Korelasyon ve Regresyon

Korelasyon Analizi

İki değişken arasında ilişki olup olmadığını belirlemek için yapılan analize korelasyon analizi denir. Korelasyon; doğrusal yada doğrusal olmayan diye ikiye ayrılır.

Korelasyon

İki değişken arasında

- **bir ilişki var mıdır?**
- **ilişki doğrusal mıdır, değil midir?**
- **(varsa) ilişkinin yönü nedir?**
- **ilişkinin gücü nedir?**
- **ilişkinin büyüklüğü nedir?**

Varsayımlar

1. (X, Y) sürekli tesadüfi değişkenlerdir.

2. X ve Y'lerin dağılımı normal olmalıdır.

Serpilme Diyagramı

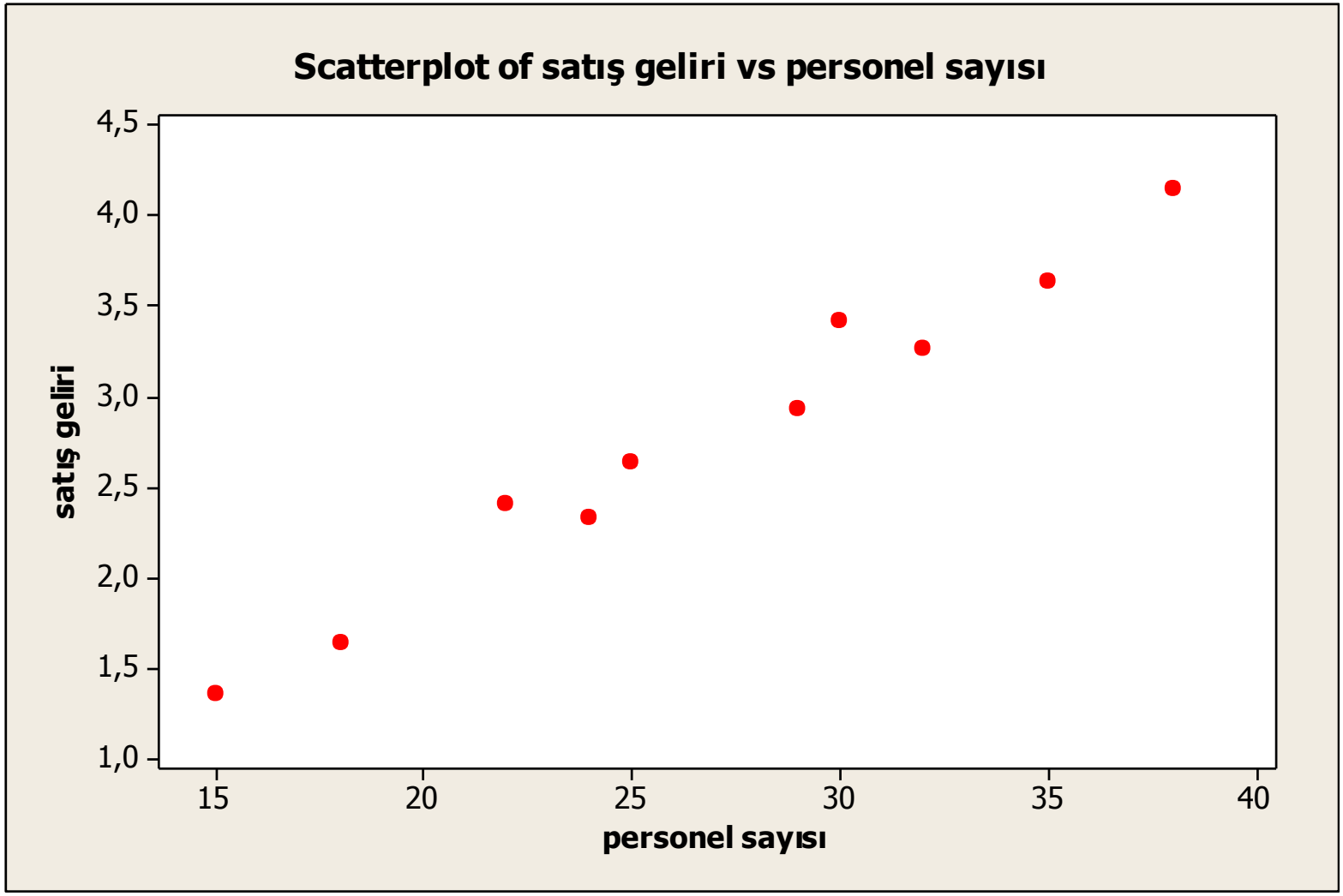
İki değişken arasındaki ilişkinin; olup olmadığını, biçimin (doğrusal mı değil mi), yönünü ve gücünü belirlemenin en kolay yolu serpilme diyagramını çizektir.

Örnek

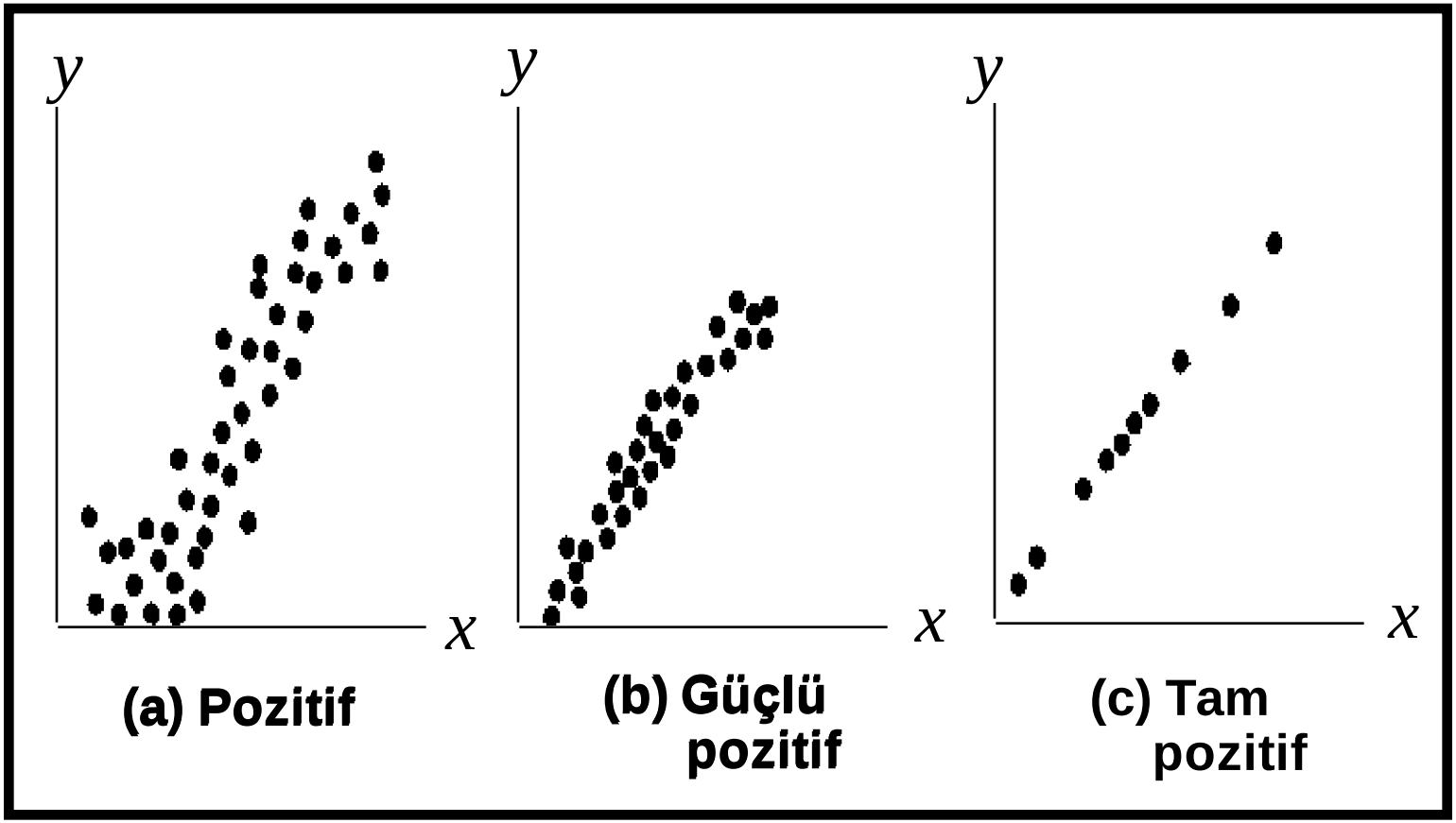
Bir firma bünyesindeki satış personeli sayısı ile satış gelirleri arasındaki ilişkiyi bilmek istemektedir.

Yıllar	*Satış Personeli Sayısı (X_i)*	*Satış Gelirleri (yüz bin \$) (Y_i)*
1999	15	1,35
2000	18	1,63
2001	24	2,33
2002	22	2,41
2003	25	2,63
2004	29	2,93
2005	30	3,41
2006	32	3,26
2007	35	3,63
2008	38	4,15

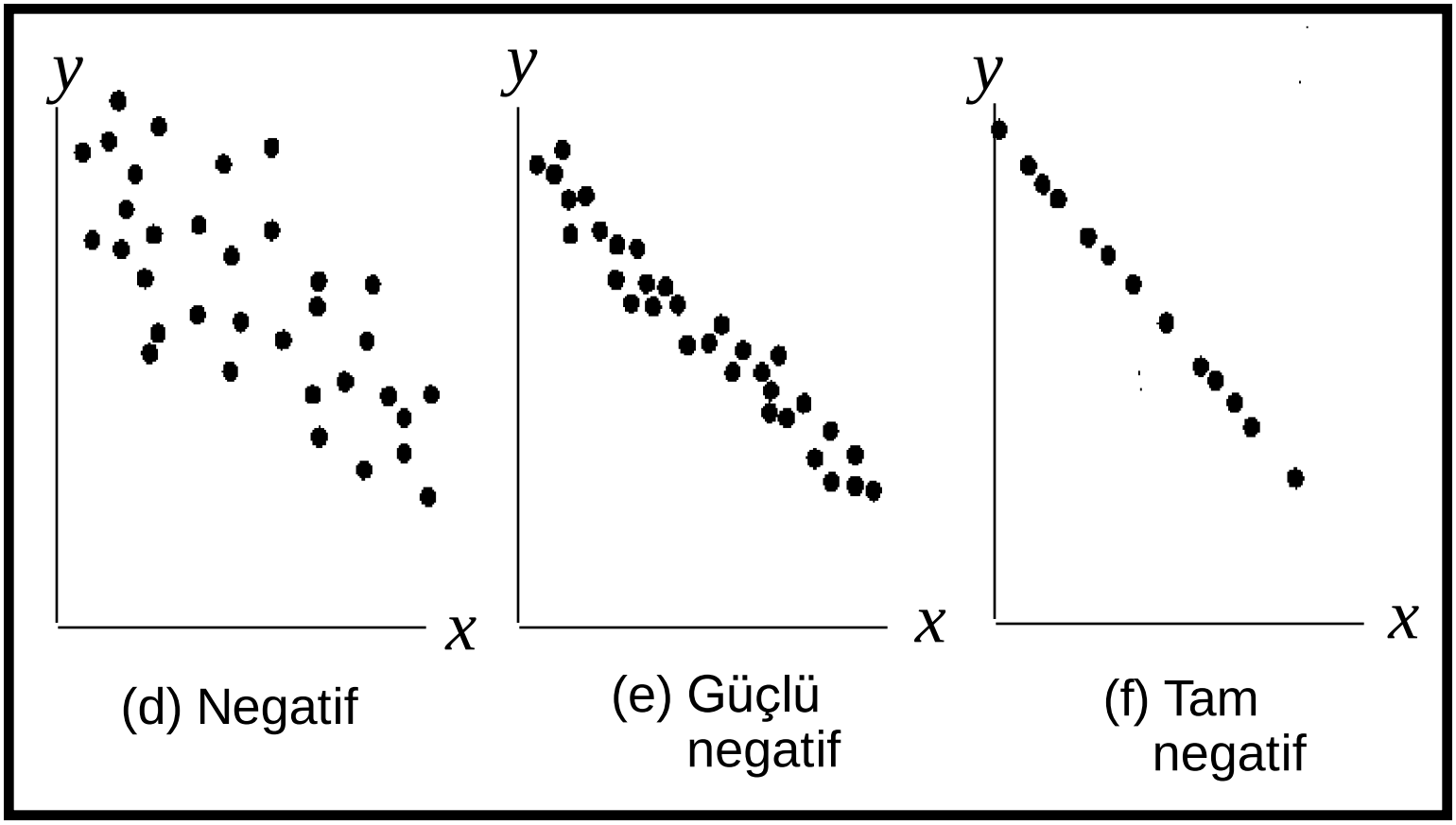
Serpilme Diyagramı



Pozitif Korelasyon



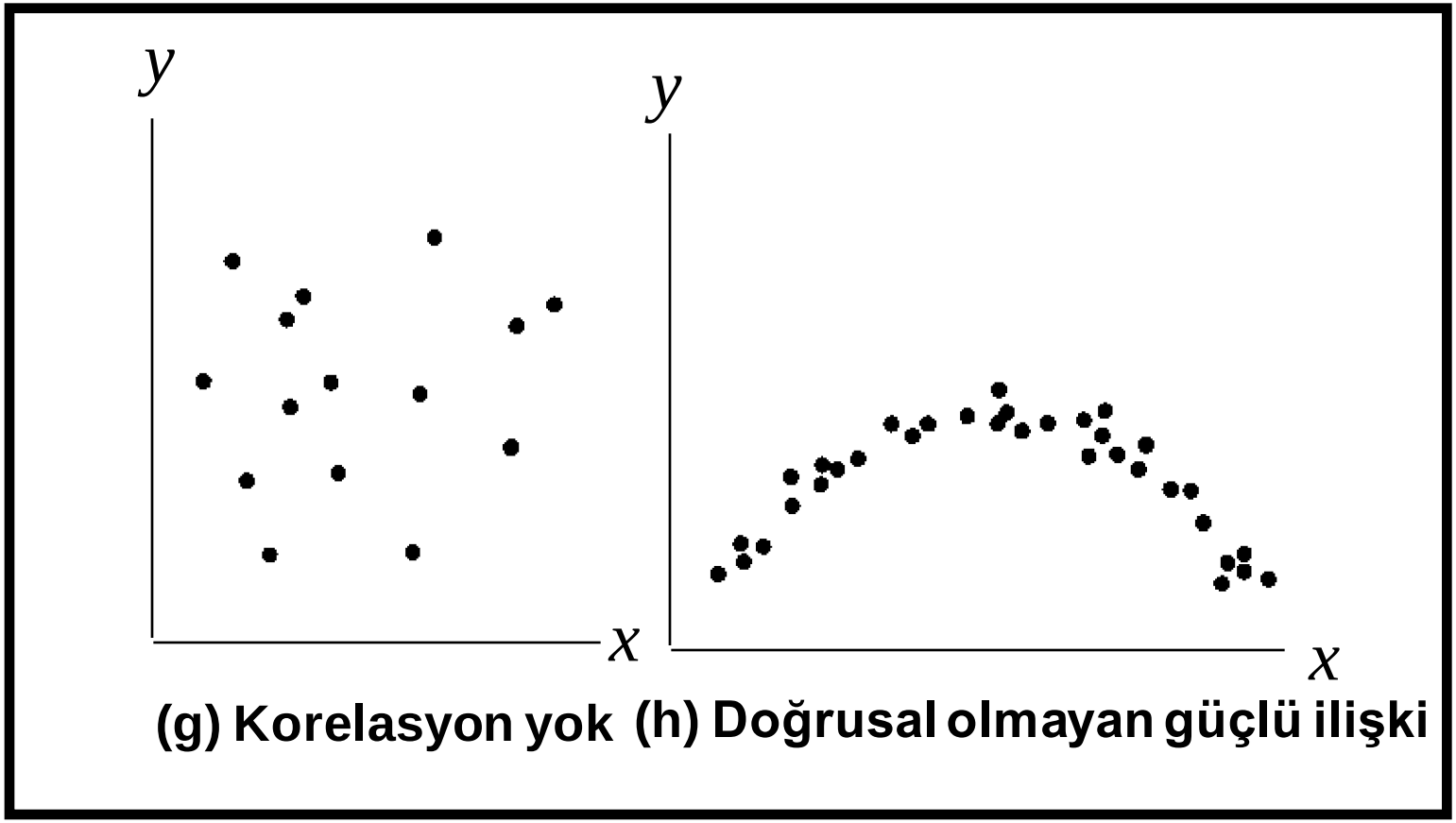
Negatif Korelasyon



(d) Negatif

(e) Güçlü negatif

(f) Tam negatif



(g) Korelasyon yok (h) Doğrusal olmayan güçlü ilişki

Doğrusal Korelasyon Katsayısı r

Bir örnekteki X_i ve Y_i gibi iki değişken arasındaki doğrusal ilişkinin büyüklüğünü ölçmektedir.

Doğrusal Korelasyon Katsayısı *r*'nin Özellikleri

$-1 \leq r \leq 1$

- ***r*** **= 1 Tam pozitif doğrusal ilişki**
- ***r*** **= -1 Tam negatif doğrusal ilişki**
- ***r*** **= 0 Doğrusal**
- **1,00-0,90 Çok kuvvetli**
- **0,70-0,89 Kuvvetli**
- **0,50-0,69 Orta**
- **0,30-0,49 Düşük**
- **0,00-0,29 Zayıf**

Korelasyon ile ilgili hatalar

1. **Nedensellik**: Korelasyon değişkenler arasındaki sebep sonuç ilişkilerini açıklamaz.

2. **Doğrusallık**: X ile Y değişkenleri arasında anlamlı bir doğrusal korelasyon olmadığı halde, aralarında doğrusal olmayan ya da farklı bir ilişki olabilir.

Örnek Veriler İçin Korelasyon Hesaplamaları

Yıllar	*Satış Personeli Sayısı (X_i)*	*Satış Gelirleri (yüz bin $) ($Y_i$)*
1999	15	1,35
2000	18	1,63
2001	24	2,33
2002	22	2,41
2003	25	2,63
2004	29	2,93
2005	30	3,41
2006	32	3,26
2007	35	3,63
2008	38	4,15
Toplamlar	**268**	**27,73**

Örnek Veriler İçin Korelasyon Hesaplamaları

Yıllar	Satış Personeli Sayısı (X_i)	Satış Gelirleri (yüz bin \$) (X_i)	$(X_i - \overline{X})$	$(Y_i - \overline{Y})$	$(X_i - \overline{X})(Y_i - \overline{Y})$	$(X_i - \overline{X})^2$	$(Y_i - \overline{Y})^2$
1999	15	1,35	-11,8	-1,42	16,76	139,24	2,02
2000	18	1,63	-8,8	-1,14	10,03	77,44	1,3
2001	24	2,33	-2,8	-0,44	1,23	7,84	0,19
2002	22	2,41	-4,8	-0,36	1,73	23,04	0,13
2003	25	2,63	-1,8	-0,14	0,25	3,24	0,02
2004	29	2,93	2,2	0,16	0,35	4,84	0,03
2005	30	3,41	3,2	0,64	2,05	10,24	0,41
2006	32	3,26	5,2	0,49	2,55	27,04	0,24
2007	35	3,63	8,2	0,86	7,05	67,24	0,74
2008	38	4,15	11,2	1,38	15,46	125,44	1,9
Toplamlar	268	27,73			57,46	485,6	6,98

Örnek Verileri İçin Korelasyon Hesaplamaları

$$r = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2 \sum (Y_i - \bar{Y})^2}} = \frac{\sum x_i y_i}{\sqrt{\sum x_i^2 \sum y_i^2}}$$

r = 0,98 Personel sayısı ile satış gelirleri arasında pozitif yönlü 0,98 büyüklüğün güçlü korelasyon vardır.

Regresyon

X_i **bağımsız değişken (açıklayıcı değişken, etkileyen)**

Y_i **bağımlı değişken (cevap, yanıt değişkeni, etkilenen)**

$$Y_i = \beta_0 + \beta_1 X_i + e_i$$ **Basit doğrusal regresyon modeli**

β_1 **= eğim katsayısı**

β_0 **= sabit (kesen) katsayı**

Regresyon Modeli Tahmini

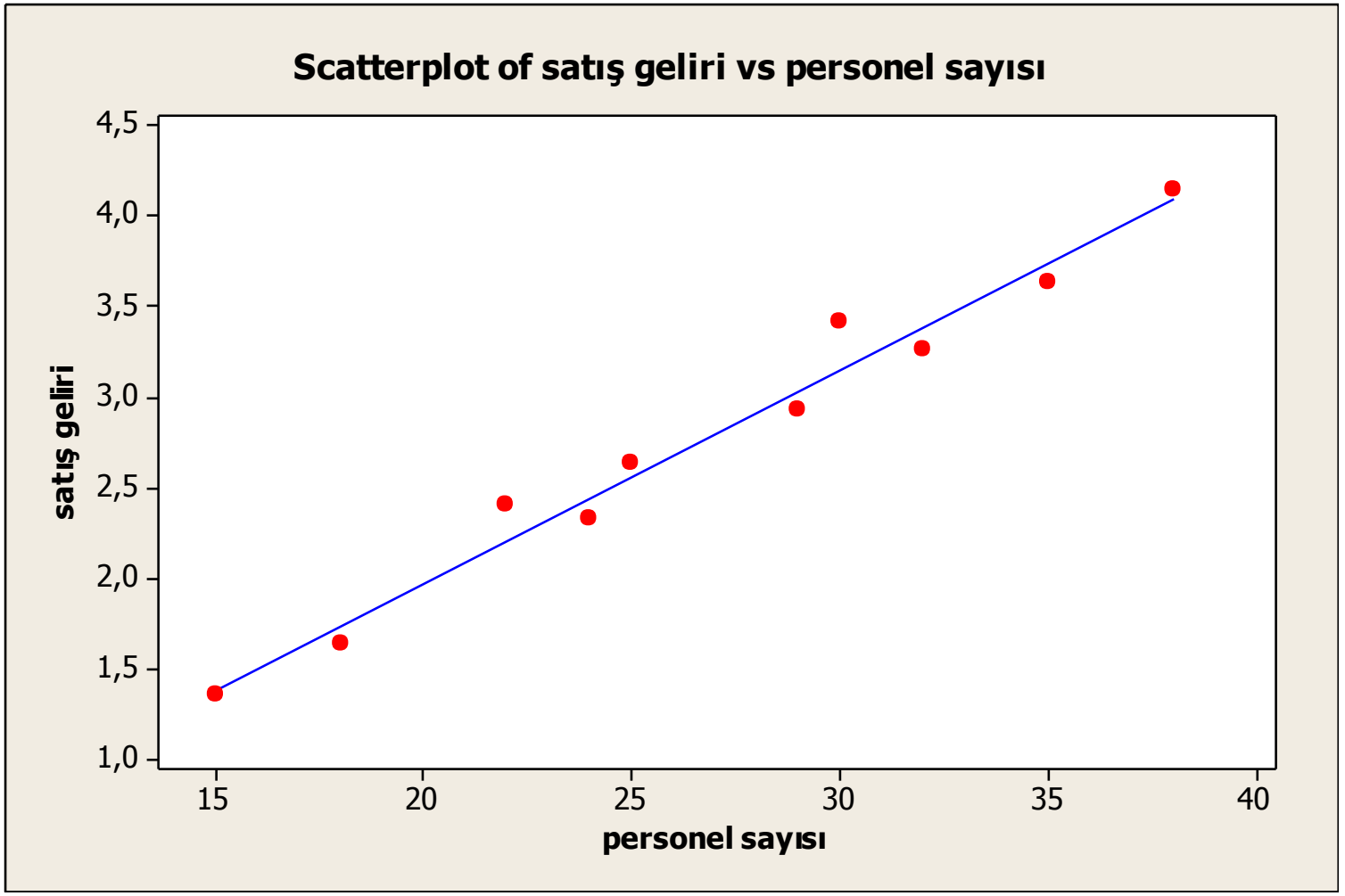
Basit doğrusal regresyon denklemi,

$$\hat{Y}_i = a + bX_i$$

Bağımsız değişkenin bağımlı değişken üzerindeki etkisini gösterir.

a = **sabit** b = **eğim**

Regresyon Doğrusu



Notasyon

	Anakütle Parametresi	Örnek istatistiği
Regresyon denkleminde sabit	β_0	a
Regresyon denkleminde eğim	β_1	b
Regresyon modeli ve eşitliği	$Y_i = \beta_0 + \beta_1 X_i + e_i$	$\hat{Y}_i = a + bX_i$

$$Y_i = a - bX_i + \hat{e}_i$$

Hata terimleri (Artıklar) ve En Küçük Kareler Yöntemi

- **Hata terimleri (Artıklar)**

$$\hat{e}_i = (Y_i - \hat{Y}_i)$$

En Küçük Kareler Yöntemi

$\sum \hat{e}_i^2$ 'yi minimum yapan a ve b değerlerinin bulunmasıdır.

β_0 ve β_1 için En Küçük Kareler Tahminleyicileri

$$b = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2} = \frac{\sum x_i y_i}{\sum x_i^{\,2}}$$

$$\mathrm{a} = \overline{\mathrm{Y}} - \mathrm{b}\overline{\mathrm{X}}$$

Örnek Veriler İçin Regreyon Katsayılarının Hesaplanması

Yıllar	*Satış Personeli Sayısı (X_i)*	*Satış Gelirleri (yüz bin $) ($Y_i$)*
1999	15	1,35
2000	18	1,63
2001	24	2,33
2002	22	2,41
2003	25	2,63
2004	29	2,93
2005	30	3,41
2006	32	3,26
2007	35	3,63
2008	38	4,15
Toplamlar	**268**	**27,73**

Regreyon Katsayılarının Hesaplanması

Yıllar	Satış Personeli Sayısı (X_i)	Satış Gelirleri (yüz bin \$) (X_i)	$(X_i - \bar{X})$	$(Y_i - \bar{Y})$	$(X_i - \bar{X})(Y_i - \bar{Y})$	$(X_i - \bar{X})^2$	
1999	15	1,35	-11,8	-1,42	16,76	139,24	
2000	18	1,63	-8,8	-1,14	10,03	77,44	
2001	24	2,33	-2,8	-0,44	1,23	7,84	
2002	22	2,41	-4,8	-0,36	1,73	23,04	
2003	25	2,63	-1,8	-0,14	0,25	3,24	
2004	29	2,93	2,2	0,16	0,35	4,84	
2005	30	3,41	3,2	0,64	2,05	10,24	
2006	32	3,26	5,2	0,49	2,55	27,04	
2007	35	3,63	8,2	0,86	7,05	67,24	
2008	38	4,15	11,2	1,38	15,46	125,44	
Toplamlar	268	27,73			57,46	485,6	

Satış gelirinin personel sayısı ile açıklandığı regresyon denklemi katsayılarının (a, b) tahmin edilmesi

$$Y_i = -0{,}17 + 0{,}11\ X_i$$

b = 0,11 Personel sayısında bir birimlik bir artış olduğunda satış gelilerinde 0,11 (xYüzbin Dolar) birimlik artış olur.

a = - 0,17 Personel sayısı sıfır olduğunda satış gelirleri -0,17 (xYüzbin Dolar) olur. Yani 17000 Dolarlık bir zarar olur.

Tahmin

Verilen bir X_i değeri için denklemden tahmin edilen $\hat{Y}_i$ nin (teorik, tahmin edilen) değeri ne olur?..

Eğer anlamlı bir korelasyon varsa, en iyi tahmin edilen $\hat{Y}_i$ değeri, X_i değerinin regresyon denkleminde yerine konulmasıyla bulunur.

Denklemden satış gelirinin tahmin edilmesi

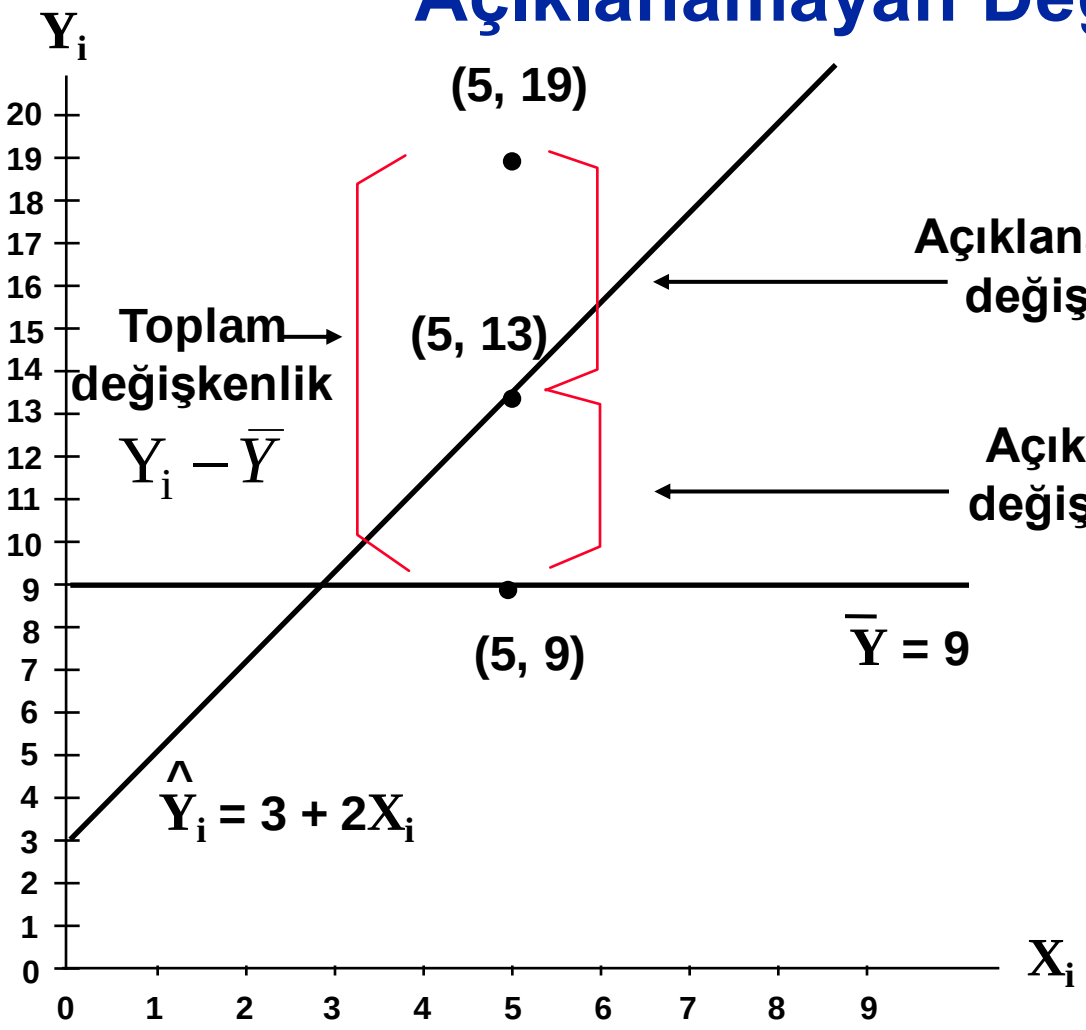
$$\hat{Y}_i = -0{,}17 + 0{,}11\, X_i$$

$$\hat{Y}_i = ?$$

Denklemden Hata terimlerin (Artıklar) tahmin edilmesi

$$\hat{e}_i = (Y_i - \hat{Y}_i) = ?$$

Toplam Değişkenlik, Açıklanan Değişkenlik ve Açıklanamayan Değişkenlik



(Toplam değişkenlik) = (Açıklanan değişkenlik) + (Açıklanamayan değişkenlik)

$$Y_i - \bar{Y} = (\hat{Y}_i - \bar{Y}) + (Y_i - \hat{Y}_i)$$

(Genel kareler toplamı) = (Regresyon kareler toplamı) + (Artık kareler toplamı)

$$\sum (Y_i - \bar{Y})^2 = \sum (\hat{Y}_i - \bar{Y})^2 + \sum (Y_i - \hat{Y}_i)^2$$

Tahmin Edilen Teorik $\hat{Y}_i$ ve $\hat{e}_i$ değerleri

$\hat{Y}_i$	$\hat{e}_i$	$\hat{e}_i^2$
1,48	-0,13	0,0169
1,81	-0,18	0,0324
2,47	-0,14	0,0196
2,25	0,16	0,0256
2,58	0,05	0,0025
3,02	-0,09	0,0081
3,13	0,28	0,0784
3,35	-0,09	0,0081
3,68	-0,05	0,0025
4,01	0,14	0,0196
Toplam		0,2137

Belirlilik Katsayısı

Y_i'deki (bağımlı değişkendeki) değişkenliğin ne kadarının bağımsız değişkenlerdeki (regresyon doğrusu) değişim tarafından açıklanabildiğini gösterir.

Basit doğrusal regresyon modellerinde belirlilik katsayısı, doğrusal korelasyon katsayısının r'nin karesine eşittir. r^2=Belirlilik katsayısı.

Çoklu regresyon modellerinde belirlilik katsayısı aşağıdaki formülle hesaplanır.

$$r^2 = \frac{\sum(\hat{Y}_i - \bar{Y})^2}{\sum(Y_i - \bar{Y})^2} = 1 - \frac{\sum e_i^2}{\sum(Y_i - \bar{Y})^2} = \frac{RKT}{GKT}$$

Örnek Veriler İçin Belirlilik Katsayısı

Satış gelirlerindeki (Y_i'deki) değişimin %97,4'ü, personel sayısındaki (X_i'deki) değişim tarafından açıklanabilmektedir.

$$r^2 = \frac{\sum(\hat{Y}_i - \bar{Y})^2}{\sum(Y_i - \bar{Y})^2} = 1 - \frac{\sum e_i^2}{\sum(Y_i - \bar{Y})^2} = \frac{RKT}{GKT}$$

$$\mathbf{r^2 = \%96,04}$$

Korelasyon Katsayısının Anlamlılığının Testi

- ρ = Anakütle korelasyon katsayısı
- $H_0: \rho = 0$
  (anlamlı bir korelasyon yoktur)

  $H_1: \rho \neq 0$
  (anlamlı bir korelasyon vardır)

Test İstatistiği *t*

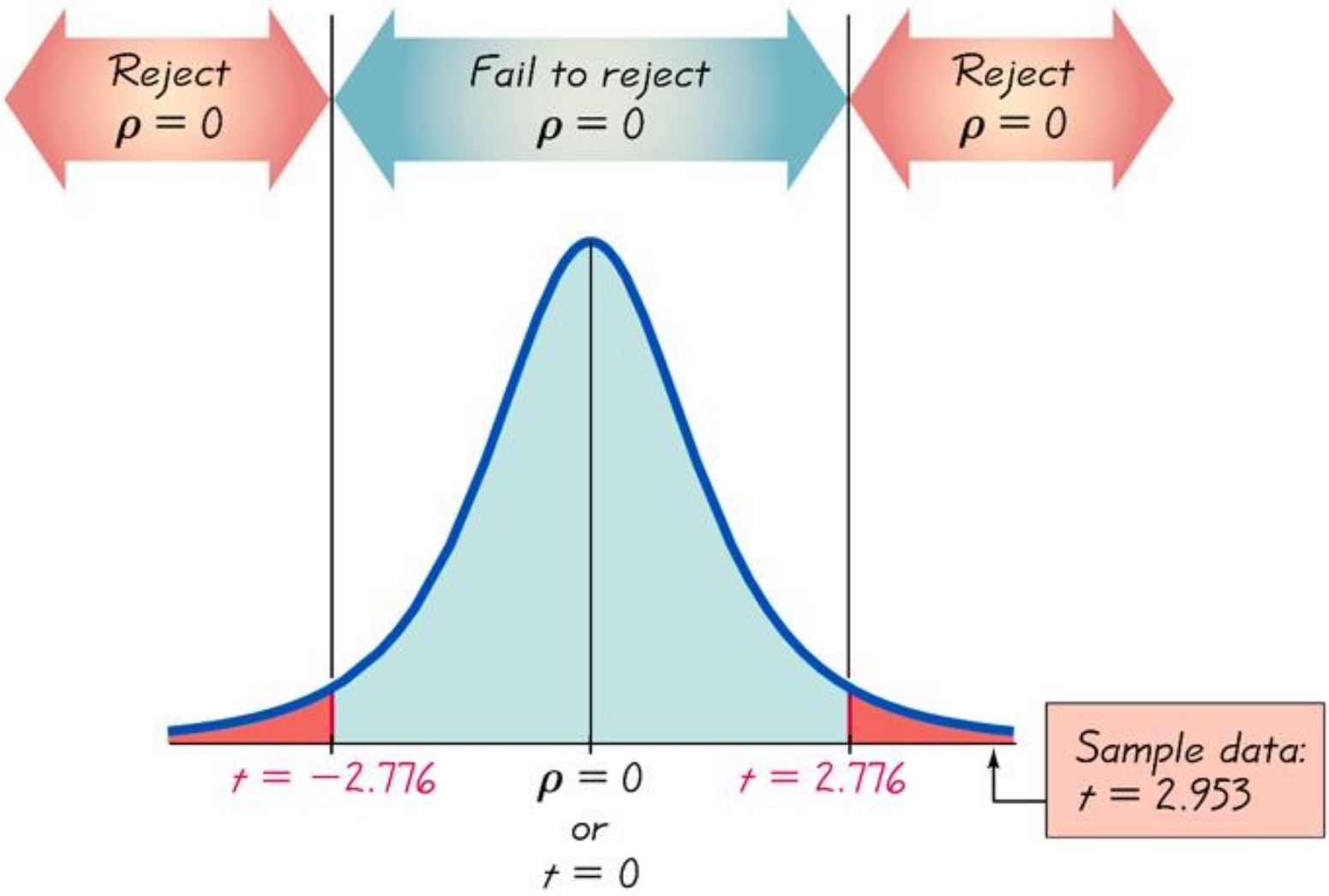
Test istatistiği:

$$t_{hesap} = \frac{r-\rho}{\sqrt{\frac{1 - r^2}{n - 2}}}$$

Kritik değerler

serbestlik derecesi = n - 2 olan tablo değerleri dikkate alınarak karar verilir.

Ret Bölgeleri



Örnek Verileri İçin Anakütle Korelasyon Katsayısının Testi (t Testi)

- ρ = Anakütle korelasyon katsayısı
- $H_0: \rho = 0$

(satış personeli sayısı ile satış gelirleri arasında anlamlı bir korelasyon yoktur)

$H_1: \rho \neq 0$

(satış personeli sayısı ile satış gelirleri arasında anlamlı bir korelasyon vardır)

Test istatistiği:

$$t_{hesap} = \frac{r-\rho}{\sqrt{\frac{1-r^2}{n-2}}} = \frac{0{,}987-0}{\sqrt{\frac{1-0{,}987^2}{10-2}}} = 17{,}39$$

Kritik değer

serbestlik derecesi = $n - 2 = 10 - 2 = 8$, $\alpha = 0{,}05$ için $t_{0{,}025,\,8} = 2{,}31 < 17{,}39$

Karar: H_0 red. Korelasyon anlamlıdır.

Regreyon Katsayılarının ve Regreyon Modelinin Anlamlılığının Testi

- Regreyon katsayılarının (t testi) ve regreyon modelinin anlamlılığının testi (F testi) ni yapabilmek için öncelikle standart hataların hesaplanması gerekmektedir.

Standart Hataların Hesaplanması

Tahminin Standart Hatası

$$S_{\hat{e}_i} = \sqrt{\frac{\sum(\hat{e}_i)^2}{(n - \mathrm{k})}}$$

Sabit Katsayının (a) Standart Hatası

$$S_a = \sqrt{S_{\hat{e}_i}\left[\frac{1}{n} \cdot \frac{\overline{X}^2}{\sum(X_\mathrm{i} - \overline{\mathrm{X}})^2}\right]}$$

Eğim Katsayının (b) Standart Hatası

$$S_b = \sqrt{\frac{S_{\hat{e}_i}^2}{\sum(X_\mathrm{i} - \overline{\mathrm{X}})^2}}$$

Regresyon Katsayısılarının Testi (t Testi)

β_1 ve β_0 **Anakütle regresyon katsayıları**

β_1 için

H_0: $\beta_1 = 0$
(β_1 anlamsızdır)
H_1: $\beta_1 \neq 0$
(β_1 anlamlıdır)

$$t_{hes} = \frac{b - \beta_1}{S_b}$$

β_0 için

H_0: $\beta_0 = 0$
(β_0 anlamsızdır)
H_1: $\beta_0 \neq 0$
(β_0 anlamlıdır)

$$t_{hes} = \frac{a - \beta_0}{S_a}$$

Kritik değerler
serbestlik derecesi = n - k olan tablo değerleri dikkate alınarak karar verilir. (modelde hesaplanacak katsayı adedi)

$$|t_{hesap}| > t_{\alpha/2,\, n-k} \text{ ise } H_0 \text{ Red.}$$

Standart Hatalar

$S_{\hat{e}_i}$ = Tahminin Standart Hatası

$$S_{\hat{e}_i} = \sqrt{\frac{\sum(\hat{e}_i)^2}{(n - k)}} = 0{,}10685$$

S_b = b_1'in standart hatasıdır.

$$S_b = \sqrt{\frac{S_{\hat{e}_i}^2}{\sum(X_i - \overline{X})^2}}$$

S_a = a'nın standart hatasıdır.

$$S_a = \sqrt{S_{\hat{e}_i}\left[\frac{1}{n} \cdot \frac{\overline{X}^2}{\sum(X_i - \overline{X})^2}\right]}$$

Örnek Veriler ile Regresyon Katsayısılarının Testi (t Testi)

- β_1 = Anakütle regresyon katsayısı (X_1 için)

- H_0: $\beta_1 = 0$
  (β_1 anlamsızdır)
  H_1: $\beta_1 \neq 0$
  (β_1 anlamlıdır)

Test İstatistiği *t*

Test istatistiği:

$$t = \frac{b - \beta_1}{S_b} = \frac{0,11}{0,006804} = 16,16$$

$$S_b = \sqrt{\frac{S^2_{\hat{e}_i}}{\sum (X_i - \overline{X})^2}} = 0,006804$$

Kritik değerler

serbestlik derecesi = *n* - k olan tablo değerleri dikkate alınarak karar verilir. α = 0,05 olsun.

$|16,16| > t_{\alpha/2,\, n-2} = t_{0,025,\, 8} = 2,306$

H_0 Red. β_1 anlamlıdır.

B_0 için

- β_0 = Anakütle regresyon modelinde sabit terim

- H_0: $\beta_0 = 0$
  (β_0 anlamsızdır)
  H_1: $\beta_0 \neq 0$
  (β_0 anlamlıdır)

Test İstatistiği *t*

Test istatistiği:

$$t = \frac{a - \beta_0}{S_a} = \frac{-0{,}17}{0{,}1884} = -0{,}902$$

$$S_a = \sqrt{S_{\hat{e}_i}\left[\frac{1}{n} \cdot \frac{\overline{X}^2}{\sum (X_i - \overline{X})^2}\right]} = 0{,}1884$$

Kritik değerler

serbestlik derecesi = *n* - 2 olan tablo değerleri dikkate alınarak karar verilir. α = 0,05 olsun.

$|-0{,}902| < t_{\alpha/2,\, n-2} = t_{0{,}025,\, 8} = 2{,}306$

H_0 REDDEDİLEMEZ. β_0 anlamsızdır.

F - Testi

- $H_0: \beta_1 = \beta_2 = \ldots = \beta_k = 0$

  (Model anlamsızdır)

  H_1: en az bir i için $\beta_i \neq 0$

  (Model anlamlıdır)

Test İstatistiği = ***F*** **– oranı** (F_{hesap}) = $$\frac{\sum(\hat{Y}_i - \bar{Y})}{\sum(Y_i - \hat{Y}_i)} = \frac{RKO}{AKO}$$

Basit doğrusal regresyonda $t^2 = F$ **olmaktadır.**

Ret Bölgesi = $F_{hesap} > F_{\alpha, k-1, (n-k)}$ **ise** H_0 **RET. (k modelde hesaplanacak katsayı adedi)**

F – Testi (Satış Gelirleri Örneği İçin)

- $H_0: \beta_0 = \beta_1 = 0$

  (Model anlamsızdır)

  H_1: **En az birisi sıfırdan farklı** β

  (Model anlamlıdır)

Test İstatistiği $F_{hes} = \dfrac{6,7982}{0,0225} = 302,41$

Karar = $F_{hes} = 302,41 > F_{0,05,\,1,\,8} = 5,32$ H_0 **RET.**