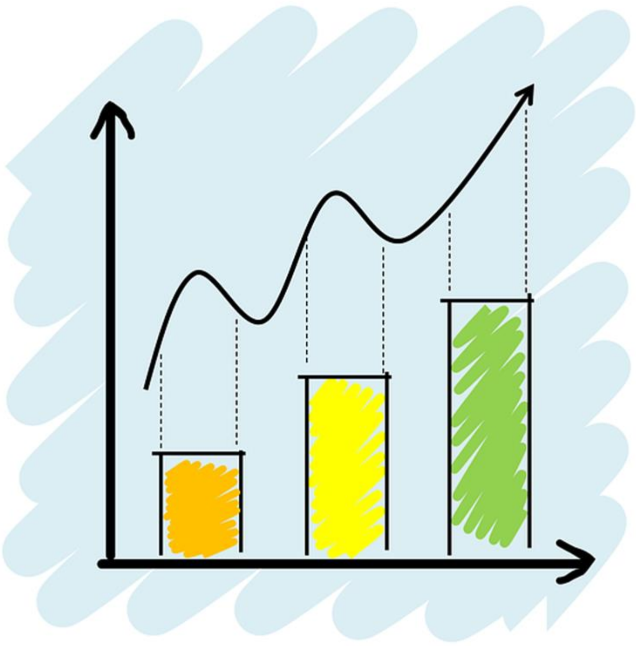
İstatistik
14. Hafta

23.12.2024



Mühendislik Fakültesi
Bilgisayar Mühendisliği

Hazırlayan: Dr. Ercan Ezin

REGRESYON VE KORELASYON ANALİZİ

KORELASYON VE REGRESYON ANALİZİNİN TANIMLARI

- **Korelasyon Analizi**: İki değişken arasındaki ilişkiyi tanımlama ve ölçmeye yönelik analizdir.
- **Regresyon Analizi**: Bir veya daha fazla bağımsız değişkenin bağımlı değişken üzerindeki etkisini açıklamaya yönelik analizdir.
- **Farkları**: Her ne kadar tanımları benzer görünse de, korelasyon analizi ve regresyon analizi bir madalyonun iki farklı yüzü gibidir. Korelasyon sadece ilişkinin derecesini gösterirken, regresyon ilişkiyi modellemeye çalışır.
- **Tarihçesi**: Regresyon terimi, 19. yüzyılda İngiliz istatistikçi **Francis Galton** tarafından ortaya atılmıştır. Galton, babaların boyları ile oğullarının boyları arasındaki ilişkiyi incelemiş ve ortalamaya doğru bir eğilim olduğunu fark etmiştir. Bu gözlemine "regresyon" adını vererek, terimin temelini atmıştır.

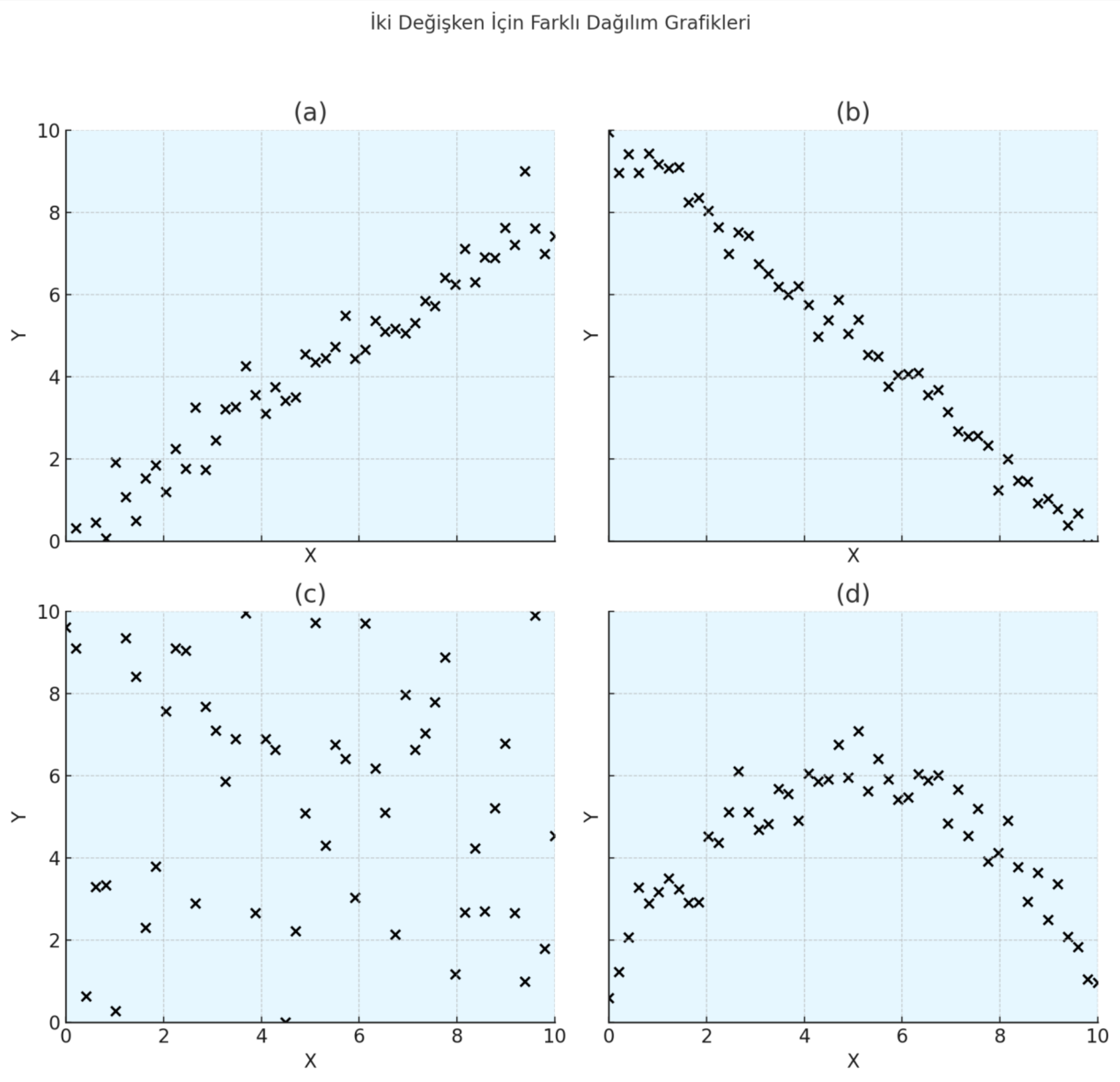
REGRESYON VE KORELASYON ANALİZİNİN ÖZELLİKLERİ VE UYGULAMALARI

- **Regresyon Analizinde Değişkenler**:
    - **Bağımlı Değişken**: Araştırmanın odaklandığı ve açıklanmaya çalışılan değişkendir.
    - **Bağımsız Değişken(ler)**: Bağımlı değişken üzerindeki etkisi incelenen değişken veya değişkenlerdir.
- **Örnek**: Bağımlı değişken trafik akım hızı ise, bağımsız değişkenler şunlar olabilir:
    - Trafik hacminin kapasiteye oranı
    - Trafik hacmi içindeki bisiklet sayısı
    - Yoldaki ticari yoğunluk oranı
    - Önemli kavşak sayısı
- **Basit Doğrusal Regresyon**: Tek bağımsız değişken içerir.
- **Çoklu Doğrusal Regresyon**: Birden fazla bağımsız değişken içerir (bu ünite kapsamında ele alınmayacaktır).

KORELASYON ANALİZİNİN ÖZELLİKLERİ:

- İki değişken arasındaki ilişkinin derecesini belirler.
- Korelasyon, nedenselliği göstermez; yüksek korelasyon, bir değişkenin diğerinin nedeni olduğu anlamına gelmez.
- **Nedensellik Araştırması**: Korelasyon analizinden farklı istatistiksel teknikler gerektirir.

KOREASYON ANALİZİ



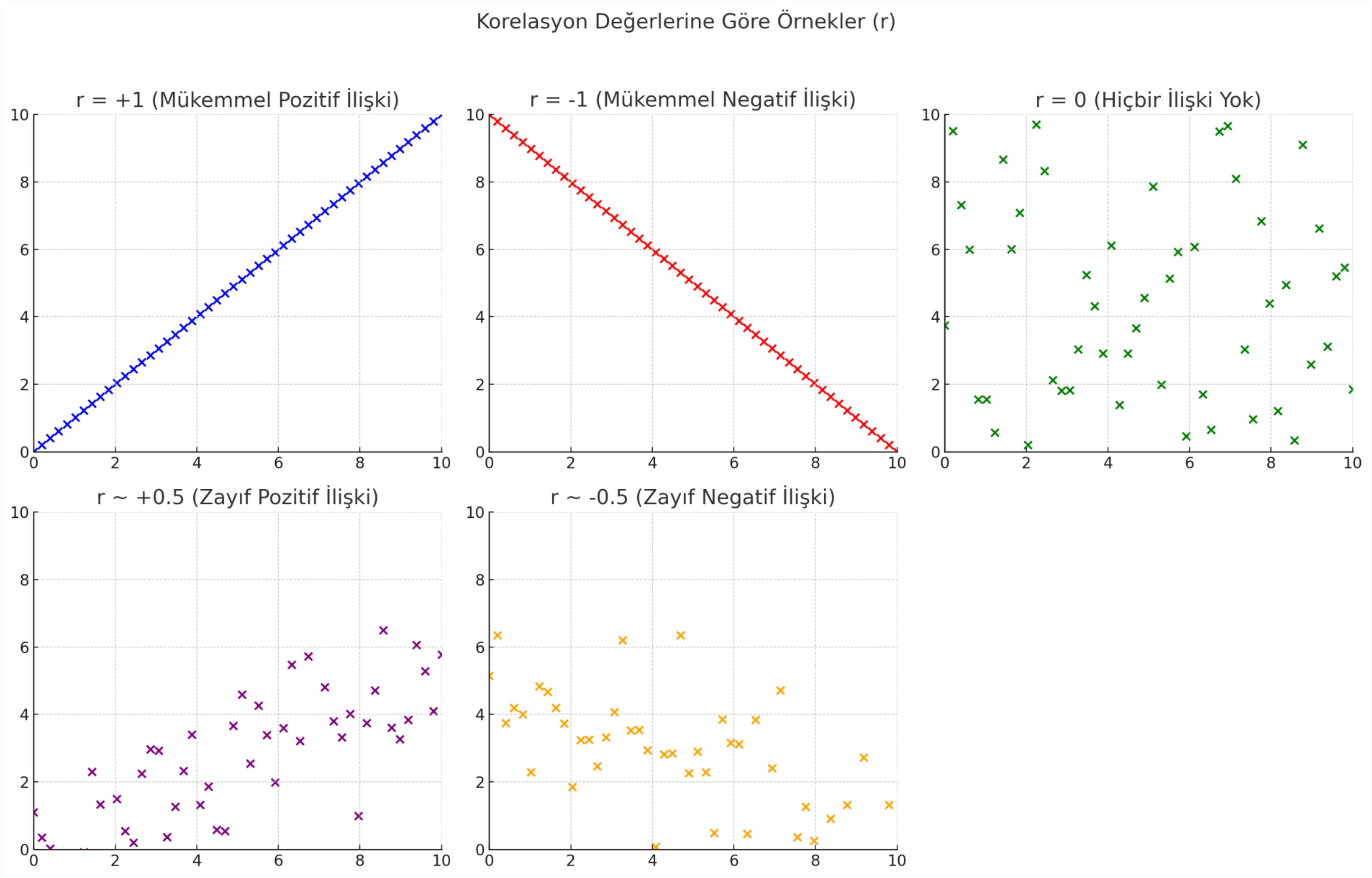
X ve Y değişkenleri arasındaki ilişkiyi yorumlayınız

A) X ve Y değişkeni pozitif ilişkili.
B) Negatif ilişkili
C) İlişki yok
D) Eğrisel ilişki var.

PEARSON KORELASYON KATSAYISI

- **Pearson Korelasyon Katsayısı**, iki ya da daha fazla, oranlı ve eşit aralıklı ölçeğe uygun şekilde ölçülmüş değişkenler arasındaki ilişkinin derecesini belirlemek için kullanılır.
- Evrendeki **x** ve **y** değişkenlerinin korelasyon değeri, **ρ** ile sembolize edilir ve bu sembol "ro" şeklinde okunur.
- Örneklemdeki **x** ve **y** değişkenlerinin korelasyon değeri, **r** simgesi ile gösterilir.
- Pearson korelasyon katsayısının değerleri **-1** ile **+1** arasında değişir:
    - **r = +1**: Mükemmel/tam pozitif ilişki
    - **r = -1**: Mükemmel/tam negatif ilişki
    - **r = 0**: Hiçbir ilişki yok (bağımsız değişkenler)
- **Yorumlama:**
    - **Negatif r Değerleri**: Değişkenler arasında ters yönlü ilişki vardır (biri artarken diğeri azalır).
    - **Pozitif r Değerleri**: Değişkenler arasında aynı yönlü ilişki vardır (biri artarken diğeri de artar).
- **Korelasyon Ölçeği:**
    - -1 -0.5 0 +0.5 +1
    - <-------- Ters Yönlü --------|Bağımsız|-------- Aynı Yönlü -------->
    - **r 'ye Yakın Değerler**: Güçlü bir ilişkiyi temsil eder.
    - **r = 0'a Yakın Değerler**: Zayıf veya hiçbir ilişkiyi temsil eder.

PEARSON KORELASYON KATSAYISI



PEARSON KORELASYON KATSAYISI HESABI

x ve y: Korelasyonunu hesaplamak istediğimiz iki değişken olmak üzere pearson korelasyon değeri r şu formül ile hesaplanır:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}$$

x_i **ve** y_i: x ve y değişkenlerinin bireysel veri noktaları.

$\bar{x}$ **ve** $\bar{y}$: x ve y değişkenlerinin aritmetik ortalamaları.

s_x **ve** s_y: x ve y değişkenlerinin standart sapmaları.

n: Veri noktalarının toplam sayısı.

$\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$: x ve y arasındaki **kovaryansı** hesaplar. Kovaryans, x ve y değişkenlerinin ortalamaları etrafında birlikte nasıl değiştiğini ölçer.

$(n-1)s_x s_y$: Kovaryansı normalize etmek için x ve y değişkenlerinin standart sapmalarının çarpımına bölünür. Bu işlem, korelasyon katsayısının boyutsuz olmasını ve -1 ile +1 arasında bir değer almasını sağlar.

PEARSON KORELASYON KATSAYISI HESABI-2

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}$$

r değeri hesaplarken elimizde tüm örneklem verileri varsa r değerini aşağıdaki formül yardımıyla da bulabiliriz.

$$r = \frac{n\sum_{i=1}^{n} x_i y_i - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{\sqrt{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}\sqrt{n\sum_{i=1}^{n} y_i^2 - (\sum_{i=1}^{n} y_i)^2}}$$

PEARSON KORELASYON ÖRNEK

Çalışma Saatleri Ve Test Puanları Verisi

	Aylar	Çalışılan Saat (Saat/Hafta)	Test Puanı
1	1	10	50
2	2	12	55
3	3	15	65
4	4	9	45
5	5	18	72
6	6	20	78
7	7	5	30
8	8	6	40
9	9	7	42
10	10	10	55
11	11	14	70
12	12	25	95

Yanda verilen aylık çalışma Saati ve alınan test puanı tablosuna göre bu iki değişken arasındaki pearson korelasyon katsayısını hesaplayınız?

ÇÖZÜM

Çalışılan Saatler Ve Test Puanları Hesaplama Tablosu

	Aylar	Çalışılan Saat (Saat/H	Test Puanı	Çalışılan Saat (Saat/Hafta)^2	Test Puanı^2	Çalışılan Saat x Test
1	1	10	50	100	2500	500
2	2	12	55	144	3025	660
3	3	15	65	225	4225	975
4	4	9	45	81	2025	405
5	5	18	72	324	5184	1296
6	6	20	78	400	6084	1560
7	7	5	30	25	900	150
8	8	6	40	36	1600	240
9	9	7	42	49	1764	294
10	10	10	55	100	3025	550
11	11	14	70	196	4900	980
12	12	25	95	625	9025	2375
13	Toplam	151	697	2305	44257	9985

Aritmetik ortalama ve Standart sapma değerlerini hesaplamadan da r değerini hesaplayabiliriz. 2. Formülde ilgili yerlere tabloda bulunan değerleri yazarsak.

$$r = \frac{12(9985) - (151)(697)}{\sqrt{[12(2305) - (151)^2][12(44257) - (697)^2]}}$$

$$r = \frac{46440}{\sqrt{419 \cdot 159875}} = 0.982$$

Bu değeri yorumlayınız!

BELİRLİLİK KATSAYISI

Belirlilik katsayısı (r^2), Pearson korelasyon katsayısı r'nin karesi alınarak hesaplanan bir ölçüdür ve bağımlı değişkendeki (örneğin test puanı) değişimlerin bağımsız değişken (örneğin çalışma saatleri) tarafından ne kadar açıklandığını gösterir.

1. **Özellikleri**:
   - Belirlilik katsayısı r^2, 0 ile 1 arasında değer alır.
   - $r^2 = 1$: Bağımlı değişkendeki değişimlerin tamamı bağımsız değişken tarafından açıklanır.
   - $r^2 = 0$: Bağımlı değişkendeki değişimlerin hiçbir kısmı bağımsız değişken tarafından açıklanmaz.
2. **Hesaplama**:
   - r^2, r'nin karesidir:

$$r^2 = r^2$$

   - Bu oran, bağımlı değişkende gözlemlenen değişimlerin bağımsız değişkene atfedilebilecek kısmını ifade eder.

ÇALIŞMA SAATLERİ TABLOSUNUN BELİRLİLİK KATSAYISI

Çalışma saatleri tablosu için belirlilik tablosunu hesaplayıp yorumlayınız.

Korelasyon katsayısı $r = 0.98$ olarak hesaplanmıştı. Belirlilik katsayısı aşağıdaki şekilde hesaplanır:

$$r^2 = (0.98)^2$$

Şimdi bu değeri hesaplayalım.

Tablo verilerinizden hesaplanan belirlilik katsayısı (r^2) **0.97** olarak bulunmuştur. Bu, bağımlı değişkendeki (örneğin test puanı) değişimlerin %97'sinin bağımsız değişken (örneğin çalışma saatleri) tarafından açıklandığını göstermektedir.

BASİT DOĞRUSAL REGRESYON ANALİZİ

- Regresyon, bir bağımsız değişken (x) ile bir bağımlı değişken (y) arasındaki ilişkiyi modelleyen bir tekniktir. Değişkenler arasındaki bu ilişki bir regresyon denklemi ile ifade edilir.
- Önemli Noktalar:
  - Regresyon denklemi, değişkenler arasındaki ilişkiyi matematiksel olarak açıklar.
  - Bağımsız değişken değiştirildiğinde bağımlı değişken değişir.
  - Basit doğrusal regresyonda, sadece bir bağımlı ve bir bağımsız değişken kullanır.

BASİT DOĞRUSAL REGRESYON ANALİZİ-2

- Doğrusal regresyon denklemi:
  - : Bağımlı değişken-y
  - : Bağımsız değişken -x
  - : Sabit terim-α
  - : Eğim-β

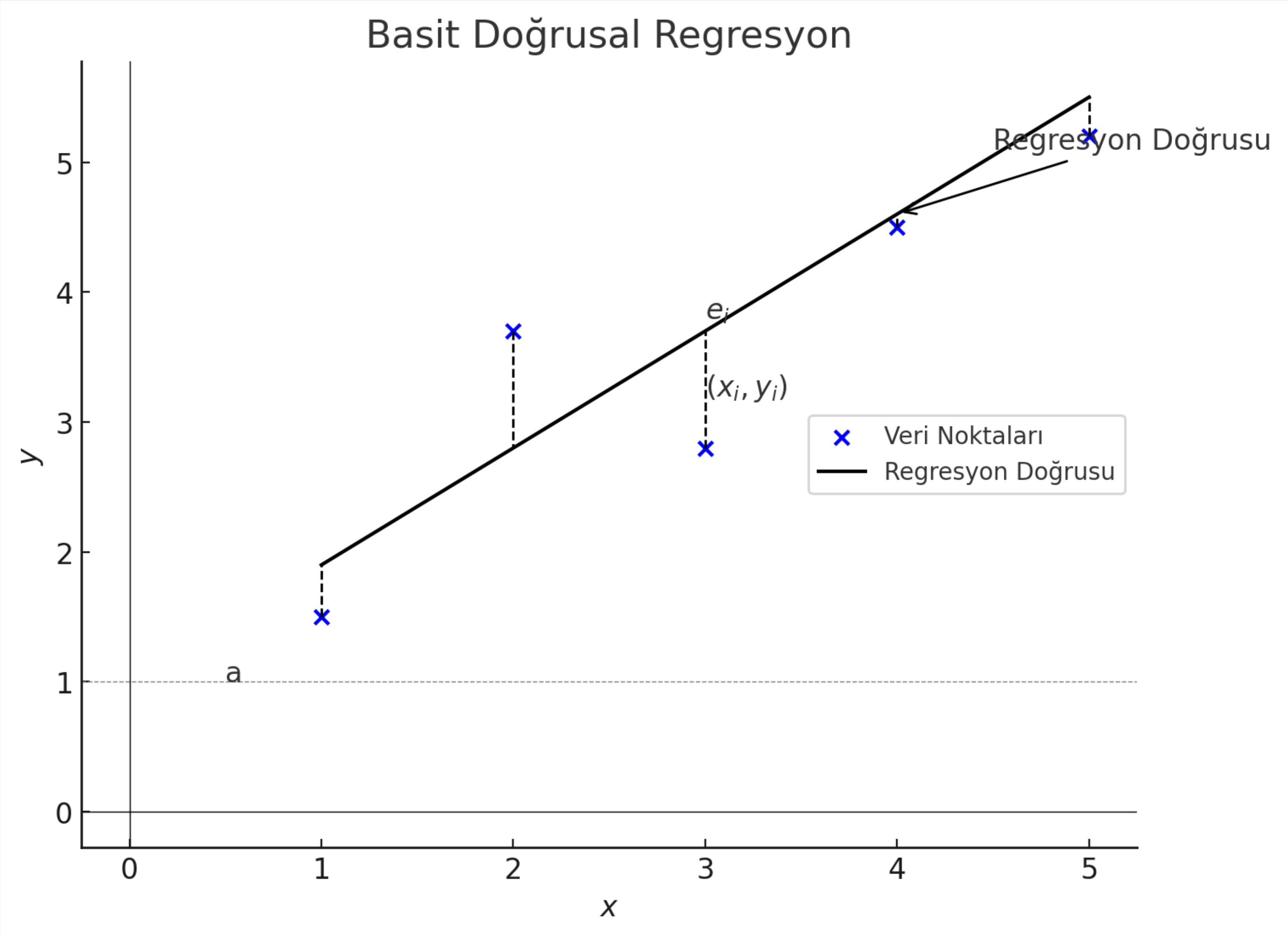
$$y_i = \alpha + \beta x_i + \varepsilon_i$$

- Tahminlerde en küçük kareler tekniği kullanılır. Bu teknik, tahmin edilen ve gerçek değerler arasındaki farkların karelerinin toplamını en aza indirmeyi amaçlar.

$$y_i = a + bx_i + e_i$$

Herhangi bir (x,y) ikilisinin değerini hesaplarken kullanacağımız formül ve değişkenler yukarıdaki gibidir. **Önemli not: Bu dersin içeriğinde yukarıdaki formülde yer alan Epsilon veya e değerleri hesaplamalarda kullanılmayacaktır.**

BASİT DOĞRUSAL REGRESYONUN GRAFİKSEL GÖSTERİMİ



$$b = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

En küçük kareler yöntemiyle eğim hesaplanabilir. Veya ilgili x ve y değişkenlerinin standar sapmaları biliniyorsa eğim şöyle hesaplanır:

$$b = r\frac{s_y}{s_x}$$

ve

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

BASİT DOĞRUSAL REGRESYON ÖRNEK

Eğitim Harcamaları Ve Mezun Öğrenci Sayısı

	Yıllar	Eğitim Harcamaları (Milyon ₺) (x_i)	Mezun Öğrenci Sayısı (Bin) (y_i)
1	2015	320	120
2	2016	410	150
3	2017	530	200
4	2018	470	170
5	2019	620	250
6	2020	580	230

Yukarıdaki tabloda verilen değerlere göre basit doğrusal regresyon denklemini tahminleyiniz ve grafiğini çiziniz.

ÇÖZÜM

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{320 + 410 + 530 + 470 + 620 + 580}{6} = \frac{1465}{3} = 488.33$$

Eğitim Ortalaması

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{120 + 150 + 200 + 170 + 250 + 230}{6} = \frac{560}{3} = 186.67$$

Mezun Öğrenci Sayısı Ortalaması

Bu ortalamaları en küçük kareler tekniğinde kullanalım.
Önce yandaki gibi bir tablo çıkarmak gerekir. Daha sonra formülde ilgili yerlere Toplam satırındaki sonuçlar yazılır.

Yıllar	Eğitim I	Mezun	(x_i - x̄)	(y_i - ȳ)	(x_i - x̄)(y_i - ȳ)	(x_i - x̄)²	ŷ
2015	320	120	-168.3	-66.7	11222.2	28336.1	113.0
2016	410	150	-78.3	-36.7	2872.2	6136.1	152.4
2017	530	200	41.7	13.3	555.6	1736.1	204.9
2018	470	170	-18.3	-16.7	305.6	336.1	178.6
2019	620	250	131.7	63.3	8338.9	17336.1	244.3
2020	580	230	91.7	43.3	3972.2	8402.8	226.8
Toplam	2930	1120			27266.7	62283.3	

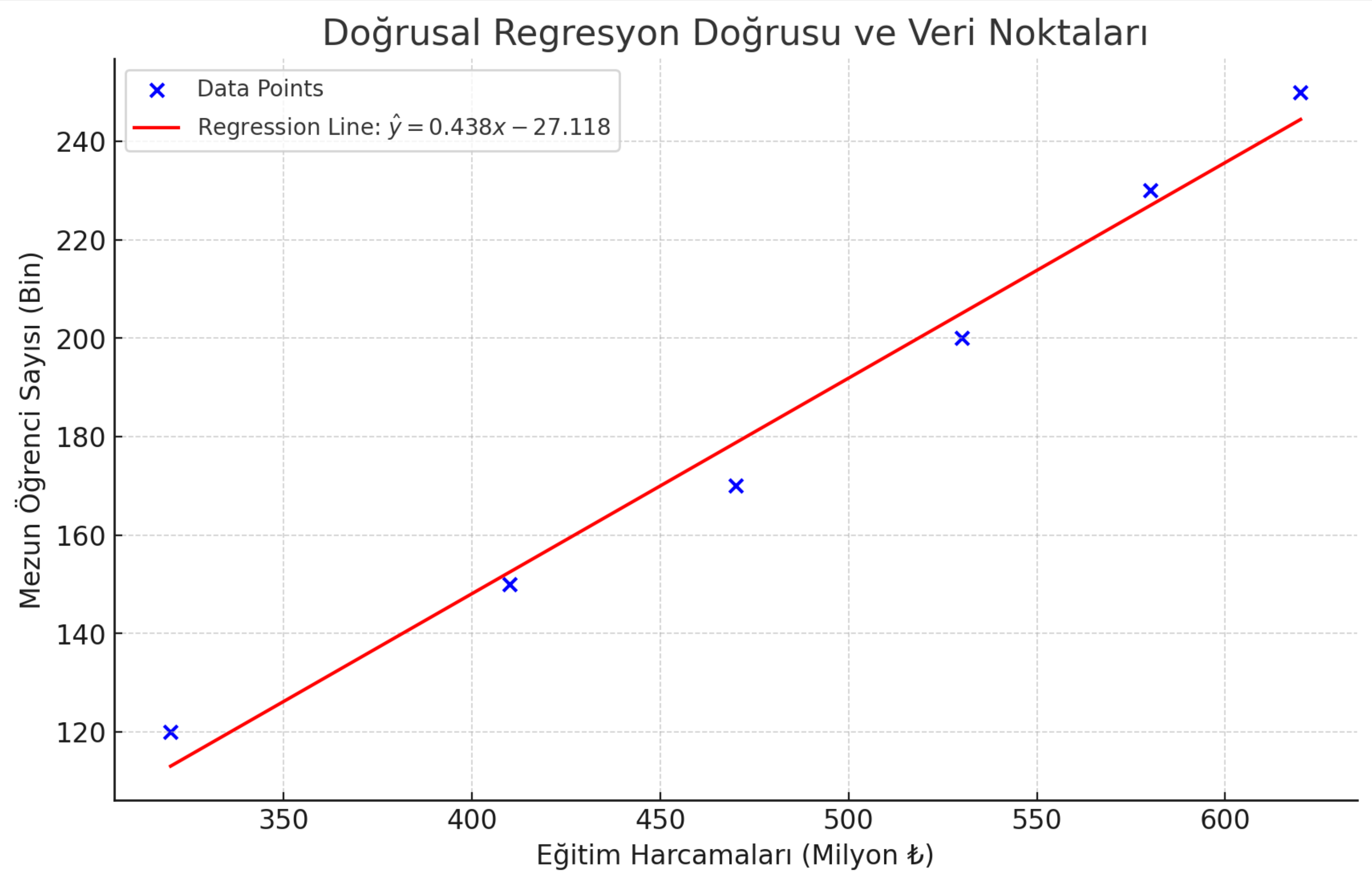
ÇÖZÜM

$$b = \frac{\sum((x_i - \bar{x})(y_i - \bar{y}))}{\sum((x_i - \bar{x})^2)} = \frac{27266.7}{62283.3} = 0.438$$

$$a = \bar{y} - b \cdot \bar{x} = 186.7 - 0.438 \cdot 488.3 = -27.118$$

$$\overline{y} = 0.438.\overline{x} - 27.118$$

ÇÖZÜM



$$\overline{y} = 0.438.\overline{x} - 27.118$$

ÖZET

- **İki değişken arasındaki ilişkiyi analiz etmek**: Araştırmacılar, iki değişken arasındaki ilişkinin yönü ve derecesini belirlemek için Pearson korelasyon katsayısını hesaplar. Bu katsayı, değişkenlerin birbiri üzerindeki etkisini anlamak için kullanılır.
- **Bağımlı değişkendeki değişimin açıklanma oranını belirlemek**: Bağımlı değişkendeki değişimin yüzde kaçının bağımsız değişken tarafından açıklandığını göstermek için belirlilik katsayısı (r^2) kullanılır. Bu katsayı 0 ile 1 arasında bir değer alır.
- **Bağımlı ve bağımsız değişkenlerle model kurmak**: Bir bağımlı değişkendeki değişimi bir bağımsız değişkenle açıklamak için basit doğrusal regresyon analizi uygulanır. Örneğin, işçilerin kıdemleri ile haftalık üretim arasındaki ilişki bu yöntemle modellenebilir.

FİNAL SINAVI

Tarih: 07.01.2025, Saat:10:30-12:00, Derslik: F-301

- 40 Soru.
- Çoktan seçmeli ve 5 seçenekli.
- Soruları iyi okuyunuz.
- Tüm konular dahil.
- Final sınav ağırlığı % 70.
- Derse katılım ekstra %10.
- Hesap makinesi getirebilirsiniz.
- Optik alana kodlama dışında hiçbir şey yazmayın, karalama için kullanmayın.
- Optik alana kodladığınız cevapların yeterince koyu olduğuna dikkat edin. Ayrıca değiştirdiğiniz kodlamaların iyice silindiğinden emin olun.
- Özellikle öğrenci numarası 0 ve 1 ile bitenler, son hanenin doğru kodlandığından emin olun.
- Sınav ile ilgili sorunuz olursa dr.ercan.ezin@gmail.com adresinden iletişime geçebilirsiniz.

BAŞARILAR

EOF

REFERANSLAR

- Bu dersin içeriği oluşturulurken **T.C. Anadolu Üniversitesi Yayını No: 2590, Açıköğretim Fakültesi Yayını No: 1559 - İstatistik** kitabından faydalanılmıştır.