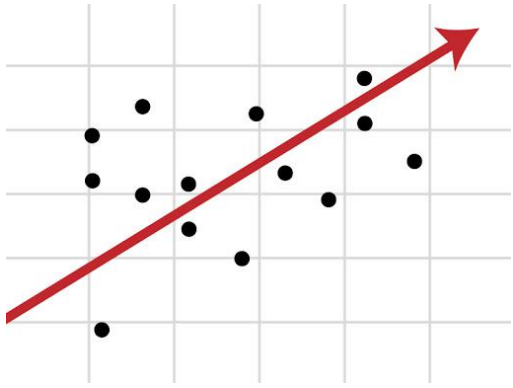
HACETTEPE
ÜNİVERSİTESİ

İSTATİSTİK BÖLÜMÜ

İST156 İSTATİSTİĞE GİRİŞ II

DERS 10
REGRESYON ÇÖZÜMLEMESİ

Ders sorumluları: Prof.Dr.Serpil AKTAŞ ALTUNAY (01 Şubesi)
Doç.Dr. Ayten YİĞİTER (02 Şubesi)



REGRESYON ÇÖZÜMLEMESİ

Aralarında sebep-sonuç ilişkisi bulunan değişkenler arasındaki ilişkinin bir istatistiksel model ile ifade edilmesidir.

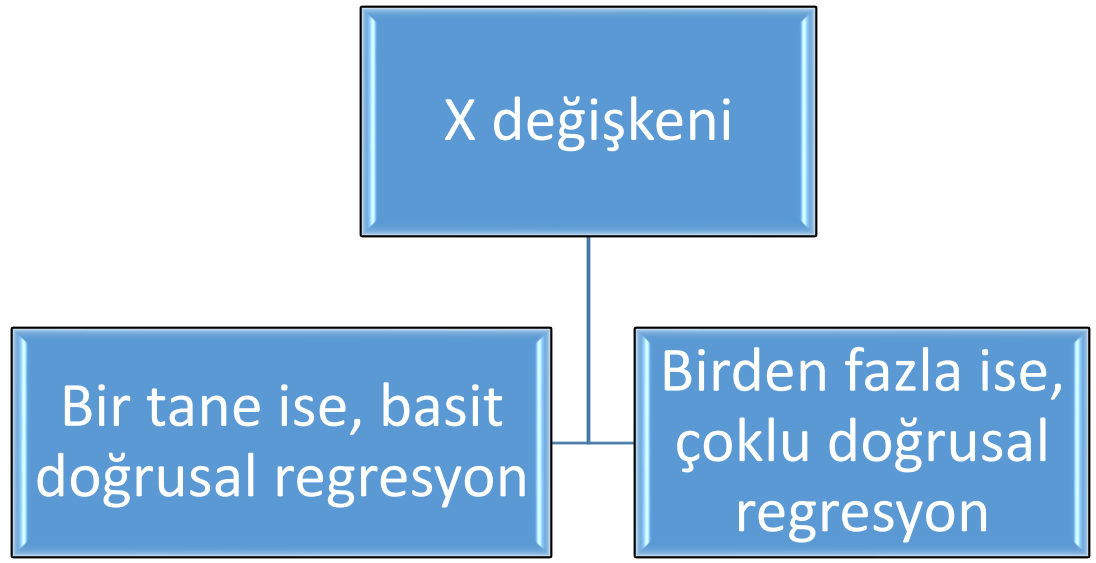
$$Y=\beta_0+\beta_1 X+\varepsilon$$

ile ifade edilen denkleme, basit doğrusal regresyon denklemi denir.

Y değişkeni: bağımlı değişken, açıklanan değişken, yordanan değişken, çıktı değişkeni (dependent variable)

X değişkeni : bağımsız değişken, açıklayıcı değişken, yordayıcı değişken, girdi değişkeni (independent variable, regressor variable)

gibi adlarla bilinir.

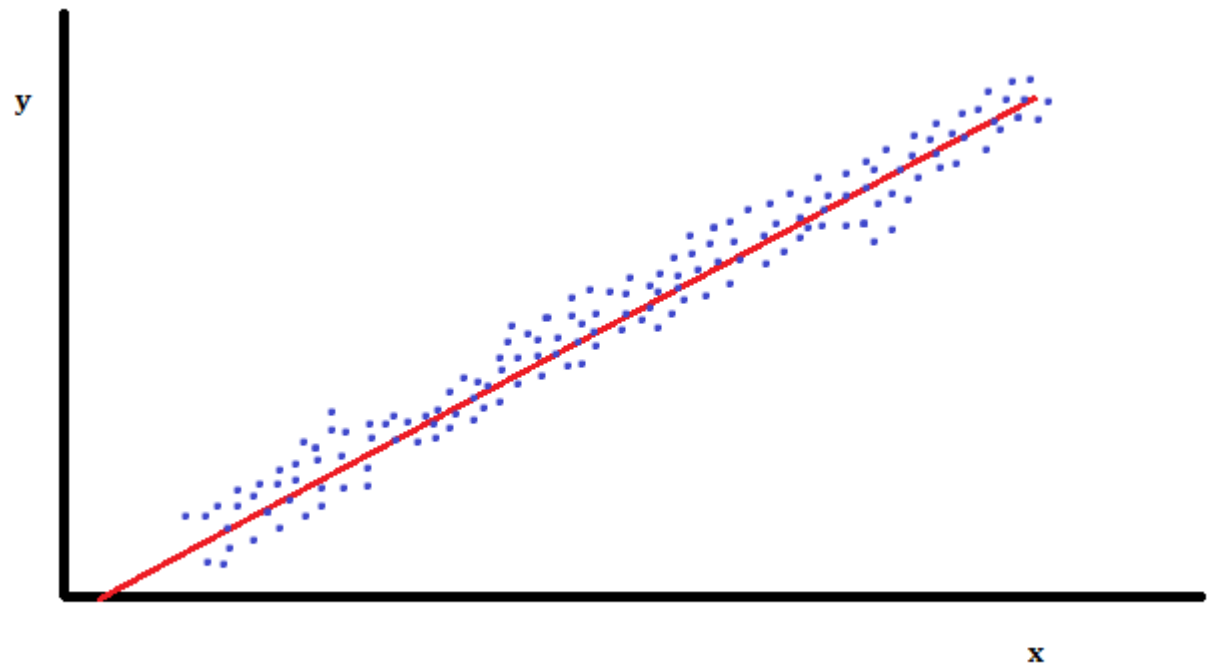


Basit doğrusal regresyonda,

$\beta_{0:}$ kesim noktası (genel ortalama) (intercept)
X'in değeri sıfır olduğunda Y'nin aldığı ortalama değerdir.

$\beta_{1:}$ Eğim (slope)
X'deki 1 birimlik değişimin Y'de yapmış olduğu değişikliği gösterir.
Negatif değer alırsa Y'deki azalışı gösterir.

ε : hata terimidir.



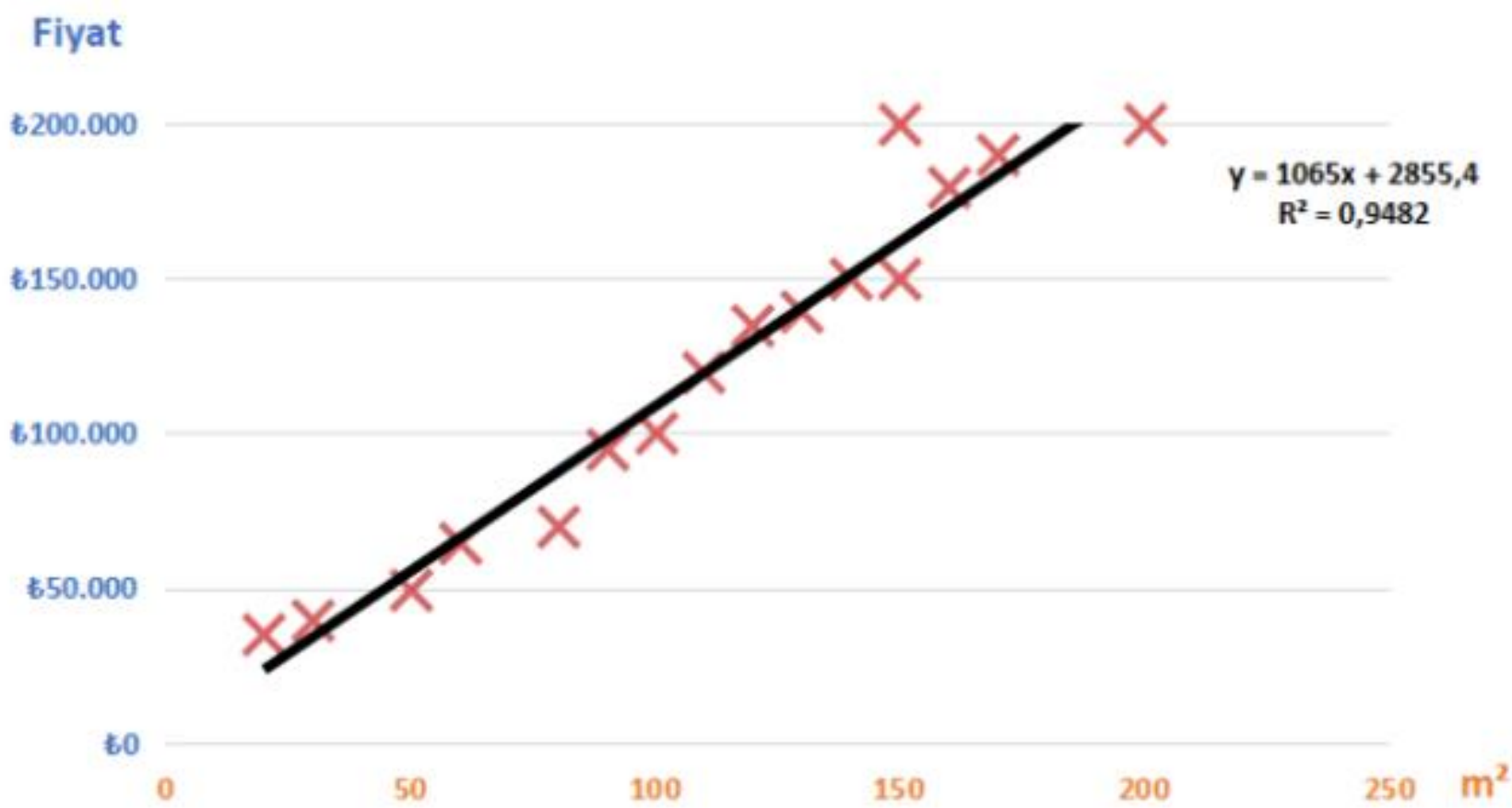
Amaç: X ve Y değişkenlerine ait noktaları temsil eden en iyi doğru denklemi bulunmaya çalışmaktır.

Örneğin: Reklam Harcaması ile Satışlar

Çalışma süresi ile başarı notu

Yağış ile verim

Fiyat ile talep


Fiyat
₺200.000
₺150.000
₺100.000
₺50.000
₺0
0
50
100
150
200
250
m²
y = 1065x + 2855,4
R² = 0,9482

VARSAYIMLAR:

- ✓ Y'nin dağılımı ya da hataların dağılımı normal dağılıma sahip olmalıdır.
- ✓ Hatalar ilişkisiz olmalıdır, $Cov(\varepsilon_i,\varepsilon_j)=0$.
- ✓ Sabit varyans koşulu sağlanmalıdır.
- ✓ Gözlemler bağımsız olmalıdır.
- ✓ Hata terimi ile bağımsız değişken arasında ilişki olmamalıdır, $Cov(X_i,\varepsilon_i)=0$.

EN KÜÇÜK KARELER YÖNTEMİ (EKK)

Hata terimlerinin karelerinin toplamının minimize ile yapılan bir tahmin yöntemidir. Amaç regresyon denklemindeki katsayıları hata kareler toplamını en küçük yapacak şekilde tahmin etmektir. Hata kareler toplamı (HKT),

$$\text{HKT}=\sum_{i=1}^{n}\varepsilon_i^2=\sum_{i=1}^{n}(Y_i-\beta_0-\beta_1 X_i)^2$$

ile tanımlanırsa, buna göre türevleri sıfır yapan değerler b_0 ve b_1 ile gösterilir

$$\left.\frac{\partial \text{HKT}}{\partial \beta_0}\right|_{b_0,b_1} = -2\sum_{i=1}^{n}(y_i-b_0-b_1x_i)^2=0$$

$$\left.\frac{\partial \text{HKT}}{\partial \beta_1}\right|_{b_0,b_1} = -2\sum_{i=1}^{n}(y_i-b_0-b_1x_i)\,x_i=0$$

olur.

Denklemler yeniden düzenlendiğinde,

$$\sum_{i=1}^{n} y_i = nb_0 + b_1 \sum_{i=1}^{n} x_i$$

$$\sum_{i=1}^{n} y_i x_i = b_0 \sum_{i=1}^{n} x_i + b_1 \sum_{i=1}^{n} x_i^2$$

elde edilir ve bu denklemlere normal denklemler adı verilir. Buradan β_0'ın tahmin edicisi b_0, β_1'in tahmin edicisi b_1 elde edilir.

$$ÇT_{XY} = \sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n}$$

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

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

($ÇT_{XY}$: X ve Y çarpımlar toplamı, KT_X: X'e ait kareler toplamı, KT_Y: Y'ye ait kareler toplamı)

$$b_1 = \frac{ÇT_{XY}}{KT_X}$$

$$b_1 = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}$$

$$b_1 = \frac{\sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n}}{\sqrt{\left(\sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}\right)}}$$

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

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

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$$

Tahmin edilen «Basit Doğrusal Regresyon» model denklemi:

$$\hat{y}_i = b_0 + b_1 x_i \quad i=1,2,\ldots,n$$

KATSAYILARIN ÖNEM KONTROLÜ

Hesaplanan katsayıların ayrı ayrı önem kontrolleri yapılır.

$H_0: \beta_0 = 0$
$H_s: \beta_0 \neq 0$

b_0'nın standart hatası,

$$S_{b_0} = \sqrt{\sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{KT_x} \right)}$$

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

$|t_{hesap}| \geq t_{tablo(\alpha/2,\, n-2)}$ ise H_0 reddedilir. Yani β_0 katsayısı anlamlıdır.

$H_0: \beta_1 = 0$
$H_s: \beta_1 \neq 0$

b_1'nın standart hatası,

$$S_{b_1} = \sqrt{\frac{\sigma^2}{KT_x}}$$

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

$|t_{hesap}| \geq t_{tablo(\alpha/2,\, n-2)}$ ise H_0 reddedilir. Yani β_1 katsayısı anlamlıdır.

β_1 ve β_0 için $(1-\alpha)\%100$ Güven Aralığı

Katsayılar için (1- α) güven düzeyinde güven aralığı hesaplanır.

β_1 için 1- α güven aralığı:

$$P\left(b_1 - t_{\left(\frac{\alpha}{2},n-2\right)} S_{b_1} \leq \beta_1 \leq b_1 + t_{\left(\frac{\alpha}{2},n-2\right)} S_{b_1}\right) = 1-\alpha$$

β_0 için 1- α güven aralığı:

$$P\left(b_0 - t_{\left(\frac{\alpha}{2},n-2\right)} S_{b_0} \leq \beta_0 \leq b_0 + t_{\left(\frac{\alpha}{2},n-2\right)} S_{b_0}\right) = 1-\alpha$$

Güven aralıkları eğer «0» kapsıyorsa o katsayılar anlamlı değil yorumu yapılır.

MODELİN ANLAMLILIĞI

Kurulan basit doğrusal regresyon denkleminin verileri temsil edip etmediği, yani önem kontrolü yapılır.

H_0: Model önemsizdir (anlamsızdır).
H_1: Model önemlidir (anlamlıdır).

hipotezi F testi ile test edilir.

$$F = \frac{RAKO}{AKO}$$

değeri $F_{Tablo(\alpha,1,n-2)}$ ile karşılaştırılır.

$F_{Hesap} \geq F_{tablo}$ ise H_0 reddedilir. Model anlamlıdır.

F hesaplamada kullanılan terimler:

RKT: Regresyon Kareler Toplamı

$RKT=b_1ÇT_{xy}$

RAKT: Regresyon Artık Kareler Toplamı

RAKT=KTy-RKT

RAKO: Regresyon Kareler Ortalaması

$$\text{RAKO} = \frac{\text{RKT}}{1}$$

AKO: Artık Kareler Ortalaması

$$\text{AKO} = \frac{\text{AKT}}{\text{n} - 2}$$

Basit doğrusal regresyonda, modelin geçerliliğinin test edilmesi ile,
H_0: $\beta_1 = 0$ hipotezinin test edilmesi aynı anlama gelir.

RAKO aynı zamanda σ^2'dir.
σ^2'nin tahmin edicisi $\hat{\sigma}^2$:

$$\hat{\sigma}^2 = \frac{\text{RAKT}}{n-2} = \frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{n-2}$$

$$= \frac{\sum_{i=1}^{n}(y_i-\underbrace{(b_0+b_1x_i)}_{\hat{y}_i})^2}{n-2}$$

ANOVA TABLOSU

Değişim Kaynağı (Model)	Kareler Toplamı:KT (Sum of Squares)	Serbestlik derecesi:Sd (df)	Kareler Ortalaması.KO (Mean Square)	F testi
Regresyon (Regression)	RKT=b_1ÇT$_{xy}$	1	$\text{RAKO} = \frac{\text{RKT}}{1}$	$\text{F} = \frac{\text{RAKO}}{\text{AKO}}$
Artık (Residuals)	RAKT=KTy-RKT	(n-2)	$\text{AKO} = \frac{\text{AKT}}{\text{n} - 2}$	
Toplam (Total)	KTy	(n-1)	-	

BELİRTME (DETERMİNASYON) KATSAYISI

Bağımlı değişken içindeki değişimin, bağımsız değişken tarafından açıklama yüzdesidir. Örneğin başarıyı, çalışma saati ne kadar açıklıyor sorusuna belirtme katsayısı cevap verir.

Belirtme katsayısının karekökü Pearson'ın korelasyon katsayısını verir.

$$R^2 = \frac{RKT}{KT_y}$$

$$\sqrt{R^2} = r \Rightarrow \text{korelasyon katsayısı}$$

TAHMİN (PREDICTION)

Tahmin edilen model üzerinden $\hat{y}_i=b_0+b_1x_i$, X yerine bir değer konularak Y değerinin tahmini elde edilir.

Örneğin, haftalık çalışma saati ile matematik dersinden aldığı puan için kurulmuş basit doğrusal regresyon denklemi

$$\hat{y}_i=55,06+0,71saat$$

olsun.

O zaman haftada ders çalışma saati 10 saat olan bir öğrencinin puanının

55,06+0,71 x 10=62,16

notunun 62,16 olması beklenir.

Yağış (X)	Verim (Y)
52	146
43	128
72	160
35	122
60	150
45	130
56	152
49	147
40	118
65	152

ÖRNEK : Bir bölgede m^2 başına düşen yağış miktarı ile o bölgede yetiştirilen bir bitkiden elde edilen verimler 10 yıl boyunca gözlenmiştir.

a) Yağış ile verim arasındaki basit doğrusal regresyon denklemini kurunuz.

b) Katsayıların ve modelin önem kontrolünü %5 anlamlılık seviyesinde yapınız.

c) Katsayıların %95 güven aralıklarını hesaplayınız.

d) Eğer m^2 başına düşen yağış miktarı 80 kg. olursa verim ne kadar olacaktır?

a) Yağış ile verim arasındaki basit doğrusal regresyon denklemini kurunuz.

$ÇT_{xy}$=1412,6 KT_y=1922,5 KT_x=1220,1

$\bar{x} = 51,7$ $\bar{y} = 140,5$

$$b_1 = \frac{1412,6}{1220,1} = 1,1577$$

$b_0 = \bar{y} - b_1\bar{x}$ =140.5-1,1577(51,7)=80,647

$\Longrightarrow$ $\hat{y}_i$=80,647+1,1577x_i

b_0: Hiç yağış yokken alınacak ortalama verim 80,647'dir.

b_1: Yağış miktarındaki bir birimlik değişiklik verim üzerinde ortalama 1,577 birimlik bir artış yaratacaktır.

b) Katsayıların ve modelin önem kontrolünü %5 anlamlılık seviyesinde yapınız.

H_0: Model anlamlı değildir.

$RKT=b_1ÇT_{xy}$ =1,1577x 1412,6 =1635,367

RAKO=RKT/1=1635,367

AKT=KTy-RKT =1922,5 - 1635,367=287,133

AKO=AKT/8=287,133/8=35,89

F=1635,367/35,89=45,57

$F_{Tablo(0,05,1,8)}$= 5,32

Fhesap > Ftablo olduğundan model anlamlıdır.

Belirtme katsayısı,

$$R^2 = \frac{1635,367}{1922,5} = 0,8506$$

elde edilir yani verim içindeki değişimin %85'ini, yağış ile açıklamaktadır. R^2'nin karekökü korelasyon katsayısını vermektedir.

$$\sqrt{R^2} = r = \sqrt{0,8506} = 0,9222$$

$$r = \frac{ÇT_{XY}}{\sqrt{KT_X KT_Y}} = \frac{1412,6}{\sqrt{1220,1x1922,5}} = 0,92$$

Yağış ile verim arasında %92'lik pozitif yönlü bir ilişki vardır.

$H_0: \beta_0 = 0$
$H_S: \beta_0 \neq 0$

b_0'nın standart hatası,

$$S_{b_0} = \sqrt{35{,}89\left(\frac{1}{10} + \frac{51{,}7^2}{1220{,}1}\right)} = 9{,}0694$$

$$t = \frac{80{,}674}{9{,}0694} = 8{,}895$$

$|t_{hesap}| \geq t_{tablo(0{,}05/2,\ 8)}$=2,306 olduğundan H_0 reddedilir. Yani β_0 katsayısı anlamlıdır.

$H_0: \beta_1 = 0$
$H_s: \beta_1 \neq 0$

b_1'in standart hatası,

$$S_{b_1} = \sqrt{\frac{35,89}{1220,1}} = 0,1714$$

$$t = \frac{b_1}{S_{b_1}} = \frac{1,1577}{0,1714} = 6,7544$$

$|t_{hesap}| \geq t_{tablo(0,05/2,\ 8)} = 2,306$

olduğundan H_0 reddedilir. Yani β_1 katsayısı anlamlıdır.

c) Katsayıların %95 güven aralıklarını hesaplayınız.

β_0 için %95 güven aralığı :

$80{,}647-2{,}306(9{,}0694) < \beta_0 < 80{,}647 + 2{,}306(9{,}0694)$

$P(59{,}733 < \beta_0 < 101{,}561)=0{,}95$

güven aralığı «0» ı içermediği için β_0 anlamlıdır.

β_1 için %95 güven aralığı :

$1{,}1577-2{,}306(0{,}1714) < \beta_1 < 1{,}1577+2{,}306(0{,}1714)$

$P(0{,}762<\beta_1<1{,}553)=0{,}95$

güven aralığı «0» ı içermediği için β_1 anlamlıdır.

d) Yağış eğer 80 kg. olursa verim ne kadar olacaktır?

$$\hat{y}_i = 80{,}647 + 1{,}1577 x_i$$

Denklemde X=80 yerine konulursa,

80,647+1,1577 (80) = 173,263 elde edilir.

Yağış 80kg. olursa ortalama 173.263 kg verim beklenir.

Soru: Bağımsız değişken sayısı birden fazla olamaz mı?

Yanıt: Bağımsız değişken sayısının birden çok olduğu durumda «Çoklu Regresyon Çözümlemesi» geçerlidir.

ÇALIŞMA SORUSU

Bir şirketin 12 aylık reklam geliri ile ürün satış miktarları arasındaki basit doğrusal regresyon denklemini kurarak modelin geçerliliğini %10 anlamlılık düzeyinde test ediniz.

Aylar	Reklam Harcamaları (000 USD)	Satışlar (000 USD)
Ocak	100	5,5
Şubat	110	5,8
Mart	112	6
Nisan	115	5,9
Mayıs	117	6,2
Haziran	116	6,3
Temmuz	118	6,5
Ağustos	120	6,6
Eylül	121	6,4
Ekim	120	6,5
Kasım	117	6,7
Aralık	123	6,8

PYTHON
ÇALIŞMA SORUSU

Bir şirkette çalışanların işte kaç yıldır çalıştıkları ve yıllık maaşları verilmiştir.

Python Uygulaması

Öncelikle aşağıdaki tablodaki gibi bir csv dosyası oluşturalım.,

TecrubeYili	Maas
1.1	39343
1.3	46205
1.5	37731
2	43525
2.2	39891
2.9	56642
3	60150
3.2	54445
3.2	64445
3.7	57189
3.9	63218
4	55794
4	56957
4.1	57081
4.5	61111
4.9	67938
5.1	66029
5.3	83088
5.9	81363
6	93940
6.8	91738
7.1	98273
7.9	101302
8.2	113812
8.7	109431
9	105582
9.5	116969
9.6	112635
10.3	122391
10.5	121872

http://farukciftler.com/?p=378

```
#Kütüphaneleri ekleme
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd

#Verisetini ekleme
dataset = pd.read_csv('maasverisi.csv')
X = dataset.iloc[:, :-1].values
y = dataset.iloc[:, 1].values

# Veri setinin bağımsız değişken obje dizisini(yani X) ve bağımlı değişken
obje dizisini (yani y) ayrı ayrı eğitim seti ve test seti olarak ikiye bölme
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 1/3,
random_state = 0)
```

```
#Lineer regresyon modelini eğitim setine göre eğitme
from sklearn.linear_model import LinearRegression
regressor = LinearRegression()
regressor.fit(X_train,y_train)

#Test setine göre tahminde bulunma
y_pred = regressor.predict(X_test)
#Test sonuçlarını görselleştirme
plt.scatter(X_test, y_test, color ='blue')
plt.plot(X_train, regressor.predict(X_train), color='orange')
plt.title('Maaş ve Tecrübe')
plt.xlabel('Tecrübe Yılı')
plt.ylabel('Maaş')
plt.show()
```

KAYNAKLAR

H.Demirhan, C.Hamurkaroğlu, “İstatistiksel Yöntemlere Giriş”, H.Ü.Yayınları, 2011.

S. Erbaş, Olasılık ve İstatistik, Gazi Kitapevi, 2019.