# BASİT REGRESYON VE KORELASYON ANALİZİ

İstatistiksel araştırmalarda iki yada daha çok değişken arasındaki ilişkinin incelenmesi için en çok kullanılan yöntemlerden birisi regresyon analizidir. Değişkenler arasındaki ilişki matematiksel bir modelle açıklanabileceği gibi, ilişkinin derecesi ve yönü bir bir katsayı ile de ortaya koyulabilir. Bu da korelasyon analizi ile sağlanabilir.

Değişkenler arasındaki ilişkilere bazı örnekler vermek gerekirse;

-İnsanların boyları ile kiloları

-Futbol takımlarının çalışma süreleri ve maç skorları toplamları

-Öğrencilerin çalışma miktarları ve sınav notları

-Bir malın fiyatı ve talep miktarı

-Bir ürünün verimi ve verilen gübre miktarı, vb.

Değişkenler arasındaki ilişkiler aşağıdaki gibi sınıflandırılabilir:

i) Belirleyici (deterministik) ilişkiler

ii) Yarı belirleyici ilişkiler

iii) Deneysel (ampirik) ilişkiler

Yarı belirleyici ve deneysel ilişkilerin incelenmesi regresyon analizinin kapsamına girmektedir.

Regresyon analizinde değişkenler iki grup altında incelenir:

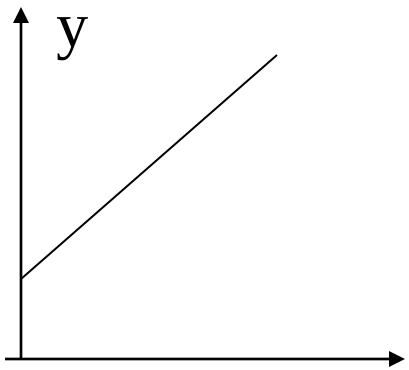
- Bağımsız değişkenler (açıklayıcı değişkenler)
- Bağımlı değişkenler

Bizim kontrol edebildiğimiz yada edemediğimiz bağımsız değişkenlerde meydana gelen değişiklikler, bağımlı değişkenlere etki ederek onların değer değiştirmesine neden olurlar. Örneğin kişilerin gelirlerinin değişmesi, harcama miktarlarının da değişmesine neden olur. Bu durumda gelir bağımsız değişken, harcama miktarı ise bağımlı değişkendir.

Regresyon analizinde genellikle bağımsız değişkenler (X) , bağımlı değişkenler (Y) ile gösterilirler.

Basit doğrusal regresyondaki basit kelimesi iki değişken arasındaki ilişkiyi açıklamak için kullanılmasından, doğrusal kelimesi ise kurulan modelin parametreleri açısından doğrusal bir model olmasındandır.

İki değişken arasındaki en basit ilişki, bir doğru ile açıklanabilen ilişkidir.




Genel olarak bir doğrunun matematik gösterimi:

$Y=\beta_0+\beta_1 X$ şeklindedir. Burada $\beta_1$ ,

eğimdir ve X'teki 1 birimlik değişmenin Y'de yaptığı değişikliği gösterir.

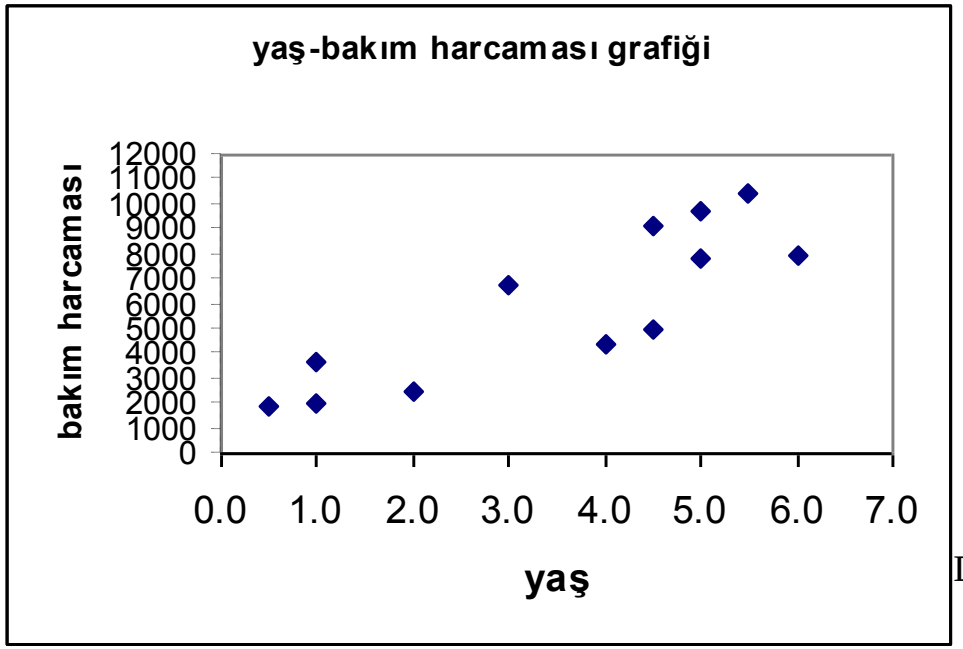
$\beta_0$ ise X'in değeri 0 olduğunda Y'nin almış olduğu değerdir ve Y ekseninin kesme noktası olarak isimlendirilir.

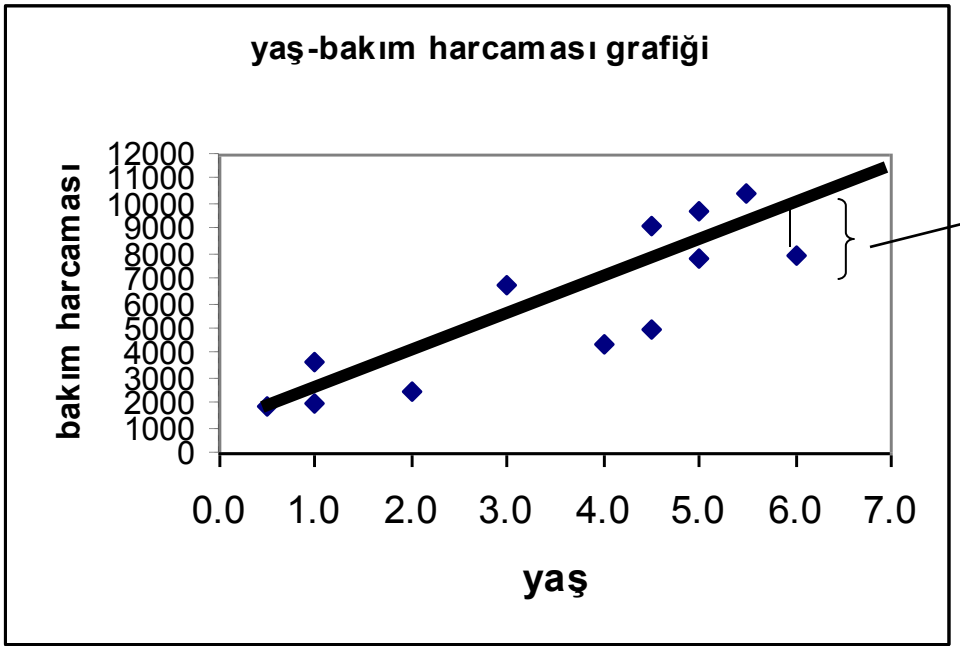
Bir fabrikada taşıma işleri için kullanılan tırların yaşı ile bakım harcamaları arasındaki ilişkiyi ele alalım. Verilerin grafiği çizildiğinde tam olarak düz bir doğrunun üzerinde olmadıkları, fakat tırlar eskidikçe bakım harcamalarının da arttığı görülmektedir. Burada bağımsız değişken yaş, bağımlı değişken ise bakım harcamalarıdır, çünkü yaş değiştikçe bakım harcamaları değişiklik göstermektedir. Pratiklik olması açısından yaş ve bakım harcaması arasındaki ilişkinin bir doğru şeklinde olduğunu varsayarsak, bu modelin matematik gösterimi:

Bakım harcaması ← $Y = \beta_0 + \beta_1 X + e$ → Hata terimi

yaş

| yaş (yıl) | bakım harcaması |
|---|---|
| 2.0 | 2500 |
| 4.5 | 9200 |
| 4.5 | 4950 |
| 4.0 | 4400 |
| 5.0 | 7900 |
| 5.5 | 10500 |
| 5.0 | 9700 |
| 0.5 | 1950 |
| 6.0 | 8000 |
| 1.0 | 2025 |
| 1.0 | 3700 |
| 3.0 | 6800 |






e hata terimi, traktörler için yapılan harcamanın, ilişkiyi açıklayan doğrudan ne kadar saptığını gösterir.

Tırların yaşı ile yapılan bakım harcamaları arasındaki gerçek ilişkiyi belirleyen model henüz belirlenmiş değildir. Bunun için modelde bulunan parametrelerin ($\beta_0$ ve $\beta_1$) bilinmesi gerekir.

$\beta_0$ ve $\beta_1$ birer parametre olduklarından, gerçek değerlerinin bulunması için taşıma işinde kullanılan tüm tırların (populasyonun) bakım harcamaları ve yaşlarının bilinmesi gerekmektedir. Bu da çoğu zaman imkansız olduğundan elimizdeki örneği kullanarak parametreleri tahminleriz veya başka bir ifade şekliyle grafikteki noktalara en iyi uyan bir doğruyu buluruz.

# EN KÜÇÜK KARELER (EKK) YÖNTEMİ İLE BİR DOĞRUNUN UYUMU

Gözlemleri en iyi açıklayan doğrunun belirlenmesi için çeşitli yöntemler ileri sürülebilir fakat günümüzde en çok kullanılan yöntem "En Küçük Kareler" adı verilen yöntemdir. Bu yöntem gözlemlerin belirlenen doğrudan uzaklıklarının (hata terimlerinin) karelerinin toplamının en küçük yapılmasına dayanır.

$Y = \beta_0 + \beta_1 X + e$ modelinde hata terimi:

$e = Y - \beta_0 - \beta_1 X$ olarak yazılabilir. Bu ifadenin karesi alınıp tüm gözlemler için toplanırsa:

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(Y - \beta_0 - \beta_1 X\right)^2$$

İfadesi elde edilir. EKK yöntemine göre bu ifadeyi minimize eden $b_0$ ve $b_1$ değerleri $\beta_0$ ve $\beta_1$ 'in tahmincileri olur.

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(Y - \beta_0 - \beta_1 X\right)^2$$

İfadesini minimize eden parametre tahmincilerinin değerlerini bulabilmek için eşitliğin $\beta_0$ ve $\beta_1$ 'e göre türevleri alınıp 0'a eşitlenir.

$\beta_0$'a göre türev alınırsa;

$$\frac{\partial}{\partial \beta_0} \sum_{i=1}^{n} e_i^2 = \frac{\partial}{\partial \beta_0} \sum_{i=1}^{n} \left(Y - \beta_0 - \beta_1 X\right)^2$$

$$= -2 \sum_{i=1}^{n} \left(Y - \beta_0 - \beta_1 X\right)$$

$\beta_1$'e göre türev alınırsa;

$$\frac{\partial}{\partial \beta_1} \sum_{i=1}^{n} e_i^2 = \frac{\partial}{\partial \beta_1} \sum_{i=1}^{n} \left(Y - \beta_0 - \beta_1 X\right)^2$$

$$= -2X \sum_{i=1}^{n} \left(Y - \beta_0 - \beta_1 X\right)$$

Her iki denklemi de 0'a eşitlersek;

$$-2 \sum_{i=1}^{n} \left(Y - b_0 - b_1 X\right) = 0$$

$$\sum_{i=1}^{n} \left(Y - b_0 - b_1 X\right) = 0$$

$$-2. \sum_{i=1}^{n} X.\left(Y - b_0 - b_1 X\right) = 0$$

$$\sum_{i=1}^{n} X.\left(Y - b_0 - b_1 X\right) = 0$$

$$-2\sum_{i=1}^{n}\left(Y-b_0-b_1X\right)=0 \qquad -2.\sum_{i=1}^{n}X.\left(Y-b_0-b_1X\right)=0$$

$$\sum_{i=1}^{n}\left(Y-b_0-b_1X\right)=0 \qquad \sum_{i=1}^{n}X.\left(Y-b_0-b_1X\right)=0$$

Parantezleri açarsak;

$$\sum Y-n.b_0-b_1\sum X=0 \qquad \sum XY-b_0\sum X-b_1\sum X^2=0$$

Bu denklemlere doğrunun NORMAL DENKLEMLERİ denir.
Normal denklemler alt alta yazılıp birlikte çözüldüklerinde $b_0$ ve $b_1$ tahmincileri bulunur.

$$\left.\begin{array}{c}\sum Y=n.b_0+b_1\sum X\\ \sum XY=b_0\sum X+b_1\sum X^2\end{array}\right\} \quad b_1=\frac{\sum XY-\frac{(\sum X).(\sum Y)}{n}}{\sum X^2-\frac{(\sum X)^2}{n}}$$

$$b_0=\overline{Y}-b_1\overline{X}$$

şeklindeki formüller yardımıyla da tahminciler bulunabilir.

Böylece veri noktalarımızdan geçen en iyi doğru denklemi:

$$\hat{Y} = b_0 + b_1 X$$

Gerçek Y'nin tahmincisi

Traktör örneğimiz için gereken hesaplamaları yapıp normal denklemleri oluşturalım:

| | yaş (yıl) (x) | bakım harcaması (y) | x | y | xy |
|---|---|---|---|---|---|
| | 2.0 | 2500 | 4 | 6250000 | 5000 |
| | 4.5 | 9200 | 20.25 | 84640000 | 41400 |
| | 4.5 | 4950 | 20.25 | 24502500 | 22275 |
| | 4.0 | 5500 | 16 | 30250000 | 22000 |
| | 5.0 | 7900 | 25 | 62410000 | 39500 |
| | 5.5 | 10500 | 30.25 | 110250000 | 57750 |
| | 5.0 | 9700 | 25 | 94090000 | 48500 |
| | 0.5 | 1950 | 0.25 | 3802500 | 975 |
| | 6.0 | 8000 | 36 | 64000000 | 48000 |
| | 1.0 | 2025 | 1 | 4100625 | 2025 |
| | 1.0 | 3700 | 1 | 13690000 | 3700 |
| | 3.0 | 6800 | 9 | 46240000 | 20400 |
| toplam | 42.0 | 72725.0 | 188.0 | 544225625.0 | 311525.0 |
| ortalama | 3.5 | 6060.4 | | | |

$$\sum Y = n.b_0 + b_1 \sum X$$

$$\sum XY = b_0 \sum X + b_1 \sum X^2$$

$$72725 = 12b_0 + 42b_1$$

$$311525 = 42b_0 + 188b_1$$

$$35*(72725 = 12b_0 + 42b_1)$$

$$311525 = 42b_0 + 188b_1$$

$$254537.5 = 42b_0 + 147b_1$$

$$- \quad 311525 = 42b_0 + 188b_1$$

$$-56988 = -41b_1$$

**$b_1 = 1390$**

$72725 = 12b_0 + 42b_1$

$72725 = 12b_0 + 42*1390$

**$b_0 = 1195$**

Tahmincileri elde etmek için normal denklemler yerine formüller kullanılırsa da aynı sonuçlar elde edilir.

Doğrunun denklemi:

$$\hat{Y} = 1195 + 1390X$$

Hesaplanan bu denklem kullanılarak yaşını bildiğimiz bir traktör için yapılacak ortalama bakım masrafını tahmin edebiliriz. Örneğin x=4 yaşındaki bir traktör için bakım masrafları:

$$\hat{Y} = 1195 + 1390X$$

$$\hat{Y} = 1195 + (1390)(4) = 6755$$

olarak bulunur.

# REGRESYON DENKLEMİNİN İNCELENMESİ

Regresyon denklemini incelerken genellikle bizi en çok ilgilendiren soru incelediğimiz iki değişken arasında gerçekten bir ilişki olup olmadığı sorusudur. Bu soru aslında basit doğrusal regresyonda $\beta_1$ 'in değerinin 0 olup olmadığının araştırılmasıdır. Bu araştırmayı yaparken istatistiksel testle kullanmak gerektiğinden hata terimi ve parametre tahmincilerinin dağılışları hakkında bazı varsayımlarda bulunmak gerekir.

Hata terimi e'ler, ortalaması 0 ve varyansı olan birbirinden bağımsız normal dağılışlar gösterirler.

$E(e)=0 \quad Var(e)= s^2$

- ***Tahminin Standart Hatası ve Varyansı***

Tahminin standart hatası s, noktaların regresyon doğrusu etrafındaki dağılımlarının ortalama bir ölçüsünü verir.

$$s = \sqrt{\frac{\sum e^2}{n-k}} \qquad s^2 = \frac{\sum e^2}{n-k}$$

# Korelasyon Katsayısı

Korelasyon katsayısı, regresyon modeli ile bulunan tahmini Y değerlerinin, gerçek değerlere uygunluğunu ölçmede kullanılır.

- Korelasyon katsayısı -1 ile 1 arasında değişir.
- Katsayının -1 çıkması, iki değişken arasında ters yönlü tam bir ilişkinin olduğunu, 1 çıkması ise doğru yönlü tam bir ilişkinin olduğunu ifade eder.
- Katsayının -1'e doğru yaklaşması ,değişkenler arasında ters yönlü kuvvetli bir ilişkiyi gösterirken, 1'e yaklaşması değişkenler arasında doğru yönlü kuvvetli bir ilişkiyi ifade eder.
- Korelasyon katsayısının işareti, regresyon doğru veya eğrisine ait eğim katsayısının işaretidir.
- Korelasyon katsayısının karesi, belirleme katsayısını determinasyon katsayısını) verir.

Sınırlı sayıda veri üzerinden hesaplanan korelasyon katsayısı bir istatistiktir ve r ile gösterilir.Bu istatistiğin anakütle parametresi olarak karşılığı $\rho$ 'dur.

Korelasyon katsayısı için genel formül; $r = \pm\sqrt{\frac{\sum(\hat{Y} - \bar{Y})^2}{\sum(Y - \bar{Y})^2}}$

yada $r = \frac{\sum xy}{\sqrt{(\sum x^2)(\sum y^2)}}$

Bu formülde;

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

$$\sum y^2 = \sum Y^2 - \frac{(\sum Y)^2}{n}$$

Bütün bu değerler n katsayısı ile çarpılırsa sonuç değişmez ve korelasyon katsayısı;

$$r = \frac{n\sum XY - (\sum X)(\sum Y)}{\sqrt{\left[n\sum X^2 - (\sum X)^2\right]\left[n\sum Y^2 - (\sum Y)^2\right]}}$$

**ÖRNEK**

Bir süper market yöneticisi tesadüfi olarak seçilen bir saatlik sürelerde kasaya gelen müşteri sayısını ve ödedikleri toplam para miktarını aşağıdaki gibi kaydetmiştir.

| | | | | | |
|---|---|---|---|---|---|
| Müşteri Sayısı | 25 | 20 | 50 | 35 | 40 |
| Ödenen Para (10000 TL) | 12.5 | 10.4 | 25.3 | 20.2 | 24.1 |

Müşteri sayısını bağımsız (X), kasalara ödenen para miktarını bağımlı değişken olarak kabul ederek, doğrusal korelasyon katsayısı;

$$r = \frac{n\sum XY - (\sum X)(\sum Y)}{\sqrt{\left[n\sum X^2 - (\sum X)^2\right]\left[n\sum Y^2 - (\sum Y)^2\right]}}$$

formülü ile kolayca hesaplanabilir.

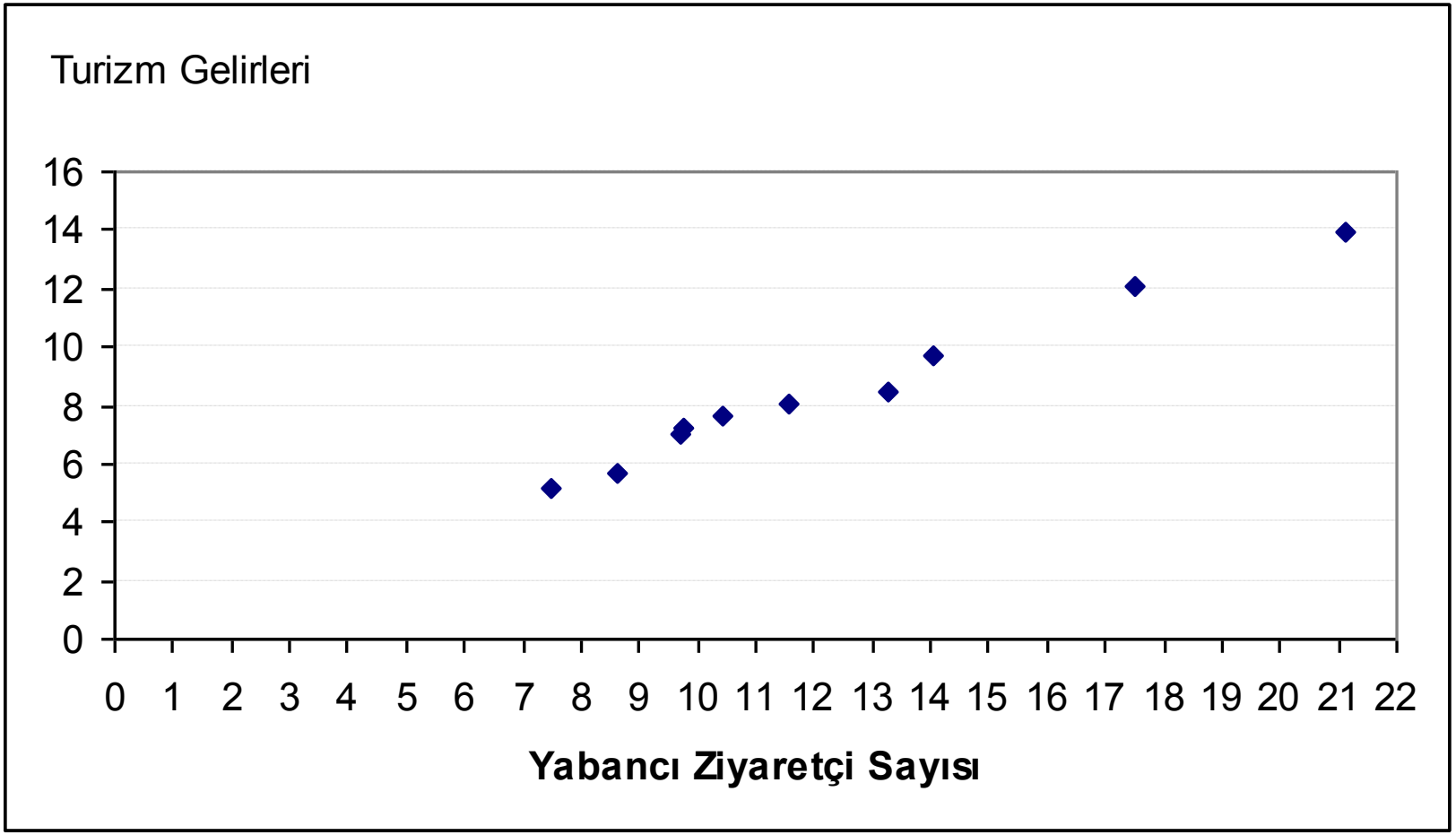
| X | Y | XY | $X^2$ | $Y^2$ |
|---|---|---|---|---|
| 25 | 12.5 | 312.5 | 625 | 156.2 |
| 20 | 10.4 | 208 | 400 | 108.1 |
| 50 | 25.3 | 1265 | 2500 | 640.09 |
| 35 | 20.2 | 707 | 1225 | 408.04 |
| 40 | 24.1 | 964 | 1600 | 580.81 |
| Toplam 170 | 92.5 | 3456.5 | 6350 | 1893.3 |

$$r = \frac{5(3456.5) - 170(92.5)}{\sqrt{\left[5(6350 - 170^2\right]\left[5(1893.3) - (92.5)^2\right]}} = 0.9669$$

***Örnek***:1996-2005 yıllarındaki Türkiye'nin turizm gelirleri ile Türkiye'ye gelen turist sayısı tabloda verilmiştir.

| Yıllar | Turizm Gelirleri | Yabancı Ziyaretçi Sayısı |
|---|---|---|
| 1996 | 5.650 | 8.614 |
| 1997 | 7.008 | 9.689 |
| 1998 | 7.177 | 9.752 |
| 1999 | 5.193 | 7.464 |
| 2000 | 7.636 | 10.412 |
| 2001 | 8.090 | 11.569 |
| 2002 | 8.481 | 13.247 |
| 2003 | 9.677 | 14.030 |
| 2004 | 12.125 | 17.517 |
| 2005 | 13.929 | 21.122 |

## *Turizm Gelirleri ile Yabancı Ziyaretçi Sayısı verileri arasındaki dağılma diyagram*



# ***Doğrusal tüketim fonksiyonunun normal denklemler yoluyla tahmini:***

**Tablo 2**: Verilerin normal denklemler ile çözüm için düzenlenmesi

| Y | X | Y*X | $X^2$ |
|---|---|---|---|
| 5.650 | 8.614 | 48.6691 | 74.201 |
| 7.008 | 9.689 | 67.9005 | 93.8767 |
| 7.177 | 9.752 | 69.9901 | 95.1015 |
| 5.193 | 7.464 | 38.7605 | 55.7113 |
| 7.636 | 10.412 | 79.5060 | 108.4097 |
| 8.090 | 11.569 | 93.5932 | 133.8418 |
| 8.481 | 13.247 | 112.3478 | 175.4830 |
| 9.677 | 14.030 | 135.7683 | 196.8409 |
| 12.125 | 17.517 | 212.3936 | 306.8452 |
| 13.929 | 21.122 | 294.2083 | 446.1388 |
| $\Sigma Y=84.966$ | $\Sigma X=123.416$ | $\Sigma YX=1153.138$ | $\Sigma X^2=1686.4501$ |

## ***Doğrusal tüketim fonksiyonunun normal denklemler yoluyla tahmini:***

$$\Sigma Y = b_0.n + b_1.\Sigma X$$
$$\Sigma YX = b_0.\Sigma X + b_1.\Sigma\ X2$$

$$84.96 = b_0.10 + b_1.\ 123.4$$
$$1153.13 = b_0.123.4 + b_1.\ 1686.4$$

**$b_0$=0.597 $b_1$=0.640**

$$\hat{Y} = 0.597 + 0.640X$$

Yabancı ziyaretçi sayısı arttıkça turizm geliri artmaktadır.

## ***Doğrusal tüketim fonksiyonunun formülden tahmini:***

$$\hat{b}_0 = \frac{\sum X^2 \sum Y - \sum X \sum YX}{n\sum X^2 - \left(\sum X\right)^2}$$

$$= \frac{(1686.45)*(84.966) - (123.416)*(1153.138)}{10*(1686.45) - (123.416)^2} = 0.597$$

$$\hat{b}_1 = \frac{n\sum YX - \sum X \sum Y}{n\sum X^2 - \left(\sum X\right)^2}$$

$$= \frac{10*(1153.138) - (123.416)*(84.966)}{10*(1686.45) - (123.416)^2} = 0.640$$

## *Tahminin standart hatası ve varyansı:*

$$s = \sqrt{\frac{\sum e^2}{n-k}} \qquad s^2 = \frac{\sum e^2}{n-k}$$

| Y | $Y^2$ | $\hat{Y} = 0.597 + 0.640X$ | $e = Y - \hat{Y}$ | $e^2$ |
|---|---|---|---|---|
| 5.65 | 31.92 | 0.597 + 0.640(**8.614**) = 6.1099 | -0.460 | 0.2115 |
| 7.008 | 49.11 | 0.597 + 0.640(**9.689**) = 6.7979 | 0.210 | 0.0441 |
| 7.177 | 51.51 | 0.597 + 0.640(**9.752**) = 6.8382 | 0.339 | 0.1147 |
| 5.193 | 26.96 | 0.597 + 0.640(**7.464**) = 5.3739 | -0.181 | 0.0327 |
| 7.636 | 58.31 | 0.597 + 0.640(**10.412**) = 7.2606 | 0.375 | 0.1408 |
| 8.09 | 65.45 | 0.597 + 0.640(**11.569**) = 8.0011 | 0.089 | 0.0078 |
| 8.481 | 71.93 | 0.597 + 0.640(**13.247**) = 9.0750 | -0.594 | 0.3529 |
| 9.677 | 93.65 | 0.597 + 0.640(**14.030**) = 9.5762 | 0.101 | 0.0101 |
| 12.125 | 147.02 | 0.597 + 0.640(**17.517**) = 11.8078 | 0.317 | 0.1005 |
| 13.929 | 194.02 | 0.597 + 0.640(**21.122**) = 14.1150 | -0.186 | 0.0346 |
| | $\Sigma Y^2 = 789.8721$ | $\sum \hat{Y} = 84.966$ | 0.010 | $\Sigma e^2 = 1.0501$ |

$$s = \sqrt{\frac{\sum e^2}{n-k}} = \sqrt{\frac{1.0501}{10-2}} = 0.362$$

$$s^2 = (0.362)^2 = 0.131$$