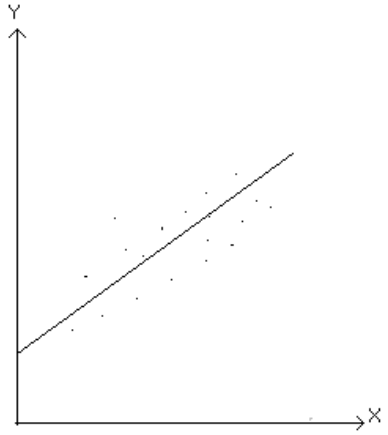
BÖLÜM 6. REGRESYON ve KORELASYON ANALİZİ

İlgilenilen olayı tanımlayan rasgele değişken bağımlı değişken, bu olayla ilgili ya da olayı etkileyen değişken ise bağımsız değişken olarak tanımlandığında bağımlı değişken ile bir ya da birkaç bağımsız değişken arasında kurulan modeldeki parametreleri tahmin ederek bağımsız değişkenlerin belirlenen değerleri için, bağımlı değişkenin alacağı değeri belirleme problemine regresyon problemi denir.

Burada, bağımlı değişken Y ile gösterilir ve açıklanan değişken olarak tanımlanır. Bağımsız değişken X ile gösterilir ve açıklayıcı değişken olarak tanımlanır.

6.1. Basit Doğrusal Regresyon

n tane denek üzerinden alınan (x, y) verileri kullanılarak serpilme diyagramı çizilebilir.



(x_1, y_1), (x_2, y_2), … ,(x_n, y_n)

Serpilme diyagramı veriye yaklaşık nasıl bir eğrinin uyduğunu gösterir. Böyle bir serpilme diyagramı için aralarında doğrusal bir ilişki vardır denir.

X ile Y arasındaki gerçek bağıntı:

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

doğru denklemi ile gösterilir. Burada amaç β_0 ve β_1 parametrelerini tahmin etmektir.

β_0: doğrunun y eksenini kestiği nokta

β_1: doğrunun eğimi

} β_0 ve β_1 bilinmeyen regresyon katsayılarıdır

ε : gerçek hata (bağımlı değişkenin gerçek değeri-gözlenen değeri)

Kitleden seçilen n birimlik örneklem için doğrusal regresyon denklemi;

$\hat{y} = b_0 + b_1 x_j + e_j \quad , \quad j = 1,2, \dots, n$

biçiminde tanımlanır. Bilinen (verilen) bir x_j değeri için y_j değeri kestirilir(öngörülür).

Tahmini doğrusal regresyon denklemi;

$\hat{y} = b_0 + b_1 x_j \quad , \quad j = 1,2, \dots, n$

ile gösterilir. Genel olarak

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

biçiminde yazılır ve bu denkleme X üzerinde Y' nin regresyonu denir.

y_j: j. Gözleme ilişkin gerçek y değeri

$\hat{y}_j$: j. Gözleme ilişkin y_j'nin tahmin değeri

x_j: j. Gözleme ilişkin bağımsız değişkenin aldığı değer

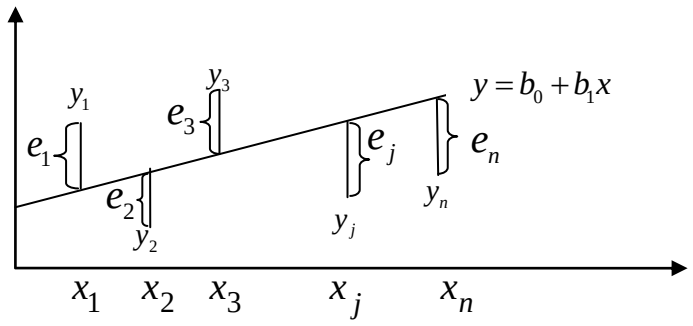
b_0: β_0'ın tahmini

b_1: β_1'in tahmini

e_j: j. Gözlemin hata terimidir.(gözlenen değer ile tahmini değer arasındaki fark)

$e_j = y_j - \hat{y}_j \quad e \sim N(0, \sigma^2)$

6.2. β_0 ve β_1 Parametreleri için En Küçük Kareler Tahmin Edicileri



$\min \sum_{j=1}^{n} e_j^2 \longrightarrow$ Hata kareleri toplamının minimum olması istenir

$\sum_{j=1}^{n} e_j^2 = \sum_{j=1}^{n} \left(y_j - \hat{y}_j^2\right)^2 = \sum_{j=1}^{n} (y_j - b_0 - b_1 x_j)^2 \rightarrow$ Bu ifadeyi minimum yapacak b_0 ve b_1 değerleri bulunmalıdır.

$\min \sum_{j=1}^{n} e_j^2$

$$\frac{\partial \sum e_j^2}{\partial b_0} = -2 \sum_{j=1}^{n} \left(y_j - b_0 - b_1 x_j\right) = 0$$

$$\sum_{j=1}^{n} y_j = n b_0 + b_1 \sum_{j=1}^{n} x_j \qquad (1)$$

$$\frac{\partial \sum e_j^2}{\partial b_1} = -2 \sum x_j \left(y_j - b_0 - b_1 x_j\right) = 0$$

$$\sum_{j=1}^{n} x_j y_j = b_0 \sum_{j=1}^{n} x_j + b_1 \sum_{j=1}^{n} x_j^2 \qquad (2)$$

(1) denklemi $\sum x_j$, (2) denklemi n ile çarpılıp toplanırsa;

(1) $\quad -\sum x_j \sum y_j = -n b_0 \sum x_j - b_1 (\sum x_j)^2$

(2) $\quad n \sum x_j y_j = n b_0 \sum x_j + n b_1 \sum x_j^2$

$$n \sum x_j y_j - \sum x_j \sum y_j = n b_1 \sum x_j^2 - b_1 (\sum x_j)^2$$

$$= b_1 (n \sum x_j^2 - (\sum x_j)^2)$$

$$\Longrightarrow \quad b_1 = \frac{n \sum x_j y_j - \sum x_j \sum y_j}{n \sum x_j^2 - (\sum x_j)^2}$$

$$= \frac{n(\sum x_j y_j - \frac{\sum x_j \sum y_j}{n})}{n(\sum x_j^2 - \frac{(\sum x_j)^2}{n})}$$

$$= \frac{\sum x_j y_j - \frac{\sum x_j \sum y_j}{n}}{\sum x_j^2 - \frac{(\sum x_j)^2}{n}}$$

$$= \frac{\text{XYOAÇT}}{\text{XOAKT}}$$

(1) denklemi ($\sum y_j - b_1 \sum x_j = nb_0$) n ile bölünürse;

$$b_0 = \frac{\sum y_j}{n} - \frac{b_1 \sum x_j}{n} = \bar{y} - b_1 \bar{x}$$

olmak üzere,

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

olarak bulunur. Tahmini regresyon denklemi,

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

olmak üzere,

$b_1 > 0$ → iki değişken birlikte artıyor ya da azalıyor

$b_1 < 0$ → değişkenlerden biri artıyor diğeri azalıyor

6.3. Açıklanan ve Açıklanamayan Değişim

1) Örneklem ortalaması etrafındaki değerlerin değişimi

$\sum_{j=1}^{n}(y_j - \bar{y})^2$;toplam değişim ya da genel kareler toplamı (YOAKT)

2) Regresyon doğrusu etrafındaki değerlerin değişimi

$\sum(y_j - \hat{y}_j)^2$;açıklanamayan değişim ya da hata kareler toplamı (RAKT)

3) Ortalama etrafındaki tahmini değerlerin değişimi

$\sum(\hat{y}_j - \bar{y})^2$;açıklanan değişim ya da regresyon kareler toplamı (RKT)

$$\sum_{j=1}^{n}(y_j - \bar{y})^2 = \sum(y_j - \hat{y}_j)^2 + \sum(\hat{y}_j - \bar{y})^2$$

Toplam değişim=Açıklanamayan Değişim+Açıklanan Değişim

$$YOAKT = RAKT + RKT$$

$$YOAKT = \sum_{j=1}^{n} y_j^2 - (\frac{\sum y_j^2}{n})$$

$$RKT = \frac{[\sum x_j y_j - \frac{\sum x_j \sum y_j}{n}]^2}{\sum x_j - \frac{(\sum x_j)^2}{n}} = \frac{XYOAÇT^2}{XOAKT} = b_1 XYOAÇT$$

$$RAKT = YOAKT - RKT$$

$R_{sd} = 2 - 1 = 1$ (Regresyon serbestlik derecesi)

$\text{YOA}_{\text{sd}} = \text{n} - 1$ (Y ortalamadan ayrılış serbestlik derecesi)

$\text{RA}_{\text{sd}} = \text{YOA}_{\text{sd}} - \text{R}_{\text{sd}} = \text{n} - 1 - 1 = \text{n} - 2$ (Regresyondan ayrılış serbestlik derecesi)

Her bir kare toplamı kendi serbestlik derecesine bölündüğünde kare ortalamaları bulunur.

$$\text{RKO} = \frac{\text{RKT}}{\text{R}_{\text{sd}}} = \frac{\text{RKT}}{1} = \text{RKT}$$

$$\text{RAKO} = \frac{\text{RAKT}}{\text{RA}_{\text{sd}}} = \frac{\text{RAKT}}{\text{n} - 2} = \text{s}^2 = \hat{\sigma}^2 \rightarrow$$ x üzerinde Y′ninregresyon doğrusunun varyansının yansız tahmin edicisi

6.4. Belirtme Katsayısı

Belirtme katsayısı, bağımsız değişkenin bağımlı değişkendeki değişimin yüzde kaçını açıkladığını gösterir. Açıklanan değişimin toplam değişime oranıdır ve R^2 ile gösterilir.

$$\text{R}^2 = \frac{\text{Açıklanan değişim}}{\text{Toplam değişim}} = \frac{\text{RKT}}{\text{YOAKT}}$$

$1 - \text{R}^2$; toplam değişimin açıklanamayan yüzdesi

$$1 - \text{R}^2 = \frac{\text{RAKT}}{\text{YOAKT}}$$

Örnek 6.1. 12 kadına ilişkin sistolik kan basıncı (Y) ve yaşları (X) aşağıdaki tabloda verilmiştir.

a) Regresyon doğrusunu tahmin ediniz.
b) 45 yaşında olan bir kadının sistolik kan basıncını tahmin ediniz
c) Bağımsız değişkenin bağımlı değişkende değişimin yüzde kaçını açıkladığı bulunuz.

Yaş(X)	56	42	72	36	63	47	55	49	38	42	68	60
Kan Basıncı (Y)	147	125	160	118	149	128	150	145	115	140	152	155

Burada;

$$\sum \text{x}_\text{j} = 628 \, , \sum \text{x}_\text{j}^2 = 34416 \, , \sum \text{y}_\text{j} = 1684 \, , \sum \text{y}_\text{j}^2 = 238822 \, , \sum \text{x}_\text{j}\text{y}_\text{j} = 89894$$

a) Tahmini regresyon denklemi:

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

olmak üzere b_0 ve b_1' in hesaplanması gerekir.

$$b_1 = \frac{XYOAÇT}{XOAKT} = \frac{\sum x_j y_j - \frac{\sum x_j \sum y_j}{n}}{\sum x_j^2 - \frac{(\sum x_j)^2}{n}} = \frac{89894 - \frac{(628)(1684)}{12}}{34416 - \frac{(628)^2}{12}} = \frac{1764.67}{1550.67} = 1.138$$

$$\bar{x} = \frac{628}{12} = 52.33 \qquad \bar{y} = \frac{1684}{12} = 140.33$$

$$b_0 = \bar{y} - b_1 \bar{x} = 140.33 - (1.138)(52.33) = 80.778$$

Tahmini regresyon denklemi:

$$\hat{y} = 80.778 + 1.138x$$

olarak bulunur.

b) 45 yaşında olan bir kadının sistolik kan basıncı,

$$\hat{y} = 80.778 + 1.138(45) = 131.988$$

olarak tahmin edilir.

c) Bağımsız değişkenin bağımlı değişkende değişimin yüzde kaçını açıkladığı R^2 ile hesaplanır.

$$R^2 = \frac{RKT}{YOAKT} = \frac{b_1 XYOAÇT}{YOAKT} = \frac{(1.138)(1764.67)}{238822 - \frac{(1684)^2}{12}} = \frac{2008.194}{2500.667} = 0.80 \rightarrow \%80$$