Makine Öğrenmesi

Doğrusal Bağlanım

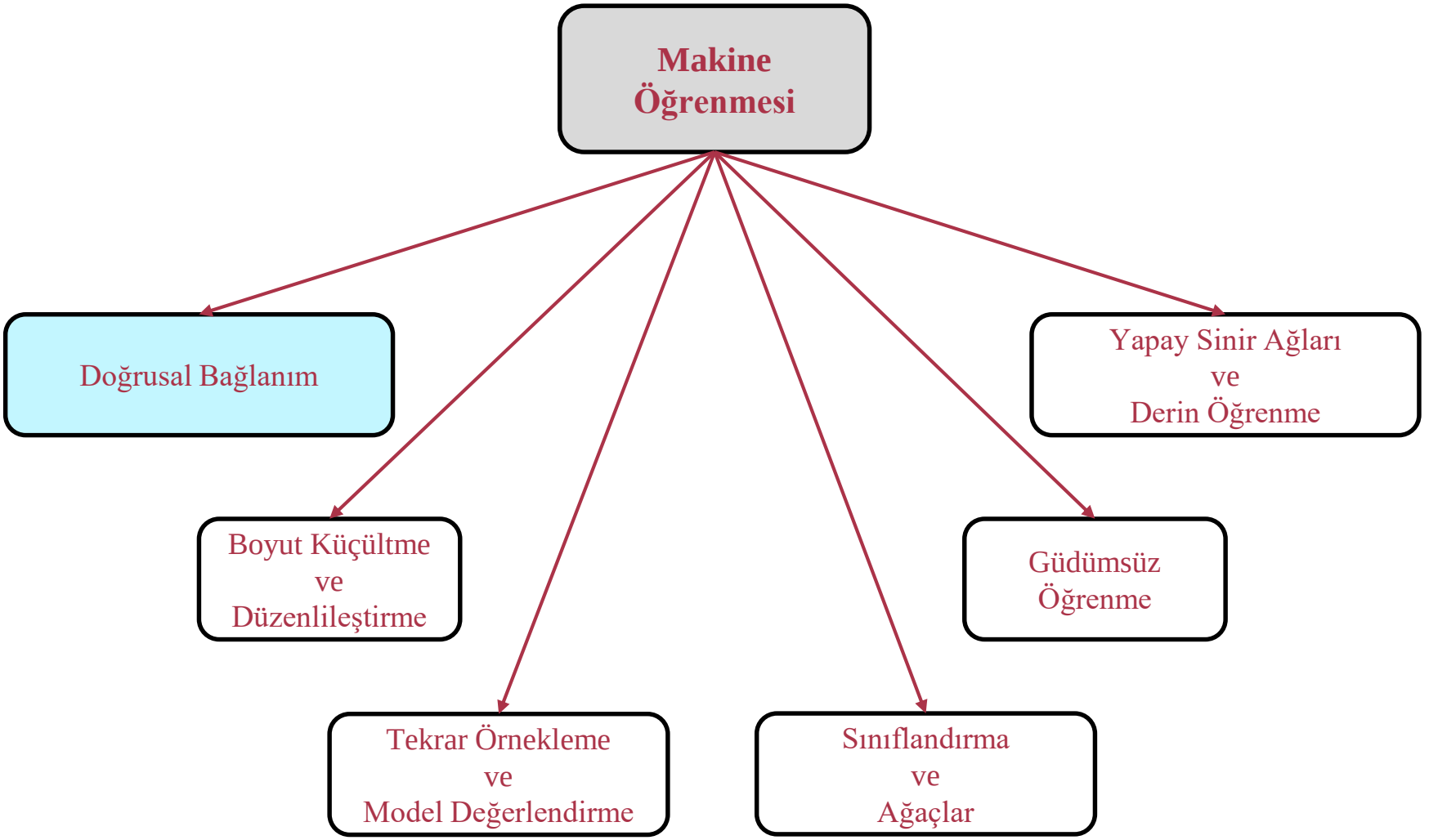
İlker Birbil ve Utku Karaca

Erasmus Üniversitesi Rotterdam

İstanbul'da Makine Öğrenmesi

27 Ocak – 2 Şubat, 2020

EUR
D4C
Bol Bilim
Veri Defteri
@sibirbil


Makine
Öğrenmesi
Doğrusal Bağlanım
Yapay Sinir Ağları
ve
Derin Öğrenme
Boyut Küçültme
ve
Düzenlileştirme
Güdümsüz
Öğrenme
Tekrar Örnekleme
ve
Model Değerlendirme
Sınıflandırma
ve
Ağaçlar

$$Y = f(X) + \epsilon \xrightarrow{\text{yaklaşık?}} \hat{Y} = \hat{f}(X)$$

$$Y \approx \hat{f}(X)$$

Basit Bağlanım (Simple Regression)

$$Y \approx \beta_0 + \beta_1 X$$

$$\text{satış} \approx \beta_0 + \beta_1 \text{TV}$$

Çoklu Bağlanım (Multiple Regression)

$$Y \approx \beta_0 + \beta_1 X_1 + \beta_1 X_2 + \cdots + \beta_p X_p$$

$$\text{satış} \approx \beta_0 + \beta_1 \text{TV} + \beta_2 \text{radyo} + \cdots + \beta_p \text{gazete}$$

Basit Bağlanım

$$Y \approx \beta_0 + \beta_1 X$$

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

?

eğitim verisi

$$\{(x_i, y_i) : 1, \ldots, n\}$$

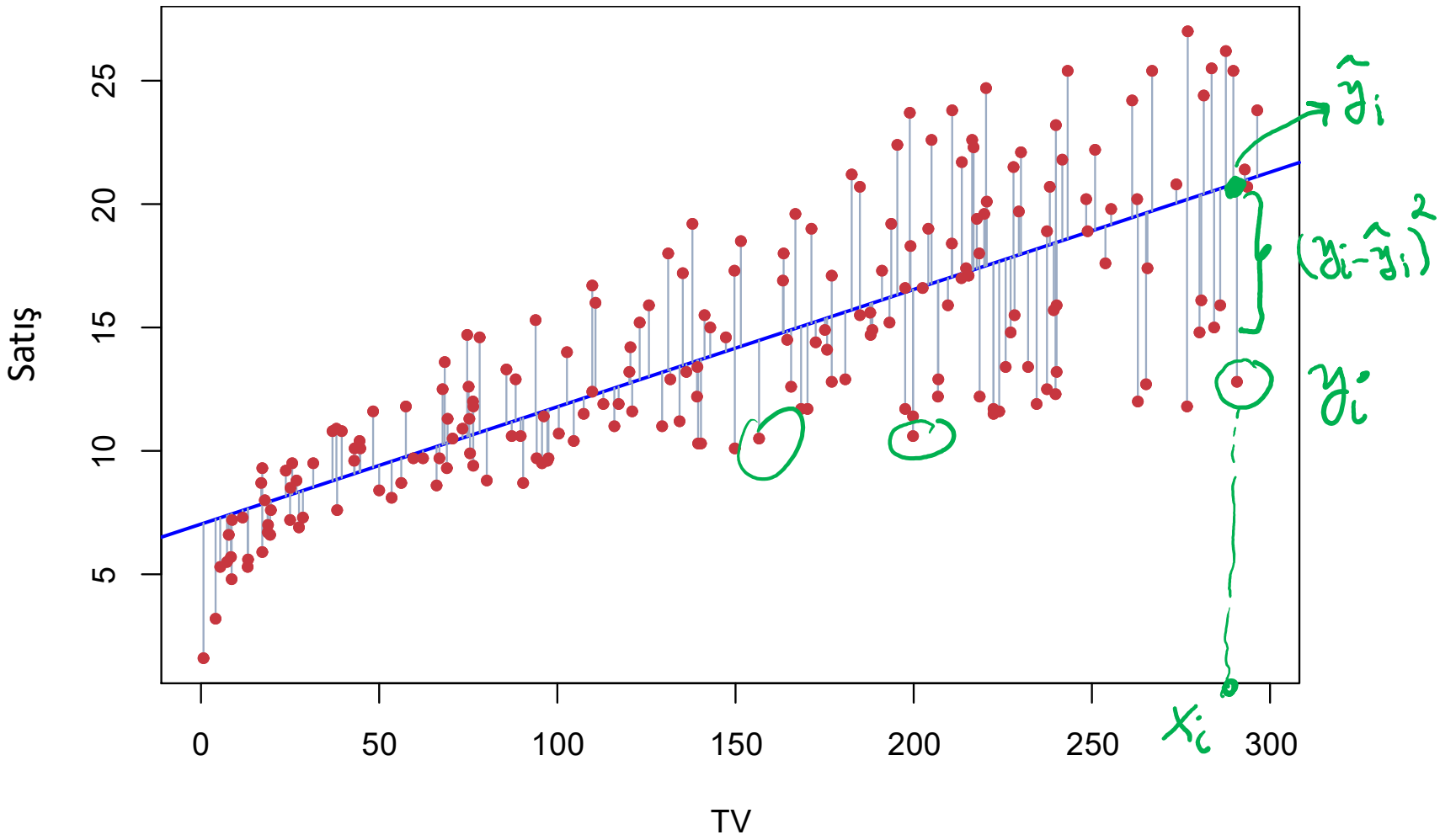
$$y_i \approx \underbrace{\hat{\beta}_0 + \hat{\beta}_1 x_i}_{\hat{y}_i}, \quad i = 1, \ldots, n$$

$$y_i \approx \underbrace{\hat{\beta}_0 + \hat{\beta}_1 x_i}_{\hat{y}_i}, \quad i = 1, \ldots, n$$

$$e_i = y_i - \hat{y}_i$$ *i*. kalıntı (*i*th residual)

Kalıntı Kareler Toplamı (Residual Sum of Squares)

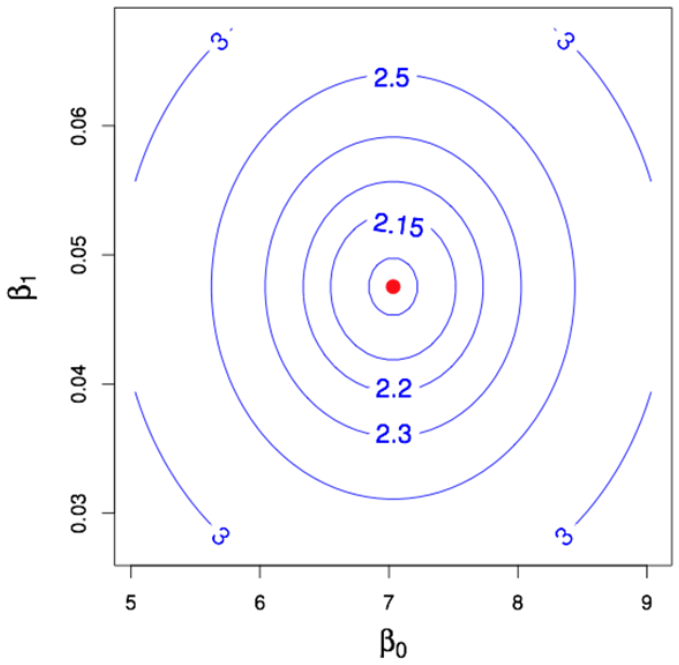
$$\text{KKT} = e_1^2 + e_2^2 + \cdots + e_n^2$$

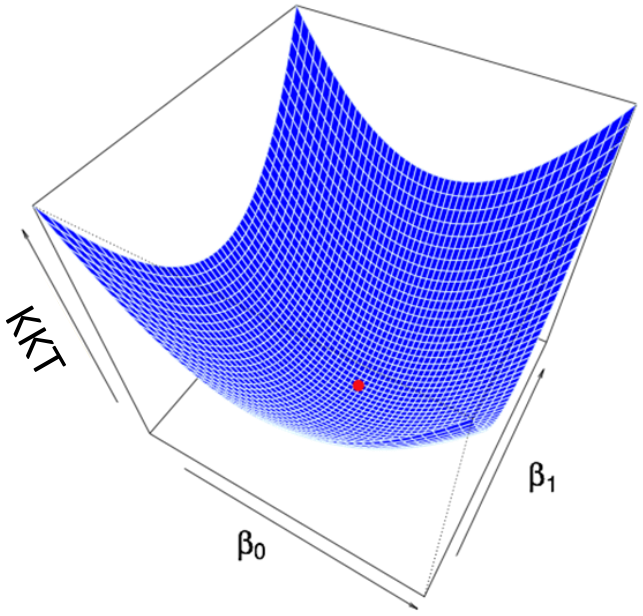

Satış
5
10
15
20
25
0
50
100
150
200
250
300
TV
$\tilde{y}_i$
$(y_i - \tilde{y}_i)^2$
y_i
x_i

En Küçük Kareler (EKK) Yöntemi (Least Squares Method)

$$\min_{\beta_0,\beta_1} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

Dışbükey Fonksiyon





En Küçük Kareler (EKK) Yöntemi (Least Squares Method)

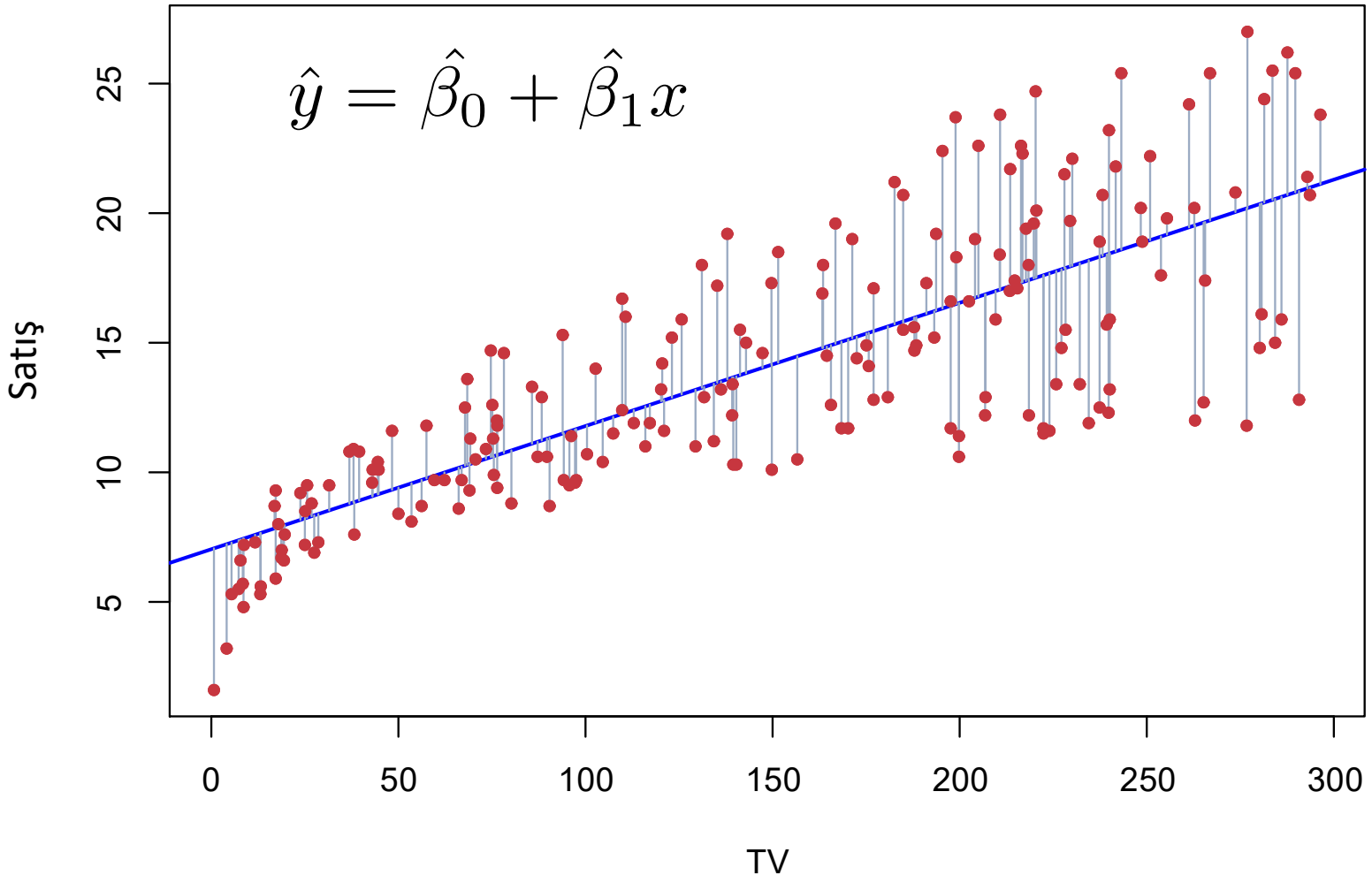
$$\min_{\beta_0,\beta_1} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

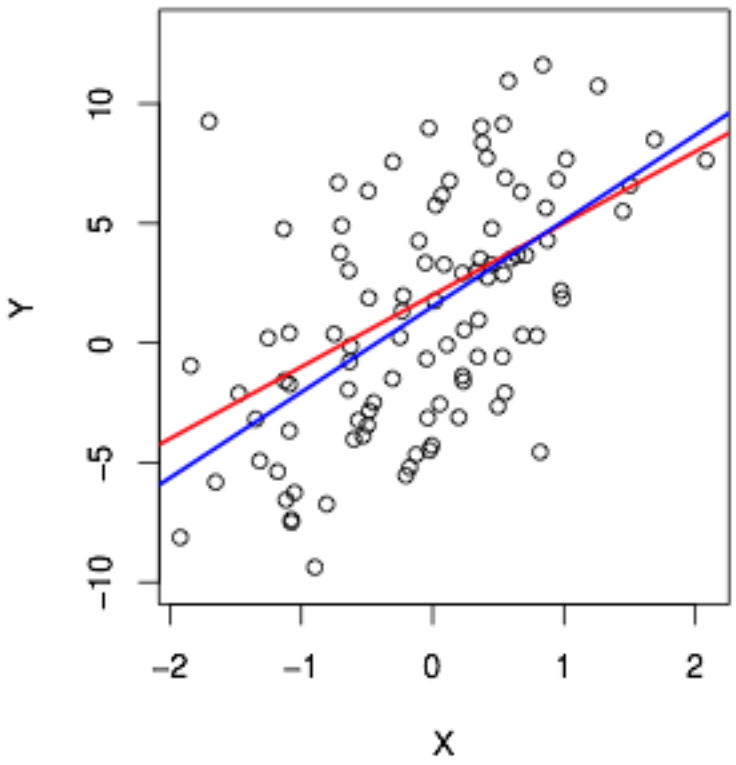

$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$
Satış
5
10
15
20
25
0
50
100
150
200
250
300
TV

$$Y = f(X) + \epsilon$$

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

anakütle bağlanım doğrusu
(population regression line)

$Y = 2 + 3X + \epsilon$

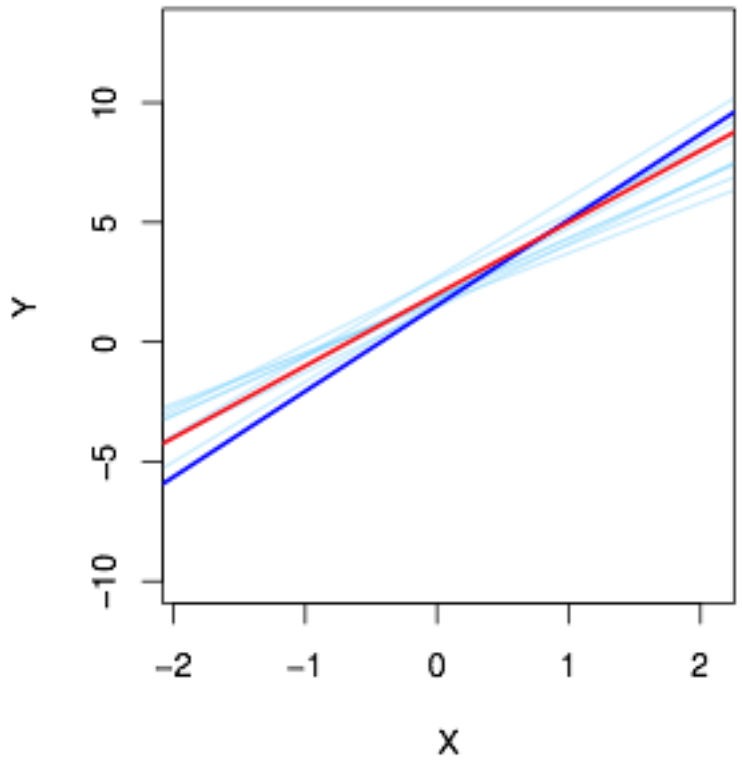


$$\{(x_i^{(1)}, y_i^{(1)}) \ : \ i = 1, \ldots, n\} \quad \overset{\text{EKK}}{\Longrightarrow} \quad \hat{\beta}_0^{(1)}, \hat{\beta}_1^{(1)}$$

$$\{(x_i^{(2)}, y_i^{(2)}) \ : \ i = 1, \ldots, n\} \quad \overset{\text{EKK}}{\Longrightarrow} \quad \hat{\beta}_0^{(2)}, \hat{\beta}_1^{(2)}$$

$$\vdots$$

$$\{(x_i^{(N)}, y_i^{(N)}) \ : \ i = 1, \ldots, n\} \quad \overset{\text{EKK}}{\Longrightarrow} \quad \hat{\beta}_0^{(N)}, \hat{\beta}_1^{(N)}$$



$$\{(x_i^{(1)}, y_i^{(1)}) \ : \ i = 1, \ldots, n\} \quad \overset{\text{EKK}}{\Longrightarrow} \quad \hat{\beta}_0^{(1)}, \hat{\beta}_1^{(1)}$$

$$\{(x_i^{(2)}, y_i^{(2)}) \ : \ i = 1, \ldots, n\} \quad \overset{\text{EKK}}{\Longrightarrow} \quad \hat{\beta}_0^{(2)}, \hat{\beta}_1^{(2)}$$

$$\vdots$$

$$\{(x_i^{(N)}, y_i^{(N)}) \ : \ i = 1, \ldots, n\} \quad \overset{\text{EKK}}{\Longrightarrow} \quad \hat{\beta}_0^{(N)}, \hat{\beta}_1^{(N)}$$

örneklem

İSTATİSTİK!

Parametreler

SH: Standart Hata (Standard Error)

$$\text{SH}(\hat{\beta}_0)^2 = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right]$$

$$\text{SH}(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

$\sigma^2 = \text{Var}(\epsilon)$

~ %95 Güven Aralığı (Confidence Interval)

$$\left[\hat{\beta}_0 - 2\ \mathrm{SH}(\hat{\beta}_0), \hat{\beta}_0 + 2\ \mathrm{SH}(\hat{\beta}_0)\right]$$

$$\left[\hat{\beta}_1 - 2\ \mathrm{SH}(\hat{\beta}_1), \hat{\beta}_1 + 2\ \mathrm{SH}(\hat{\beta}_1)\right]$$

[6,130; 7,935]

[0,042; 0,053]

~ %95 Güven Aralığı (Confidence Interval)

$$\left[\hat{\beta}_0 - 2\ \mathrm{SH}(\hat{\beta}_0), \hat{\beta}_0 + 2\ \mathrm{SH}(\hat{\beta}_0)\right]$$

$$\left[\hat{\beta}_1 - 2\ \mathrm{SH}(\hat{\beta}_1), \hat{\beta}_1 + 2\ \mathrm{SH}(\hat{\beta}_1)\right]$$

Hipotez Testi (Hypothesis Testing)

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

$H_0 : \beta_1 = 0$ sıfır hipotezi (null hypothesis)

$H_a : \beta_1 \neq 0$ karşı hipotez (alternative hypothesis)

t testi, *p* değeri ...

Model

KSH: Kalıntı Standart Hata (Residual Standard Error)

$$\text{KSH} = \sqrt{\frac{\text{KKT}}{n-2}}$$

R^2 İstatistiği (R^2 Statistics)

$$R^2 = 1 - \frac{\text{KKT}}{\text{TKT}}$$

$$\text{TKT} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

TKT: Tüm Kareler Toplamı

Ölçüm	Değer
KSH	3,26
R^2 İstatistiği	0,61

Çoklu Bağlanım

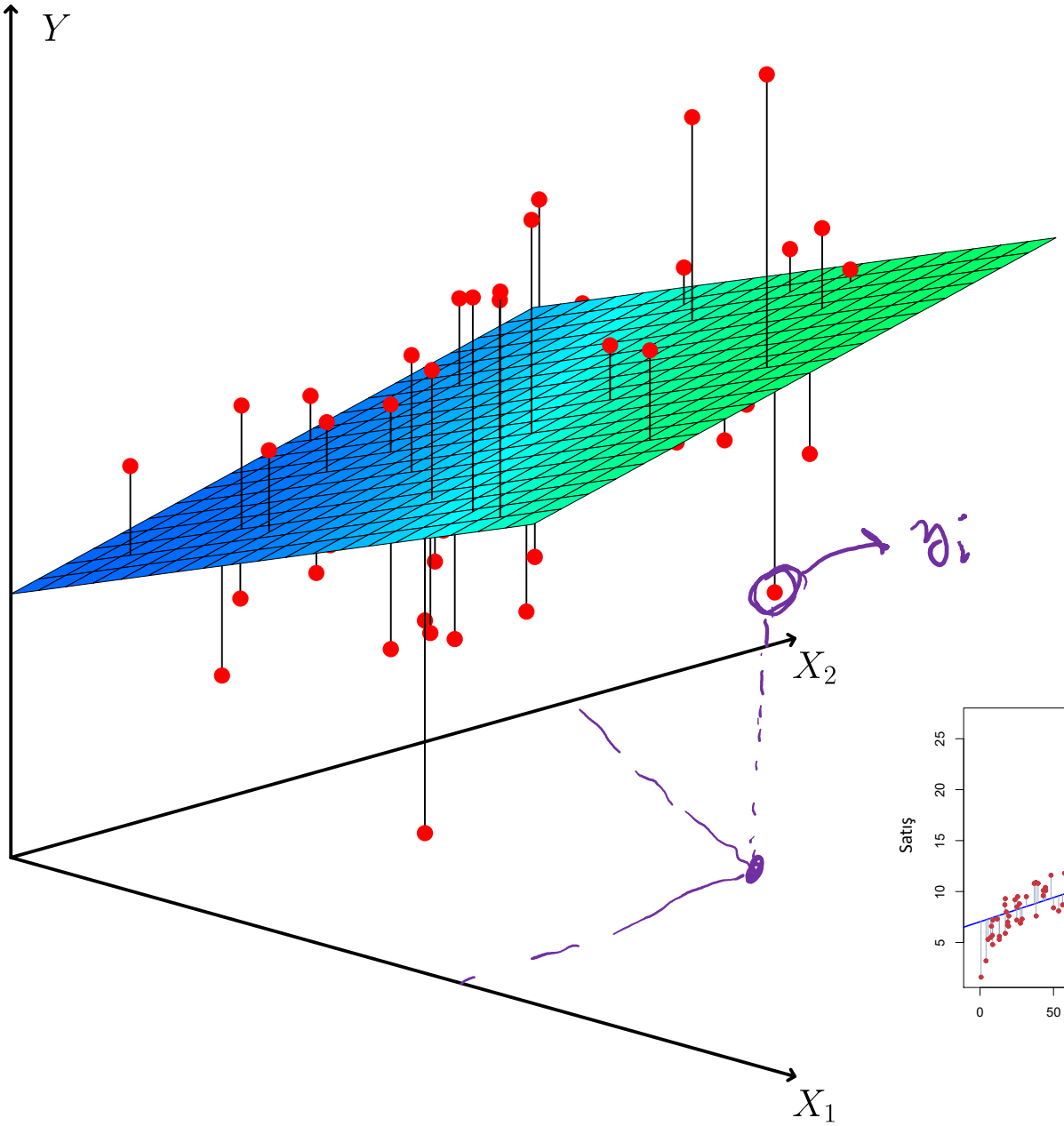
$$Y \approx \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$

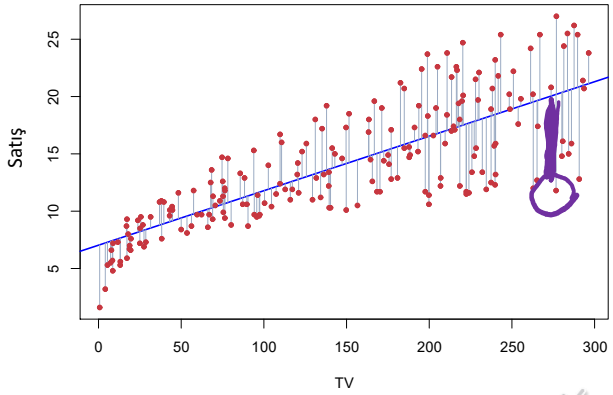
eğitim verisi

$$\{(x_i, y_i) : 1, \ldots, n\}$$

$$y_i \approx \underbrace{\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \cdots + \hat{\beta}_p x_{ip}}_{\hat{y}_i}, \quad i = 1, \ldots, n$$

$$\text{KKT} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$


Y
X2
X1
yi


Satış
TV
5
10
15
20
25
0
50
100
150
200
250
300

EKK

$$\min_{\beta_0,\beta_1,\ldots,\beta_p} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2$$

Dışbükey Fonksiyon

$$\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$$

$$\min_{\beta_0,\beta_1,\ldots,\beta_p} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2$$

$$\mathbf{X} = \begin{bmatrix} 1 & x_1^\intercal \\ 1 & x_2^\intercal \\ \vdots & \vdots \\ 1 & x_n^\intercal \end{bmatrix}_{n \times (p+1)}$$

$$\mathbf{y}^\intercal = (y_1, \ldots, y_n)$$

$$\boldsymbol{\beta}^\intercal = (\beta_0, \beta_1, \ldots, \beta_p)$$

$$\hat{\boldsymbol{\beta}}_{\text{LS}} = \arg\min_{\boldsymbol{\beta}} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\intercal (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

(tam kerte (full rank) varsayımı ile)

$$\hat{\boldsymbol{\beta}}_{\text{LS}} = (\mathbf{X}^\intercal \mathbf{X})^{-1} \mathbf{X}^\intercal \mathbf{y}$$

$$\hat{y}_0 = (1 \; x_0^\intercal) \hat{\boldsymbol{\beta}}_{\text{LS}}$$

Basit Bağlanım (satış – gazete)

	Katsayı	St. Hata	*t* testi	*p* değeri
Kesme nok.	12,351	0,621	19,88	< 0,0001
Gazete	0,055	0,017	3,30	0,00115

Çoklu Bağlanım (satış – TV, radyo, gazete)

	Katsayı	St. Hata	*t* testi	*p* değeri
Kesme nok.	2,939	0,3119	9,42	< 0,0001
TV	0,046	0,0014	32,81	< 0,0001
Radyo	0,189	0,0086	21,89	< 0,0001
Gazete	-0,001	0,0059	-0,18	0,8599

	Katsayı	St. Hata	*t* testi	*p* değeri
Kesme nok.	2,939	0,3119	9,42	< 0,0001
TV	0,046	0,0014	32,81	< 0,0001
Radyo	0,189	0,0086	21,89	< 0,0001
Gazete	-0,001	0,0059	-0,18	**0,8599**

İlinti Matrisi

	TV	Radyo	Gazete	Satış
TV	1,0000	0,0548	0,0568	0,7822
Radyo		1,0000	**0,3541**	0,5762
Gazete			1,0000	0,2283
Satış				1,0000

Parametreler

Hipotez Testi

$$H_0 \ : \ \beta_1 = \beta_2 = \cdots = \beta_p = 0$$

$$H_a \ : \ \text{En az bir } \beta_j \neq 0$$

F-İstatistiği (*F*-Statistics)

$$F = \frac{(\text{TKT} - \text{KKT})/p}{\text{KKT}/(n-p-1)}$$

Ölçüm	Değer
KSH	1,69
F İstatistiği	570

Model

KSH

$$\mathrm{KSH} = \sqrt{\frac{\mathrm{KKT}}{n-p-1}}$$

R^2 İstatistiği

$$R^2 = 1 - \frac{\mathrm{KKT}}{\mathrm{TKT}}$$

Kategorik Değişkenler

İki Seviyeli
(kadın, erkek)

$$x_i = \begin{cases} 1, & \text{eğer } i. \text{ kişi kadın ise;} \\ 0, & \text{eğer } i. \text{ kişi erkek ise.} \end{cases}$$

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i, & \text{eğer } i. \text{ kişi kadın ise;} \\ \beta_0 + \epsilon_i, & \text{eğer } i. \text{ kişi erkek ise.} \end{cases}$$

Kategorik Değişkenler

Çok Seviyeli
(esmer, kumral, sarışın)

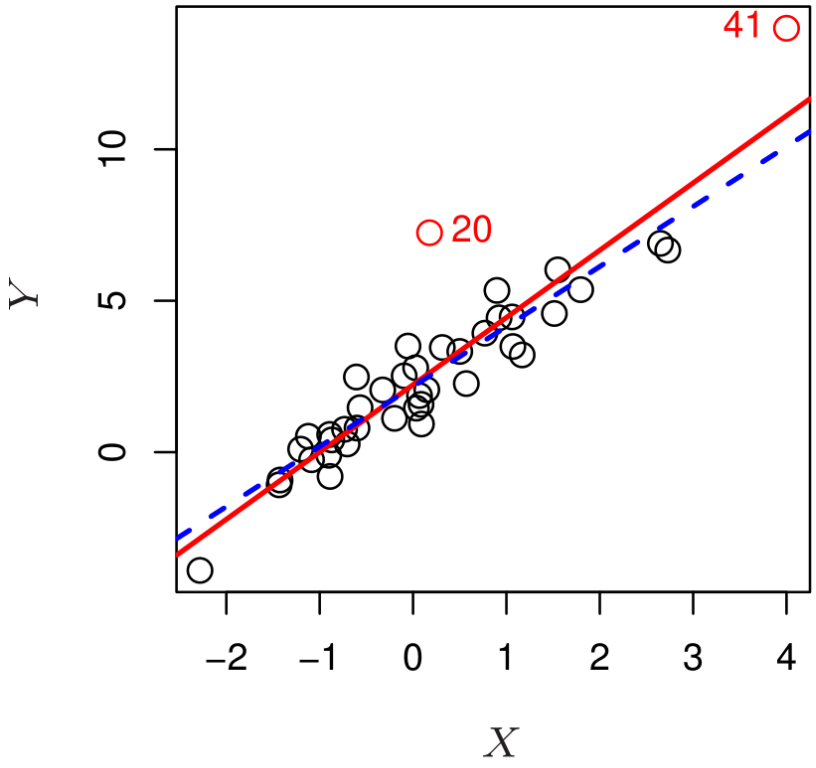
$$x_{i1} = \begin{cases} 1, & \text{eğer } i. \text{ kişi esmer ise;} \\ 0, & \text{eğer } i. \text{ kişi esmer değilse.} \end{cases}$$

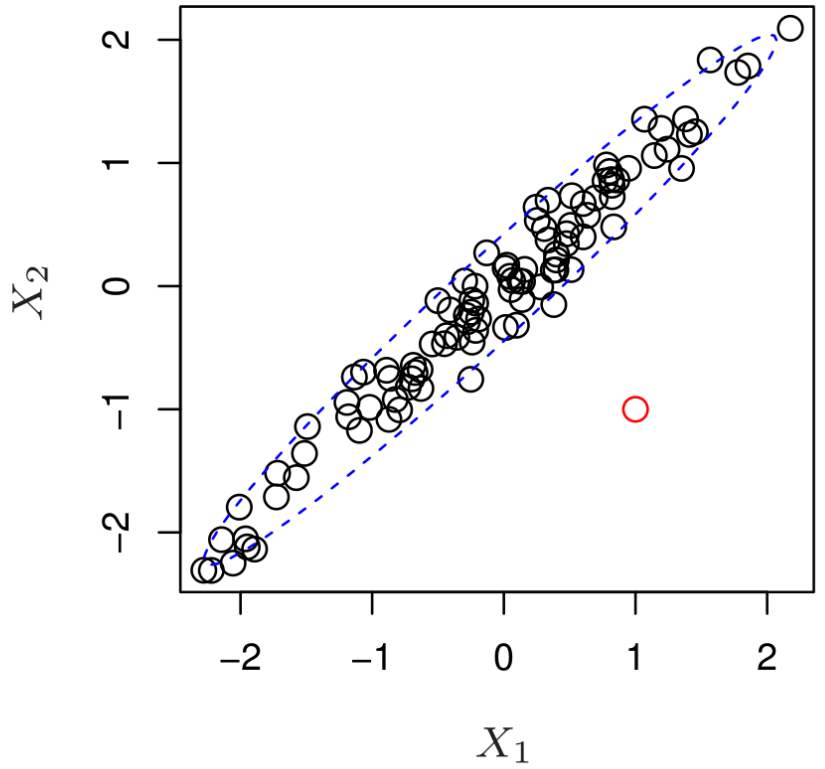
$$x_{i2} = \begin{cases} 1, & \text{eğer } i. \text{ kişi kumral ise;} \\ 0, & \text{eğer } i. \text{ kişi kumral değilse.} \end{cases}$$

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i, & \text{eğer } i. \text{ kişi esmer ise;} \\ \beta_0 + \beta_2 + \epsilon_i, & \text{eğer } i. \text{ kişi kumral ise;} \\ \beta_0 + \epsilon_i, & \text{eğer } i. \text{ kişi sarışın ise.} \end{cases}$$

Dikkat!

- Doğrusal olmayan girdi-çıktı ilişkisi
- İlintili hata terimleri (correlated error terms)
- Hata terimlerinin sabit olmayan varyansı (nonconstant variance)
- Aykırı değerler (outliers) ve *ayrıksı* değerler
- Doğrudaşlık (collinearity)


41
20
Y
X
10
5
0
−2
−1
0
1
2
3
4


X_2
X_1
2
1
0
−1
−2
−2
−1
0
1
2

Özet

- Basit Bağlanım
- En Küçük Kareler Yöntemi
- Parametre ve model istatistikleri
- Çoklu Bağlanım
- Kategorik Değişkenler
- Dikkat edilecek birkaç nokta