Basit Doğrusal Regresyon Analizi

İstatistikte değişkenler arasındaki ilişkinin derecesini gösteren katsayıya korelasyon, değişkenler arasındaki ilişkinin fonksiyonel şeklini belirleyen denkleme ise regresyon denklemi denir.

Regresyon analizi yapılırken kurulan matematiksel modelde bir bağımlı değişken ile bir veya birden çok bağımsız değişken bulunur.

Bağımsız değişken sayısı bir tane ise basit doğrusal regresyon modeli, birden fazla bağımsız değişken olması durumunda çoklu regresyon modeli kurulmuş olur.

Y : Bağımlı değişken

X_i : Bağımsız değişkenler ; $\mathrm{i} = 1, 2, \ldots, \mathrm{k}$

$\varepsilon : Hata\ terimi \quad , \quad \varepsilon \sim N(0, \sigma^2)$

$\beta_0, \beta_1, \ldots, \beta_k : Parametreler$

olmak üzere,

Basit doğrusal regresyon modeli; $\mathrm{Y} = \beta_0 + \beta_1 X + \varepsilon$

Çoklu doğrusal regresyon modeli; $\mathrm{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon$

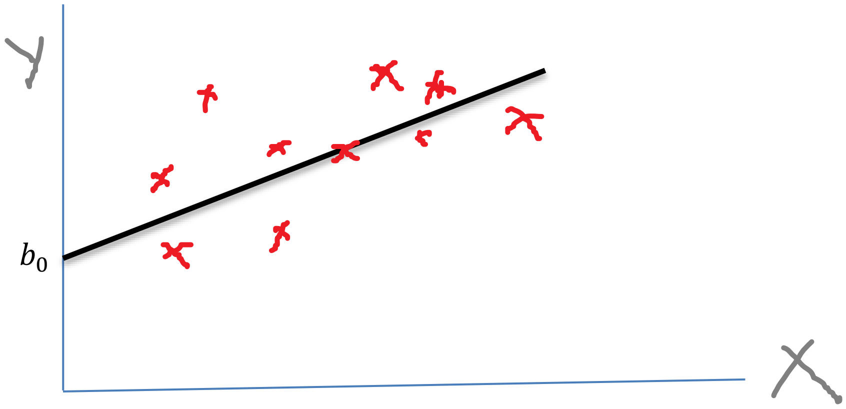
biçimindedir.

Regresyon modelinde X_i 'lerin hatasız ölçüldüğü, Y 'nin ise belli bir hata (ε) miktarı ile ölçüldüğü varsayılır.

n çaplı tesadüfi bir örneklem için regresyon denklemi

$$Y_i = b_0 + b_1 X_i + e \qquad \equiv \qquad Y_i = \widehat{\beta_0} + \widehat{\beta_1} X_i + \varepsilon$$

biçimindedir.



b_0 : $Doğrunun\ Y\ eksenini\ kestiği\ nokta$

b_1 : $Eğim$

anlamındadır.

Basit Doğrusal Regresyon modelinin parametreleri olan β_0 ve β_1 bilinmeyenlerdir. Gözlenen veri yardımıyla bunların tahmin edilmesi gerekir.

Parametre tahmini en küçük kareler metodu kullanılarak yapılır.

β_0 ve β_1 paremetrelerinin tahminleri olan b_0 ve b_1 'in son hesaplama formülleri aşağıda verilmiştir:

$$b_1 = \frac{\sum XY - \frac{(\sum X)(\sum Y)}{n}}{\sum X^2 - \frac{(\sum X)^2}{n}}$$

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

Basit Doğrusal Regresyon Modeli Varsayımları

1. Bağımsız değişkenin değerleri sabit kabul edilir. Bağımlı değişkenin değerleri ise rasgeledir.

2. Değişkenler hatasız ölçülmüştür.

3. Her X_i değeri için;

Y_i değerleri birbirinden bağımsızdır,

Y_i gözlemlerinin tüm dağılımları normaldir,

Y_i gözlemlerinin tüm dağılımları aynı varyansa sahiptir.

4. Bağımlı ve bağımsız değişkenler arasındaki ilişki doğrusaldır.

Hata Terimi (ε_i) Varsayımları

1. Rassal hataların beklenen değeri sıfırdır. $E(\varepsilon)=0$
2. ε'ların olasılık dağılımının varyansı sabittir.
3. Hata değerleri birbirinden bağımsızdır.
4. Rassal hataların dağılımı normaldir.
5. Hatalar ile bağımlı değişken arasında korelasyon yoktur.

   $Cov(\varepsilon,Y)=0$

6. Hatalar ile bağımsız değişkenler birbirinden bağımsızdır.

   $Cov(\varepsilon,X_i)=0$

Determinasyon (Belirleme, Belirlilik) katsayısı (R^2)

Bağımlı değişkendeki değişimin ne kadarının (yüzde kaçının) bağımsız değişken tarafından açıklanabildiğini belirtir ve

$$R^2 = (r_{XY})^2$$

formülü ile hesaplanır.

$0 \leq R^2 \leq 1$ dir.

Ayrıca, Regresyon Modelinin performansı R^2 ile ölçülür.

R^2 1'e ne kadar yakınsa regresyon denklemi o kadar anlamlıdır, belirleyicidir.

Örnek. Bir ürüne ait reklam harcamaları ve yapılan satış miktarları aşağıdaki gibi bulunmuştur.

Reklam harcama (bin TL):	10	20	30	40	50
Satış (bin adet) :	3	4	6	7	10

a) Veri seti için matematiksel modeli tahmin ediniz.

b) Belirlilik katsayısını bulup yorumlayınız.

c) 60 (bin TL) reklam harcaması yapılırsa tahmini satış miktarını tahmin ediniz.

Çözüm:

a) Reklam harcaması bağımsız(X), satış miktarı bağımlı(Y) değişkendir.

$n = 5$, $\sum X = 150$, $\sum Y = 30$, $\sum XY = 1070$, $\sum X^2 = 5500$, $\bar{x} = 30$, $\bar{y} = 6$

$$b_1 = \frac{\sum XY - \frac{(\sum X)(\sum Y)}{n}}{\sum X^2 - \frac{(\sum X)^2}{n}} = \frac{1070 - \frac{(150)(30)}{5}}{5500 - \frac{(150)^2}{5}} = \frac{1070 - 900}{5500 - 4500} = \frac{170}{1000} = 0{,}17$$

$$b_0 = \bar{y} - b_1\bar{x} = 6 - (0{,}17)30 = 0{,}9$$

olup

$$\hat{Y} = 0{,}9 + 0{,}17X$$

elde edilir.

b) $\mathrm{R}^2 = (\mathrm{r_{XY}})^2$

$$r_{XY} = \frac{\sum XY - \frac{(\sum X)(\sum Y)}{n}}{\sqrt{(\sum X^2 - \frac{(\sum X)^2}{n})(\sum Y^2 - \frac{(\sum Y)^2}{n})}} = \frac{1070 - \frac{(150)(30)}{5}}{\sqrt{(5500 - \frac{(150)^2}{5})(210 - \frac{(30)^2}{5})}}$$

$$= \frac{1070 - 900}{\sqrt{(5500 - 4500)(210 - 180)}} = \frac{170}{\sqrt{(1000)(30)}} = \frac{170}{173{,}205} = 0{,}9814$$

$\mathrm{R}^2 = (0{,}9814)^2 = 0{,}9633$

c) $\hat{Y} = 0{,}9 + 0{,}17X$

X=60 ise $\hat{Y} = 0{,}9 + 0{,}17(60) = 11{,}1$ bulunur.

Örnek:

Dekara atılan gübre miktarı	Dekar başına verim			
8	15			
12	18			
14	25			
16	35			
22	48			
28	55			
35	60			
41	64			

a) Serpme diyagramını çiziniz.
b) Regresyon denklemini tahmin ediniz.
c) 50 birim gübre atıldığında tahmini olarak ne kadar verim elde edilir?
d) 70 birim verim alabilmek için ne kadar gübre atılmalıdır?
e) Verimdeki değişimin % kaçı gübre miktarı ile açıklanabilmektedir?

Çözüm.

a)

b)

$\hat{Y} = 5{,}1084 + 1{,}58628X$

c) 84,4158
d) 40,9121

e)

$r = 0,96047$

$R^2 = (0,96047)^2 = 0,9225$

Varyans Analizi Tablosu (ANOVA)

Değişim Kaynağı	Serbestlik derecesi	Kareler Toplamı	Kareler Ortalaması	F değeri
Regresyon	m-1	$RKT = \sum(\widehat{Y_i} - \bar{y})^2$	$RKO = RKT/sd$	$F = \frac{RKO}{AKO}$
Artık	n-m	$AKT = \sum(Y_i - \widehat{Y_i})^2$	$AKO = AKT/sd$	
Toplam	n-1	$TKT = \sum(Y_i - \bar{y})^2$		

$TKT = RKT + AKT$

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

$n : Gözlem\ sayısı$

$m : Regresyon\ modelinde\ tahmin\ edilen\ parametre\ sayısı$

Örnek: Bir regresyon probleminde aşağıdaki veri toplanmıştır.

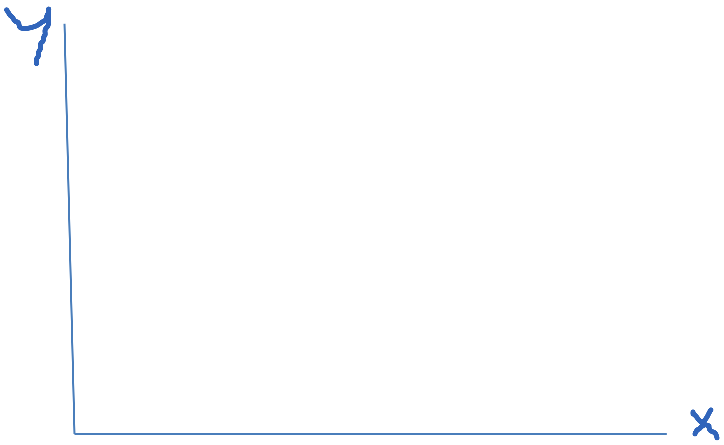
$X_i : 1, 3, 5, 7, 9, 11, 13$

Y_i : $20, 16, 14, 13, 12, 10, 6$

a) Bu veriye uygun olabilecek regresyon modelini yazınız.
b) Yazılan modelin parametrelerini tahmin ediniz.
c) Tahmin denklemini yazınız.
d) Varyans analizi tablosunu oluşturunuz.
e) Bağımlı değişkendeki değişimin ne kadarı bağımsız değişken tarafından açıklanmaktadır?

Çözüm.

a) $\mathrm{Y} = \beta_0 + \beta_1 X + \varepsilon$



b)

$n = 7$, $\sum X = 49$, $\sum Y = 91$, $\sum XY = 525$, $\sum X^2 = 455$, $\bar{x} = 7$, $\bar{y} = 13$

$\sum Y^2 = 1301$

$$b_1 = \frac{\sum XY - \frac{(\sum X)(\sum Y)}{n}}{\sum X^2 - \frac{(\sum X)^2}{n}} = -1$$

$$b_0 = \bar{y} - b_1\bar{x} = 20$$

c)

$$\hat{Y} = 20 - 1X$$

d) **Varyans Analizi Tablosu (ANOVA)**

Değişim Kaynağı	Serbestlik derecesi	Kareler Toplamı	Kareler Ortalaması	F değeri
Regresyon	m-1	$RKT = \sum(\hat{Y}_i - \bar{y})^2$	$RKO = RKT/sd$	$F = \frac{RKO}{AKO}$
Artık	n-m	$AKT = \sum(Y_i - \hat{Y}_i)^2$	$AKO = AKT/sd$	
Toplam	n-1	$TKT = \sum(Y_i - \bar{y})^2$		

X_i	Y_i	$\hat{Y}_i$	$Y_i - \bar{y}$	$(Y_i - \bar{y})^2$	$Y_i - \hat{Y}_i$	$(Y_i - \hat{Y}_i)^2$	$\hat{Y}_i - \bar{y}$	$(\hat{Y}_i - \bar{y})^2$
1	20							
3	16							
5	14							
7	13							
9	12							
11	10							
13	6							
				118		6		112

Varyans Analizi Tablosu (ANOVA)

Değişim Kaynağı	Serbestlik derecesi	Kareler Toplamı	Kareler Ortalaması	F değeri
Regresyon	1	112	112	$F = 93{,}3$
Artık	5	6	1,2	
Toplam	6	118		

e)

$$R^2 = \frac{RKT}{TKT} = \frac{112}{118} = 0{,}949$$