BATCH: 223

DT/NT : DATA SCIENCE

LESSON : STATISTICS -2

SUBJECT : LINEAR REGRESSION R^2

TECHPRO EDUCATION

techproeducation.com

+1 (585) 304 29 59

STATISTICS - 2

Data Science Program

Session -4

RECAP

Herkes önceki dersten hatırladığı 1 cümle yazabilir mi?

Session - 4 Content

Content

- Linear Regression
- Regression Equation
- Coefficient of Determination
- R^2

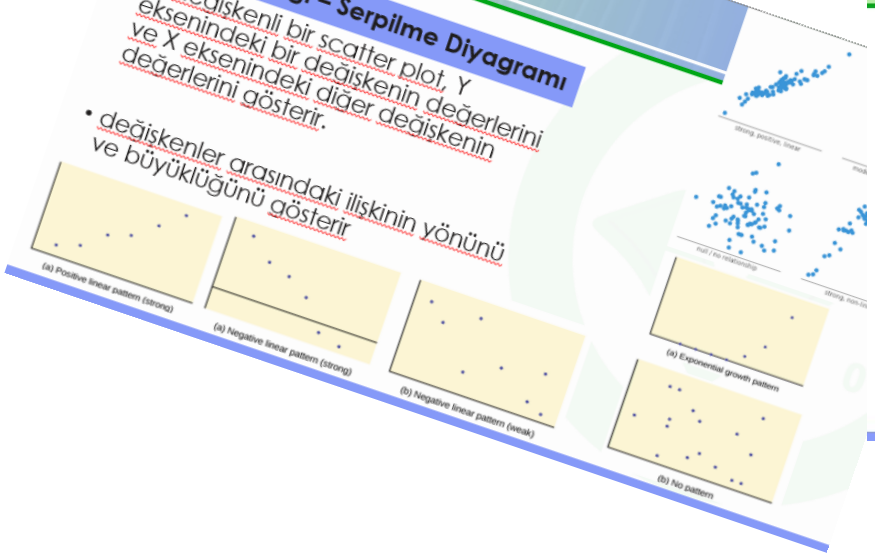
LMS Pre-Class’ta bu dersle ilgili kısma çalıştım

Recap – Previous Lesson

Scatter Plot

Saçılım Grafiği – Serpilme Diyagramı

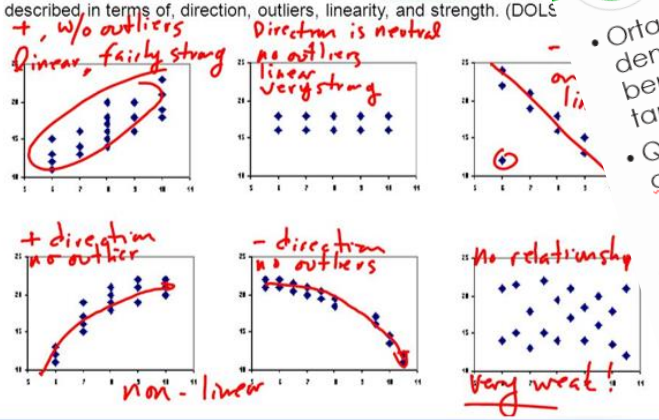
- İki değişkenli bir scatter plot, Y eksenindeki bir değişkenin değerlerini ve X eksenindeki diğer değişkenin değerlerini gösterir.
- değişkenler arasındaki ilişkinin yönünü ve büyüklüğünü gösterir



Pattern of Data in Scatterplot

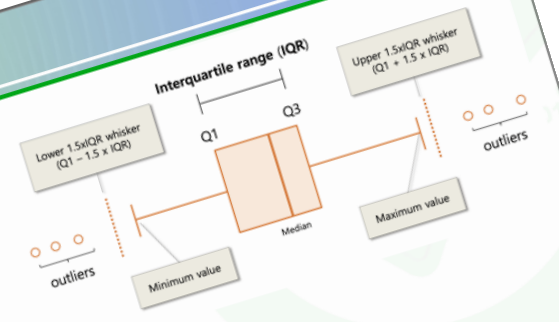
described in terms of, direction, outliers, linearity, and strength. (DOLS

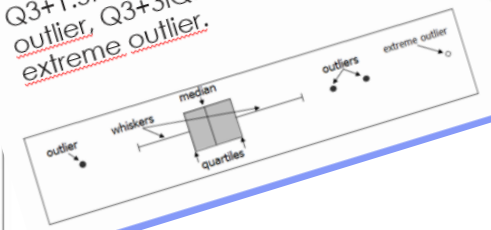
- Meti



Box Plot

- Ortadaki quartiles denen kısım benim datamın tamamının %50'si
- Q3+1.5IQR sonrası outlier, Q3+3IQR extreme outlier.





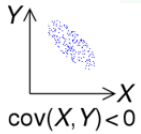
Covariance

Cov (x,y) > 0

- İlişki pozitiftir.
- X artarken Y de artar

Cov (x,y) < 0

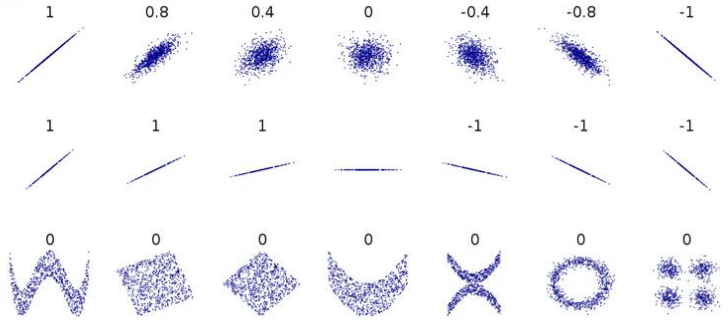
- İlişki negatiftir.
- X artarken Y azalır



Cov (x,y) = 0

- İki değişkenin arasında ilişki yoktur, birbirinden bağımışızlar.

Correlation



Correlation - r Calculation

Cigarette (X)	Lung Capacity (Y)
0	45
5	42
10	33
15	31
20	29

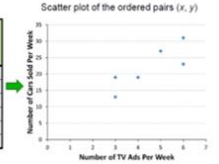
$$r = \frac{n\Sigma(xy) - (\Sigma x)(\Sigma y)}{\sqrt{\left[n\Sigma x^2 - (\Sigma x)^2\right]\left[n\Sigma y^2 - (\Sigma y)^2\right]}}$$

$$r_{xy} = \frac{(5)(1585) - (50)(180)}{\sqrt{\left[(5)(750) - 50^2\right]\left[(5)(6680) - 180^2\right]}} = \frac{7925 - 9000}{\sqrt{(3750 - 2500)(33400 - 32400)}} = \frac{-1075}{\sqrt{(1250)(1000)}} = -.9615$$

Week	Number of TV Ads x	Number of Cars Sold y	xy	x^2	y^2
1	3	13	39	9	169
2	6	31	186	36	961
3	4	19	76	16	361
4	5	27	135	25	729
5	6	23	138	36	529
6	3	19	57	9	361
	$\Sigma x = 27$	$\Sigma y = 132$	$\Sigma xy = 631$	$\Sigma x^2 = 131$	$\Sigma y^2 = 3110$

Week	Number of TV Ads x	Number of Cars Sold y
1	3	13
2	6	31
3	4	19
4	5	27
5	6	23
6	3	19

Scatter plot of the ordered pairs (x, y)



$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{\left[n\sum x^2 - (\sum x)^2\right]\left[n\sum y^2 - (\sum y)^2\right]}}$$

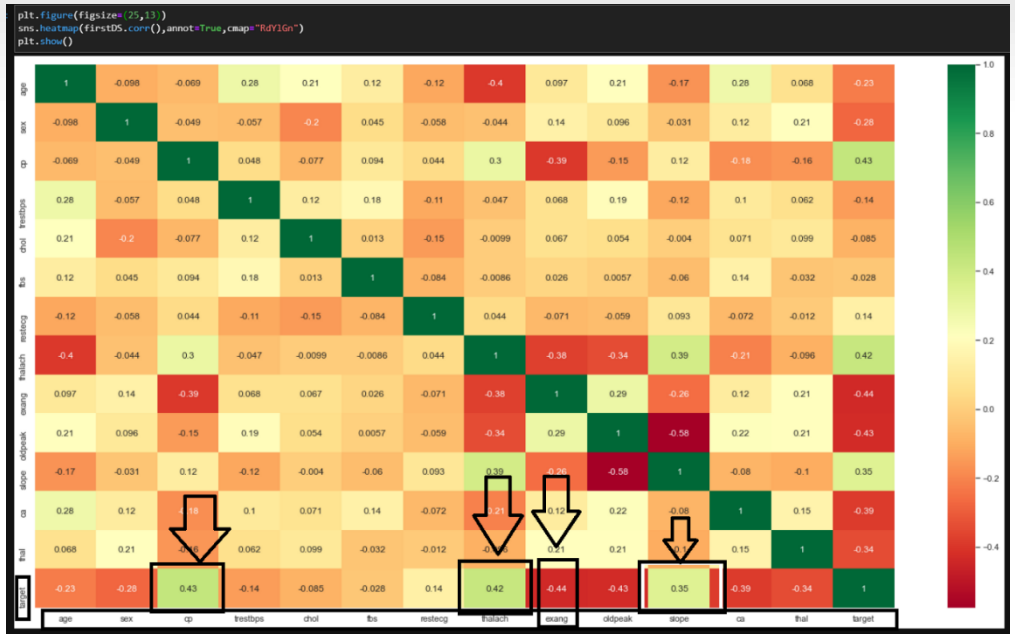
$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{\left[n\sum x^2 - (\sum x)^2\right]\left[n\sum y^2 - (\sum y)^2\right]}} = \frac{(6)(631) - (27)(132)}{\sqrt{\left[(6)(131) - (27)^2\right]\left[(6)(3110) - (132)^2\right]}} = \frac{222}{\sqrt{[57]\,[1236]}} = \frac{222}{265.43} = 0.836$$

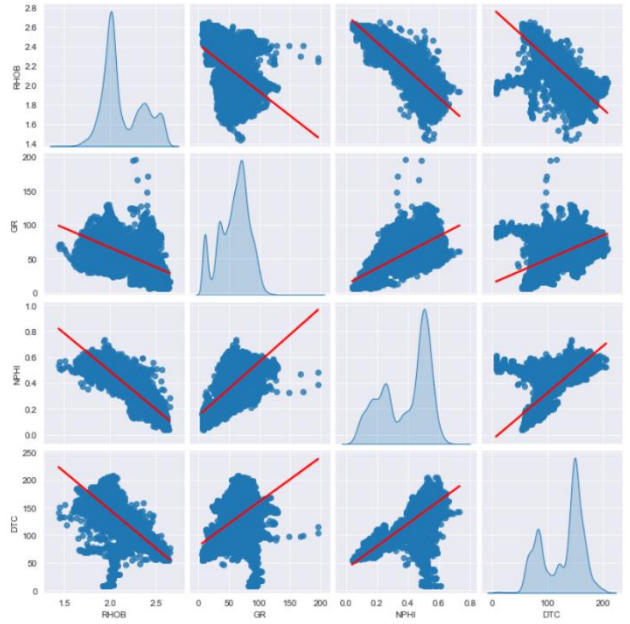
TECHPRO EDUCATION

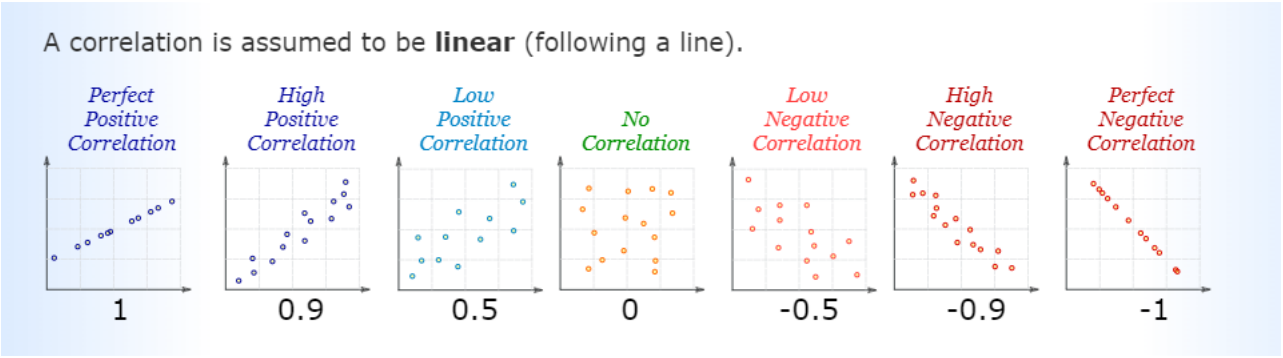
Review

Konular

- Correlation
- Pearson katsayısı
- Sample ve Population corr.
- R hesaplanması







Question14: What is the difference between Covariance and Correlation?

Covariance

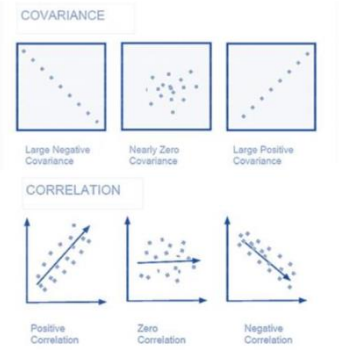
- Signifies the direction of the linear relationship between two variables
- In simple terms, It is a measure of variance between two variables
- It can take any value from positive infinity to negative infinity

Correlation

- It measures the relationship between two variables, as well as the strength between these two variables.
- It can take any value from -1 to 1

Q7. What is correlation and covariance in statistics?

Covariance and Correlation are two mathematical concepts; these two approaches are widely used in statistics. Both Correlation and Covariance establish the relationship and also measure the dependency between two random variables. Though the work is similar between these two in mathematical terms, they are different from each other.



Correlation:
Correlation is considered or described as the best technique for measuring and also for estimating the quantitative relationship between two variables. Correlation measures how strongly two variables are related.

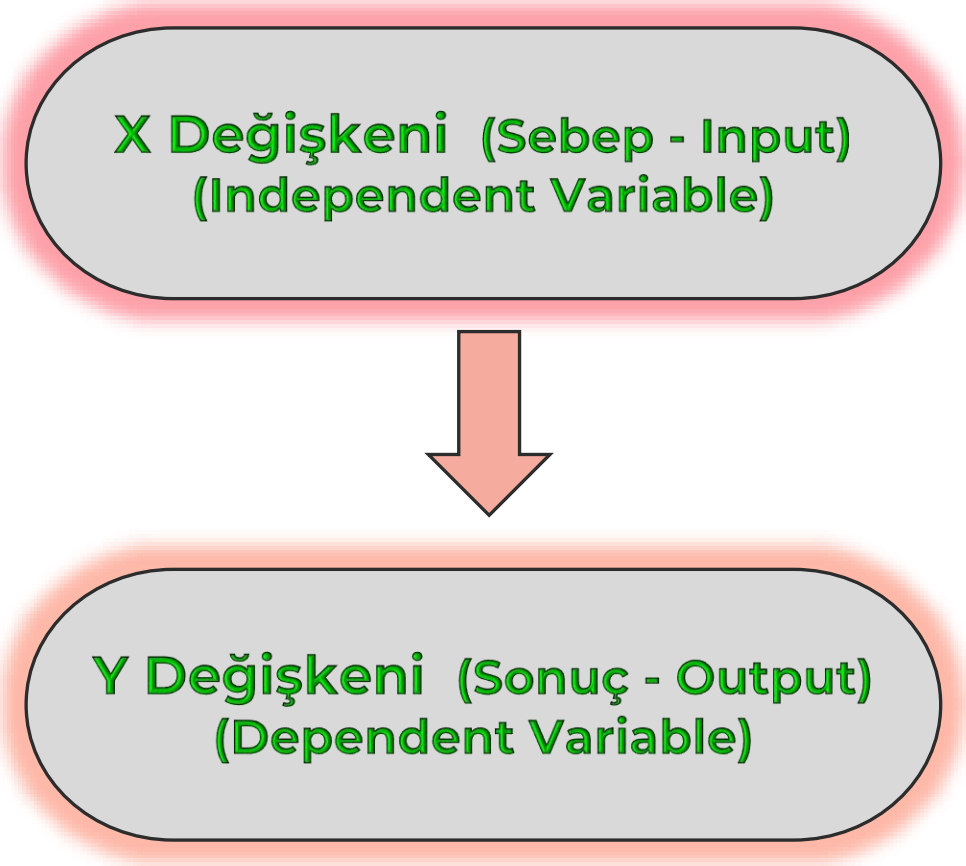
Covariance: In covariance two items vary together and it's a measure that indicates the extent to which two random variables change in cycle. It is a statistical term; it explains the systematic relation between a pair of random variables, wherein changes in one variable reciprocal by a corresponding change in another variable.

Linear Regression

Lineer Regresyon

- Amaç: İki değişken arasındaki ilişkiye dayanarak ileri dönük tahmin yapmak

- Sebep-sonuç ilişkisi içinde, Independent variable (bağımsız değişken) sebep, bağımlı değişken ise sonuçtur.



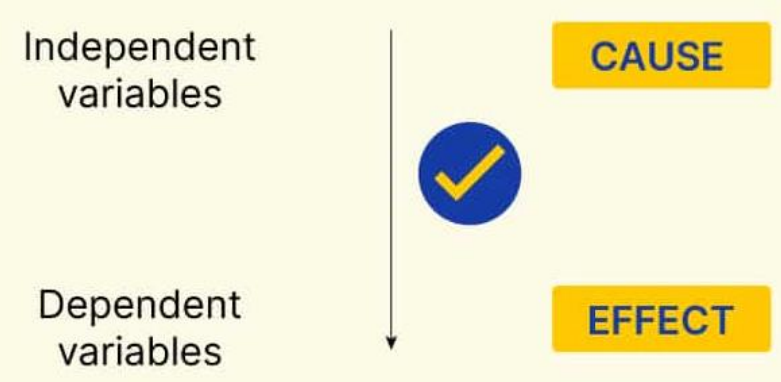
Linear Regression

Independent Variable

- Gelir
- Araç sahipliliği
- IQ değeri
- ???

Dependent Variable

- Yaşam konforu
- Trafik hacmi
- İş performansı
- ???



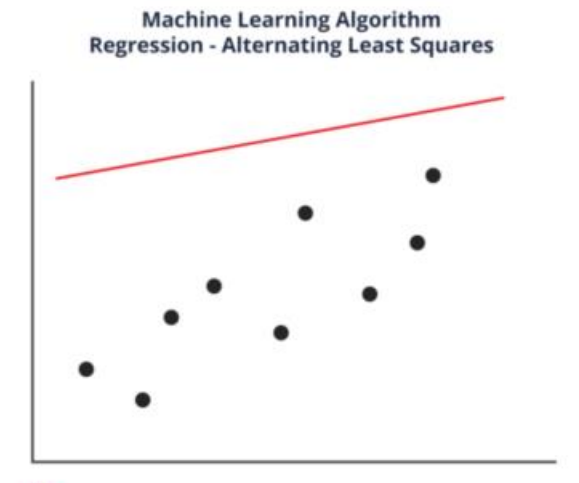
RECAP

Aklınıza gelen Linear Regresyon örneklerini yazar mısınız?

INTERVIEW QUESTION

Q52. What is Linear Regression?

Linear regression is a statistical technique where the score of a variable Y is predicted from the score of a second variable X. X is referred to as the predictor variable and Y as the criterion variable.



The regression line is the one with the least value of D

Matching on Peardeck

Daily temperature	Amount of time spent studying
Annual Salary	Electricity Consumption
Number of exams passed	GDP per Capita
Life Expectancy	Number of vacations taken

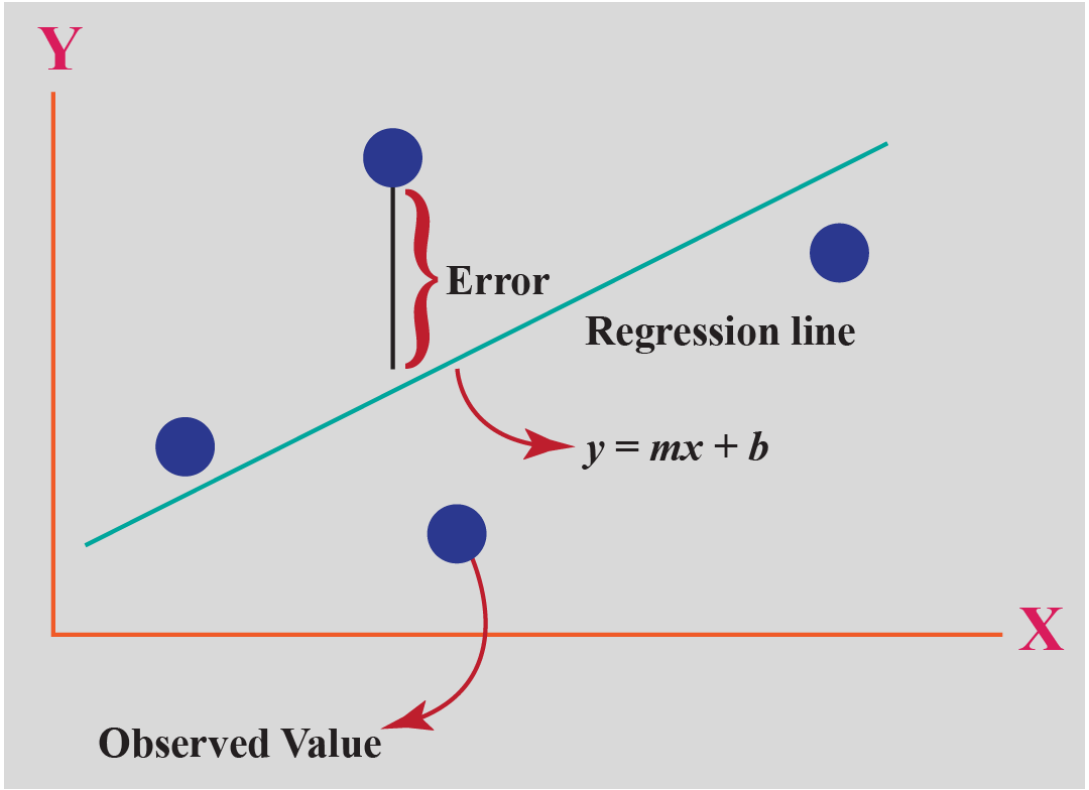
Linear Regression and Equation

En küçük Kareler Yöntemi

- The least squares (en küçük kareler) yöntemi

- X bağımsız değişkenin değerine bağlı olarak, Y bağımlı değişkenin değerini tahmin etmek için kullanılan bir yöntem

$$Y = aX + b$$



Linear Regression Requirement

Değişken Sayısı

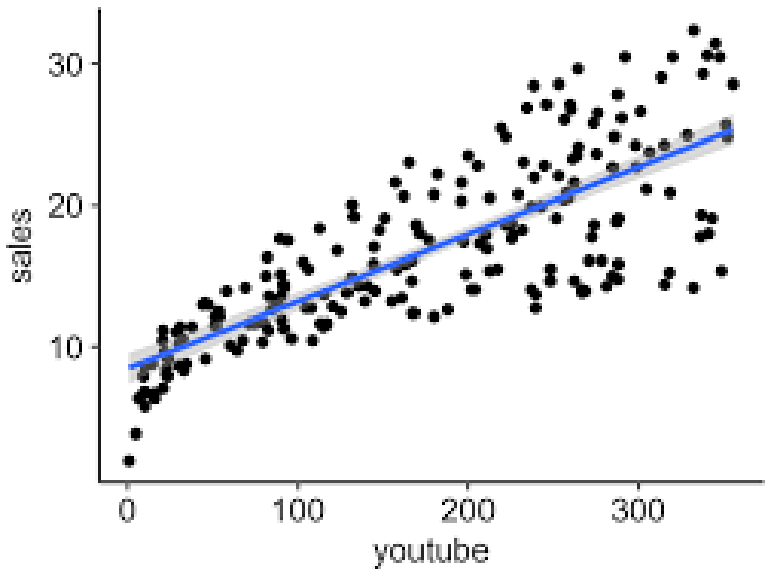
1 Bağımlı değişken
1 Bağımsız değişken

Lineerlik

Lineer ilişki olmalı, nonlineer vb. değil

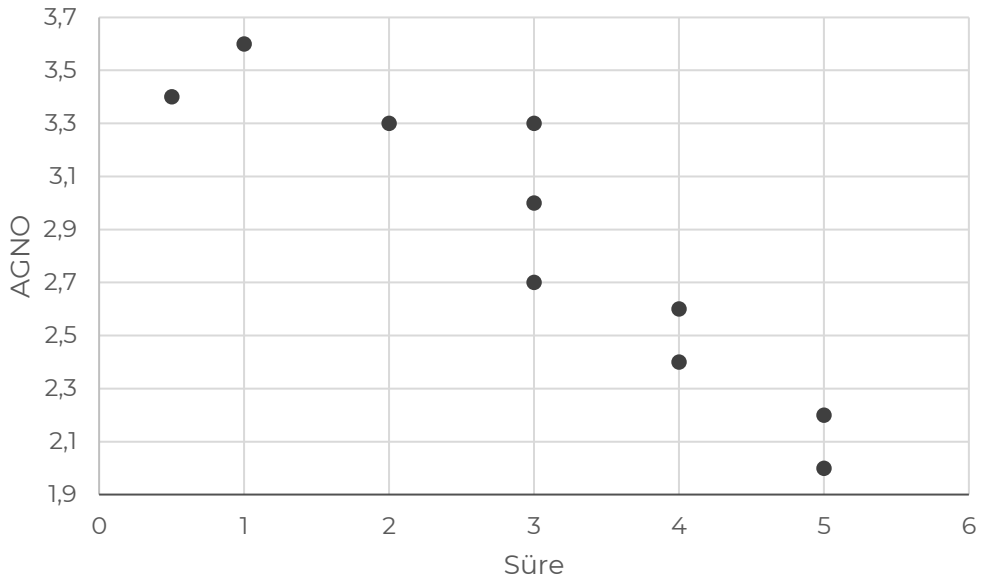
Ölçülebilirlik

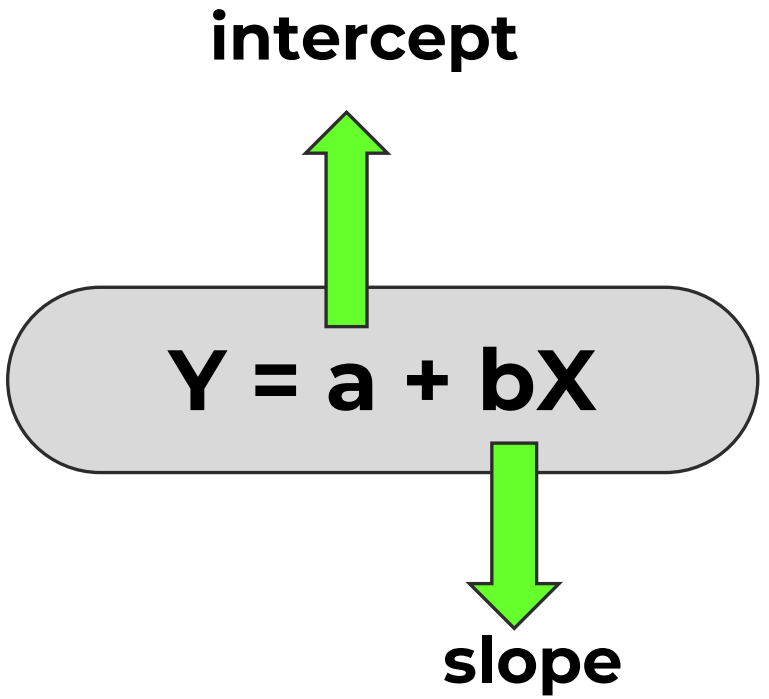
Interval veya ratio scale



Linear Regression Sample

Herhangi Ekranda geçirilen süre -X	Ağırlıklı Genel Not Ortalaması - Y
3	2,7
5	2,2
2	3,3
0,5	3,4
5	2
3	3
1	3,6
4	2,4
3	3,3
4	2,6
3,05 (ort)	2,85 (ort)





Linear Regression Sample

$x-x_{ort}$	$(x-x_{ort})^2$	$y-y_{ort}$	$(y-y_{ort})^2$	$(x-x_{ort})*(y-y_{ort})$	
-0,05	0,0025	-0,15	0,0225	0,0075	
1,95	3,8025	-0,65	0,4225	-1,2675	
-1,05	1,1025	0,45	0,2025	-0,4725	
-2,55	6,5025	0,55	0,3025	-1,4025	
1,95	3,8025	-0,85	0,7225	-1,6575	
-0,05	0,0025	0,15	0,0225	-0,0075	
-2,05	4,2025	0,75	0,5625	-1,5375	
0,95	0,9025	-0,45	0,2025	-0,4275	
-0,05	0,0025	0,45	0,2025	-0,0225	
0,95	0,9025	-0,25	0,0625	-0,2375	
	21,225		**2,725**	**-7,025**	Toplam
	SSx		**SSy**	**SP**	

b = SP / SSx = -7.025/21,225 = -0,3310

a = yort – b*xort= 2,85-(-0,3310)*3,05=3,85

Y = a + bX

Y = 3,85 - 0,331X

intercept

slope

SSx: Sum of Square for independent variable
SSy: Sum of Square for dependent variable
SP : Sum of products

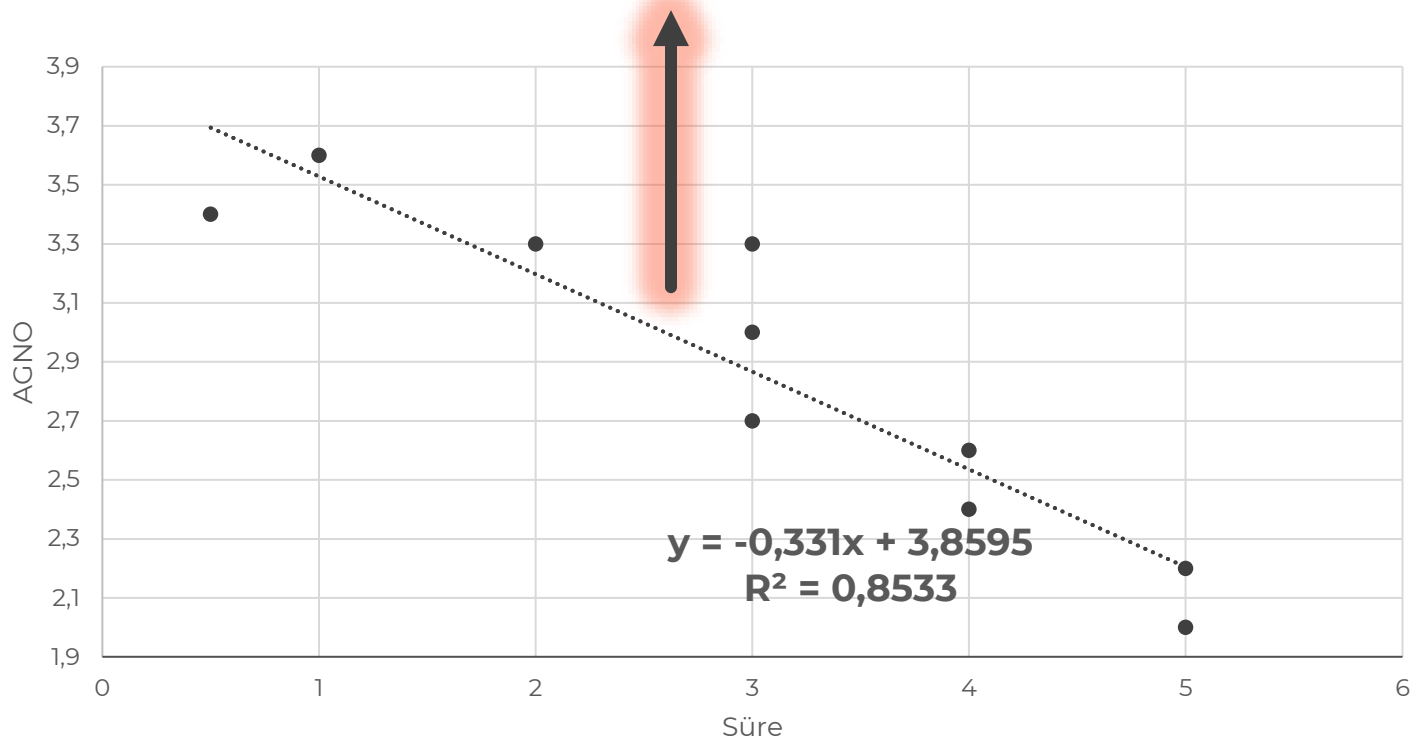
Linear Regression Sample

En iyi eğilim çizgisi (Line of Best Fit)

$$Y = 3{,}85 - 0{,}331X$$

intercept

slope



Regresyon çizgisi, gerçek değerler ve tahmin edilen değerler arasındaki 'sum of square farklarını' minimize eder

Excel

TECHPRO EDUCATION

Linear Regression Sample - Python

```
# Linear Regression Sample
```

```
In [2]: import numpy as np
        from scipy import stats

In [3]: Ekran_sure = np.array([3,5,2,0.5,5,3,1,4,3,4])

In [4]: AGNO = np.array([2.7,2.2,3.3,3.4,2,3,3.6,2.4,3.3,2.6])

In [5]: reg = stats.linregress(Ekran_sure, AGNO)

In [7]: print("a: ", reg.intercept)
        print("b: ", reg.slope)

        a:  3.859481743227327
        b:  -0.330977620730271
```

$$Y = a + bX$$

$$Y = 3{,}85 - 0{,}331X$$

intercept

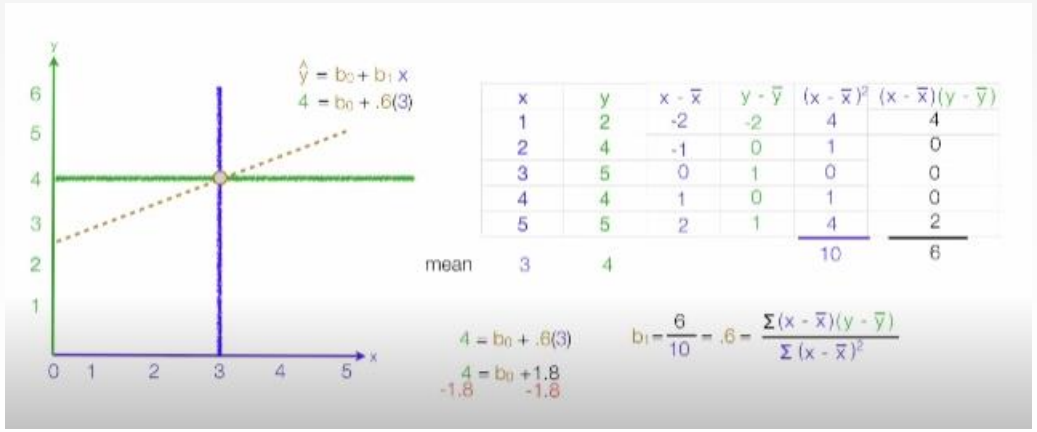
slope

TECHPRO EDUCATION

YOUTUBE VIDEO ONERI

- How to calculate linear regression using least square method

https://www.youtube.com/watch?v=JvS2triCgOY



x	y	$x - \bar{x}$	$y - \bar{y}$	$(x - \bar{x})^2$	$(x - \bar{x})(y - \bar{y})$
1	2	-2	-2	4	4
2	4	-1	0	1	0
3	5	0	1	0	0
4	4	1	0	1	0
5	5	2	1	4	2
mean	3	4		10	6

Pearson's r Calculation

$x-x_{ort}$	$(x-x_{ort})^2$	$y-y_{ort}$	$(y-y_{ort})^2$	$(x-x_{ort})*(y-y_{ort})$	
-0,05	0,0025	-0,15	0,0225	0,0075	
1,95	3,8025	-0,65	0,4225	-1,2675	
-1,05	1,1025	0,45	0,2025	-0,4725	
-2,55	6,5025	0,55	0,3025	-1,4025	
1,95	3,8025	-0,85	0,7225	-1,6575	
-0,05	0,0025	0,15	0,0225	-0,0075	
-2,05	4,2025	0,75	0,5625	-1,5375	
0,95	0,9025	-0,45	0,2025	-0,4275	
-0,05	0,0025	0,45	0,2025	-0,0225	
0,95	0,9025	-0,25	0,0625	-0,2375	
	21,225		**2,725**	**-7,025**	Toplam
	SSx		**SSy**	**SP**	

Formula of Pearson's Correlation Coefficient

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

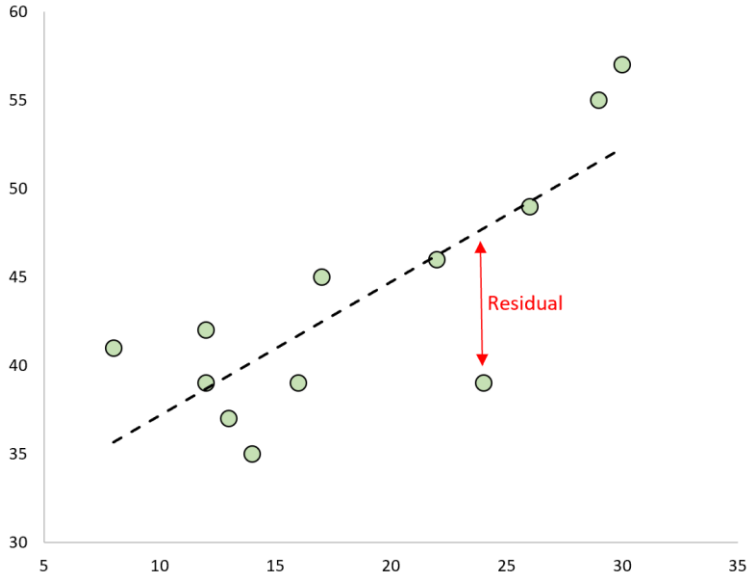
$$r = \frac{SP}{\sqrt{SS_x SS_y}}$$

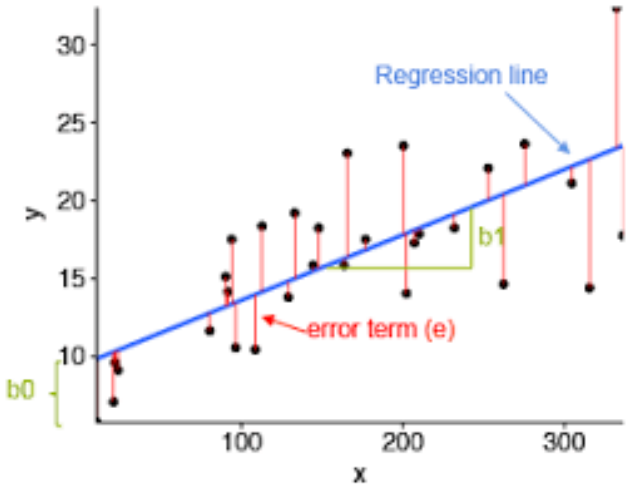
r = -0,92

TECHPRO EDUCATION

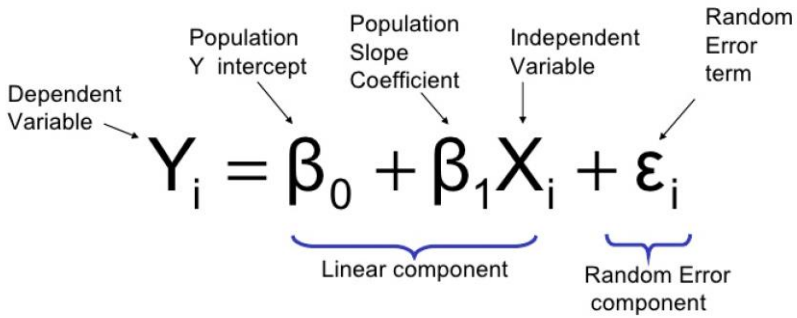
Residual term (e)

Residual = Observed value – Predicted value





Regression Model



Coefficient of Determination – R^2

Determinasyon – Belirlilik Katsayısı

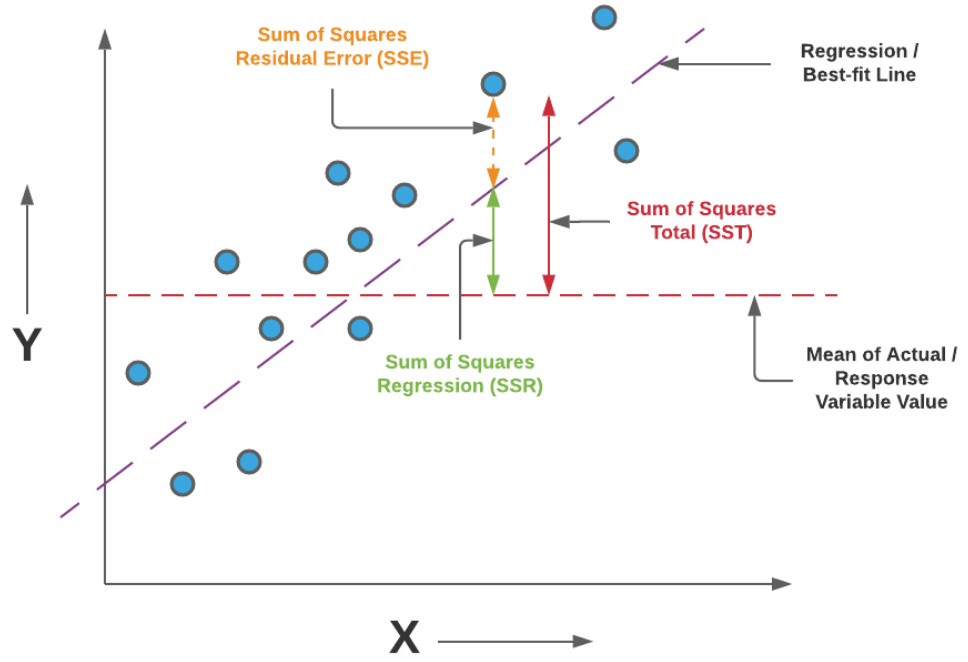
Coefficient of Determination – R^2

R^2 R-square

- Analizimizde iki değişken arasındaki ilişki hakkında fikir sunar
- R^2 değeri bize bağımlı değişkendeki toplam varyansın yüzde kaçının bağımsız değişken tarafından açıklandığını söyler.
- R^2 0 -1 arasında değişir

r = -0,92

R^2 = 0,85



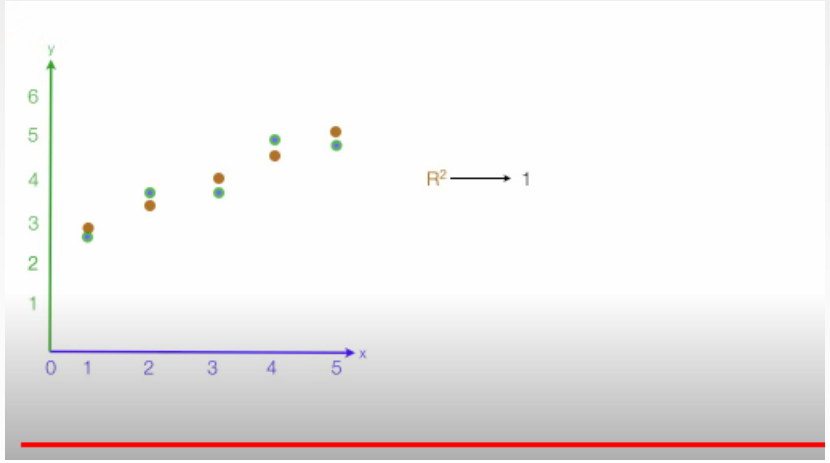
$$R^2 = \frac{SSR}{SST} = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}$$

TECHPRO EDUCATION

YOUTUBE VIDEO ONERI

- How to Calculate R Squared Using Regression Analysis

https://www.youtube.com/watch?v=w2FKXOa0HGA



Kahoot Uygulaması

Python Calculation

- It is time to code by Python...