İSTATİSTİK-II

Korelasyon ve Regresyon

Korelasyon ve Regresyon

- Genel Bakış
- Korelasyon
- Regresyon
- Belirleme katsayısı
- Varyans analizi
- Kestirimler için aralık tahminlemesi

Genel Bakış

İkili veriler

- aralarında bir ilişki var mıdır?
- varsa bu ilişki bir eşitlik ile temsil edilebilir mi?
- bu eşitliğin kestirimler (öngörümler) için kullanılması

Korelasyon

Tanım

❖**Korelasyon**

bir değişkenin değeri değişirken diğer bir değişken bununla doğrusal ilişkili olarak değişiyorsa korelasyon vardır denebilir.

Varsayımlar

1. (x,y) ikili verilerden oluşan örnek bir şans örneğidir.

2. x ve y'lerin dağılışı normaldir.

Tanım

- **Saçılma diyagramı**

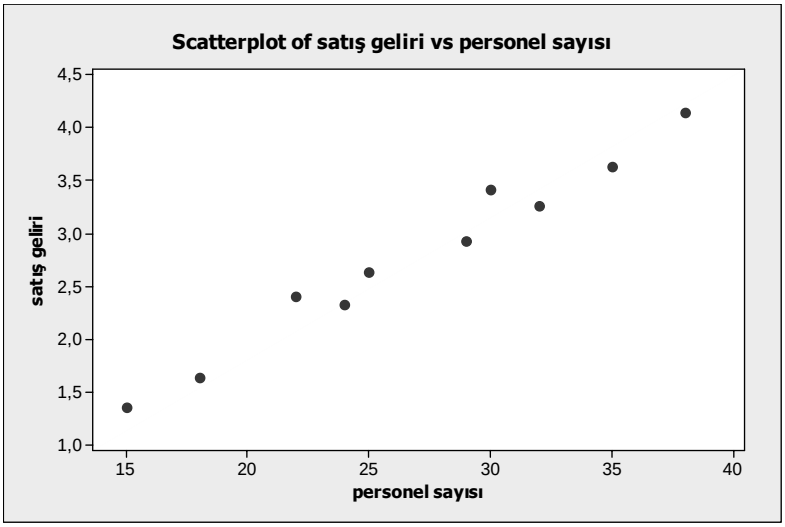
  yatay eksen x, dikey eksen y olmak üzere, (x,y) ikili örnek verilerinin işaretlendiği bir grafiktir. Her bir (x,y) ikilisi tek bir noktadır.

Örnek

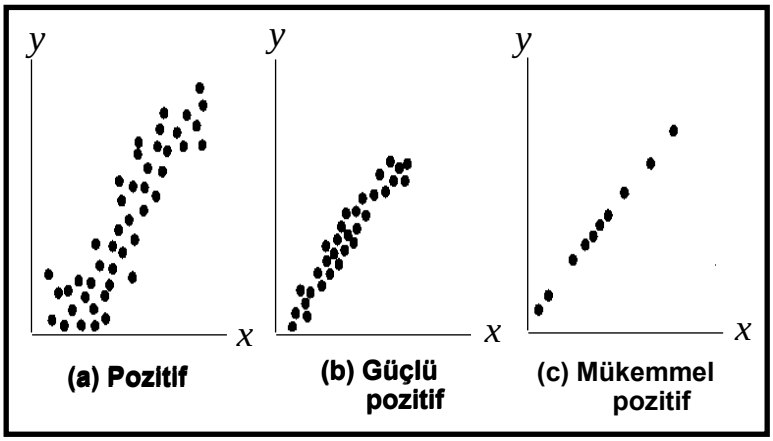
Bir firma bünyesindeki satış personeli sayısı ile satış gelirleri arasındaki ilişkiyi bilmek istemektedir.

Yıllar	*Satış Personeli Sayısı (x)*	*Satış Gelirleri (yüz bin $) (y)*
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

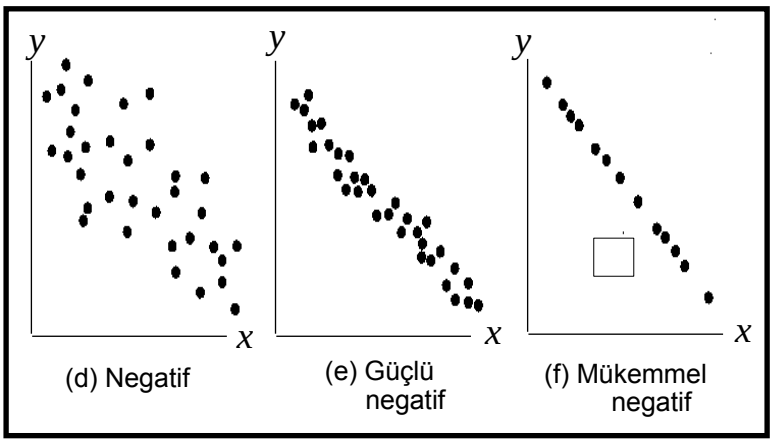
İkili Verilerin Saçılma Diyagramı



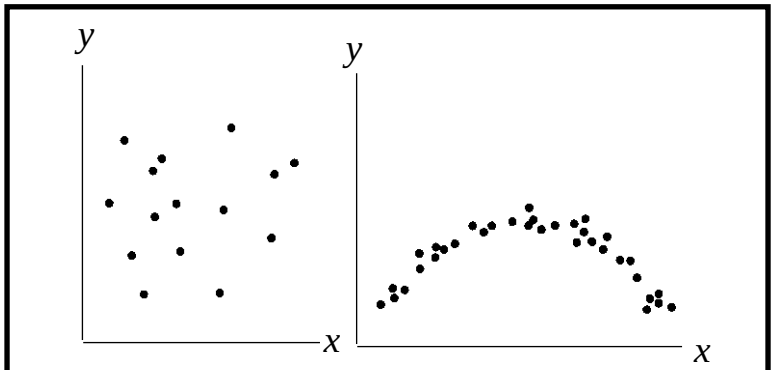
Pozitif Korelasyon



Negatif Korelasyon



(d) Negatif (e) Güçlü negatif (f) Mükemmel negatif



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

Tanım
Korelasyon Katsayısı *r*

Bir örnekteki *x* ve *y* ikili değerleri arasındaki doğrusal ilişkinin gücünü ölçmektedir.

$$r = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{n(\Sigma x^2) - (\Sigma x)^2}\sqrt{n(\Sigma y^2) - (\Sigma y)^2}}$$

Korelasyon Katsayısı *r*'nin Özellikleri

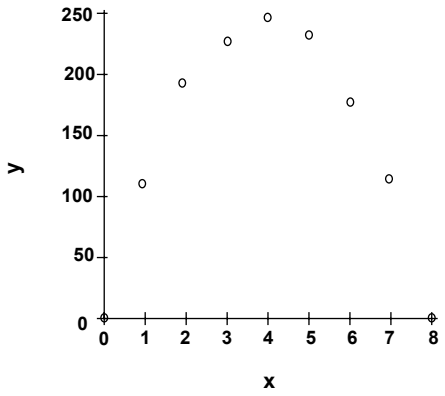
1. $-1 \leq r \leq 1$
2. **Mükemmel pozitif doğrusal ilişki olduğunda $r = 1$ olur.**
3. **Mükemmel negatif doğrusal ilişki olduğunda $r = -1$ olur.**
4. **Doğrusal ilişki yok ise $r = 0$ olur.**

Korelasyon ile ilgili hatalar

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

2. Doğrusallık: x ile y arasında anlamlı bir korelasyon olmadığı halde, aralarında farklı şekilde bir ilişki olabilir. (Bakınız izleyen slayt)

Korelasyon ile ilgili hatalar



Örnek Verileri İçin Korelasyon Hesaplamaları

Yıllar	*Satış Personeli Sayısı (x)*	*Satış Gelirleri (yüz bin $) (y)*	x^2	y^2	*xy*
1999	15	1,35	225	1,8225	20,25
2000	18	1,63	324	2,6569	29,34
2001	24	2,33	576	5,4289	55,92
2002	22	2,41	484	5,8081	53,02
2003	25	2,63	625	6,9169	65,75
2004	29	2,93	841	8,5849	84,97
2005	30	3,41	900	11,6281	102,3
2006	32	3,26	1024	10,6276	104,32
2007	35	3,63	1225	13,1769	127,05
2008	38	4,15	1444	17,2225	157,7
Toplamlar	**268**	**27,73**	**7668**	**83,8733**	**800,62**

Örnek Verileri İçin Korelasyon Hesaplamaları

$$r = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{n(\Sigma x^2) - (\Sigma x)^2}\sqrt{n(\Sigma y^2) - (\Sigma y)^2}}$$

$$r = \frac{(10)(800,62) - (268)(27,73)}{\sqrt{(10)(7668) - (268)^2}\sqrt{(10)(83,8733) - (27,73)^2}}$$

r = 0,987 Güçlü pozitif korelasyon

Anakütle Korelasyon Katsayısının 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*

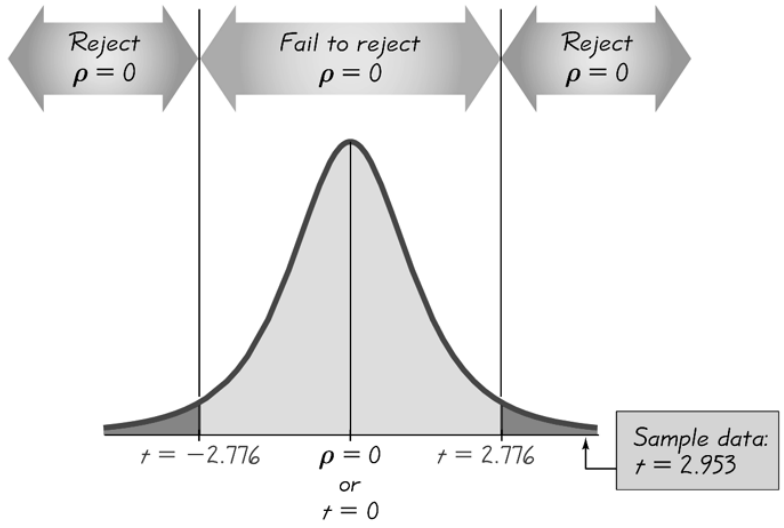
Test istatistiği:

$$t = \frac{r}{\sqrt{\frac{1 - r^2}{n - 2}}}$$

Kritik değerler

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

Ret Bölgeleri



Anakütle Korelasyon Katsayısının 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 İstatistiği *t*

Test istatistiği:

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

Kritik değer

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

Karar: H_0 ret. Korelasyon anlamlıdır.

Regresyon

x **bağımsız değişken (açıklayıcı değişken)**

y **bağımlı değişken (cevap = yanıt değişkeni)**

$$y = b_0 + b_1 x + e$$ **Basit doğrusal regresyon modeli**

b_1 **= eğim** b_0 **= kesen**

Regresyon

❖ **Regresyon Eşitliği**

Verilen bir ikili veriler topluluğu için regresyon eşitliği,

$$\hat{y} = b_0 + b_1 x$$

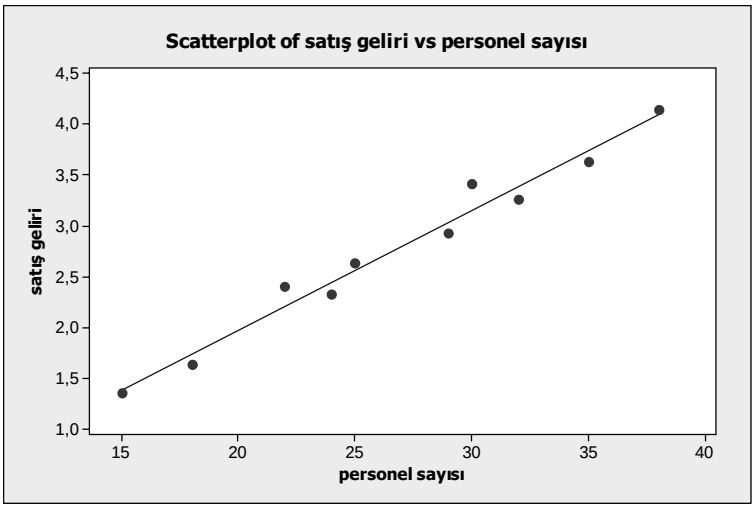
iki değişken arasındaki ilişkiyi tanımlamaktadır.

b_1 = eğim b_0 = kesen

❖ **Regresyon Doğrusu**

Regresyon eşitliğinin grafiğidir.

Regresyon Doğrusu



Notasyon

	Anakütle Parametresi	Örnek istatistiği
Regresyon eşitliğinde kesen	β_0	b_0
Regresyon eşitliğinin eğimi	β_1	b_1
Regresyon modeli ve eşitliği	$y = \beta_0 + \beta_1 x + \varepsilon$	$\hat{y} = b_0 + b_1 x$

Artıklar ve En Küçük Kareler Yöntemi

- **Artıklar**

  $e = (y - \hat{y})$

- **En Küçük Kareler Yöntemi**

  Σe^2'yi minimum yapan b_0 ve b_1 değerlerinin bulunmasıdır.

β_0 and β_1 için En Küçük Kareler Tahminleyicileri

$$b_0 = \frac{(\Sigma y)(\Sigma x^2) - (\Sigma x)(\Sigma xy)}{n(\Sigma x^2) - (\Sigma x)^2}$$

$$b_1 = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{n(\Sigma x^2) - (\Sigma x)^2}$$

Önce b_1 bulunursa, ardından

$$b_0 = \bar{y} - b_1\bar{x}$$

Satış geliri için regresyon eşitliğinin tahminlenmesi

$$b_1 = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{n(\Sigma x^2) - (\Sigma x)^2}$$

$$b_1 = \frac{10(800{,}62) - (268)(27{,}73)}{10(7668) - (268)^2}$$

$$b_1 = 0{,}118$$

$$b_0 = \overline{y} - b_1\overline{x} = 2{,}773 - (0{,}118)(26{,}8) = -0{,}398$$

Kestirimler (Öngörümler)

Verilen bir *x* değeri için *y*'nin değeri ne olur?..

Eğer anlamlı bir korelasyon varsa, en iyi öngörülen y değeri, x değerinin regresyon eşitliğinde yerine konulmasıyla bulunur.

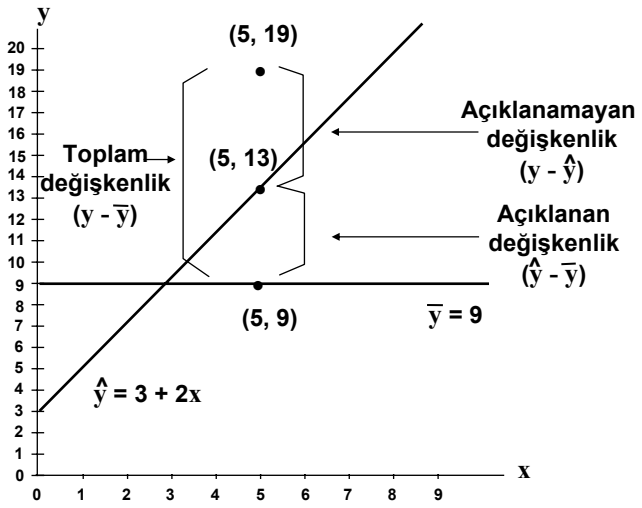
Önemli Not: Regresyon doğrusu yalnızca tahminlemede kullanılan x uzayı içinde geçerlidir. Mevcut x'lerden uzak bir noktada öngörümleme yapılmamalıdır.

30 satış personeli çalıştığında satış gelirinin kestirilmiş değeri nedir?

$\hat{y} = - 0.398 + 0.118 (30)$

$\hat{y} = 3.1516,$ 315160 $

Toplam Değişkenlik, Açıklanan Değişkenlik ve Açıklanamayan Değişkenlik



(toplam değişkenlik) = (açıklanan değişkenlik) + (açıklanamayan değişkenlik)

$$(y - \bar{y}) = (\hat{y} - \bar{y}) + (y - \hat{y})$$

(toplam değişkenlik) = (açıklanan değişkenlik) + (açıklanamayan değişkenlik)

$$\Sigma (y - \bar{y})^2 = \Sigma (\hat{y} - \bar{y})^2 + \Sigma (y - \hat{y})^2$$

(Genel kareler toplamı) = (regresyon kareler toplamı) + (artık kareler toplamı)

Tanım

Belirleme Katsayısı

y'deki değişkenliğin ne kadarının regresyon doğrusu tarafından açıklanabildiğini söyler.

$$r^2 = \frac{\text{Regresyon kareler toplamı}}{\text{Genel kareler toplamı}}$$

$$r^2 = \frac{\Sigma (\hat{y} - \bar{y})^2}{\Sigma (y - \bar{y})^2} = \frac{\text{RKT}}{\text{GKT}}$$

$$r^2 = \frac{\Sigma (\hat{y} - \bar{y})^2}{\Sigma (y - \bar{y})^2} = \frac{b_1^2(\Sigma x^2 - (\Sigma x)^2/n)}{\Sigma y^2 - (\Sigma y)^2/n}$$

$$r^2 = \frac{0{,}118^2(7668 - (268)^2/10)}{83{,}873 - (27{,}73)^2/10} = \%97{,}4$$

y'deki değişmelerin %97,4'ü regresyon doğrusu ile açıklanabilmektedir.

Varyans Analizi Tablosu (VAT)

Değişkenlik Kaynağı	*Kareler Toplamları* (KT)	*Serbestlik Derecesi*	*Kareler Ortalaması* (KO)	*F-Oranı*
Regresyon	RKT = $b_1^2(\Sigma x^2 - (\Sigma x)^2/n)$	1	Regresyon KO = RKO = RKT / 1	$F = \frac{RKO}{AKO}$
Artık	Artık Kareler Toplamı AKT = GKT - RKT	n - 2	Artık KO = AKO = AKT / (n – 2) = S^2	
Toplam (Genel)	Genel Kareler Toplamı GKT = $\Sigma y^2 - (\Sigma y)^2/n$	n - 1		

Hata Varyansının Tahmini

$$S^2 = \frac{\Sigma (y - \hat{y})^2}{n - 2}$$ = Artık Kareler Ortalaması

Tahminin Standart Hatası

$$S = \sqrt{\frac{\Sigma (y - \hat{y})^2}{n - 2}}$$

F - Testi

- H_0: $\beta_1 = \beta_2 = ... = \beta_k = 0$
  (Model anlamsızdır)
  H_1: en az bir i için $\beta_i \neq 0$
  (Model anlamlıdır)

F – Testi (Basit Doğrusal Regresyon İçin)

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

Test İstatistiği = *F* – oranı

Ret Bölgesi = $F > F_{\alpha, 1, (n-2)}$ ise H_0 RET.

Varyans Analizi Tablosu (VAT) (Satış Gelirleri Örneği)

Değişkenlik Kaynağı	***Kareler Toplamları* (KT)**	***Serbestlik Derecesi***	***Kareler Ortalaması* (KO)**	***F-Oranı***
Regresyon	RKT = $b_1^2(\Sigma x^2 - (\Sigma x)^2/n)$ = $0{,}118^2(7668 - (268)^2/10)$ = 6,7982	1	Regresyon KO = RKO = RKT / 1 = 6,7982 / 1 = 6,7982	$F = \frac{RKO}{AKO}$
Artık	Artık Kareler Toplamı AKT = GKT – RKT = 6,9780 - 6,7982 = 0,1798	n – 2 = 10 – 2 = 8	Artık KO = AKO = AKT / (n – 2) = 0,1798 / 8 = 0,0225	$F = \frac{6{,}7982}{0{,}0225}$ = 302,41
Toplam (Genel)	GKT = $\Sigma y^2 - (\Sigma y)^2/n$ = $83{,}873 - (27{,}73)^2/10$ = 6,9780	n – 1 = 10 – 1 = 9		

F – Testi (Satış Gelirleri Örneği İçin)

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

Test İstatistiği = *F* – oranı = 302,41

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

Anakütle Regresyon Katsayısılarının 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:

$$t = \frac{b_1}{S_{b_1}}$$

S_{b_1} **= b_1'in standart hatasıdır.**

$$S_{b_1} = \frac{S}{\sqrt{(\Sigma x^2 - (\Sigma x)^2/n)}}$$

Kritik değerler

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

$|t| > t_{\alpha/2,\, n-2}$ ise H_0 RET.

Anakütle Regresyon Katsayısılarının Testi (Satış Gelirleri Örneği)

- 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:

$$t = \frac{b_1}{S_{b_1}} = \frac{0,118}{0,006804} = 17,39$$

S_{b_1} = b_1'in standart hatasıdır.

$$S_{b_1} = \frac{S}{\sqrt{(\Sigma x^2 - (\Sigma x)^2/n)}} = \frac{0,1499}{\sqrt{(7668 - (268)^2/10)}} = 0,006804$$

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

$|17{,}39| > t_{\alpha/2,\, n-2} = t_{0{,}025,\, 8} = 2{,}306$
H_0 RET. β_1 anlamlıdır.

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

Anakütle Regresyon Katsayısılarının Testi

- **β_0 = Anakütle regresyon modelinde sabit terim**

- **H_0: β_0 = 0**
  (β_0 anlamsızdır)
  H_1: β_0 ≠ 0
  (β_0 anlamlıdır)

Test İstatistiği *t*

Test istatistiği:

$$t = \frac{b_0}{S_{b_0}}$$

S_{b_0} = b_0**'in standart hatasıdır.**

$$S_{b_0} = \frac{S\sqrt{\Sigma x^2}}{\sqrt{n(\Sigma x^2 - (\Sigma x)^2/n)}}$$

Kritik değerler

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

$|t| > t_{\alpha/2,\, n-2}$ ise H_0 RET.

Anakütle Regresyon Katsayısılarının Testi (Satış Gelirleri Örneği)

- β_0 = Anakütle regresyon modelindeki sabit terim
- H_0: $\beta_0 = 0$
  (β_0 anlamsızdır)
  H_1: $\beta_0 \neq 0$
  (β_0 anlamlıdır)

Test İstatistiği *t*

Test istatistiği:

$$t = \frac{b_0}{S_{b_0}} = \frac{-0,398}{0,1884} = -2,11$$

$$S_{b_1} = \frac{S\sqrt{\Sigma x^2}}{\sqrt{n(\Sigma x^2 - (\Sigma x)^2/n)}} = \frac{(0,1499)\sqrt{(7668)}}{\sqrt{(10)(7668 - (268)^2/10)}}$$

$$= 0,1884$$

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

$|- 2{,}11 | < t_{\alpha/2,\, n-2} = t_{0{,}025,\, 8} = 2{,}306$

H_0 REDDEDİLEMEZ. β_0 anlamsızdır.

E(y) Değeri İçin Kestirim Aralığı

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

Burada

$$E = t_{\alpha/2,n-2}\, s \sqrt{\frac{1}{n} + \frac{n(x_0 - \bar{x})^2}{n(\Sigma x^2) - (\Sigma x)^2}}$$

- x_0, x'in verilen bir değeridir.
- Karekök içindeki ifade ile S'nin çarpımı ise x_0'daki $\hat{y}$ değeri için standart hatadır.
- Standart hata en düşük değerini $x_0 = \bar{x}$ olduğunda alır.

E(y) Değeri İçin Kestirim Aralığı

x_0 = 30 personel için satışların beklenen değeri %95 güven ile hangi aralıkta gerçekleşir?

3.1516 - E < E(y) < 3.1516 + E

$$E = (2{,}306)(0{,}1499)\sqrt{\frac{1}{10} + \frac{(10)(30-26{,}8)^2}{(10)(7668) - (268)^2}}$$

E = (2,306)(0,01815) = 0,04186

3,1097 < E(y) < 3,1935