SAKARYA ÜNİVERSİTESİ

Olasılık ve İstatistik
HAFTA 14
Korelasyon Analizi
Regresyon Analizi

Dr. Öğretim Üyesi Burcu ÇARKLI YAVUZ

bcarkli@sakarya.edu.tr

Korelasyon Analizi & Regresyon Analizi

- **Korelasyon analizi** ile, iki yada daha çok değişken arasında ilişki olup olmadığını, ilişki varsa yönünü ve gücünü inceleriz.
- **Regresyon analizi** ile ise, değişkenlerden biri bir birim değiştiğinde, diğerinin nasıl bir değişim göstereceğini inceleriz.
- Değişkenler arasındaki ilişkilerden yararlanılarak geliştirilecek matematiksel modeller yardımıyla tahminler yapılabilir.
- Örneğin ürün reklamı ile ürün satışı arasında bir model geliştirilebilirse, ne kadarlık reklam ile ne kadar satış yapılabileceği tahmin edilebilir.
- Belirli bir hastalıkta kullanılan ilaç dozu ile iyileşme süresi arasında bir model geliştirilebilirse verilen doza göre hastanın ne kadar sürede iyileşebileceği tahmin edilebilir.

Kovaryans

➢ Kovaryans değeri değişkenler arasındaki ilişkinin yönünü belirler, -∞ ile +∞ arasında bir değer alabilir.
- Cov(x,y)=0 ise bu değişkenler arasında ilişki yoktur
- Cov(x,y)>0 ise değişkenler aynı yönde değişim gösteriyor
- Cov(x,y)<0 ise değişkenler farklı yönde değişim gösteriyor

x_i ve y_i değişkenlerin ortalamaları μ_x ve μ_y olmak üzere

kovaryans anakütle için

$$\sigma_{xy} = \frac{1}{N} \sum (x_i - \mu_x)(y_i - \mu_y)$$

anakütle ortalaması

örneklem için

$$S_{xy} = \frac{1}{n-1} \sum (x_i - \bar{x})(y_i - \bar{y})$$

örneklem ortalaması

ile hesaplanır.

Kovaryans örnek

a) verilen iki serinin kovaryansını hesaplayınız.

x:	5	6	7	8	9
y:	6	9	11	16	18

$$\mu_x = \frac{5+6+7+8+9}{5} = 7$$

$$\mu_y = \frac{6+9+11+16+18}{5} = 12$$

$x_i - \mu_x$	$y_i - \mu_y$	$(x_i-\mu_x)(y_i-\mu_y)$
$5-7=-2$	$6-12=-6$	$(-2)(-6)=12$
$6-7=-1$	$9-12=-3$	$(-1)(-3)=3$
$7-7=0$	$11-12=-1$	$(0).(-1)=0$
$8-7=1$	$16-12=4$	$(1).(4)=4$
$9-7=2$	$18-12=6$	$(2)(6)=12$

$$Cov(x,y) = \frac{1}{N}\sum (x_i-\mu_x)(y_i-\mu_y) = \frac{1}{5}(12+3+0+4+12) = \frac{31}{5} = 6,2$$

Bu değişkenler arasında aynı yönde bir ilişki var. Biri artarken diğeri de artar.

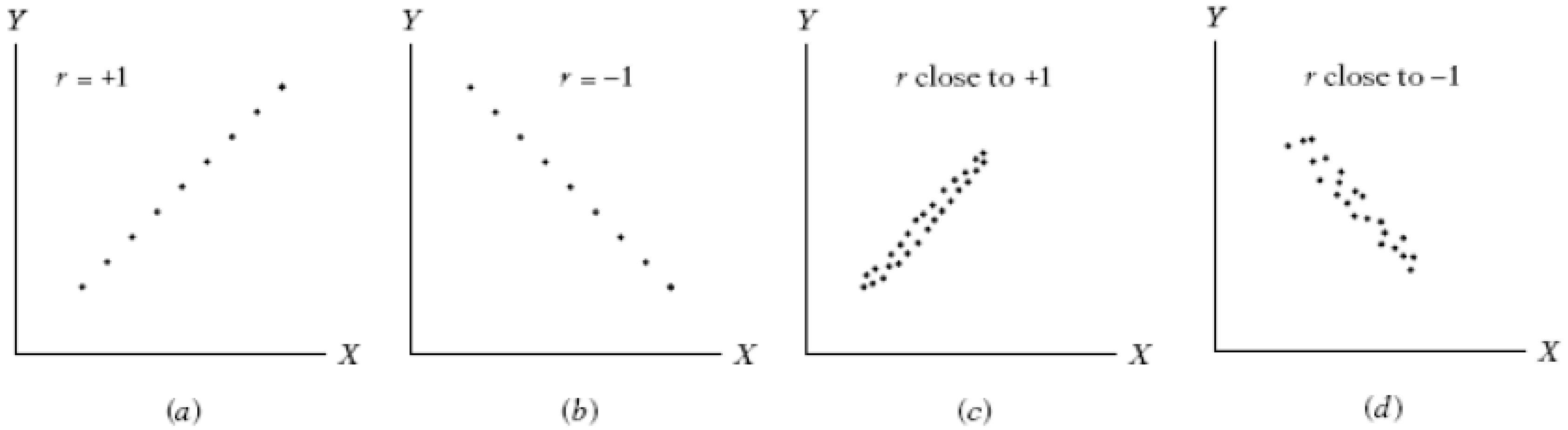
Korelasyon

- Değişkenler arasındaki ilişkinin derecesinin şiddeti (zayıf/kuvetli) hakkındaki bilgiyi **korelasyon katsayısı** verir.
- Korelasyon katsayısı anakütle için ρ (ro) ile örneklem için r ile gösterilir.
- İki değişken arasındaki ilişkinin derecesine **basit korelasyon**, ikiden fazla değişken arasındaki korelasyona ise **çoklu korelasyon** denir.
- Bir fabrikanın büyüklüğü ve kazancı, bir öğrencinin derse ilgisi ve o dersin sınavından aldığı not örneklerinde iki değişken vardır. Bu değişkenlerden biri **bağımlı değişken**, diğeri **bağımsız değişken**dir.

Korelasyon katsayısının özellikleri

- Korelasyon katsayısı $-1 \leq \rho \leq 1$ aralığında değer alır.
- Korelasyon katsayısı 0 ise **ilişki yoktur**, 1 yada -1 ise **güçlü ilişki** (**tam korelasyon**) vardır.
- Korelasyon katsayısı 1 yada -1 e ne kadar yakınsa ilişki o kadar **güçlüdür**.
- Korelasyon katsayısı 0 a ne kadar yakında ilişki o kadar **zayıftır**.
- $-0.5 \leq \rho \leq 0.5$ aralığında zayıf ilişki vardır.
- $0.5 \leq \rho \leq 1$ ve $-1 \leq \rho \leq -0.5$ aralığında güçlü ilişki vardır.
- Korelasyon katsayısının işareti (negatif/pozitif) ilişkinin yönünü belirler. Pozitif ise ilişki **aynı yönlüdür**, negatif ise ilişki **zıt yönlüdür**.

Korelasyon katsayısının özellikleri

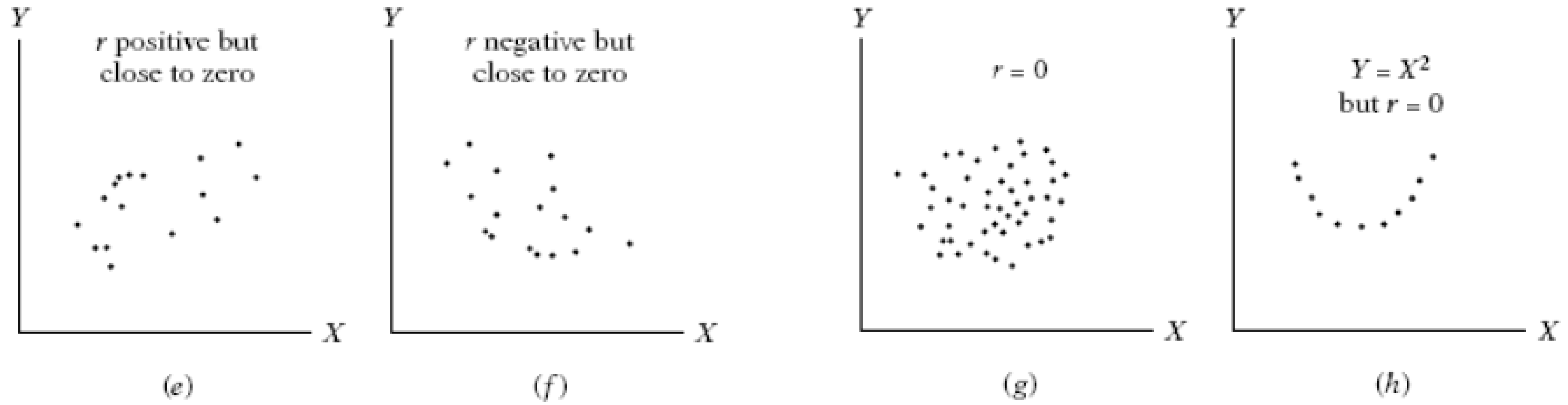


Çizim (a) da aynı yönlü tam doğrusal bir ilişki söz konsudur.

Çizim (b)'de ise ters yönlü tam doğrusal bir ilişki söz konusudur.

Çizim (c) ve (d) tama yakın sırayla aynı yönlü ve ters yönlü ilişki söz konusudur.

Korelasyon katsayısının özellikleri



(e) ve (f) çizimlerinde ise korelasyon katsayısı sıfıra yakındır. Çizim (g)'de ilişki yoktur, korelasyon katsayısı sıfırdır. Son çizim (h) de de korelasyon katsayısı sıfırdır ama bu durum değişkenler arasında ilişki olmamasından değil ilişkinin doğrusal olmamasından kaynaklanır.

Korelayon örnek

x_i	2	5	7	9	12
y_i	1	3	8	11	13

korelasyon pozitif
ilişki aynı yönlü

x_i	2	5	7	9	12
y_i	13	11	8	3	1

korelasyon negatif
ilişki zıt yönlü

x_i	2	5	7	9	12
y_i	3	3	3	3	3

korelasyon sıfır
ilişki yok

Korelasyon katsayısının hesaplanması

X_i ve Y_i gibi iki değişken arasındaki korelasyon katsayısı

ana kütle için $\rho = \dfrac{\sum (x_i-\mu_x)(y_i-\mu_y)}{\sqrt{\sum (x_i-\mu_x)^2 \sum (y_i-\mu_y)^2}}$

örneklem için $r = \dfrac{\sum (x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum (x_i-\bar{x})^2 \sum (y_i-\bar{y})^2}}$

standart sapma $\sigma = \sqrt{\dfrac{\sum_{i=1}^{N} (x_i-\mu)^2}{N}} \Rightarrow \sqrt{\sum_{i=1}^{N} (x_i-\mu)^2} = \sigma\sqrt{N}$

$s = \sqrt{\dfrac{\sum_{i=1}^{n} (x_i-\bar{x})^2}{n-1}} \Rightarrow \sqrt{\sum_{i=1}^{n} (x_i-\bar{x})^2} = s \cdot \sqrt{n-1}$

Korelasyon katsayısının hesaplanması

korelasyon katsayısı formülünde yerine koyarsak

$$\rho = \frac{\sum (x_i-\mu_x)(y_i-\mu_y)}{\sqrt{\sum (x_i-\mu_x)^2}\cdot\sqrt{\sum (y_i-\mu_y)^2}} = \frac{\sum (x_i-\mu_x)(y_i-\mu_y)}{\sigma_x\sqrt{N}\cdot\sigma_y\sqrt{N}} = \frac{\overbrace{\sum (x_i-\mu_x)(y_i-\mu_y)}^{cov\ \sigma_{xy}}}{N\cdot\sigma_x\,\sigma_y}$$

$$r = \frac{\sum (x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum (x_i-\bar{x})^2}\cdot\sqrt{\sum (y_i-\bar{y})^2}} = \frac{\overbrace{\sum (x_i-\bar{x})(y_i-\bar{y})}^{cov}}{(n-1)\,s_x\cdot s_y}$$

$$\rho = \frac{\sigma_{xy}}{\sigma_x\,\sigma_y}$$ kovaryansı standart sapmaların çarpımına böl

$$r = \frac{s_{xy}}{s_x\cdot s_y}$$

Örnek1

a) Anakütleden seçilmiş örneklem verileri için korelasyon katsayısını hesaplayınız.

x_i	10	15	16	22	25	28	29	40	45	50
y_i	5	12	15	20	10	26	25	38	42	47

$$\bar{x} = \frac{10+15+16+22+25+28+29+40+45+50}{10} = 28$$

$$\bar{y} = \frac{5+12+15+20+10+26+25+38+42+47}{10} = 24$$

Örnek1

$x_i-\bar{x}$	$y_i-\bar{y}$	$(x_i-\bar{x})^2$	$(y_i-\bar{y})^2$	$(x_i-\bar{x})(y_i-\bar{y})$
$10-28=-18$	$5-24=-19$	324	361	342
$15-28=-13$	$12-24=-12$	169	144	156
-12	-9	144	81	108
-6	-4	36	16	24
-3	-14	9	196	42
0	2	0	4	0
1	1	1	1	1
12	14	144	196	168
17	18	289	324	306
22	23	484	529	506
		+ 1600	+ 1852	+ 1653

Örnek1

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}} = \frac{1653}{\sqrt{1600 \cdot 1852}} = \frac{1653}{1721,39} = 0.96$$

Pozitif olduğu için aynı yönlü ilişki

1'e yakın bir değer olduğu için güçlü bir ilişki var

Örnek2

Bir anakitle üyelerinde yapılan iki değişken arasındaki araştırmaya göre kovaryans $\sigma_{xy} = 6$ ve değişkenlerin standart sapması $\sigma_x = 3$ ve $\sigma_y = 4$ olarak elde edilmiştir. Değişkenler arasındaki korelasyon katsayısı kaçtır?

$$\rho = \frac{\sigma_{xy}}{\sigma_x \cdot \sigma_y} = \frac{6}{3 \cdot 4} = 0.5$$

Regresyon Analizi

- Regresyon analizi değişkenler arasındaki ilişkinin araştırılmasında kullanılan istatistiksel bir araçtır. Bu araçla, bir değişkenin diğer değişken üzerindeki nedensel ilişkisi araştırılır.
- İncelenen ilişkideki değişkenler aralarındaki ilişki göz önüne alınarak değişkenler bağımlı ve bağımsız olarak isimlendirilir.
- Davranışı tahmin edilecek olan rastlantı değişkeni bir diğer değişken(ler)in fonksiyonu olarak gösterilebilir ve bu değişken bağımlı olarak isimlendirilir ve Y ile gösterilir.
- Bağımlı değişkeni etkileyen değişken ise X ile gösterilir ve bağımsız değişken olarak isimlendirilir.
- Yani, bağımlı değişken, bağımsız değişken(ler) tarafından açıklanmaya çalışılır ve açıklayıcı değişkenlerin modelde bilinen sabitler olduğu varsayılır.

Regresyon Modeli

$$Y_i = b_0 + b_1 X_i + \varepsilon_i$$

şeklinde gösterilen regresyon denkleminde,

b_0 başlangıç parametresi de denilen sabit bir katsayıdır

b_1 eğim parametresidir, X deki 1 birimlik değişmenin Y üzerinde nasıl bir değişim yaptığını gösterir.

ε_i hata terimidir

- Bağımsız değişken (X) sayısının bir tane olması basit regresyonla, birden fazla olması ise çoklu regresyonla açıklanır.
- Regresyon modeli, doğrusal yapıda olabileceği gibi parabolik, logaritmik, üstel biçimli de olabilir.

Kaynaklar

- Bilişim Teknolojileri için İşletme İstatistiği ders notları, Dr. Öğr. Üyesi Halil İbrahim Cebeci, Sakarya Üniversitesi
- İstatistik II ders notları, Yrd. Doç. Dr. Leyla İşbilen Yücel, Doç. Dr. Özlem Yorulmaz, İstanbul Üniversitesi