SPSS İLE İSTATİSTİKSEL VERİ ANALİZİ

Statistical Packages for the Social Sciences

PROF.DR.YÜKSEL TERZİ

ONDOKUZ MAYIS ÜNİVERSİTESİ
FEN-EDEBİYAT FAKÜLTESİ
İSTATİSTİK BÖLÜMÜ
SAMSUN

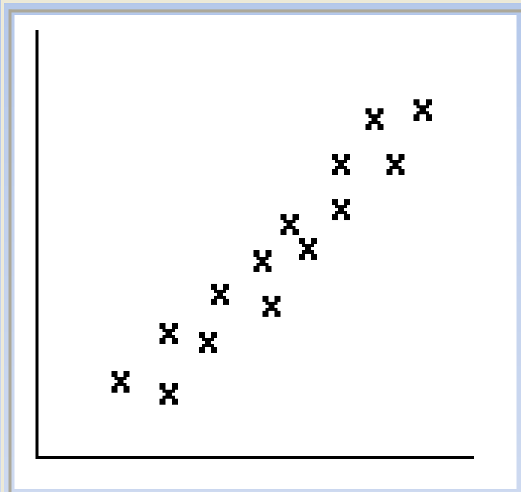
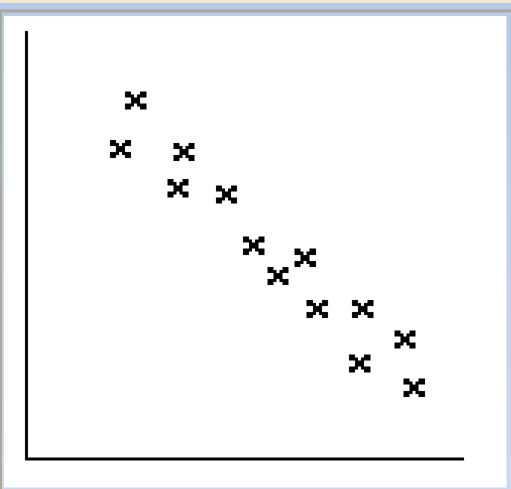
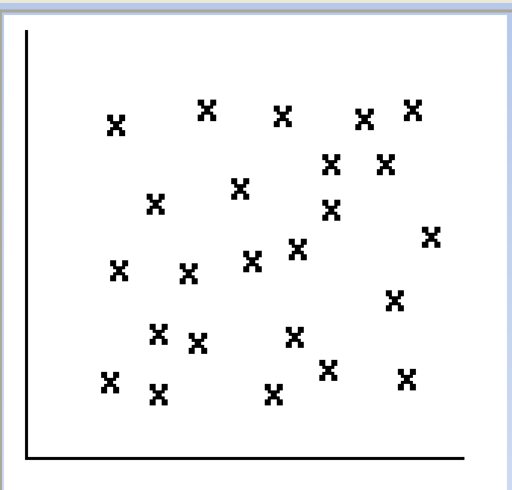
2019

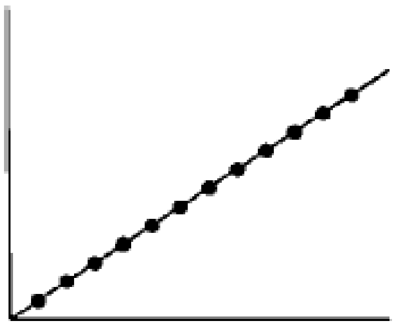
KORELASYON

Korelasyon analizinde iki veya daha çok sayıda değişken arasında bir ilişki bulunup bulunmadığı, eğer varsa bu ilişkinin derecesi ve fonksiyonel şekli belirlenmeye çalışılır. Örneğin reklamların satışı arttırdığı şeklinde bir düşünce yaygındır. Ancak satışların artışı sadece reklamlar ile açıklanamaz. Nüfus artışı, moda, fiyat rakiplerle rekabet satışları etkileyen diğer nedenler olarak düşünülebilir. Öyle ise reklamlar ile satış arasında ilişkinin olup olmadığı incelenmelidir.

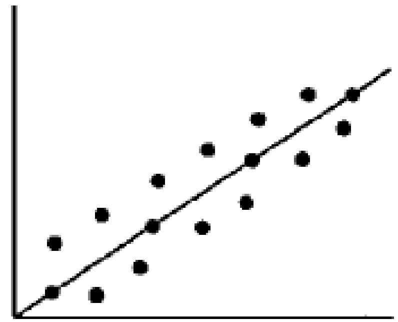
Doğrusal Korelasyon:

Bir değişkenin değeri artarken diğer değişkenin değeri düzenli artıyor veya eksiliyorsa iki değişken arasındaki ilişki doğrusaldır. İlişki grafik üzerinden de incelenebilir.

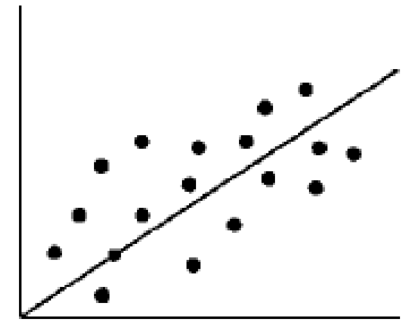
		
Korelasyon=+1	**Korelasyon=-1**	**Korelasyon=0**



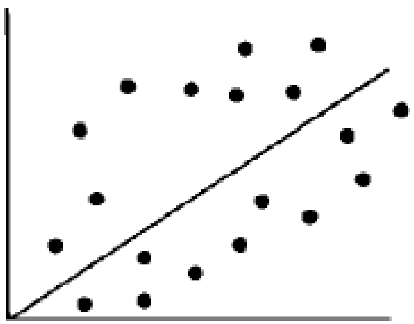
(a) $r = +1$
Maximum positive correlation



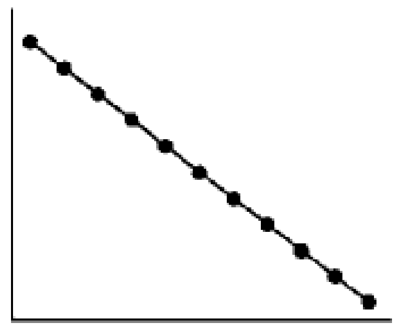
(b) $r = .80$
Strong positive correlation



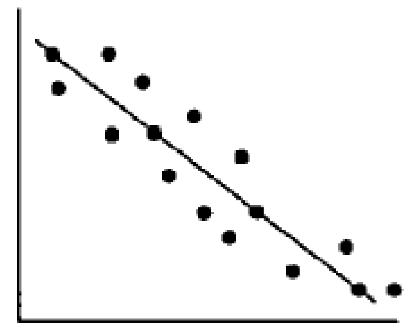
(c) $r = .50$
Moderate positive correlation



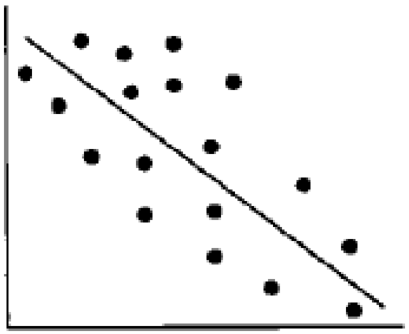
(d) $r = .20$
Weak positive correlation



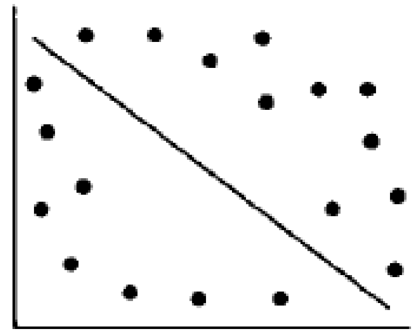
(e) $r = -1$
Maximum negative correlation



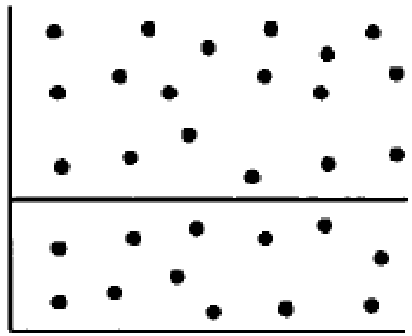
(f) $r = -.80$
Strong negative correlation



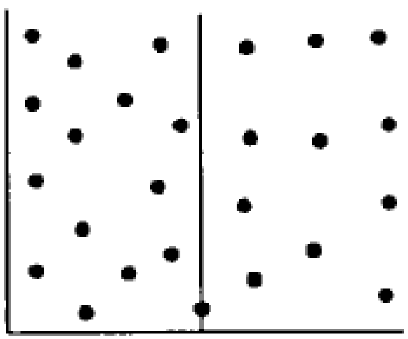
(g) $r = -.50$
Moderate negative correlation



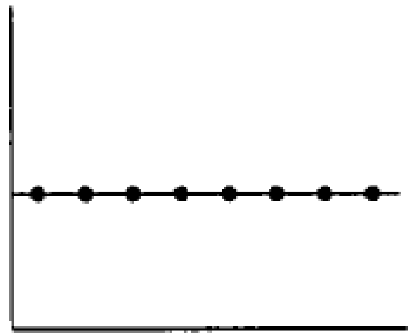
(h) $r = -.20$
Weak negative correlation



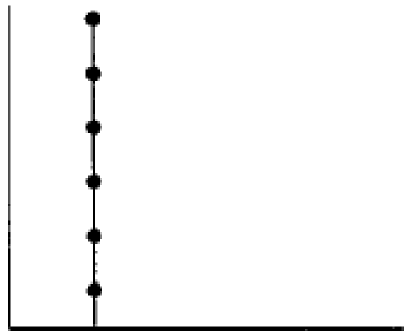
(i) $r = 0$
No correlation



(j) $r = 0$
No correlation



(k) Value of r cannot be computed



(l) Value of r cannot be computed

Doğrusal korelasyonun hesaplanmasında Pearson Momentler Çarpımı korelasyonu kullanılır. Bu formülün uygulanabilmesi için veriler en az aralıklı ölçekle toplanmalı ve süreklilik gösteren nicel bir değişken olmalıdır.

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

Korelasyon katsayısının değeri -1 ile +1 arasında değişir. Sonucun +1 çıkması iki değişken arasında kuvvetli olumlu ilişkinin bulunduğunu, -1 ise kuvvetli olumsuz ilişkinin bulunduğunu gösterir. Korelasyon katsayısı 0 'a yaklaştıkça ilişkinin kuvveti zayıflar, sıfır ise iki değişken arasında ilişkinin olmadığını gösterir.

Korelasyon katsayısının önem denetimi:

Hesaplanmış olan korelasyon katsayısının tesadüfi mi yoksa gerçek bir ilişkiyi mi gösterdiğinin belirlenmesi için denetlenmesi gerekir. Denetim için kurulan hipotezler $H_0 : \rho=0$ (doğrusal ilişki yoktur); $H_1 : \rho > 0$ (doğrusal ilişki vardır) şeklinde belirlenir. Test istatistiği şu formüle göre hesaplanır,

$$t = r \times \sqrt{\frac{(n-2)}{1-r^2}}$$

Serbestlik derecesi (n-2) olup, n çok büyük olduğunda r'nin yaklaşacağı değeri ρ ile gösterelim Bu durumda $\rho=0$ etrafında korelasyon katsayısının dağılımı simetrik olup, t dağılımına yaklaşır.

ÖRNEK..Aşağıda bir işletmede gün olarak kullanılan izin (X) ile performans puanları (Y) verilmiştir. Bu iki değişken arasında ilişki var mıdır?

X	Y	X^2	Y^2	XY
1	14	1	196	14
2	13	4	169	26
3	12	9	144	36
3	13	9	169	39
2	11	4	121	22
1	12	1	144	12
4	12	16	144	48
5	11	25	121	55
4	14	16	196	56
3	13	9	169	39
6	12	36	144	72
5	12	25	144	60
10	10	100	100	100
9	11	81	121	99
1	14	1	196	14
8	11	64	121	88
9	10	81	100	90
7	9	49	81	63
6	12	36	144	72
7	10	49	100	70
$\sum x$ 96	$\sum y$ 236	$\sum x^2$ 616	$\sum y^2$ 2824	$\sum xy$ 1075

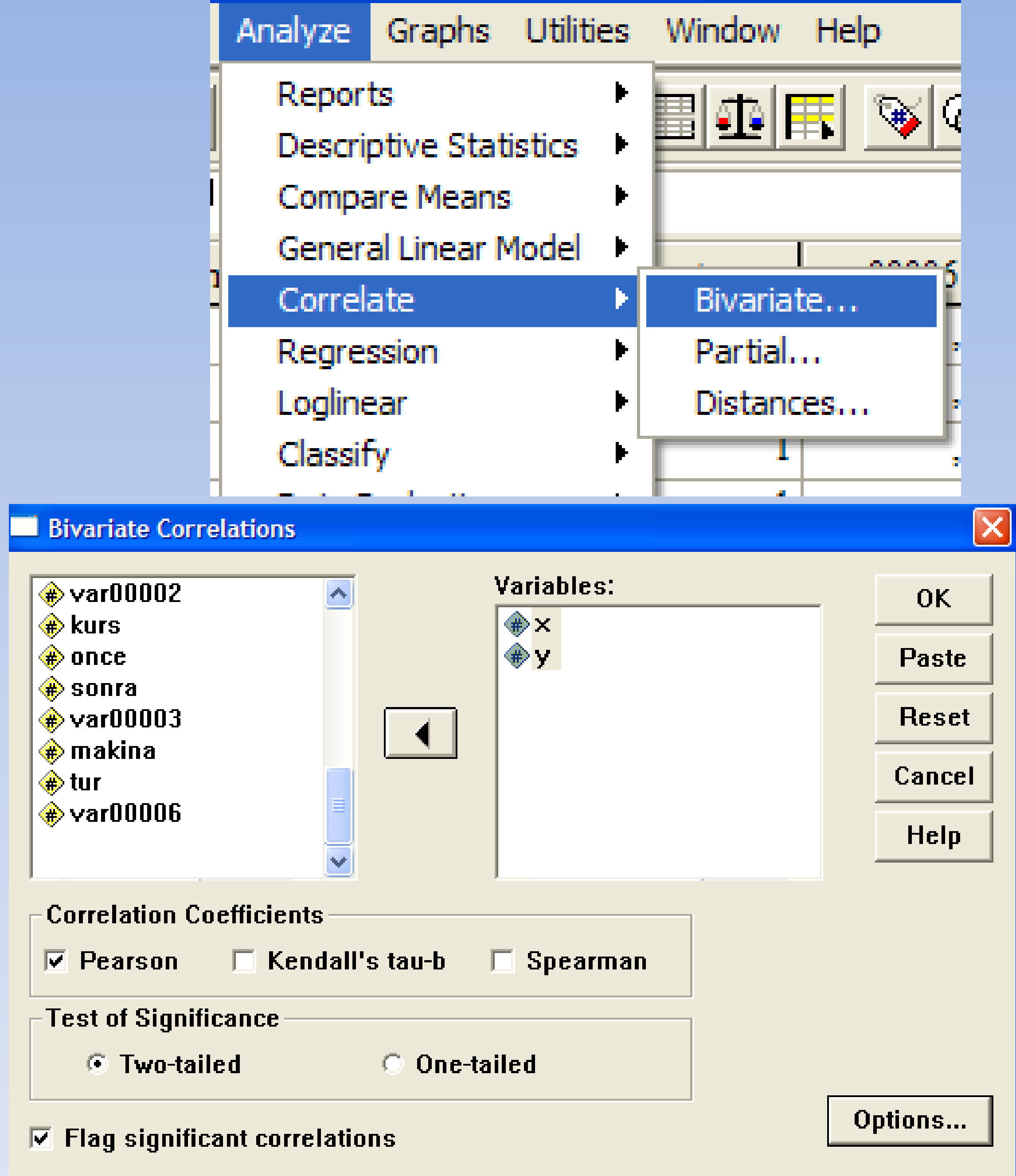
$$r=\frac{\sum XY-\frac{(\sum X)(\sum Y)}{n}}{\sqrt{\left(\sum X^2-\frac{(\sum X)^2}{n}\right)\left(\sum Y^2-\frac{(\sum Y)^2}{n}\right)}}=\frac{1075-\frac{96*236}{20}}{\sqrt{\left(616-\frac{96^2}{20}\right)\left(2824-\frac{236^2}{20}\right)}}=-0,74$$

Elde edilen sonuca göre kullanılan izin miktarı ile performans puanları arasında negatif yönlü kuvvetli ilişki vardır. Kullanılan izin miktarı arttıkça performans puanları düşmektedir. Bulunan korelasyonun gerçekten önemli olup olmadığı incelenirse

$$t=r*\sqrt{\frac{(n-2)}{1-r^2}}=-0,74*\sqrt{\frac{(20-2)}{1-(-0,74)^2}}=-4,5$$

H_0 : ρ =0 ; H_1 : ρ > 0 $t_h=|-4,5| > t_{(20-2),\,0.05}$ =1,73 olduğundan H_0 reddedilir. Bulunan korelasyon doğrusaldır-önemlidir ve tesadüfi değildir.

x	y
1	14
2	13
3	12
3	13
2	11
1	12
4	12
5	11
4	14
3	13
6	12
5	12
10	10
9	11
1	14
8	11
9	10
7	9
6	12
7	10



Correlations

Correlations

		X	Y
X	Pearson Correlation	1,000	-,741**
	Sig. (2-tailed)	,	,000
	N	20	20
Y	Pearson Correlation	-,741**	1,000
	Sig. (2-tailed)	,000	,
	N	20	20

**. Correlation is significant at the 0.01 level

KISMİ KORELASYON KATSAYISI
(Partial Correlation)

İki değişken arasında korelasyon katsayısının yüksek çıkması muhakkak o değişkenler arasında bir neden-sonuç ilişkisi bulunduğu anlamına gelmez. İki değişken arasında neden-sonuç ilişkisi olmadığı halde korelasyon katsayısı eğer yüksek çıkıyorsa, bu iki değişkenin üçüncü bir değişkenden etkilenmesi söz konusu olabilir.

Kısmi korelasyon katsayısı, iki değişken arasındaki ilişkinin bir yada daha çok değişkenin kontrol edilmesiyle hesaplanan bir değerdir.

$$r_{XY,Z} = \frac{r_{XY} - r_{XZ} r_{YZ}}{\sqrt{(1 - r_{XZ}^2)(1 - r_{YZ}^2)}}$$

$r_{XY,Z}$: Z değişkeni sabitken (etkisi giderildiğinde), X ile Y arasındaki kısmi korelasyon
r_{XY} : X ile Y arasındaki korelasyon
r_{XZ} : X ile Z arasındaki korelasyon
r_{YZ} : Y ile Z arasındaki korelasyon

Kısmi korelasyon katsayısının anlamlı olup olmadığı F testi ile incelenebilir. Bu testte n gözlem sayısını, k ise tahmin edilen parametre sayısını gösterir.

$$F = \frac{r^2 / (k-2)}{(1 - r^2)/(n-k)}$$

H_0 : $\rho = 0$ (z sabitken x ile y arasında doğrusal ilişki yoktur.)
H_1 : $\rho > 0$ (z sabitken x ile y arasında doğrusal ilişki vardır.)

Örnek. Talep (Y), gelir (X) ve fiyat (Z) serileri arasındaki ilişki araştırılıyor. Fiyatı sabit alarak talep ile gelir arasındaki kısmi korelasyonu bulup ve doğrusal ilişki olup olmadığını araştırınız?

Yıllar	Gelir(X)	Talep (Y)	Fiyat(Z)
1995	3	8	4
1996	2	9	7
1997	5	13	11
1998	6	22	18
$\sum$:	16	52	40

Yıllar	X	Y	Z	X^2	Y^2	Z^2	XY	XZ	YZ
1995	3	8	4	9	64	16	24	12	32
1996	2	9	7	4	81	49	18	14	63
1997	5	13	11	25	169	121	65	55	143
1998	6	22	18	36	484	324	132	108	396
Toplam	16	52	40	74	798	510	239	189	634

$$r_{XY} = \frac{\sum XY - \frac{(\sum X)(\sum Y)}{n}}{\sqrt{\left(\sum X^2 - \frac{(\sum X)^2}{n}\right)\left(\sum Y^2 - \frac{(\sum Y)^2}{n}\right)}} = \frac{239 - \frac{16*52}{4}}{\sqrt{\left(74 - \frac{16^2}{4}\right)\left(798 - \frac{52^2}{4}\right)}} = 0{,}89$$

$$r_{XZ} = \frac{\sum XZ - \frac{(\sum X)(\sum Z)}{n}}{\sqrt{\left(\sum X^2 - \frac{(\sum X)^2}{n}\right)\left(\sum Z^2 - \frac{(\sum Z)^2}{n}\right)}} = \frac{189 - \frac{16*40}{4}}{\sqrt{\left(74 - \frac{16^2}{4}\right)\left(510 - \frac{40^2}{4}\right)}} = 0{,}87$$

$$r_{YZ} = \frac{\sum YZ - \frac{(\sum Y)(\sum Z)}{n}}{\sqrt{\left(\sum Y^2 - \frac{(\sum Y)^2}{n}\right)\left(\sum Z^2 - \frac{(\sum Z)^2}{n}\right)}} = \frac{634 - \frac{52*40}{4}}{\sqrt{\left(758 - \frac{52^2}{4}\right)\left(510 - \frac{40^2}{4}\right)}} = 0,98$$

$$r_{XY,Z} = \frac{r_{XY} - r_{XZ}r_{YZ}}{\sqrt{(1 - r_{XZ}^2)(1 - r_{YZ}^2)}}$$

$$= \frac{0,89 - 0,98 * 0,87}{\sqrt{(1 - 0,98^2)(1 - 0,87^2)}} = 0,4$$

Z(fiyat) sabitken x (gelir) ile y(talep) arasında ki ilişki aynı yönlü fakat düşük bir ilişki söz konusudur.

H_0 : ρ =0 (z sabitken x ile y arasında doğrusal ilişki yoktur.)
H_1 : ρ > 0 (z sabitken x ile y arasında doğrusal ilişki vardır.)

$$F = \frac{r^2/(k-2)}{(1-r^2)/(n-k)} = \frac{0{,}4^2/(3-2)}{(1-0{,}4^2)/(4-3)} = 0{,}2$$

F_h=0,2<F_{tablo}=6,6 H_o kabul edilir. Yani z sabitken x ile y arasında doğrusal ilişki yoktur.

SIRA KORELASYON KATSAYISI

Doğrusal korelasyonda ilişkisi araştırılan değişkenlerin nicel ve normal olması gerekir. Bu varsayımlar sağlanmadığında Spearman Sıra Korelasyon Katsayısı kullanılır.

Sıra korelasyon katsayısının hesaplanmasında önce gözlem değerleri büyükten küçüğe veya küçükten büyüğe doğru sıralanır ve bu sıralamaya göre sıra numarası verilir.

$$r_s = 1 - \frac{6\sum D_i^2}{n(n^2-1)}$$

D_i : X ve Y'nin sıra numaraları arasındaki fark

n : Gözlem sayısı

Eğer verilerde bağlı-aynı (ties) gözlemler varsa Spearman sıra korelasyonu aşağıdaki gibi hesaplanır.

I.Yol:

$$r_s = \frac{\sum_{i=1}^{n}\left[R(x_i)-R(\bar{x})\right]\left[R(y_i)-R(\bar{y})\right]}{(n-1)\times S_{R(x)}\times S_{R(y)}}$$

$R(\bar{x})$:X değişkeninin sıra puanlarının aritmetik ortalaması

$R(\bar{y})$:Y değişkeninin sıra puanlarının aritmetik ortalaması

$S_{R(X)}$:X değişkeninin sıra puanlarının standart sapması

$S_{R(Y)}$:Y değişkeninin sıra puanlarının standart sapması

Bağlı gözlem olduğunda Spearman sıra korelasyonu
2. yol:

$$r_s = \frac{A_X + A_Y - \sum D^2}{2 \times \sqrt{A_X \times A_Y}}$$

$$A_X = \frac{n^3 - n - T_X}{12} \qquad A_Y = \frac{n^3 - n - T_Y}{12}$$

$$T_X = \sum_{i=1}^{s} \left(t_{i_{(X)}}^3 - t_{i_{(X)}} \right) \qquad T_Y = \sum_{i=1}^{s} \left(t_{i_{(Y)}}^3 - t_{i_{(Y)}} \right)$$

Bağlı gözlem olduğunda Spearman sıra korelasyonu
3. yol:

$$r_s = \frac{\sum R_X R_Y - \frac{\left(\sum R_X\right)\left(\sum R_Y\right)}{n}}{\sqrt{\left(\sum R_X^2 - \frac{\left(\sum R_X\right)^2}{n}\right) \times \left(\sum R_Y^2 - \frac{\left(\sum R_Y\right)^2}{n}\right)}}$$

Sıra Korelasyonunun Yorumu

0.90-1.00	Çok güçlü ilişki
0.70-0.89	Güçlü
0.50-0.69	Orta seviyede
0.30-0.49	Düşük
0.16 -0.29	Zayıf
<0.16	Çok zayıf

Örnek. 7 öğrencinin boy ve ağırlıklarının büyükten küçüğe doğru sıra puanları aşağıda gösterilmiştir. Veriler normal dağılım göstermediğine göre, boy ve ağırlıklar arasındaki ilişkiyi hesaplayınız?

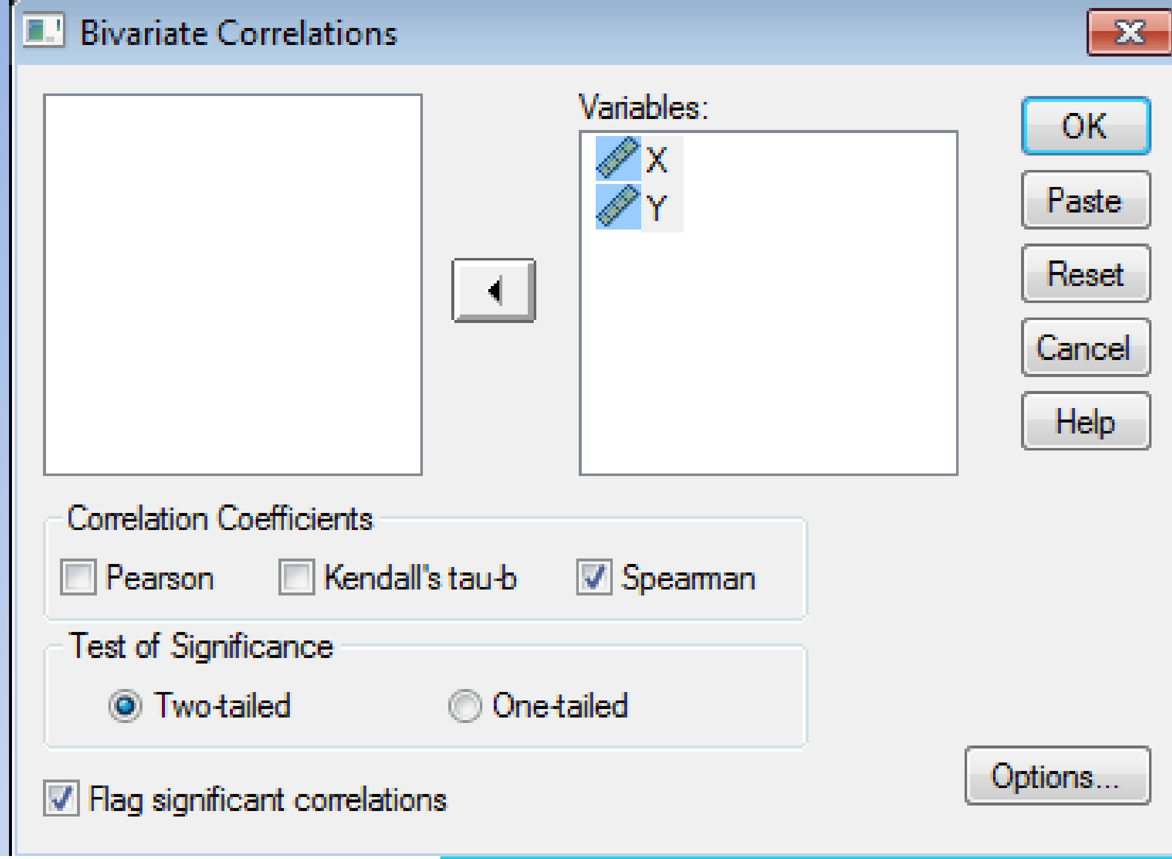
X	Y	D	D^2
2	1	1	1
4	6	-2	4
6	5	1	1
1	2	-1	1
3	3	0	0
7	7	0	0
5	4	1	1

$$r_s = 1 - \frac{6\sum D_i^2}{n(n^2-1)}$$

$$= 1 - \frac{6\times 8}{7(7^2-1)} = 0,86$$

Öğrencilerin boy uzunlukları ile ağırlıkları arasında aynı yönde önemli bir ilişki vardır.

	X	Y
1	2,00	1,00
2	4,00	6,00
3	6,00	5,00
4	1,00	2,00
5	3,00	3,00
6	7,00	7,00
7	5,00	4,00



Correlations

			X	Y
Spearman's rho	X	Correlation Coefficient	1,000	,857*
		Sig. (2-tailed)	.	,014
		N	7	7
	Y	Correlation Coefficient	,857*	1,000
		Sig. (2-tailed)	,014	.
		N	7	7

*. Correlation is significant at the 0.05 level (2-tailed).

Örnek. **Aşağıdaki verilere göre fiyat ile talep arasındaki sıra korelasyon katsayısını bulunuz?**

Fiyat (X)	Talep(Y)	Rx	Ry
60	800	8	2
100	750	5,5	4
120	700	3	5
100	800	5,5	3
80	850	7	1
120	650	3	6,5
130	650	1	6,5
120	600	3	8

	x	y	Rx	Ry
1	60	800	8,0	2,5
2	100	750	5,5	4,0
3	120	700	3,0	5,0
4	100	800	5,5	2,5
5	80	850	7,0	1,0
6	120	650	3,0	6,5
7	130	650	1,0	6,5
8	120	600	3,0	8,0

Descriptive Statistics

	N	Mean	Std. Deviation
Rank of x	8	4,500	2,3755
Rank of y	8	4,500	2,4202
Valid N (listwise)	8		

$$r_s = \frac{\sum_{i=1}^{n}\left[R(x_i)-R(\bar{x})\right]\left[R(y_i)-R(\bar{y})\right]}{(n-1)\times S_{R(x)}\times S_{R(y)}}$$

$$= \frac{\sum\left[(8-4.5)(2.5-4.5)+\ldots+(3-4.5)(8-4.5)\right]}{(8-1)\times 2.375\times 2.42}$$

$$= \frac{-34.25}{40.233} = -0.851$$

Correlations

			x	y
Spearman's rho	x	Correlation Coefficient	1,000	-,851**
		Sig. (2-tailed)	.	,007
		N	8	8
	y	Correlation Coefficient	-,851**	1,000
		Sig. (2-tailed)	,007	.
		N	8	8

**. Correlation is significant at the 0.01 level (2-tailed).

Fiyat ile talep arasında ters yönde yüksek bir ilişki vardır. (%85,1)

Örnek. **Aşağıdaki verilere göre X ile Y arasındaki sıra korelasyonu bulunuz.**

X	Y	R_X	R_Y	$D = R_X - R_Y$	D^2	R_X^2	R_Y^2	$R_X R_Y$
0	1	1.5	2	-0.5	0.25	2.25	4	3
0	0	1.5	1	0.5	0.25	2.25	1	1.5
2	2	3	3.5	-0.5	0.25	9	12.25	10.5
4	2	4	3.5	0.5	0.25	16	12.25	14
8	8	6	9.5	-3.5	12.25	36	90.25	57
8	8	6	9.5	-3.5	12.25	36	90.25	57
8	3	6	5	1	1	36	25	30
13	4	8	6	2	4	64	36	48
16	6	9.5	7	2.5	6.25	90.25	49	66.5
16	7	9.5	8	1.5	2.25	90.25	64	76
		55	55	0	39	382	384	363.5

I.Yol

Descriptive Statistics

	N	Mean	Std. Deviation
Rank of X	10	5,500	2,972
Rank of Y	10	5,500	3,009
Valid N (listwise)	10		

$$r_s = \frac{\sum_{i=1}^{n}\left[R(x_i)-R(\bar{x})\right]\left[R(y_i)-R(\bar{y})\right]}{(n-1)\times S_{R(x)}\times S_{R(y)}}$$

$$= \frac{(1.5-5.5)(2-5.5)+\ldots+(9.5-5.5)(8-5.5)}{(10-1)\times 2.972\times 3.009} = \frac{61}{80.494} = 0.758$$

II.Yol

$$T_X = \sum_{i=1}^{s}\left(t_{i_{(X)}}^3 - t_{i_{(X)}}\right) = \left(2^3-2\right)+\left(3^3-3\right)+\left(2^3-2\right)=36$$

2 tane 0 var 3 tane 8 var 2 tane 16 var

$$T_Y = \sum_{i=1}^{s}\left(t_{i_{(Y)}}^3 - t_{i_{(Y)}}\right) = \left(2^3-2\right)+\left(2^3-2\right)=12$$

2 tane 2 var 2 tane 8 var

$$A_X = \frac{n^3-n-T_X}{12} = \frac{10^3-10-36}{12} = 79.5$$

$$A_Y = \frac{n^3-n-T_Y}{12} = \frac{10^3-10-12}{12} = 81.5$$

$$r_s = \frac{A_X + A_Y - \sum D^2}{2\times\sqrt{A_X \times A_Y}} = \frac{79.5+81.5-39}{2\times\sqrt{79.5\times 81.5}} = 0.758$$

III.Yol

$$r_s = \frac{\sum R_X R_Y - \frac{\left(\sum R_X\right)\left(\sum R_Y\right)}{n}}{\sqrt{\left(\sum R_X^2 - \frac{\left(\sum R_X\right)^2}{n}\right) \times \left(\sum R_Y^2 - \frac{\left(\sum R_Y\right)^2}{n}\right)}}$$

$$= \frac{363.5 - \frac{55 \times 55}{10}}{\sqrt{\left(382 - \frac{55^2}{10}\right) \times \left(384 - \frac{55^2}{10}\right)}} = 0.758$$

Correlations

			X	Y
Spearman's rho	X	Correlation Coefficient	1,000	,758*
		Sig. (2-tailed)	.	,011
		N	10	10
	Y	Correlation Coefficient	,758*	1,000
		Sig. (2-tailed)	,011	.
		N	10	10

*. Correlation is significant at the 0.05 level (2-tailed).

Kendall Tau-b

Kendall (1938) tarafından geliştirilen bu test, ikili ve sıralı ölçekli veriler arasındaki ilişkileri belirler. İki farklı gözlemcinin yaptığı değerlendirmeler puan büyüklüğü sırası içinde verilmişse veya hakem puanları büyüklük sırasına sokmuşsa Spearman korelasyonun yanında, Kendall tau b analizi de yapılır. Spearman rho büyüklük sırasına sokulmuş değerler arasındaki Pearson momentler çarpımını temsil ederken, Kendall tau olasılığı temsil eder. Yani iki değişkende aynı sırada yer alan(uyuşan) verilerin gözlenme olasılığı ile farklı sırada yer alan (uyuşmayan) verilerin gözlenme olasılığı arasındaki farktır. İki serideki veriler uyuşuyorsa tau-b +1 değerini alır. Uyuşma yoksa tau-b -1 olur.

Kendall tau-b genelde kare tablolar için, Kendall-tau-c ise dikdörtgen tablolar için kullanılır. Tau-a ise uyuşan ve uyuşmayan çiftler arasındaki farkın toplam çift sayısına bölünmesiyle bulunur.

Kendall tau_b bağlı (ties) sıralığa sahip (aynı gözlemli olan değerlerin sıra puanlarının ortalaması alınır) ölçek verileri için uygundur. Bu analiz verilerin normal dağılış göstermediği ve n<20 olduğu durumlarda daha iyi sonuç vermektedir.

Kendall tau katsayıları aşağıdaki gibi yorumlanır:

>0,5	**Yüksek ilişki**
0,36-0,49	**Önemli ilişki**
0,20-0,35	**Orta derecede ilişki**
0,10-0,19	**Düşük ilişki**
<0,1	**İlişki yok**

- İki seri verildiğinde birinci seri doğal sırada olacak şekilde (küçükten-büyüğe) iki seri yeniden dizilir.

-İkinci serideki her bir Yi değerine bakarak bu değeri izleyen değerlerden kaç tanesi Yi'den büyük (ai) ve kaç tanesi Yi'den küçük (bi) olduğu sayılır. Bu işlem her satır için yapılarak ai ve bi serileri oluşturulur.

- Na=Σai ve Nb= Σbi değerleri el de edilir.

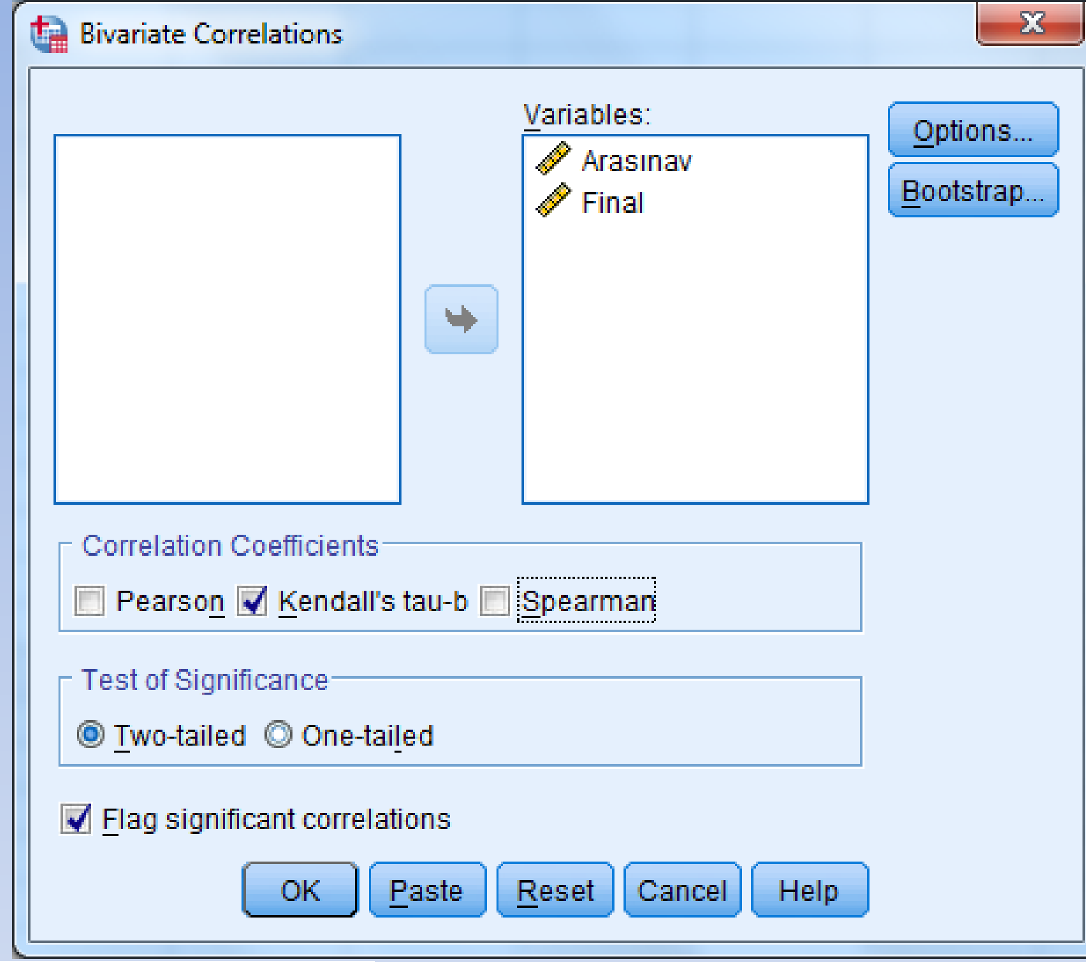
$$\tau = \frac{N_a - N_b}{N(N-1)/2}$$

Örnek. 7 öğrencinin bir dersin arasınav ve final sınavı notları aşağıdaki gibidir. Verilerin normal dağılış göstermediği varsayılarak iki sınav notu arasındaki korelasyon katsayısını bulunuz.

Arasınav	Final	Xi	Yi	ai	bi
55	65	45	55	5	1
72	68	55	65	4	1
63	54	63	54	4	0
45	55	70	75	2	1
70	75	72	68	2	0
75	90	75	90	0.5	0.5
80	90	80	90	0	0
				17.5	3.5

$$\tau = \frac{N_a - N_b}{N(N-1)/2} = \frac{17.5 - 3.5}{7(7-1)/2} = 0.67$$

	Arasınav	Final
1	55	65
2	72	68
3	63	54
4	45	55
5	70	75
6	75	90
7	80	90

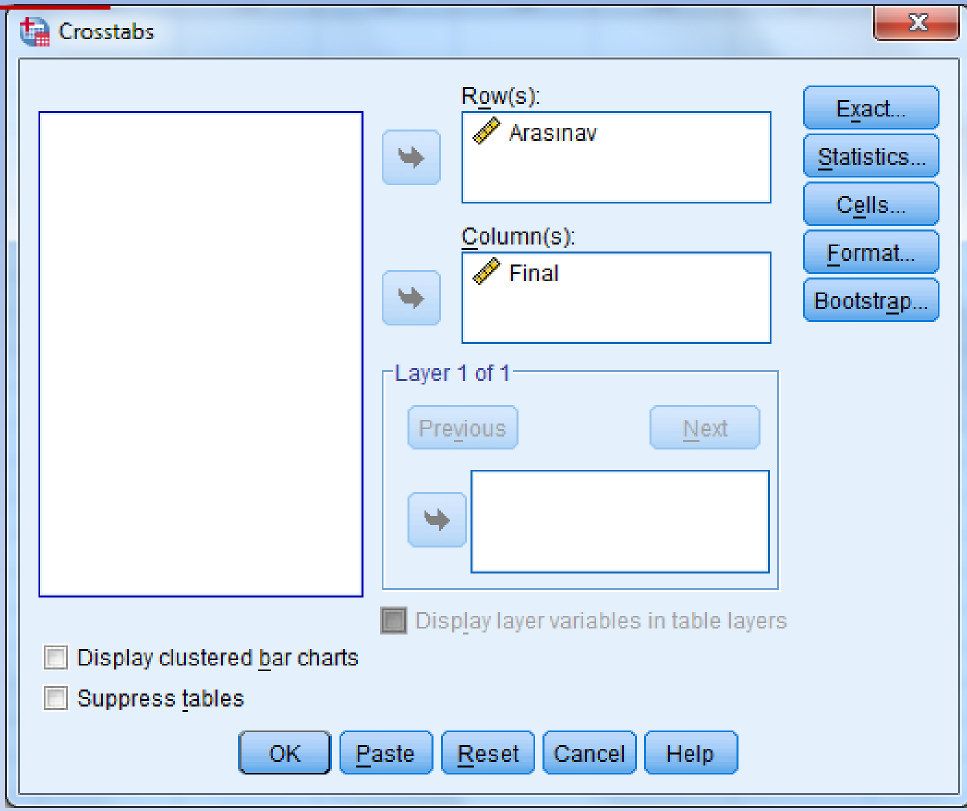


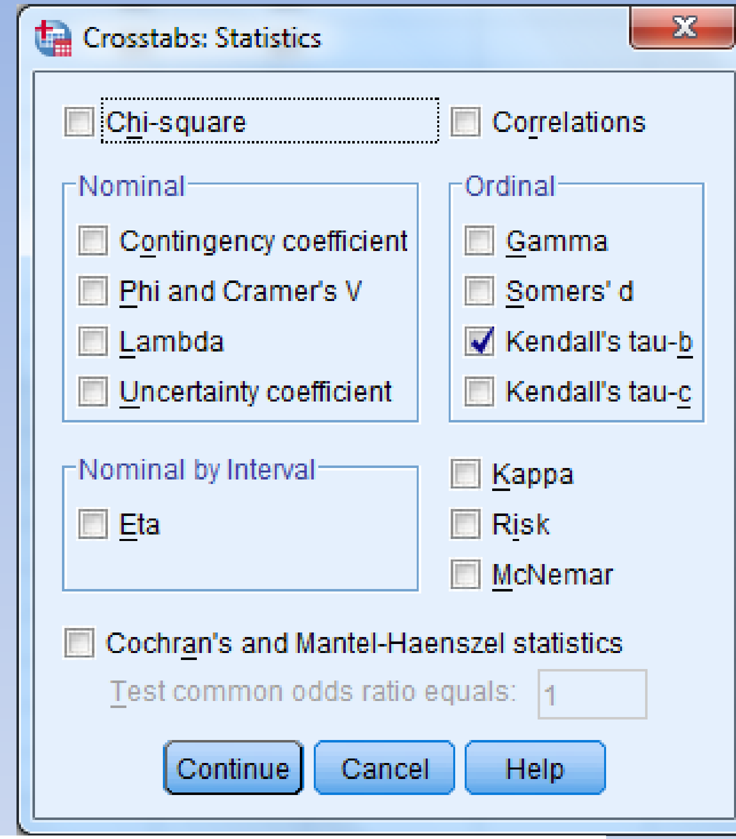
Correlations

			Arasınav	Final
Kendall's tau_b	Arasınav	Correlation Coefficient	1,000	,683*
		Sig. (2-tailed)	.	,033
		N	7	7
	Final	Correlation Coefficient	,683*	1,000
		Sig. (2-tailed)	,033	.
		N	7	7

*. Correlation is significant at the 0.05 level (2-tailed).

II.YOL





Symmetric Measures

		Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Ordinal by Ordinal	Kendall's tau-b	,683	,134	5,715	,000
N of Valid Cases		7			

a. Not assuming the null hypothesis.

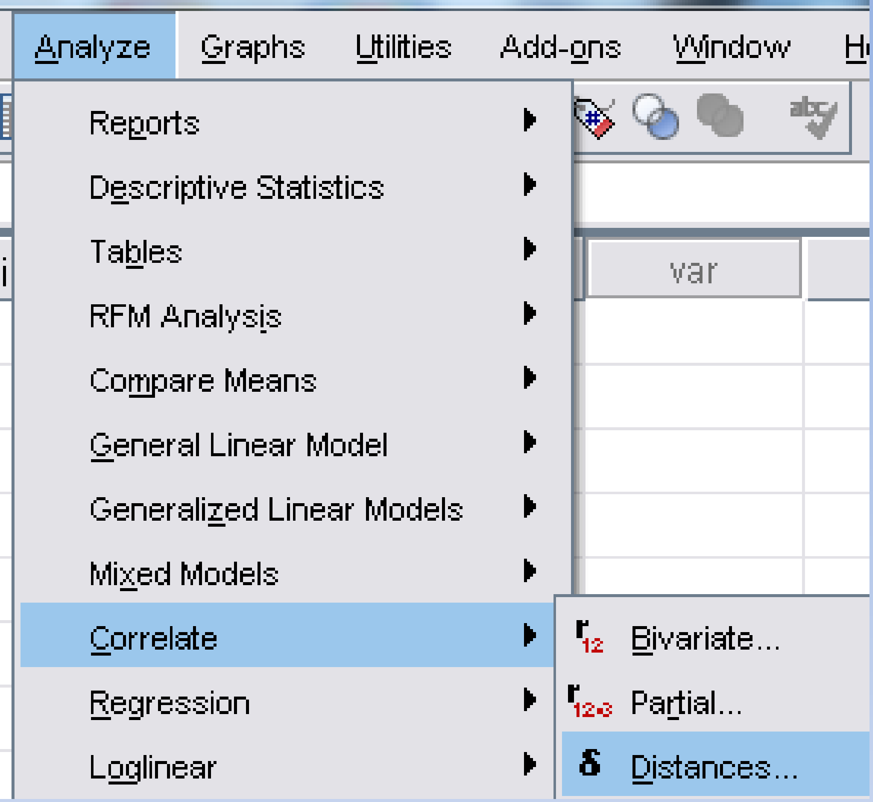
b. Using the asymptotic standard error assuming the null hypothesis.

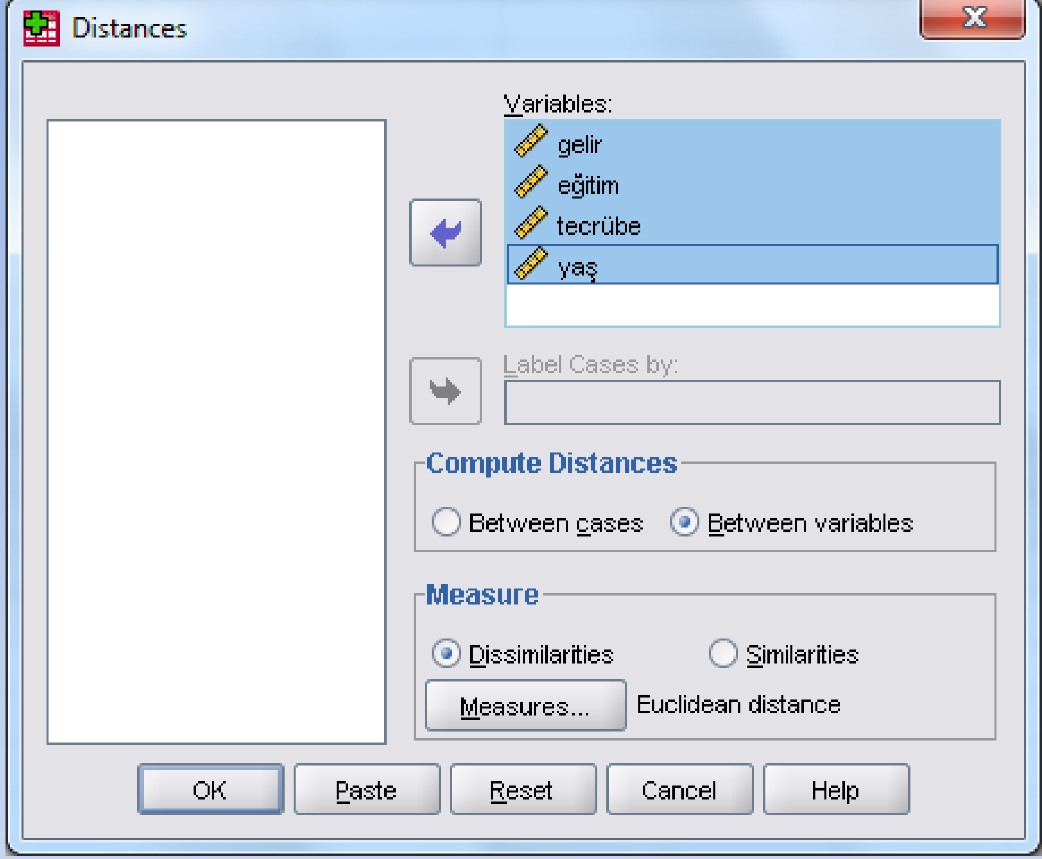
Distances (Uzaklıklar)

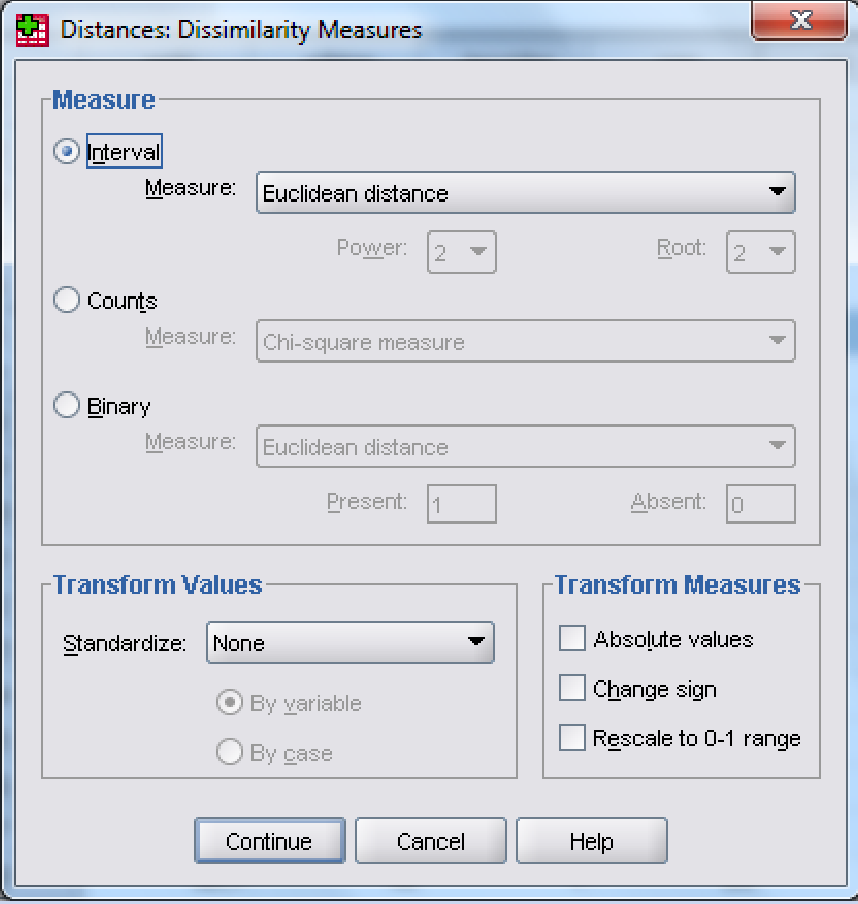
Bu yöntem değişkenler arasındaki mesafelerin ölçülmesini amaçlar. Bağımsız değişkelerin bağımlı değişken ile olan uzaklığının fazla olması onların bağımlı değişkeni açıklama oranını düşürür.

Örnek. **Yıllık gelir (bağımlı değişken) ile eğiti, tecrübe, yaş (bağımsız değişkenler) arasındaki ilişki araştırılmış ve yandaki sonuçlar elde edilmiştir.**
Uzaklık değerlerine göre hangi açıklayıcı değişken gelir üzerinde daha çok etkiye sahiptir?

	gelir	eğitim	tecrübe	yaş
1	5,0	2	9	29
2	9,7	4	18	50
3	28,4	8	21	41
4	8,8	8	12	55
5	21,0	8	14	34
6	26,6	10	16	36
7	25,4	12	16	61
8	23,7	12	9	29
9	22,5	12	18	64
10	19,5	12	5	30
11	21,7	12	7	28
12	24,8	13	9	29
13	30,1	14	12	35
14	24,8	14	17	59
15	28,5	15	19	65
16	26,0	15	6	30
17	38,9	16	17	40
18	22,1	16	1	23
19	33,1	17	10	58
20	48,3	21	17	44


Analyze
Graphs
Utilities
Add-ons
Window
var
Reports
Descriptive Statistics
Tables
RFM Analysis
Compare Means
General Linear Model
Generalized Linear Models
Mixed Models
Correlate
Regression
Loglinear
Bivariate...
Partial...
Distances...


Distances
Variables:
gelir
eğitim
tecrübe
yaş
Label Cases by:
Compute Distances
Between cases
Between variables
Measure
Dissimilarities
Similarities
Measures...
Euclidean distance
OK
Paste
Reset
Cancel
Help



Proximity Matrix

	Euclidean Distance			
	gelir	eğitim	tecrübe	yaş
gelir	,000	62,164	68,035	105,444
eğitim	62,164	,000	32,465	147,482
tecrübe	68,035	32,465	,000	139,682
yaş	105,444	147,482	139,682	,000

This is a dissimilarity matrix

Yıllık gelirle eğitim ve tecrübe arasındaki uzaklık birbirine yakın, yaş ile olan uzaklık ise oldukça yüksektir. Yaş değişkeni geliri en az etkileyen değişkendir.

Correlations

		gelir	eğitim	tecrübe	yaş
gelir	Pearson Correlation	1	,846**	,266	,102
	Sig. (2-tailed)		,000	,258	,669
	N	20	20	20	20
eğitim	Pearson Correlation	,846**	1	-,107	,098
	Sig. (2-tailed)	,000		,654	,680
	N	20	20	20	20
tecrübe	Pearson Correlation	,266	-,107	1	,676**
	Sig. (2-tailed)	,258	,654		,001
	N	20	20	20	20
yaş	Pearson Correlation	,102	,098	,676**	1
	Sig. (2-tailed)	,669	,680	,001	
	N	20	20	20	20

**. Correlation is significant at the 0.01 level (2-tailed).

Korelasyon matrisinden görüldüğü gibi yaş değişkeni ile gelir arasındaki korelasyon oldukça düşüktür.

REGRESYON ANALİZİ

Regresyon analizinin temelinde; gözlenen bir olayın değerlendirilirken, hangi olayların etkisi içinde olduğunun araştırılması yatmaktadır. Bu olaylar bir veya birden çok olacağı gibi dolaylı veya direkt etkileniyor da olabilirler.

Regresyon analizi yapılırken, gözlem değerlerinin ve etkilenilen olayların bir matematiksel gösterimle yani bir fonksiyon yardımıyla ifadesi gerekmektedir. Kurulan bu modele regresyon modeli denilmektedir. Regresyon analizi incelenirken, genellikle konusunu oluşturan, etkilendiği olaylara değişkenler adı verilir ve bu değişkenlerin yer alacağı matematiksel model incelenir.

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

Regresyon modelinin kullanılması, ilgilenilen olayla ilgili olarak, bir sebep-sonuç ilişkisi bulunması gerekmektedir. Örneğin 1990-1997 yılları arasındaki hisse senedi fiyatlarını incelersek, seçilen zaman aralığında bir matematiksel model kurma gereği vardır ve bu modelde bir sebep, sonuç ilişkisi aranmaktadır. Sebep, hisse senedinin fiyatını yükselten veya düşüren unsurlardır. Faiz oranları, ekonomik nedenler, enflasyon oranları vs. olarak incelenebilir. Sonuç ise hisse senedinin fiyatının değişmesidir.

Sebep-sonuç ilişkisi, regresyon modeli kurulurken, bağımlı ve bağımsız değişkenler olarak anlatılmaktadır. Yukarıdaki hisse senedi fiyatı sonuç olan bağımlı değişken, faiz oranları, ekonomik nedenler, enflasyon oranları vs. sebep olan bağımsız değişkenlerdir. Regresyon analizi yapılırken kurulan matematiksel modelde yer alan değişkenler bir bağımlı değişken ve bir veya birden çok bağımsız değişkenden oluşmaktadır.

Bağımsız değişkenler kurulacak modelde bir değişkenli olarak ele alınırsa basit doğrusal regresyon, birden fazla bağımsız değişkenli olarak alınırsa çoklu regresyon modeli konusunu oluşturmaktadır.

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

Regresyon modelinde Y ve ε rasgele değişkenler, X ise rasgele değişken değildir.

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

X ile Y arasında doğrusal ilişkinin olması regresyon hattının bir doğru şeklinde olmasından çok parametrelerin modelde doğrusal bir ilişki içinde olması ile ilgilidir.

$$E(Y) = \beta_0 + \beta_1 X \Rightarrow E(Y \mid X) = \beta_0 + \beta_1 X$$

$$Y \sim N(\beta_0 + \beta_1 X, \sigma^2)$$

$$\hat{Y} = E(\hat{Y}) = \hat{\beta}_0 + \hat{\beta}_1 X$$

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \qquad , \quad i = 1,2,\ldots,n$$

β_0: Doğrunun Y eksenini kestiği nokta. X'in sıfır değerine karşılık Y'nin alacağı değer.
β_1: Eğim, hız, regresyon katsayısı. X'in 1 birim değişmesi halinde Y'nin kaç birim ve ne yönde değişeceğini gösterir.

$H_0 : \beta_0 = 0$ Doğrunun merkezden (orjinden) geçmesini test eder.

$H_0 : \beta_1 = 0$ Doğrunun X eksenine paralel olup olmadığını (eğimin önemli olup olmadığını) test eder.

En Küçük Kareler Yöntemi (EKK)

$$\sum \varepsilon_i = \sum (Y_i - \hat{Y}_i) = 0$$

$$\sum \varepsilon_i^2 = \sum (Y_i - \hat{Y}_i)^2 = \sum (Y_i - \beta_0 - \beta_1 X)^2 = \min$$

$$\frac{\partial}{\partial \beta_0} \sum (Y_i - \beta_0 - \beta_1 X_i)^2 = 0 \quad \Rightarrow -2\sum (Y_i - \beta_0 - \beta_1 X_i)^2 = 0$$

$$\frac{\partial}{\partial \beta_1} \sum (Y_i - \beta_0 - \beta_1 X_i)^2 = 0 \quad \Rightarrow -2\sum X_i (Y_i - \beta_0 - \beta_1 X_i) = 0$$

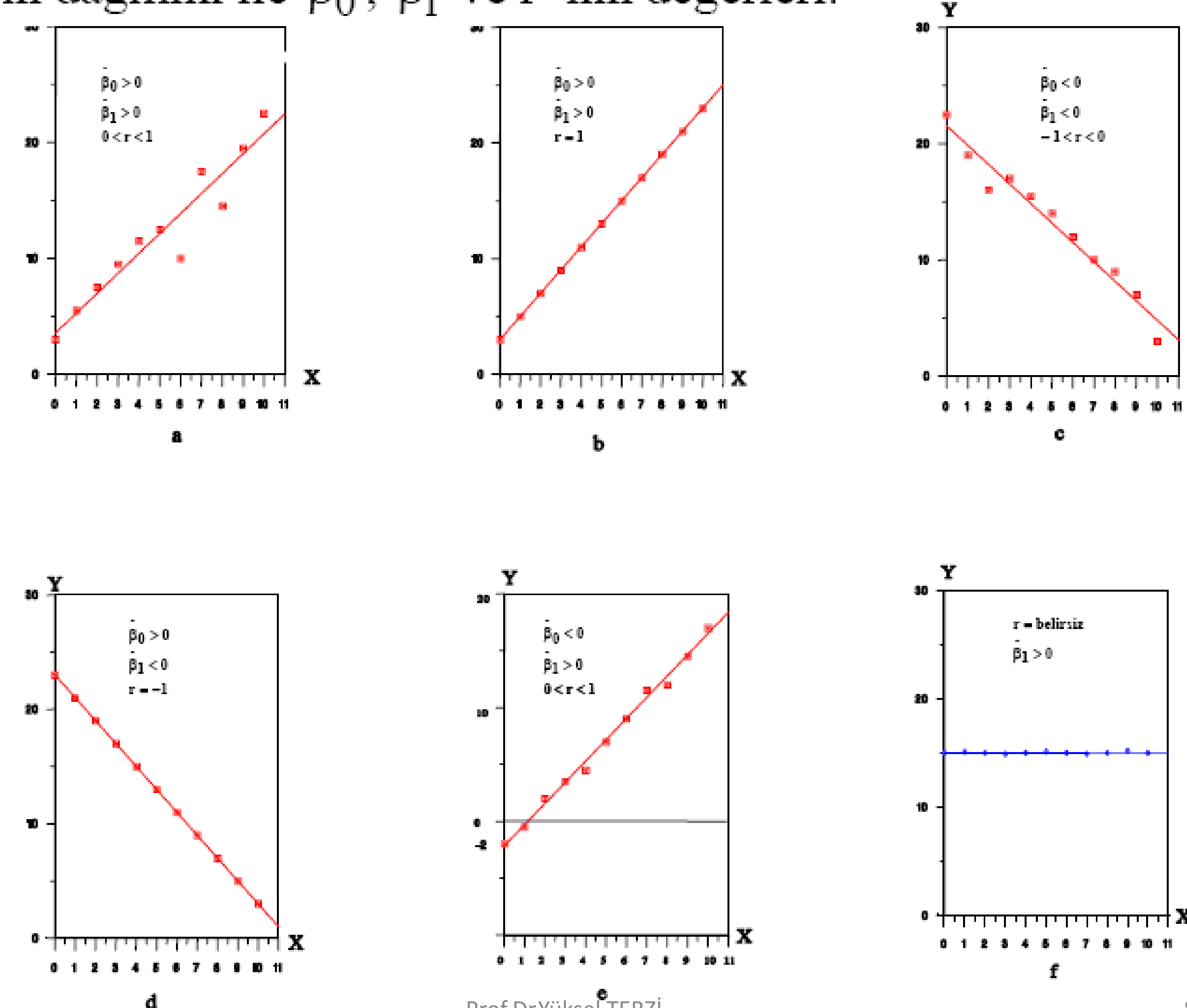
$$\boxed{\begin{aligned} \sum Y_i &= n\beta_0 + \beta_1 \sum X_i \\ \sum X_i Y_i &= \beta_0 \sum X_i + \beta_1 \sum X_i^2 \end{aligned}}$$

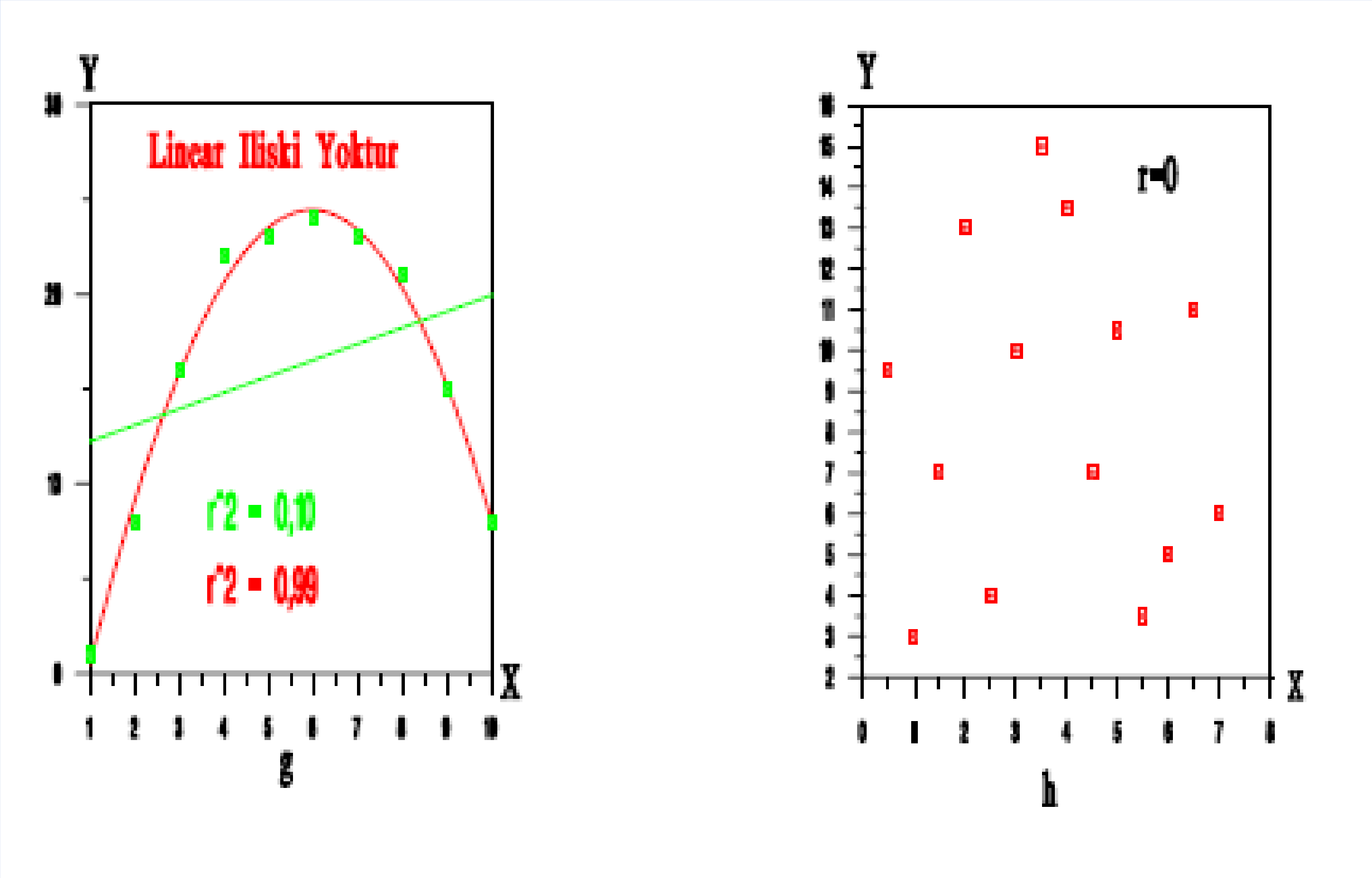
Normal denklemler

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}$$

$$\hat{\beta}_1 = \frac{\sum (Y_i - \overline{Y})(X_i - \overline{X})}{\sum (X_i - \overline{X})^2} = \frac{\sum X_i Y_i - \frac{(\sum X_i)(\sum Y_i)}{n}}{\sum X_i^2 - \frac{(\sum X_i)^2}{n}}$$

Noktaların dağılımı ile $\hat{\beta}_0$, $\hat{\beta}_1$ ve r ‘nin değerleri.




Y
Linear Iliski Yoktur
r^2 = 0,10
r^2 = 0,99
X
g
Y
r=0
X
h

Eğrisel Modeller

(1) Linear	$E(Y_t) = \beta_0 + \beta_1 t$
(2) Logarithmic	$E(Y_t) = \beta_0 + \beta_1 \ln(t)$
(3) Inverse	$E(Y_t) = \beta_0 + \beta_1 / t$
(4) Quadratic	$E(Y_t) = \beta_0 + \beta_1 t + \beta_2 t^2$
(5) Cubic	$E(Y_t) = \beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 t^3$
(6) Compound	$E(Y_t) = \beta_0 \beta_1^t$
(7) Power	$E(Y_t) = \beta_0 t^{\beta_1}$
(8) S	$E(Y_t) = \exp(\beta_0 + \beta_1 / t)$
(9) Growth	$E(Y_t) = \exp(\beta_0 + \beta_1 t)$
(10) Exponential	$E(Y_t) = \beta_0 e^{\beta_1 t}$
(11) Logistic	$E(Y_t) = \left(\frac{1}{u} + \beta_0 \beta_1^t \right)^{-1}$

BASİT DOĞRUSAL REGRESYON MODELİ VARSAYIMLARI

1. Bağımsız değişkenin değerleri sabit kabul edilir. Bağımlı değişkenin değerleri ise rasgeledir. Fakat korelasyon analizinde her iki değişkende tesadüfidir.
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. Bu alt küme değerlerinin (Y'lerin) varyansları eşittir.
5. Bağımlı ve bağımsız değişkenler arasındaki ilişki doğrusaldır.

Hata Terimi (ε_i) Varsayımları

1. Rassal hataların beklenen değeri sıfırdır. $E(\boldsymbol{\varepsilon})=0$

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.
   $Cov\ (\varepsilon,X_i)=0$