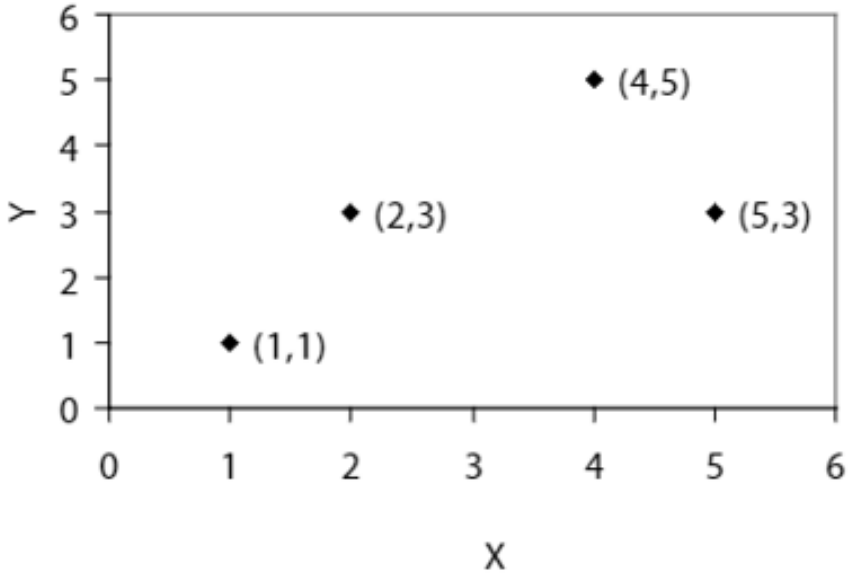
Şekil 6.1b; *Tablo 6.1b'deki Veriler İçin Saçılma Diyagramı*



Şekil 6.1c: *Tablo 6.1c'deki Veriler İçin Saçılma Diyagramı*

ÖRNEK I: Bir grup öğrencinin çalışma saati ve başarı puanı değerleri verilmiştir.

X:	2	4	1	5
Y:	3	5	1	3



- Çalışma saati (X) bağımsız değişken ve başarı puanı (Y) olarak alındığında regresyon eşitliğini hesaplayalım.

(Büyüköztürk ve diğerleri, 2018)

ÖRNEK I ÇÖZÜMÜ...

Tablo *Çalışma Saati (X) ve Başarı Puanı (Y)*

X	Y	XY	X^2	Y^2
2	3	6	4	9
4	5	20	16	25
1	1	1	1	1
5	3	15	25	9
$\Sigma X=12$	**$\Sigma Y=12$**	**$\Sigma XY=42$**	**$\Sigma X^2=46$**	**$\Sigma Y^2=44$**

$$b_{YX} = \frac{n\sum XY - (\sum X)(\sum Y)}{n\sum X^2 - (\sum X)^2}$$

$$b_{YX} = \frac{4(42) - (12)(12)}{4(46) - 144} = \frac{(168) - (144)}{(184) - (144)} = \frac{24}{20}$$

$b_{YX} = 0.60$

Şimdi de, $a_{YX} = \overline{Y} - b_{YX}\ \overline{X}$ formülünü kullanılarak sabit değeri bulalım.

$a_{YX} = 3-(0.60)3 = 3-(1.8) = 1.2$

Hesaplanan b_{YX} ve a_{YX} katsayıları regresyon eşitliğinde yerlerine koyulduğunda,

$\widehat{Y} = 1.2 + 0.6X$ bulunur.

…ÖRNEK I ÇÖZÜMÜ…

- $\hat{Y}$= 1.2+0.6X
  - Ders çalışma saati ile başarı puanı arasında pozitif yönlü bir ilişki var.
  - Çalışma saatindeki bir birimlik artış, başarı puanının 0.6'lık bir artışa neden olur.
  - Çalışma saati 0 olan birinin başarı puanı, 1.2 olacaktır.
  - İki değişken arasındaki ilişki (r=0.67), determinasyon katsayısı r^2=0.67*0.67= 0.45. İki değişken arasında pozitif ve orta düzeyde bir ilişki var. Çalışma saati, başarı puanlarındaki varyansın %45'ini açıklar.

GEÇEN HAFTADAN KORELASYON HESAPLAMA

Tablo *Çalışma Saati ve Başarı Puanları*

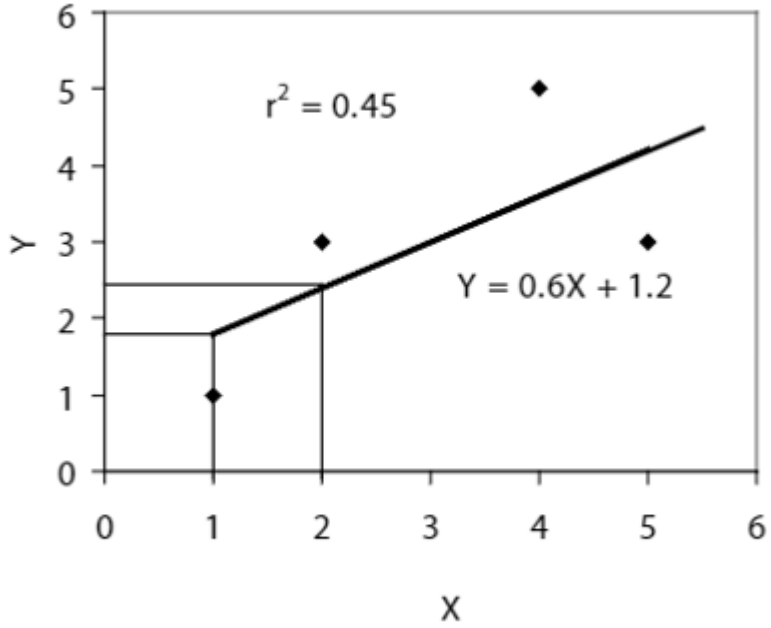
Birey	X	X^2	Y	Y^2	XY
1	2	4	3	9	6
2	4	16	5	25	20
3	1	1	1	1	1
4	5	25	3	9	15
n=4	$\sum X$=12	$\sum X^2$ =46	$\sum Y$=12	$\sum Y^2$=44	$\sum XY$=42
	$(\sum X)^2$=144		$(\sum Y)^2$ =144		

$$r = \frac{n\sum XY - \sum X \sum Y}{\sqrt{\left[n\sum X^2 - (\sum X)^2\right]\left[n\sum Y^2 - (\sum Y)^2\right]}}$$

$$r = \frac{4(42) - (12)(12)}{\sqrt{\left[4(46) - (12)^2\right]\left[4(44) - (12)^2\right]}} = \frac{24}{\sqrt{(40)(32)}} = \frac{24}{35.8} = 0.67$$

…ÖRNEK I ÇÖZÜMÜ

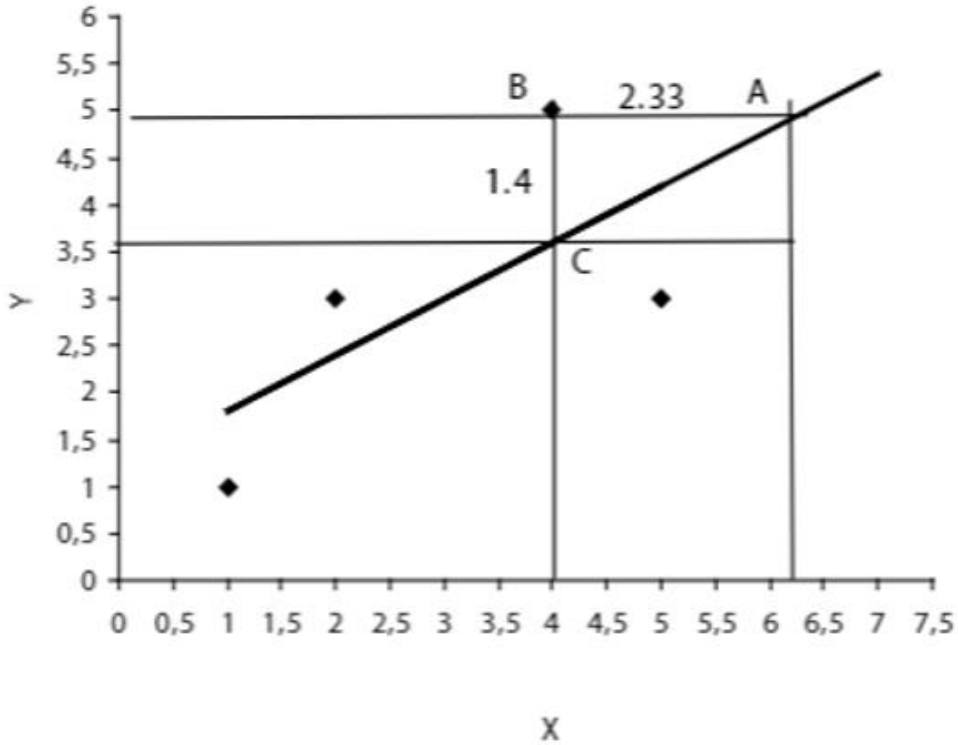
- Elde edilen regresyon denklemine ait olan doğruyu çizelim.
  - X'in katsayısı (eğim) pozitif mi negatif mi?
  - Belirlenen iki rastgele X değerine karşılık gelen y değeri hesaplanır.
    - X=1 için $\hat{Y}$ =1.2+(0.6*1)=1.8 ve
    - X=2 için $\hat{Y}$=1.2+(0.6*2)=2.4
  - Saçılma diagramının üzerine hesaplanan noktalardan geçen doğru çizilir.



$\hat{Y}$ = 1.2+0.6X (X=4 değeri için sapma tablo ve grafikten)

Tablo *Gözlenen Değer (X, Y), Tahmini Değerler ($\widehat{Y}$), Sapma Miktarları ($Y-\widehat{Y}$)*

X	Y	$\widehat{Y}$	$Y-\widehat{Y}=(e)$	$(Y-\widehat{Y})^2=e^2$
2	3	2.4	0.6	0.36
4	5	3.6	1.4	1.96
1	1	1.8	-0.8	0.64
5	3	4.2	-1.2	1.44
			$\sum e=0.0$	**$\sum e^2=4.4$**



Şekil 6.2c: *Gözlenen Değerler, X=4 için Tahmin Edilen Değer ve Sapma Miktarı*

Şekil 6.2c'de X=4 için gözlenen değer olan Y=5'ten X eksenine çizilen parelel doğrunun regresyon doğrusunu kestiği nokta A ile B ve C noktaları bir dik üçgen oluşturmaktadır. Dik üçgenin X=4 için sapma miktarını gösteren [BC] kenarının, üçgenin diğer dik kenarı [BA]'ya oranı regresyon katsayısını (b) veya regresyon doğrusunun eğimini verir. Örneğimizde, regresyon katsayısı,

$$b=\frac{[BC]}{[BA]}=\frac{1.4}{2.33}=0.6\text{'dır.}$$

Burada izlenen yöntem, En Küçük Kareler Yöntemi'dir. Bu yöntemde, gözlenen değerler ile tahmin edilen değerler arasında kalan uzaklığın, yani gözlenen değerler ile regresyon doğrusu arasındaki uzaklıkların kareler toplamını $[\sum(Y-\widehat{Y})^2]=\sum e^2$ en küçük olacak şekilde doğruya ait parametrelerin (a ve b) değerlerini bulmak söz konusudur. Anılan farkların kareleri toplamının sıfır olması durumunda, gözlenen değerler ile tahmin edilen değerler birbirine eşittir. Bu durumda $S_{YX}=0$ olur ve bu tahminin mükemmel olduğunu gösterir. Böyle bir sonuç, ancak iki değişken arasındaki korelasyonun ∓ 1.0 olduğu durumda görülür.

ÖRNEK II: Bir danışman, 20 danışanının katıldığı seans saati ile süreç sonundaki depresyon düzeylerini ölçüyor. **Buna göre uyguladığı terapi seans saati (X) ile depresyon düzeyi (Y) arasında anlamlı bir ilişki bulunmakta mıdır? Basit doğrusal regresyon modelini kurunuz.**

No	Seans Saati	Depresyon	No	Seans Saati	Depresyon
1	33	10.8	11	33	10.5
2	33	9.5	12	30	11.0
3	23	14.2	13	35	10.9
4	34	9.7	14	25	14.0
5	32	11.2	15	22	13.8
6	35	9.7	16	28	12.9
7	30	12.1	17	27	12.8
8	23	13.0	18	29	11.0
9	28	12.0	19	24	13.5
10	26	13.2	20	31	10.8

KAYNAKLAR

- Büyüköztürk, Ş., Çokluk, Ö ve Köklü, N. (2018). Sosyal bilimler için istatistik. Ankara: Pegem Akademi.