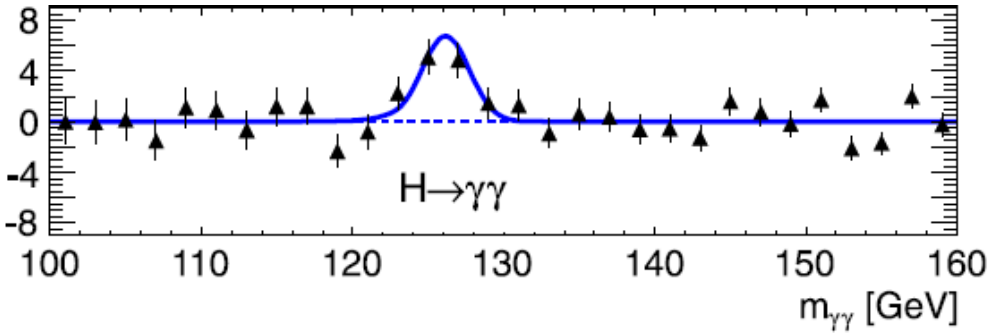
YBS514
Mühendislikte İstatistik Yöntemler

Bölüm 9

Korelasyon, Regresyon ve Eğri Uydurma



http://www1.gantep.edu.tr/~bingul/stat

Gaziantep Üniversitesi

Yönetim Bilişim Sistemleri

Tezsiz Yüksek Lisans Programı

Aralık 2020

İçerik

- Korelasyon
- Doğrusal Regresyon
- Eğri Uydurma

Korelasyon
(Correlation)

İki değişkenli veriler için **korelasyon katsayısı** (ρ), söz konusu değişkenler arasında doğrusal bir ilişki olup olmadığı konusunda bize bilgi verir.

Boyutu *n* olan iki değişkenli bir veri düşünelim:

$Z = \{ (x_1,y_1),\ (x_2,y_2),\ (x_3,y_3),\ \ldots ,(x_n,y_n)\ \}$

Korelasyon katsayısı aşağıdaki formülle tanımlıdır:

$$\rho = \frac{\overline{xy} - \bar{x}.\bar{y}}{s_x s_y}$$

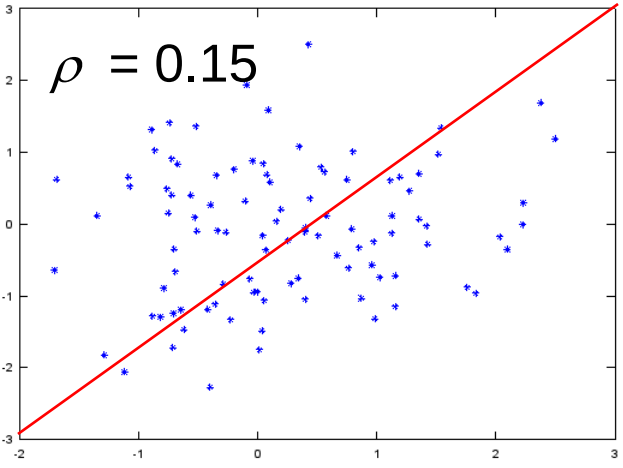
$$\overline{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i \qquad \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

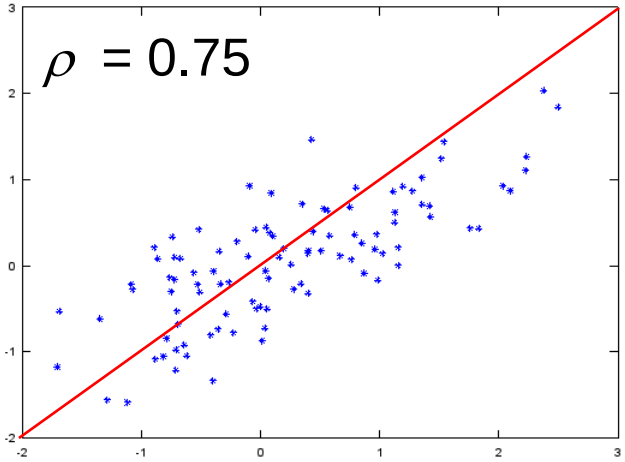
$$s_x = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad s_y = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

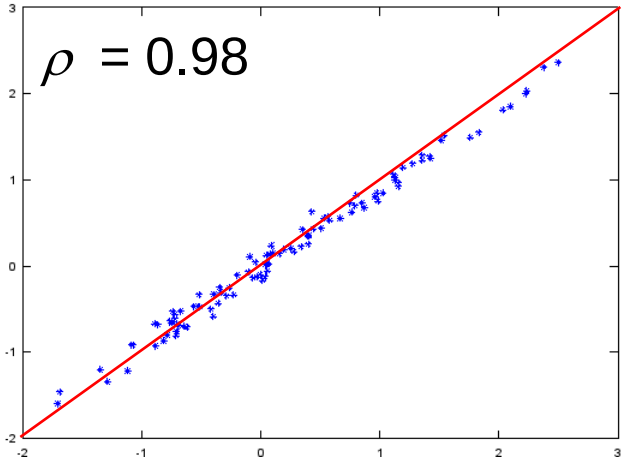
$$\rho = \frac{\overline{xy} - \overline{x}.\overline{y}}{s_x s_y}$$

Bu her zaman doğrudur.

$$-1 \leq \rho \leq 1$$



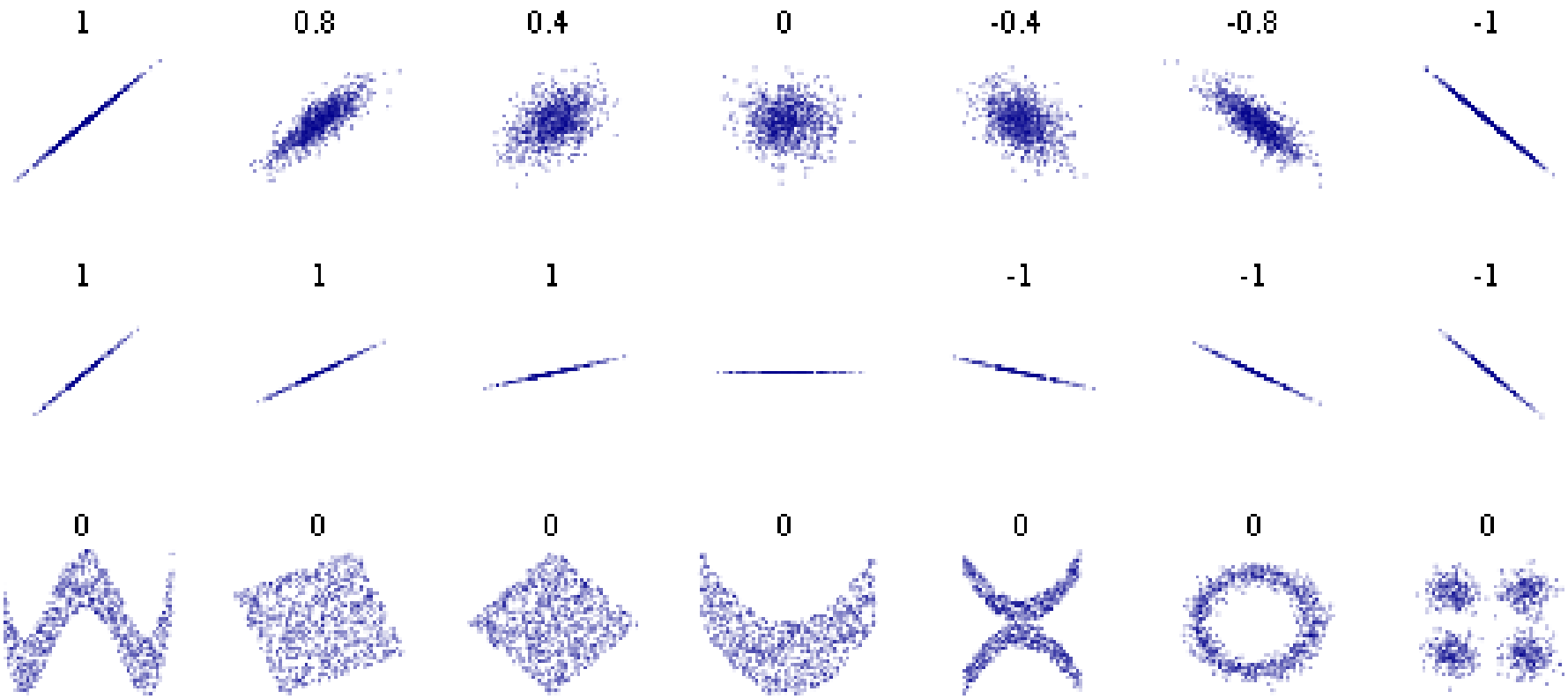




Eğer bir korelasyon(=doğrusal ilişki) yoksa $\rho = 0$

Eğer tam korelasyon(=doğrusal ilişki) varsa $\rho = \pm 1$

$-1 \leq \rho \leq 1$



Örnek 1:

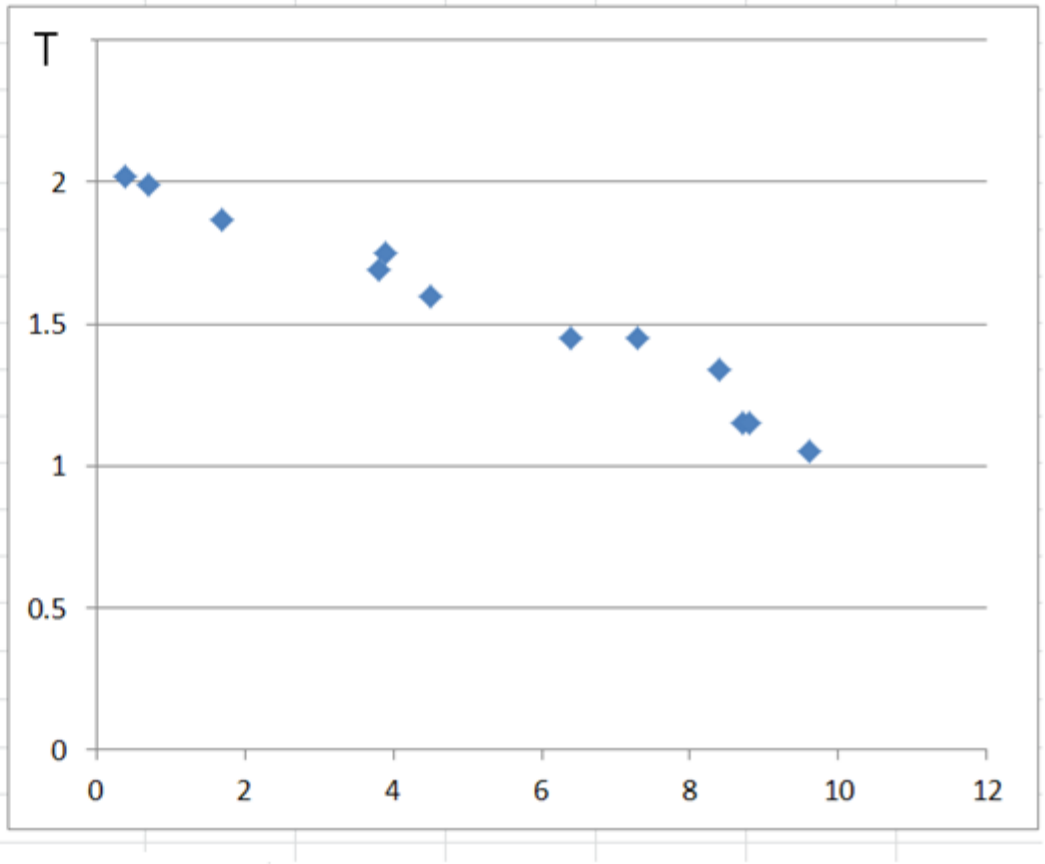
Sanayi tipi soğutucular uzun süre bozulmadan gıda ürünleri saklayabilir. Bir mühendis, soğutucu içindeki rüzgar hızı (R) ile soğutucu içindeki sıcaklık (T) arasındaki bir korelasyon olup olmadığını anlamak için yanda verilen veriyi topluyor.

(a) Verinin dağlım grafiğini çizin.
(b) Korelasyon katsayısını hesaplayın.
(c) Sonucu yorumlayın.

R (m/s)	T(°C)
3.8	1.69
8.4	1.34
7.3	1.45
3.9	1.75
1.7	1.87
9.6	1.05
4.5	1.60
6.4	1.45
0.4	2.02
8.7	1.15
8.8	1.15
0.7	1.99

Örnek1 - devam:

	A	B	C	D
1		R (m/s)	T (oC)	R*T
2		3.8	1.69	6.422
3		8.4	1.34	11.256
4		7.3	1.45	10.585
5		3.9	1.75	6.825
6		1.7	1.87	3.179
7		9.6	1.05	10.08
8		4.5	1.6	7.2
9		6.4	1.45	9.28
10		0.4	2.02	0.808
11		8.7	1.15	10.005
12		8.8	1.15	10.12
13		0.7	1.99	1.393
14	**Toplam**	**64.2**	**18.51**	**87.153**
15	**Ortalama**	**5.35**	**1.5425**	**7.26275**
16	**std.sapma**	**3.30385**	**0.331995**	**3.672633**
17				
18	rho	-0.9022		



B14 : =topla(B2:B13)

B15 : =ortalama(B2:B13)

B16 : =stdsapma(B2:B13)

rho = -0.9 R ve T arasında negatif yönde güçlü bir doğrusal korelasyon olduğunu göstermektedir.

Sen Çöz

Bir hastanede doğan 250 bebeğin ağırlıkları belli aylarda ölçülmüş ve yandaki tabloya kaydedilmiştir. Aşağıdaki ikili veriler arasındaki korelasyon katsayılarını hesaplayın.

	Ağırlık (kg)	
Ay	Erkek	Kız
0	3.4	3.2
1	4.4	4.1
2	5.5	5.0
3	6.4	5.7
4	7.1	6.4
5	7.7	6.9
6	8.3	7.5
9	9.4	8.6
12	10.2	9.4

(a) Ay ve erkek bebek

(b) Ay ve kız bebek

(c) Erkek bebek ve kız bebek

Doğrusal Regresyon
(Linear Regression)

Problem

- Gerçek veriler genellikle kesiklidir. Bazı uygulamalarda ara değerleri tahmin etmek gerekir.

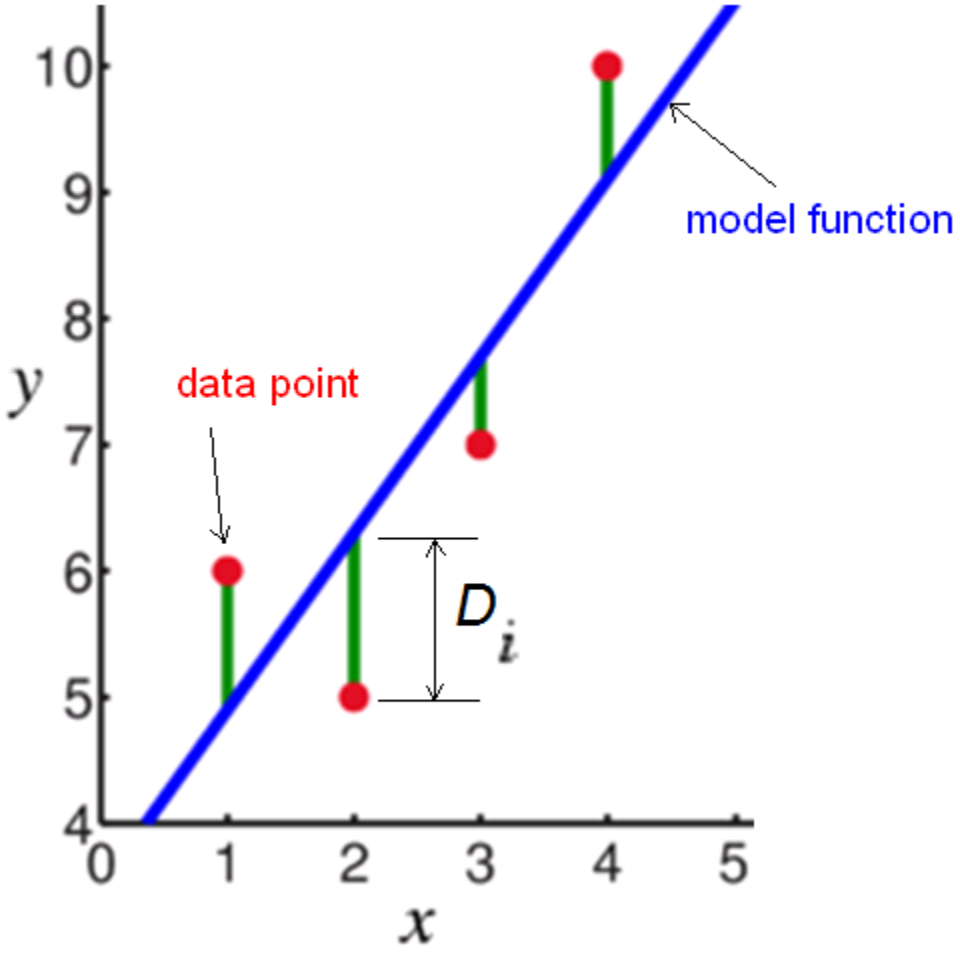
Bir otomobilin güvenli durma mesafesi (d) onun hızının (v) bir fonksiyonudur. Bu ilişkiyi belirlemek için aşağıdaki veriler toplanmıştır.

v(km/sa)	d(m)
24	4.8
32	6.0
40	10.2
48	12.0
64	18.0
80	27.0

Buna göre araç 50 km/sa hızla giderken güvenli durma mesafesi nedir?

- Bu kısımda en küçük kareler yöntemi kullanılarak doğrusal regresyon analizi ele alınacaktır.

Deneysel veriler için en uygun doğruyu bulmaya regresyon analizi denir.



Deneysel veriler

x	y
---	---
x_1	y_1
x_2	y_2
.	.
.	.
.	.
x_n	y_n

$$Hata = D_i = y_i - y_{\text{model}}$$

Amaç: deneysel veri noktalarını (x_i, y_i),

$$y = ax+b$$

formundaki bir fonksiyona uydurmak (a ve b değerleri bulmak).

Hataların karelerini toplayalım

$$S = \sum_{i=1}^{n} D_i^2 = \sum_{i=1}^{n} (y_i - y_{\text{model}})^2$$

$$S = \sum_{i=1}^{n} (y_i - ax_i - b)^2$$

Bu toplamı en küçük hale getirelim:

$$\frac{\partial S}{\partial a} = 0 \qquad \frac{\partial S}{\partial b} = 0$$

Biraz cebirden sonra, en uygun a ve b değerini elde etmiş oluruz:

$$a = \frac{n\sum x_i y_i - \left(\sum x_i\right)\left(\sum y_i\right)}{n\sum x_i^2 - \left(\sum x_i\right)^2} \qquad b = \frac{\left(\sum x_i^2\right)\left(\sum y_i\right) - \left(\sum x_i\right)\left(\sum x_i y_i\right)}{n\sum x_i^2 - \left(\sum x_i\right)^2}$$

Fit kalitesi (r^2 = regresyon katsayısı)

$$r^2 = \frac{S_t - S}{S_t}$$

$$S = \sum_{i=1}^{n} (y_i - ax_i - b)^2$$

$$S_t = \sum_{i=1}^{n} (y_i - y_m)^2 \quad \leftarrow \quad y_m = \frac{\sum_{i=1}^{n} y_i}{n}$$

İyi bir fit için

$$S \to 0$$

$$r^2 \to 1$$

Örnek 2:

Bir otomobilin güvenli durma mesafesi (d) onun hızının (v) bir fonksiyonudur. Bu ilişkiyi belirlemek için aşağıdaki veriler toplanmıştır.

v(km/sa)	d(m)
24	4.8
32	6.0
40	10.2
48	12.0
64	18.0
80	27.0

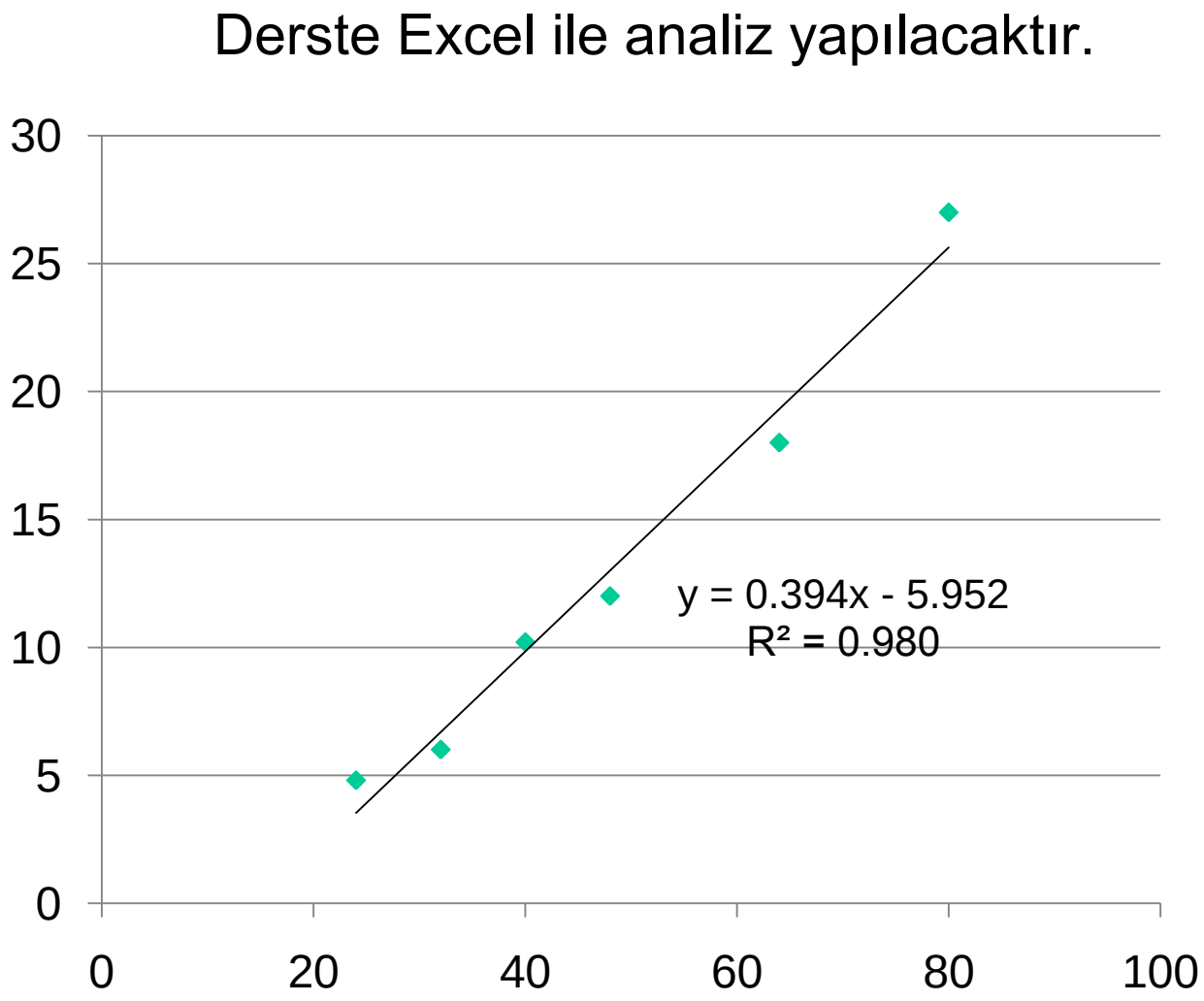
(a) Bu verileri y = ax + b biçimindeki bir fonksiyona fit edin.
(Bu veriler için hataların kare toplamlarının en küçük yapan a ve b değerlerini bulun)

(b) Fitin kalitesini belirleyen r^2 ve S değerlerini hesaplayın.

(c) Araç 50 km/sa hızla giderken güvenli durma mesafesi hesaplayın.

(c) Araç 100 km/sa hızla giderken güvenli durma mesafesi hesaplayın.

Örnek 2 -devam



Örnek 2 -devam

MATLAB'da polyfit fonksiyonu ile aynı sonuçlara ulaşılabilir.

```
>> x = [24 32 40 48 64 80];
>> y = [4.8 6 10.2 12 18 27];
>> p = polyfit(x,y,1)
p = 0.3949   -5.9529
```

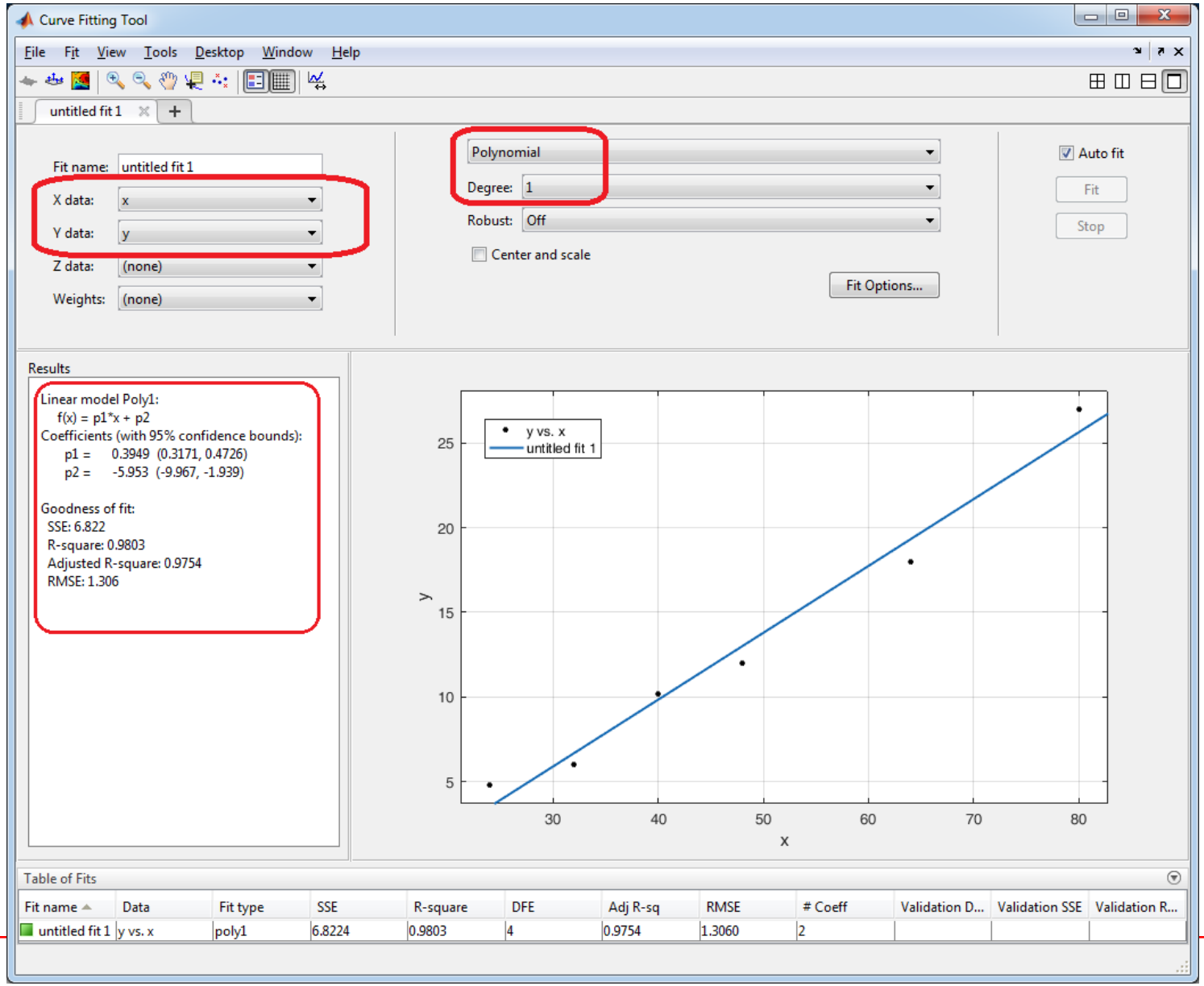
*y = p(1)*x + p(2) fonksiyonuna fit yap demek*

```
>> s = sum( (y-p(1)*x-p(2)).^2 )
s = 6.8224
>> st = sum((y-mean(y)).^2)
st = 346.0800
>> r2 = (st-s)/st
r2 = 0.9803
```

r^2 değeri. Bu fit, x-y arasında %98 doğrusal bir uygum olduğunu belirtir.

Örnek 2 –devam MATLAB'da cftool kullanmak

```
>> cftool
```



Örnek 2 -devam

V = 50 km / sa için durma mesafesi:

```
v = 50;
d = 0.394*v – 5.953 = 13.7470 m
```

V = 100 km / sa için durma mesafesi:

```
v = 100;
d = 0.394*v – 5.953 = 33.4470 m
```

Doğrusal Olmayan Regresyon (Nonlinear Regression)

Problem

Aşağıdaki problemde, doğrusal fit yaparak

$$d = 0.394*v - 5.953$$

bulduk. Ancak, v = 0 için d = −5.953 m, anlamsız bir sonuçtur!

Bir otomobilin güvenli durma mesafesi (d) onun hızının (v) bir fonksiyonudur. Bu ilişkiyi belirlemek için aşağıdaki veriler toplanmıştır.

v(km/sa)	d(m)
24	4.8
32	6.0
40	10.2
48	12.0
64	18.0
80	27.0

Buna göre araç 50 km/sa hızla giderken güvenli durma mesafesi nedir?

Çözüm

Fizik bize durma mesafesinin v^2 ile orantılı olduğunu söyler. Yani

$$\mathbf{d = a*v^2}$$

veya

$$\mathbf{d = a*v^2 + b*v}$$

olabilir. Bu iki denklemde de v = 0 olunca, d = 0 olur.
*(Son denklemdeki b*v terimi sürücünün tepki süresi ile ilgilidir. Tepki süresinde aracın alacağı yol hız ile orantılıdır).*

Bu durumda doğrusal fit yerine, doğrusal olmayan fit yöntemleri kullanılır.

Curve Fitting Tool

File Fit View Tools Desktop Window Help

untitled fit 1

Fit name: untitled fit 1

X data: x

Y data: y

Z data: (none)

Weights: (none)

Custom Equation

y = f(x)

= 1 a*x^2+b*x

Fit Options...

Auto fit

Fit

Stop

Results

General model:
f(x) = a*x^2+b*x
Coefficients (with 95% confidence bounds):
a = 0.002574 (0.001671, 0.003478)
b = 0.1275 (0.06911, 0.186)

Goodness of fit:
SSE: 2.145
R-square: 0.9938
Adjusted R-square: 0.9923
RMSE: 0.7324

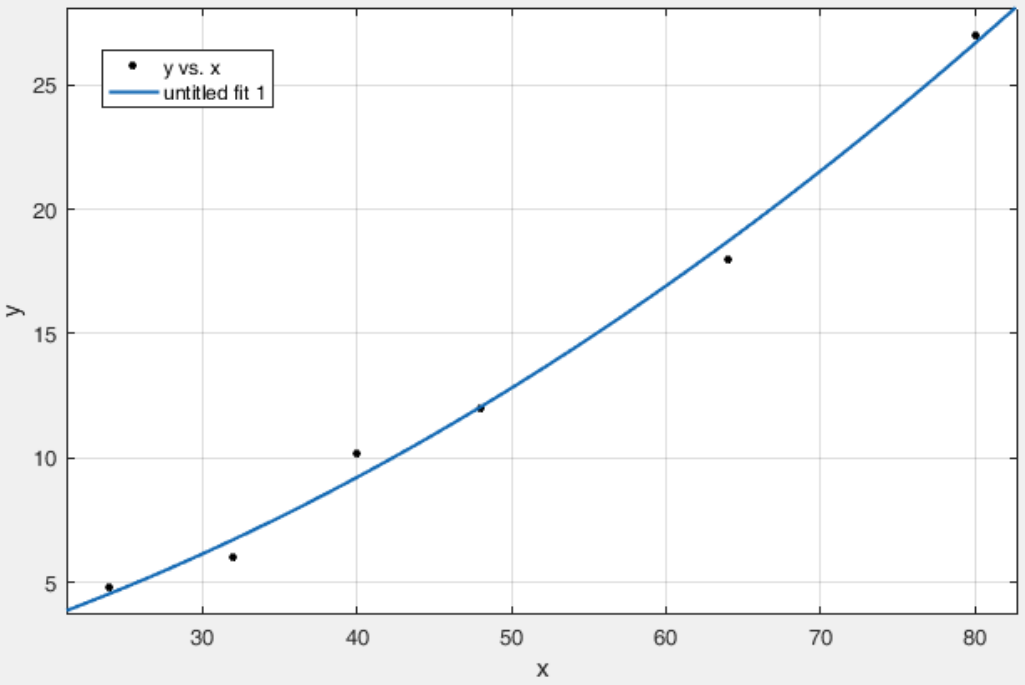


Table of Fits

Fit name	Data	Fit type	SSE	R-square	DFE	Adj R-sq	RMSE	# Coeff	Validation D...	Validation SSE	Validation R...
untitled fit 1	y vs. x	a*x^2+b*x	2.1455	0.9938	4	0.9923	0.7324	2			

Örnek 4

Aşağıdaki bilgisayar programı MATLAB'da yazılmış ve farklı n değerleri için programın çalışma süresi (*t*) not edilmiştir.

```
n = 10000;
tic; % zaman sayacını başlat
for i = 1:n-1
    for j = i+1:n
       a = i + j;
    end
end
t = toc; % sayacı bitir

fprintf('%d %f\n',n,t)
```

n	t (s)
10000	0.30
20000	1.03
30000	2.21
40000	3.93
50000	6.14
60000	8.76
70000	11.96
80000	15.62
90000	19.69

a) En ugun eğri nedir?

Excel'de 2. derece polinom kullanarak,

MATLAB'da 2. derece polinom kullanarak problemi çözün.

b) n = 200 000 için çalışma süresi nedir?