Kaynak Kitap: TEMEL BİYOMETRİ (2013) Kocabaş Z., Özkan, MM ve Başpınar E.
Ankara Üniversitesi Ziraat Fakültesi Yayın No:1606 Ders Kitabı 558

# KORELASYON VE REGRESYON

## 3.1. Giriş

Yürütülen birçok araştırmada, araştırıcı üzerinde çalıştığı örnekte birden fazla özelliğe ait veri toplayabilir. Birden fazla özelliğe ait veri toplandığı zaman sadece bu özellikler için tanıtıcı istatistiklerin hesaplanması ve bu özelliklere ilişkin populasyonların tanımlanması yeterli değildir. Birden fazla özelliğe ait veri toplanması durumunda araştırıcı, bu özellikler arasında bir ilişki olup olmadığı, varsa bu ilişkinin derecesini, özelliklerden birinin değişmesine karşılık diğer özelliğin ne kadar değiştiği vb. soruları araştırmak isteyebilir.

Eğer söz konusu deney ünitelerinin iki özelliği üzerinde duruluyorsa bu iki özellik arasındaki ilişkinin belirlenmesi amacıyla hesaplanan korelasyon katsayısı basit korelasyon katsayısı (Pearson korelasyon katsayısı) olarak, hesaplanan regresyon katsayısı da basit regresyon katsayısı olarak adlandırılır. Özellikler arasındaki ilişkiler doğrusal olabildiği gibi doğrusallığın dışındaki çeşitli eğrisel ilişkiler şeklinde de olabilir.

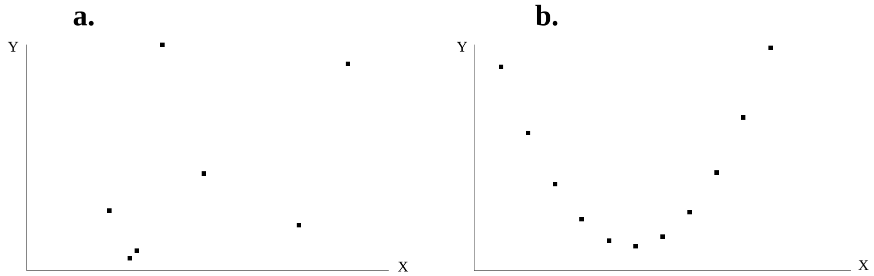
Bu kitabın kapsamında iki özellik ve bu özellikler arasındaki ilişkinin doğrusal olduğu durum dikkate alınacaktır.

## 3.2. Korelasyon Katsayısı

Korelasyon katsayısı, iki özellik arasındaki ilişkinin doğrusallık derecesini verir. Bu sebeple birimi yoktur. Korelasyon katsayısı -1 ile +1 arasındadır. -1'den küçük ve +1'den büyük olamaz. Korelasyon katsayısının negatif işaretli olması, iki özellik arasında azalan (ters) ilişki olduğunu yani özelliklerden birinin artarken diğerinin azaldığını gösterir. Pozitif işaretli korelasyon katsayısı ise iki özellik arasında artan bir ilişki olduğunu gösterir.

Örnekteki bireylerin iki özelliğine ait veriler toplandıktan sonra, bu verilerin koordinat sisteminde işaretlenmesi, korelasyon katsayısını hesaplamaya başlamadan önce özellikler arasında doğrusal ilişki olup olmadığı hakkında bir fikir verebilir.

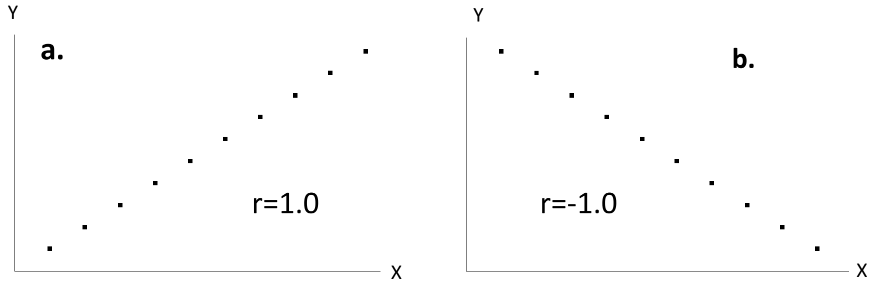
Elde edilen veriler koordinat düzleminde işaretlendiği zaman, Şekil 3.1a'da görüldüğü gibi bir grafik elde ediliyorsa, bu iki özellik arasındaki korelasyon katsayısının 0 veya 0'a çok yakın olduğunu yani değişkenlerden biri artarken diğer değişkenin artıp yada azalmadığını gösterir. İki özelliğe ait veriler koordinat düzleminde işaretlendiği zaman Şekil 3.1b'deki gibi bir grafik elde ediliyorsa, bu iki özellik arasında doğrusal bir ilişkinin olmadığını gösterir ve bu durumda yine korelasyon katsayısı 0 veya 0'a çok yakın bir değer alır.



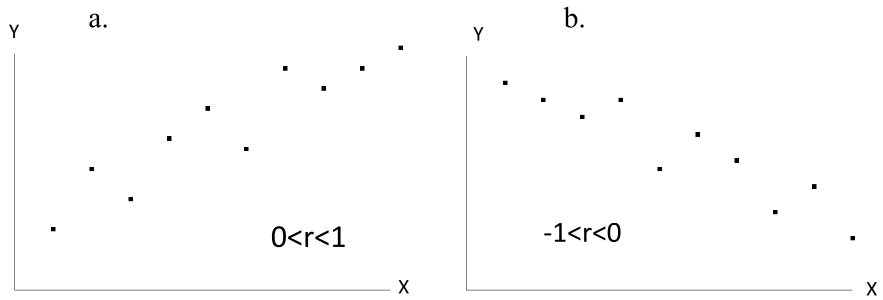

ŞEKİL 3.1. X ve Y özellikleri arasındaki ilişkinin 0 veya 0'a yakın olabileceği durumların ikisi

Her X değerine karşılık bir Y değerinin karşılık geldiği özelliklere ait veriler koordinat ekseninde işaretlendiği zaman eğer Şekil 3.2a ve b'deki grafikler elde ediliyorsa bu durumda iki özellik arasında tam, %100'lük, bir ilişki vardır ve X ve Y özellikleri arasındaki korelasyon katsayısı hesaplandığında sırasıyla +1.0 ve -1.0 olarak bulunur. Korelasyon katsayısının işareti ilişkinin yönü hakkında bilgi verir. Eğer korelasyon katsayısı pozitif işaretli ise iki özellik arasında artan (Şekil 3.1a), negatif işaretli ise azalan (ters) bir ilişki (Şekil 3.1b) vardır.

Özelliklere ait veriler koordinat ekseninde işaretlendiği zaman Şekil 3.3a ve b'deki gibi grafikler elde ediliyorsa bu iki özellik arasında tam bir ilişkinin olmadığını gösterir. Elde edilen grafik Şekil 3.3a'daki gibi ise korelasyon katsayısı 0 ila +1.0 arasında, Şekil 3.3b'deki gibi ise -1.0 ila 0 arasındadır.




ŞEKİL 3.2. X ve Y özellikleri arasında (a.) artan ve (b.) azalan tam doğrusal ilişki




ŞEKİL 3.3. X ve Y özellikleri arasında tam olmayan a. artan ve b. azalan doğrusal ilişki

Elde edilen veriler koordinat ekseninde işaretlendiği zaman şekil 3.1 veya 3.2'deki gibi grafikler elde edilirse korelasyon katsayısının hangi değeri aldığı hakkında yorum yapmak mümkündür. Fakat Şekil 3.3'de görüldüğü gibi grafikler elde ediliyorsa korelasyon katsayısının bir formül yardımıyla hesaplanması gerekir.

Korelasyon katsayısı populasyonda hesaplanıyorsa parametredir ve ρ ile gösterilir. X ve Y gibi iki özellik arasındaki korelasyon katsayısı, eşitlik (3.1) kullanılarak hesaplanır. Örnekte hesaplanıyor ise istatistiktir ve r ($r=r_{xy}=r_{yx}$) ile gösterilir. Eşitlik (3.2) kullanılarak hesaplanır.

Populasyonda:

$$\rho = \frac{\sum(X_i-\mu_x)(Y_i-\mu_y)}{\sqrt{\sum(X_i-\mu_x)^2\sum(Y_i-\mu_y)^2}} \qquad \ldots(3.1)$$

Örnekte:

$$r = \frac{\sum(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum(X_i - \overline{X})^2 \sum(Y_i - \overline{Y})^2}} \qquad \ldots(3.2)$$

(3.1) ve (3.2) numaralı eşitliklerde $\sum(X_i - \mu_x)(Y_i - \mu_y)$ ve $\sum(X_i - \overline{X})(Y_i - \overline{Y})$ çarpımlar toplamını, paydadaki karekök içindeki terimler ise Bölüm II'de açıklandığı gibi kareler toplamlarını göstermektedir. Çarpımlar toplamı kısaca $\sum d_x d_y$, X'e ve Y'ye ait kareler toplamları da sırasıyla $\sum d_x^2$ ve $\sum d_y^2$ şeklinde gösterilir. Bu durumda (3.1) ve (3.2) numaralı eşitlikler (3.3) numaralı eşitlikte görüldüğü gibi yazılabilir.

$$r = \frac{\sum d_x d_y}{\sqrt{\sum d_x^2 \sum d_y^2}} \qquad \ldots(3.3)$$

Çarpımlar toplamı ($\sum d_x d_y$), X özelliğine ilişkin gözlem değerlerinin kendi ortalamalarından olan farkları ile Y özelliğine ilişkin gözlem değerlerinin kendi ortalamalarından olan farklarının çarpımlarının toplamıdır. Çarpımlar toplamının daha kolay bir şekilde hesaplanması için $\sum d_x d_y = \sum(X_i - \overline{X})(Y_i - \overline{Y})$ ifadesinin açılarak sadeleştirilmesiyle elde edilen (3.4) numaralı eşitlik kullanılır. (3.4) numaralı eşitlikte, her birey için X ve Y özelliklerine ilişkin gözlem değerlerinin teker teker çarpımlarının toplamından X ve Y değerlerinin toplamlarının çarpımının birey sayısına bölümü çıkarılır. X ve Y özellikleri arasındaki ilişkinin yönüne bağlı olarak çarpımlar toplamı negatif veya pozitif işaret alır. Kareler toplamı hiçbir zaman negatif olamayacağından, Korelasyon katsayısının işaretini Çarpımlar Toplamının işareti belirlemektedir.

$$\sum d_x d_y = \sum X_i Y_i - \frac{(\sum X_i)(\sum Y_i)}{n} \qquad \ldots(3.4)$$

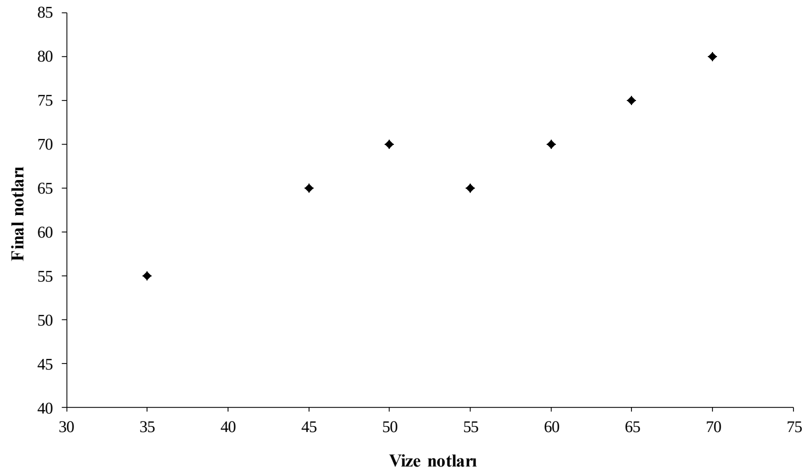
**ÖRNEK 1:**

Herhangi bir dersten 7 adet öğrencinin vize ve final sınavından aldıkları notlar Tablo 3.1'deki gibidir. Vize ve final notları arasındaki doğrusal ilişkinin derecesini (korelasyon katsayısı) hesaplayınız.

Tablo 3.1. Herhangi bir dersten 7 adet öğrencinin vize ve final notları

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Vize notu (X) | 35 | 45 | 50 | 70 | 65 | 60 | 55 |
| Final notu (Y) | 55 | 65 | 70 | 80 | 75 | 70 | 65 |

Öğrencilerin vize ve final notları belirlendikten sonra Şekil 3.4'de görüldüğü gibi notlar koordinat ekseninde işaretlenir. Şekil 3.4'de elde edilen grafik, vize ve final notları arasında artan bir ilişki olduğunu, yani öğrencilerin vize notları arttıkça final notlarının da arttığını ve bu iki özellik arasındaki korelasyon katsayısının da 0 ila +1.0 arasında bir değer alacağını göstermektedir.




Şekil 3.4. Herhangi bir dersten 7 adet öğrencinin vize notlarına karşılık gelen final notlarının koordinat eksenindeki görünümü

Korelasyon katsayısının hesaplanması için gerekli adımlar aşağıdaki gibidir.

| Vize notu (X) | Final notu (Y) | XY | $X^2$ | $Y^2$ |
|---|---|---|---|---|
| 35 | 55 | 1925 | 1225 | 3025 |
| 45 | 65 | 2925 | 2025 | 4225 |
| 50 | 70 | 3500 | 2500 | 4900 |
| 70 | 80 | 5600 | 4900 | 6400 |
| 65 | 75 | 4875 | 4225 | 5625 |
| 60 | 70 | 4200 | 3600 | 4900 |
| 55 | 65 | 3575 | 3025 | 4225 |
| $\sum X = 380$ | $\sum Y = 480$ | $\sum XY = 26600$ | $\sum X^2 = 21500$ | $\sum Y^2 = 33300$ |

Yukarıdaki hesaplamalar yapıldıktan sonra (2.15) numaralı eşitlik kullanılarak X ve Y'ye ait kareler toplamı ve (3.4) numaralı eşitlik kullanılarak ta çarpımları toplamı hesaplanır:

$$\sum d_x^2 = 21500 - \frac{(380)^2}{7} \cong 871.429$$

$$\sum d_y^2 = 33300 - \frac{(480)^2}{7} \cong 385.71$$

$$\sum d_x d_y = 26600 - \frac{(380)(480)}{7} \cong 542.86$$

(3.3) numaralı eşitlikten korelasyon katsayısı

$r = \dfrac{(542.86)}{\sqrt{(871.429)(385.71)}} \cong +0.936$ olarak hesaplanır. Hesaplanan korelasyon katsayısı öğrencilerin vize ve final notları arasında %93.6'lık artan doğrusal bir ilişki olduğunu göstermektedir.

**ÖRNEK 2:**

Bir gölde sonbaharda belirli derinliklerdeki (m) oksijen miktarı mg/L olarak Tablo 3.2'deki gibi ölçülmüştür. Göl derinliği ve söz konusu derinliklerde ölçülen oksijen miktarları arasındaki korelasyon katsayısının hesaplanması için gerekli işlemler Tablo3.2'de verilmiştir.

Tablo 3.2. Bir gölde sonbaharda belirli derinliklerde oksijen miktarı

| Derinlik (m) (X) | Oksijen miktarı (mg/L) (Y) | XY | $X^2$ | $Y^2$ |
|---|---|---|---|---|
| 10 | 7.0 | 70.0 | 100 | 49 |
| 15 | 6.5 | 97.5 | 225 | 42.25 |
| 20 | 5.6 | 112.0 | 400 | 31.36 |
| 30 | 5.4 | 162.0 | 900 | 29.16 |
| 40 | 5.1 | 204.0 | 1600 | 26.01 |
| 50 | 4.9 | 245.0 | 2500 | 24.01 |
| 60 | 3.9 | 234.0 | 3600 | 15.21 |
| 70 | 2.1 | 147.0 | 4900 | 4.41 |
| 80 | 1.6 | 128.0 | 6400 | 2.56 |
| $\sum X = 375$ | $\sum Y = 42.10$ | $\sum XY = 1399.5$ | $\sum X^2 = 20625$ | $\sum Y^2 = 223.97$ |

$$\sum d_x^2 = 20625 - \frac{(375)^2}{9} \cong 5000$$

$$\sum d_y^2 = 223.97 - \frac{(42.10)^2}{9} \cong 27.04$$

$$\sum d_x d_y = 1399.5 - \frac{(375)(42.10)}{9} \cong -354.67$$

(3.3) numaralı eşitlikten korelasyon katsayısı

$r = \frac{(-354.67)}{\sqrt{(5000)(27.04)}} = -0.965$ olarak hesaplanır. Hesaplanan korelasyon katsayısı negatif işaretli çıktığından göl derinliği ile söz konusu derinliklerde ölçülen oksijen miktarları arasında %96.5'lik azalan bir ilişki olduğu yani, göl derinliği artıkça oksijen miktarının azaldığı sonucuna varılır.

## 3.3. Regresyon Katsayısı

Yapılan bir çalışmada iki özellik üzerinde çalışılıyorsa, özelliklerden (değişkenlerden) biri diğerine bağlı olarak değişiyor olabilir. Dolayısıyla özelliklerden biri diğerinin fonksiyonu olarak dikkate alınabilir. Örneğin, öğrencilerin boy ve ağırlıkları üzerinde yapılan bir çalışmada ağırlık, boyun bir fonksiyonu olarak dikkate alınabilir ya da bir araç fren yaparak durduğu zaman, yolda bıraktığı iz (metre olarak) aracın hızının bir fonksiyonu olarak yazılabilir. Eğer üzerinde durulan özelliklerden biri, örneğin Y, diğer özelliğin, örneğin X'in, fonksiyonu olarak yazılabiliyorsa bu durum **Y=f(X)** şeklinde ifade edilir. Yukarıda verilen örneklerde **ağırlık=f(boy)** ve **fren izi=f(hız)** şeklinde gösterilir.

Üzerinde durulan X ve Y gibi iki özellik **Y=f(X)** şeklinde ifade ediliyorsa, X değişkeni **bağımsız değişken**, Y değişkeni ise, X'e bağlı olarak değişen yani **bağımlı değişken** olarak ifade edilir. Bu durumda X bağımsız değişkeninin kendi birimi cinsinden bir birim değişmesine karşılık Y bağımlı değişkeninin kendi birimi cinsinden ortalama olarak değişeceği miktar araştırılır. İki özellik arasındaki ilişkinin miktar olarak ifadesi, regresyon katsayısı olarak bilinen bir katsayının hesaplanmasını gerektirir. Hesaplanan regresyon katsayısının bir birimi vardır ve bu birim, bağımlı değişkenin ölçü birimiyle aynıdır.

X ve Y gibi iki özellik üzerinde çalışılıyorsa $b_{yx}$ ve $b_{xy}$ olmak üzere iki regresyon katsayısı hesaplanabilir:

$b_{yx}$, Y'nin X'e göre regresyon katsayısıdır, X bağımsız değişkeninin kendi birimi cinsinden bir birim artmasına karşılık Y bağımlı değişkeninin kendi birimi cinsinden ortalama olarak değişeceği miktarı gösterir ve (3.5) numaralı eşitlik kullanılarak hesaplanır.

$$b_{yx} = \frac{\sum d_x d_y}{\sum d_x^2} \qquad \ldots(3.5)$$

$b_{xy}$, X'in Y'ye göre regresyon katsayısıdır, Y bağımsız değişkeninin kendi birimi cinsinden 1 birim artmasına karşılık X bağımlı değişkeninin kendi birimi cinsinden ortalama olarak değişeceği miktarı gösterir ve (3.6) numaralı eşitlik kullanılarak hesaplanır.

$$b_{xy} = \frac{\sum d_x d_y}{\sum d_y^2} \qquad \ldots(3.6)$$

Korelasyon katsayısı, sadece $-1 \le r \le 1$ aralığındaki değerleri alabilirken, regresyon katsayısı için böyle bir sınırlama yoktur ve $-\infty \le b \le +\infty$ aralığındaki tüm reel sayıları değer olarak alabilir. Regresyon katsayısı da korelasyon katsayısında olduğu gibi pozitif veya negatif işaretlidir. Regresyon katsayısının negatif işaretli olması iki özellik arasında ters (azalan), pozitif işaretli olması ise iki özellik arasında artan bir ilişki olduğunu gösterir. Bir örnek için hesaplanan regresyon ve korelasyon katsayılarının işaretinin farklı olması söz konusu değildir. Çünkü her iki katsayının hesaplanması için aynı çarpımlar toplamı kullanılır ve katsayıların işaretini belirleyen çarpımlar toplamıdır.

**ÖRNEK 1:**

Tablo 3.1'de 7 öğrencinin vize (X) ve final notlarına (Y) ait korelasyon katsayısı hesaplanmıştı. Öğrencilerin vize notunun 1 not artmasına karşılık, final notunun ortalama olarak ne kadar değiştiği görülmek istense, Y'nin (final notunun) X'e (vize notuna) göre regresyon katsayısının (3.5) numaralı eşitlik kullanılarak hesaplaması gerekir.

$$\sum d_x^2 = 21500 - \frac{(380)^2}{7} \cong 871.429$$

$$\sum d_x d_y = 26600 - \frac{(380)(480)}{7} \cong 542.86$$ olarak hesaplanmıştı.

Bu durumda (3.5) numaralı eşitlikten regresyon katsayısı;

$$b_{yx} = \frac{542.86}{871.429} \cong 0.62$$ olarak bulunur.

Hesaplanan regresyon katsayısı pozitif işaretli olduğundan değişimin artış şeklinde ifade edileceğini ve öğrencilerin vize notunun 1 not artmasına karşılık, final notunun ortalama olarak 0.62 not artacağını gösterir.

**ÖRNEK 2:**

Bir gölde sonbaharda 9 farklı derinlikte (X, m) ölçülen oksijen miktarları (Y, mg/L) Tablo 3.2'de verilmiş ve iki özellik arasındaki korelasyon katsayısı r=-0.965 olarak hesaplanmıştı. Eğer araştırıcı göl derinliğinin 1 metre artmasına karşılık oksijen miktarının mg/L olarak ortalama ne kadar değiştiğini araştırmak isterse oksijen miktarının derinliğe göre regresyon katsayısını hesaplamalıdır. Derinliklere ait kareler toplamı ve iki özellik için çarpımlar toplamı;

$$\sum d_x^2 = 20625 - \frac{(375)^2}{9} \cong 5000$$

$\sum d_x d_y = 1399.5 - \dfrac{(375)(42.10)}{7} \cong -354.67$ olarak hesaplanmıştı. (3.5) numaralı eşitlik kullanılarak regresyon katsayısı ise;

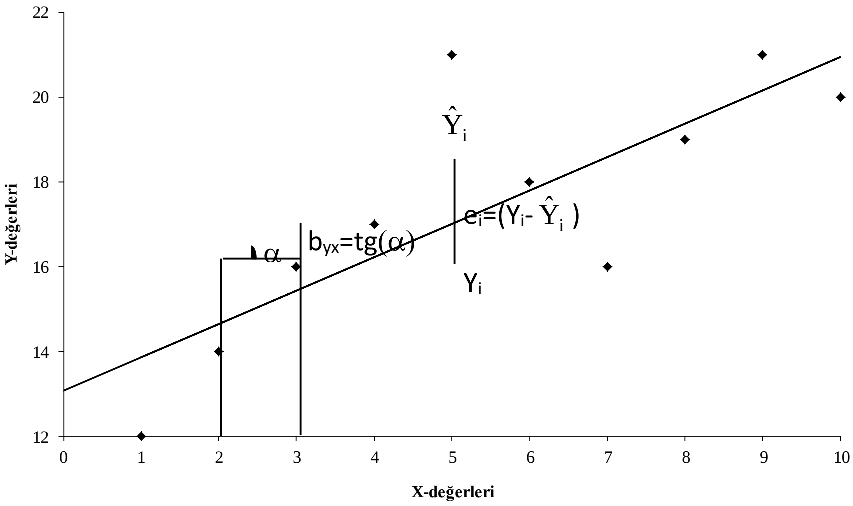
$$b_{yx} = \frac{-354.67}{5000} \cong -0.071 \quad \text{mg/L olarak bulunur.}$$

Hesaplanan regresyon katsayısı negatif işaretli olduğundan değişimin azalış şeklinde ifade edileceğini ve göl derinliğinin 1 m artmasına karşılık oksijen miktarının ortalama olarak 0.071 mg/L azalacağını gösterir.

### 3.4. Regresyon Denklemi

Yapılan bir çalışmada üzerinde durulan X ve Y özellikleri birbirlerinin fonksiyonu ise ve Y=f(X) şeklinde ifade ediliyorsa, X bağımsız değişkeninin değerlerinden Y bağımlı değişkeninin değerleri tahmin edilebilir. İki özellik arasında tam bir ilişki varsa her X değerine farklı bir Y değeri karşılık gelir ve (X,Y) gözlem çiftleri koordinat düzleminde hepsi bir doğru üzerinde yer alacak şekilde dağılırlar. Bu durumda doğrunun denklemi kolay bir şekilde oluşturulabilir. Fakat biyolojik bilimlerde çalışılan özelliklerin arasında çoğunlukla tam bir ilişki yoktur. İki özellik arasında tam bir ilişki olmadığı zaman işaretlenen noktaların hepsine birden en yakın geçen doğrunun oluşturulması gerekir. Bu doğruya "**regresyon doğrusu**", bu doğruyu çizmek için oluşturulan denkleme de "**regresyon denklemi**" veya "**önceden tahmin denklemi**" denir. Regresyon denklemi (veya önceden tahmin denklemi) $\mathbf{\hat{Y} = a + b_{yx}X}$ veya $\mathbf{\hat{X} = a + b_{xy}Y}$ şeklinde oluşturulur. Şekil 3.5'te görüldüğü gibi denklemdeki "a" değeri, regresyon doğrusunun Y-eksenini kestiği noktanın ordinatı, $b_{yx}$ ise regresyon doğrusunun eğimi, yani regresyon katsayısıdır.

a




Şekil 3.5. X ve Y değerlerine ait gözlemler ve regresyon doğrusu

Oluşturulan regresyon denklemi kullanılarak yapılan tahminler ($\hat{Y}$ değerleri) ile gözlenen Y-değerleri arasındaki fark, rasgele hata olarak bilinir ve "e" ile gösterilir. Aynı zamanda rasgele hata regresyondan sapma olarak da ifade edilir. Regresyondan sapmaların, ortalaması sıfır, varyansı $\sigma^2$ olan normal dağılım gösterdiği varsayılır.
Regresyon denklemindeki a ve b katsayıları, bu sapma kareler toplamını (yanı hata kareler toplamını) ($\sum e^2 = \sum (Y - \hat{Y})^2$) minimum yapacak şekilde hesaplanırlar. Böylece koordinat düzlemine işaretlenen gözlem değerlerinin hepsine birden en yakın geçen doğru denklemi bulunmuş olur. Yani, regresyondan sapma kareler toplamında $\hat{Y}$ yerine $(a + b_{yx}X)$ yazılarak $\sum(Y - (a + b_{yx}X))^2$ eşitliği elde edilip, bunun a'ya ve $b_{yx}$'e göre kısmi türevleri alınıp sıfıra eşitlenerek elde edilen; $\sum Y_i = na + \sum b_{yx} X_i$ ve $\sum X_i Y_i = a \sum X_i + b_{yx} \sum X_i^2$ denklem sisteminden yararlanılır. Bu denklem sisteminde $X_i$ ve $Y_i$ doğrudan gözlem değerlerinden ölçüldüklerinden bilinmeyenler sadece "a" ve "$b_{yx}$" dir. Elimizde iki denklem iki de bilinmeyen olduğundan, bu denklem sistemini çözersek (birinden a'yı çekip diğerinde yerine yazarak) "a" ve "$b_{yx}$" için (3.5) ve (3.7) numaralı eşitlikler elde edilir.

$$b_{yx} = \frac{\sum d_x d_y}{\sum d_x^2} \qquad \ldots(3.5)$$

$$a = \overline{Y} - b_{yx}\overline{X} \qquad \ldots(3.7)$$

**ÖRNEK 1:**

Göl derinlikleri ve söz konusu derinliklerde ölçülen oksijen miktarları Tablo 3.2'de verilmiş ve bu özellikler için korelasyon ve regresyon katsayıları hesaplanmıştı. Bu katsayılar hesaplandıktan sonra göl derinliklerinden oksijen miktarlarının tahmin edilmesi istenebilir. Bu durumda regresyon denkleminin oluşturulması gerekir.

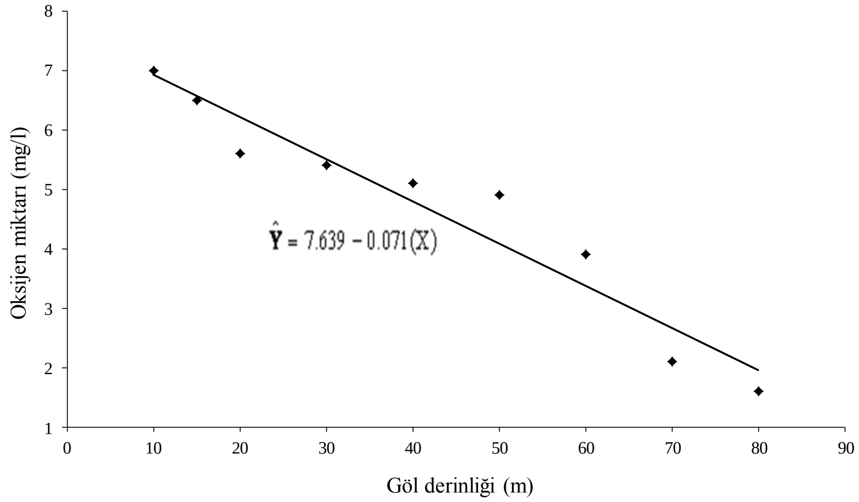
Regresyon katsayısı, $b_{yx}$=-0.071 mg/L olarak bulunmuştu. Regresyon denkleminin oluşturulması için gerekli olan "a" katsayısı ise Tablo 3.2'de verilen özelliklere ait toplamlardan, X ve Y değerlerine ait ortalamalar hesaplandıktan sonra (3.7) numaralı eşitlik kullanılarak hesaplanır.

$$\overline{X} = \frac{375}{9} \cong 41.67 \text{ ve } \overline{Y} = \frac{42.1}{9} \cong 4.68$$

a=4.68-(-0.071)(41.67)=7.639 olarak bulunur. Bu katsayılar yerine konularak, regresyon denklemi;

$$\hat{Y} = 7.639 - 0.071X \text{ yani,}$$
$$\text{Oksijen miktarı} = 7.639 - 0.071(\text{Derinlik})$$

şeklinde elde edilir. Göl derinlikleri ve oksijen miktarlarına ait gözlemler, derinliklerden oksijen miktarını tahmin etmek için oluşturulan regresyon denklemi ve bu denkleme göre çizilen regresyon doğrusu Şekil 3.6'da gösterilmiştir.




Şekil 3.6. Göl derinlikleri ve oksijen miktarlarına ait gözlemler ve regresyon doğrusu

Regresyon denklemi oluşturulduktan sonra istenen derinliklerdeki oksijen miktarı tahmin edilebilir. Örneğin; X=0 m için oksijen miktarı, $\hat{Y} = 7.639 - 0.071(X) = 7.639$mg/L, yani X=0 olduğu zaman $\hat{Y}$, a'ya, yani regresyon doğrusunun Y-eksenini kestiği noktanın ordinatına eşittir. X=35 m için oksijen miktarı, $\hat{Y} = 7.639 - 0.071(35) = 5.154$ mg/L olarak ve X=55 m için oksijen miktarı, $\hat{Y} = 7.639 - 0.071(55) = 3.734$ mg/L olarak tahmin edilir. X=85m için oksijen miktarı tahmin edilse bile yapılan tahmin yanıltıcıdır. Çünkü oluşturulan regresyon denklemi, bağımsız değişkenin (derinliğin) değişim sınırları içinde, 10-80 m aralığında, geçerlidir. Bu derinliklerin dışında göl derinliği ile oksijen miktarı arasında nasıl bir ilişki olduğu bilinmediğinden, bağımsız değişkenin değişim aralığı dışında tahmin yapılamaz.

Araştırmanın yapıldığı gölden bilinmeyen bir derinlikten alınan örnekte oksijen miktarı Y=4.7 mg/L olarak ölçülmüş olabilir. İstenirse ölçülmüş olan oksijen miktarından göl derinliği de tahmin edilebilir.

Tahmin için regresyon denklemi, derinlik bağımlı değişken, oksijen miktarı bağımsız değişken olmak üzere $\hat{X} = a + b_{xy}Y$ şeklinde oluşturulur. Bu durumda, $b_{xy}$, (3.6) numaralı eşitlik ve "a" değeri de (3.8) numaralı eşitlik kullanılarak hesaplanır.

$$b_{yx} = \frac{\sum d_x d_y}{\sum d_x^2} \qquad (3.6)$$

$$a = \overline{X} - b_{xy}\overline{Y} \qquad (3.8)$$

$b_{yx} = \frac{-354.67}{27.04} = -13.12\,\mathrm{m}$ ve $a = 41.67 - (-13.12)(4.68) = 103.07$ olarak bulunduktan sonra regresyon denklemi;

$$\hat{X} = 103.07 - 13.12Y$$

Derinlik = 103.07 – 13.12 ( Oksijen miktarı ) şeklinde oluşturulur.

Oluşturulan regresyon denkleminde Y=4.7mg/L yerine yazılarak, göl derinliği, $\hat{X} = 103.07 - 13.12(4.7) \cong 41.41\mathrm{m}$ olarak tahmin edilir.

### 3.5. Tahmindeki İsabet (Doğruluk) Derecesi

Üzerinde çalışılan iki özelliğe ait verilerden yararlanarak regresyon denklemi oluşturulduğu zaman bağımlı değişkene ait değerler, bağımsız değişkenin değerlerinden yararlanılarak tahmin edilebilir. Oluşturulan regresyon denklemi ile yapılan tahminlerin doğruluk (isabet) derecesi, $r^2$ ile gösterilir. $r^2$, korelasyon katsayısının karesine eşittir. İki özellik arasındaki ilişkinin derecesi arttıkça, yani korelasyon katsayısı büyüdükçe doğruluk derecesi de büyür ve yapılan tahminlerin güvenilirliği artar. Doğruluk derecesi, bağımlı değişkendeki varyasyonun % ne kadarının çalışılan bağımsız değişken ile açıklandığını gösterir.

Örneğin, bir göldeki derinlikler ve söz konusu derinliklerde ölçülen oksijen miktarları arasında korelasyon katsayısı, r=-0.965 ve regresyon katsayısı ise $b_{yx}$=-0.071 mg/L olarak hesaplanmıştı. Derinliklerden yararlanarak oksijen miktarlarının tahmin edilmesi için regresyon denklemi ise $\hat{Y} = 7.639 - 0.071(X)$ şeklinde oluşturulmuştu. Oluşturulan bu regresyon denklemi kullanılarak derinliklerden oksijen miktarı tahmin edildiği zaman, yapılan tahminlerin doğruluk derecesi, $r^2=(-0.965)^2 \cong 0.931$ olarak bulunur. Hesaplanan doğruluk derecesi, oksijendeki değişimin %93.1'inin derinlikteki değişim ile açıklanabildiğini gösterir.

### 3.6. KORELASYON KATSAYISI ile REGRESYON KATSAYISI ARASINDAKİ İLİŞKİLER

1. X ve Y gibi iki özellik arasındaki korelasyon katsayısı, Y'nin X'e regresyon katsayısı ile X'in Y'ye regresyon katsayısının GEOMETRİK ORTALAMASINA eşittir. Yani;

$$r_{yx} = r_{xy} = r = \sqrt{b_{yx}.b_{xy}} \text{ dir.}$$

2. $r_{yx} = r_{xy} = r = b_{yx}\frac{S_x}{S_y} = b_{xy}\frac{S_y}{S_x}$ dir.

**İpucu:** Yukarıdaki ifadelerde $b_{yx}$, $b_{xy}$, $S_x$ ve $S_y$ değerleri yerine bunların eşit olduğu ifadeler konulup, gerekli sadeleştirilmeler yapılarak bu sonuçlar kolaylıkla elde edilebilir.

### 3.7. Sorular

1. Korelasyon kavramını açıklayınız.
2. Regresyon kavramını açıklayınız.
3. Korelasyon ve regresyon arasındaki farklılıkları belirtiniz.
4. Korelasyon ve regresyon katsayıları arasında ne gibi ilişkiler vardır? Açıklayınız.
5. Regresyon denklemi ne amaçla oluşturulur? Açıklayınız.
6. Doğruluk derecesi (isabet derecesi) ne demektir? Açıklayınız.
7. Belirli bir yörede yetişen Titrek kavaklardan tesadüfen seçilen 10 tanesinin yaş ve boylarına ilişkin aşağıdaki ölçümler yapılmıştır.

| | |
|---|---|
| Yaş (yıl) | 23.0, 20.0, 19.0, 23.0, 18.0, 16.0, 19.0, 21.0, 24.0, 23.0 |
| Boy (m) | 19.5, 18.0, 17.0, 19.0, 17.5, 16.0, 17.0, 16.5, 16.4, 16.2 |

a. Gözlemleri koordinat ekseninde işaretleyiniz.
b. Yaş ve boy arasındaki korelasyon katsayısını hesaplayıp anlamını açıklayınız.
c. Yaş'ın boy'a olan regresyon katsayısını hesaplayıp anlamını açıklayınız.
d. Boylara bakarak yaşları tahmindeki isabet (doğruluk) derecesini bulunuz.

8. Belirli bir firmanın ürettiği vitamin mineral karması drajelerde depolama süresi ile (X) vitamin miktarı (Y) arasındaki ilişkiyi araştırmak üzere 2., 3., 4., 5., 6., 7., 8., ve 10. aylarda drajelerdeki vitamin miktarı ölçülmüş ve aşağıdaki değerler hesaplanmıştır.
   X (ay) değerlerinin toplamı: 45
   X (ay) değerlerinin karelerinin toplamı: 303
   Y (mg) değerlerinin toplamı: 262
   Y (mg) değerlerinin karelerinin toplamı:8770
   X ve Y değerlerinin çarpımlarının toplamı:1380

   Regresyon denklemini oluşturarak 9. ayda ölçülen vitamin miktarını tahmin ediniz ve yapılabilecek tahminlerdeki isabet derecesini bulunuz.
9. 20 bireylik bir örnekte $b_{yx}$ = -0.9, $b_{xy}$ = -0.4 olarak hesaplansa, X ve Y özellikleri arasındaki korelasyon katsayısı kaçtır?
10. X ve Y gözlem çiftlerine ilişkin, $S_x$=2.5, $S_y$=1.9 ve r = 0.63 olarak hesaplanmıştır. Y'nin X'e olan regresyon katsayısını hesaplayınız.
11. Belirli bir ırka ait 6 adet fareye verilen ilaç dozları (X) ile dakikadaki kalp atışı sayıları (Y) arasındaki ilişki araştırılmış ve bu araştırma sonucunda aşağıdaki değerler hesaplanmıştır.
    X (mg) değerlerinin toplamı : 18.2
    X değerlerinin karelerinin toplamı: 57.74
    Y değerlerinin toplamı: 511
    Y değerlerinin karelerinin toplamı: 43657
    X ve Y değerlerinin çarpımlarının toplamı: 1566.5

a. İlaç dozlarından yararlanarak kalp atış sayılarını tahmin etmek için gerekli regresyon denklemini oluşturunuz.
b. Oluşturduğunuz regresyon denklemini kullanarak tahmin yaptığınız zaman doğruluk dereceniz nedir? Hesaplayarak açıklayınız.

12. Altı adet kronik böbrek yetmezliği hastasına ait serum fosfat ve serum protein değerleri aşağıdaki tabloda verilmiştir.

| | | | | | | |
|---|---|---|---|---|---|---|
| Protein(g/dl) | 1.00 | 1.05 | 1.40 | 1.65 | 1.70 | 2.80 |
| Fosfat(mg/dl) | 2.02 | 3.80 | 4.44 | 6.52 | 7.10 | 11.85 |

a. Hastaların protein ve fosfat düzeyleri arasındaki doğrusal ilişkinin derecesini hesaplayarak anlamını açıklayınız.

b. Protein değeri 1.43, 2.3, 3.7 olan hastalara ait tahmini fosfat değerlerini ve bu tahminlerdeki isabet derecesini hesaplayınız.

13. X özelliğinden yararlanarak Y özelliğine ilişkin değerleri tahmin etmedeki isabet derecesi (doğruluk derecesi) 0.81 olarak hesaplandığına göre, iki özellik arasındaki korelasyon katsayısı kaçtır? Hesaplayarak anlamını açıklayınız.

14. Uzun yıllar itibariyle domates fiyatı (x 1000 TL) ile pazarlandığı aylar arasındaki ilişkinin $\text{DomatesFiyatı} = 15.9 - 0.983\text{Ay}$ şeklinde olduğu bilindiğine göre, oluşturulan bu regresyon denklemi ne ifade eder? Açıklayınız.

15. $Y = a + b_{YX}X$ regresyon denklemindeki $b_{YX}$ (regresyon katsayısı) ne ifade eder? Açıklayınız.

16. İstatistik sınav notu (Y) ile istatistik sınavına çalışmak için ayrılan zamana (X) ilişkin regresyon denklemi Y = 55 + 7.5 X olarak bulunmuştur. İstatistik sınavına çalışmak için 4 saat zaman ayıran bir öğrencinin sınav notunu tahmin ediniz.

17. 7 adet öğrencinin belirli bir dersin vize ve final sınavından aldıkları notlar aşağıdaki gibidir.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Vize notu | 35 | 45 | 50 | 70 | 65 | 60 | 55 |
| Final notu | 55 | 65 | 70 | 80 | 75 | 70 | 65 |

a. Vize ve final notları arasındaki korelasyon katsayısını hesaplayarak anlamını açıklayınız.

b. Final notlarının vize notlarına göre regresyon katsayısını hesaplayarak anlamını açıklayınız.

c. Bu örnekte vize notlarından yararlanarak final notlarının tahmin edilmesi için bir regresyon denklemi oluşturulsa bu denklem kullanılarak yapılacak tahminlerin doğruluk (isabet) derecesi ne kadardır? Hesaplayarak açıklayınız.

18. $\sum d_x d_y = \sum (X_i - \overline{X})(Y_i - \overline{Y}) = \sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n} =$ olduğunu ispat ediniz.