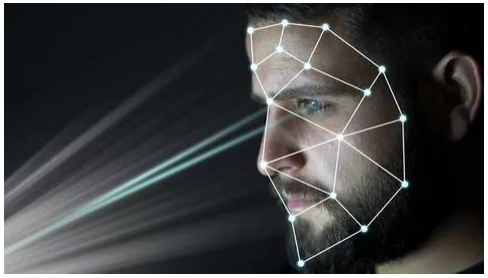


Makine Öğrenmesi ve İmge Tanıma

Altıncı hafta
Regresyon Analizi ve Naive Bayes

Dersin Kaynakları:

Makine Öğrenmesi Ders Notları, Doç. Dr. Hidayet Takcı, Sivas Cumhuriyet Üniversitesi

İSTATİSTİKSEL ANALİZ, İÜ AUZEF DERS NOTLARI, DOÇ.DR. SEMA ULUTÜRK AKMAN http://auzefkitap.istanbul.edu.tr/kitap/kok/istanaau216.pdf

https://acikders.ankara.edu.tr/pluginfile.php/169660/mod_resource/content/0/9_REGRESYON.pdf

Regresyon

Regresyon, bağımsız değişkenlerdeki (x) değişime dayalı olarak bağımlı değişkendeki (y) değişimi açıklama girişimidir.

Regresyon, nedenselliğin (sonuçların neden böyle çıktığının) bir açıklamasıdır.

Eğer bağımsız değişken(ler) bağımlı değişkendeki değişimi yeteri kadar açıklayabiliyorsa, model tahmin için kullanılabilir.

Basit doğrusal regresyon

- Bağımlı değişkeni etkileyen sadece bir adet bağımsız değişken olduğu durumda kurulan model bir **basit regresyon** modelidir.
- Tek değişkenli bir model olmakla birlikte modelde bağımlı değişkeni etkileyen; başlangıç değeri ve hata oranı da bulunur.

Bir regresyon modeli katsayılar değişkenler ile ifade edilebilen matematiksel bir fonksiyondur.

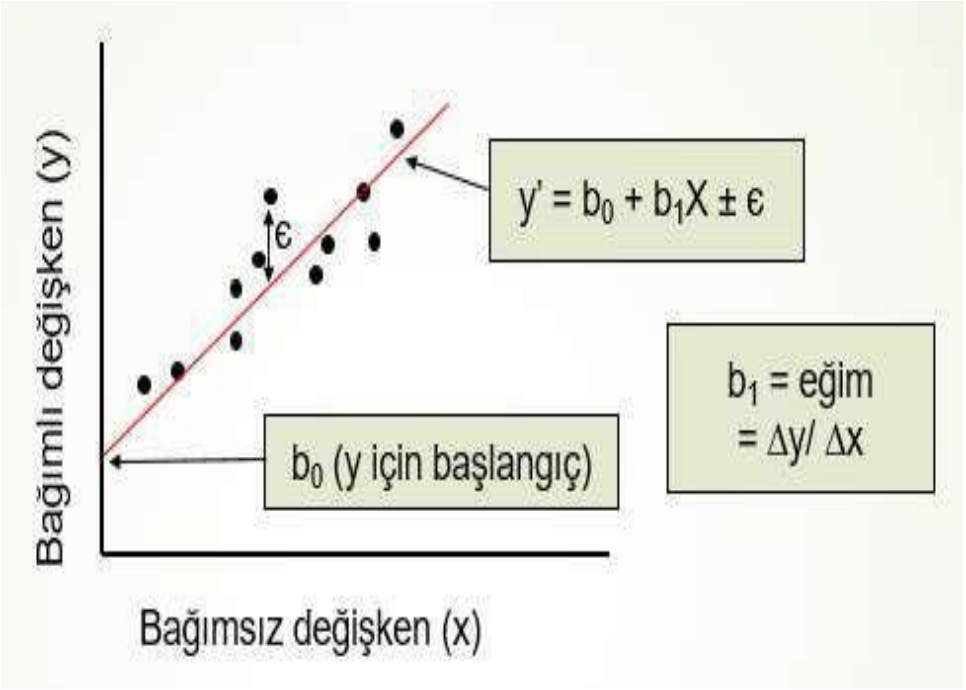
Basit regresyon veri için düz bir çizgi elde etme işlemidir.

Basit doğrusal regresyon

- Bağımlı değişkeni etkileyen sadece bir adet bağımsız değişken olduğu durumda kurulan model bir **basit regresyon** modelidir.
- Tek değişkenli bir model olmakla birlikte modelde bağımlı değişkeni etkileyen; başlangıç değeri ve hata oranı da bulunur.

Bir regresyon modeli katsayılar değişkenler ile ifade edilebilen matematiksel bir fonksiyondur.

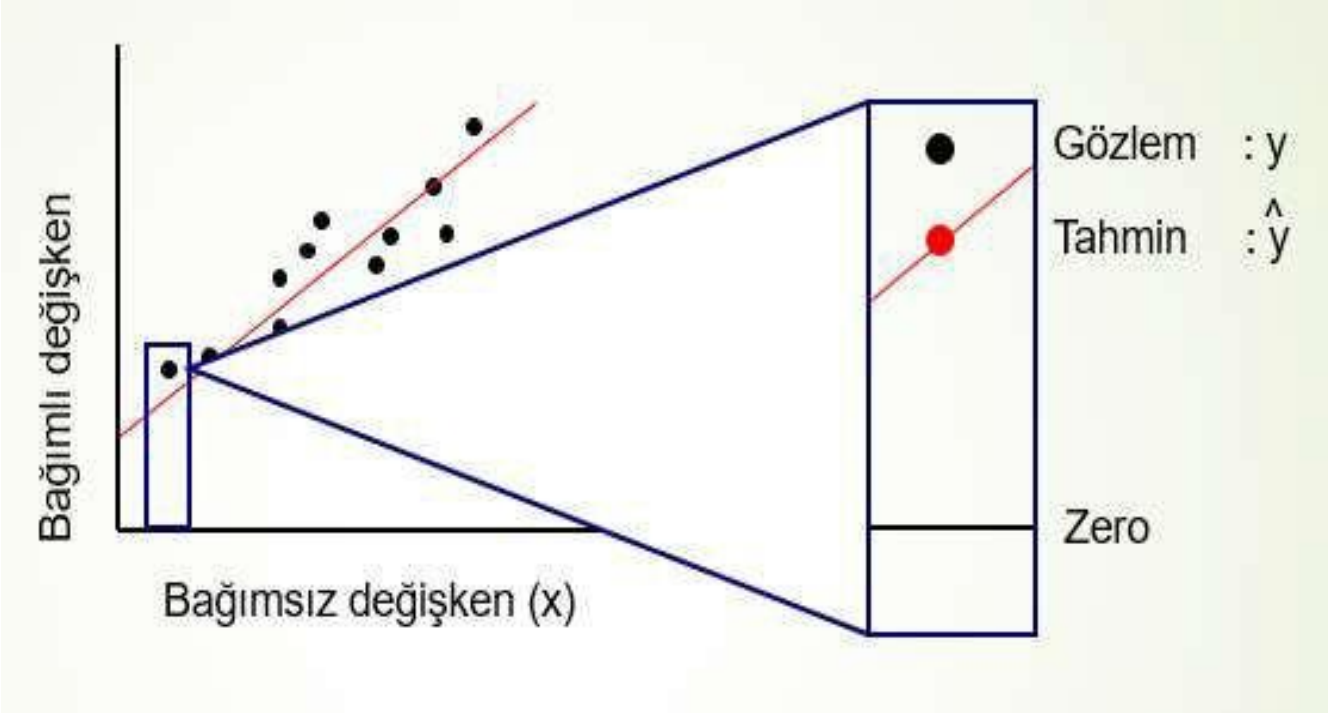
Basit regresyon veri için düz bir çizgi elde etme işlemidir.



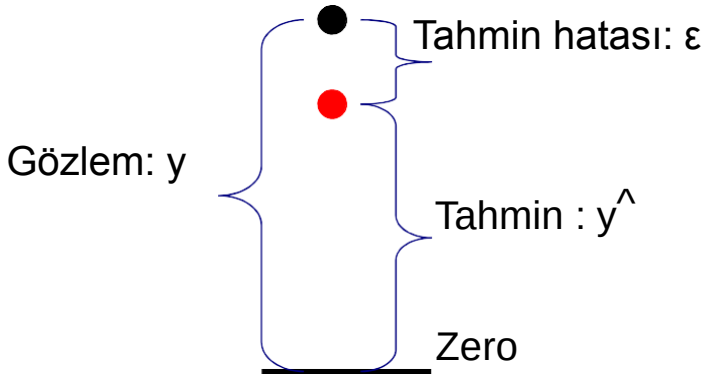
Basit doğrusal regresyon

Fonksiyonun görevi her bir veri noktası için bir tahmin yapmaktır.

Gözlem değeri y ile ifade edilirken tahmin değeri $\hat{y}$ ile ifade edilir.



Basit doğrusal regresyon



Her bir gözlem için, hata oranı aşağıdaki gibi açıklanabilir:

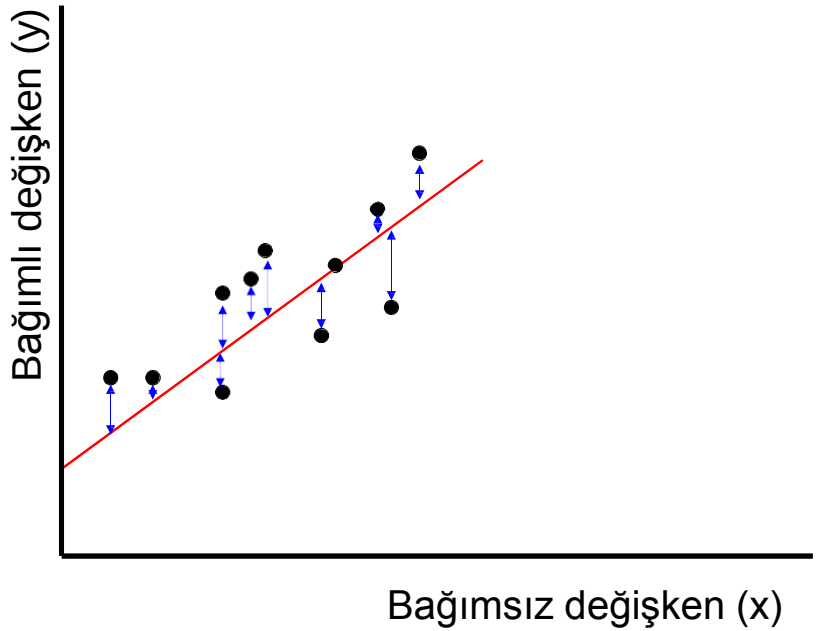
$$y = \hat{y} + \varepsilon$$

Gerçek = Tahmin + Hata

En küçük kareler yöntemi

- Bağımlı değişken ile bağımsız değişken arasında çok sayıda regresyon modeli ortaya konabilir. Bunların en iyisinin hangisi olduğunu seçmede **en küçük kareler** yöntemi kullanılır.
- Gözlem değerleri (gerçek değerler) ile tahmin değerleri arasındaki farkların karesi toplamı hangi modelde en küçük ise o model seçilir.
- En küçük kareler yönteminin esası hatayı minimize eden modelin seçimidir.

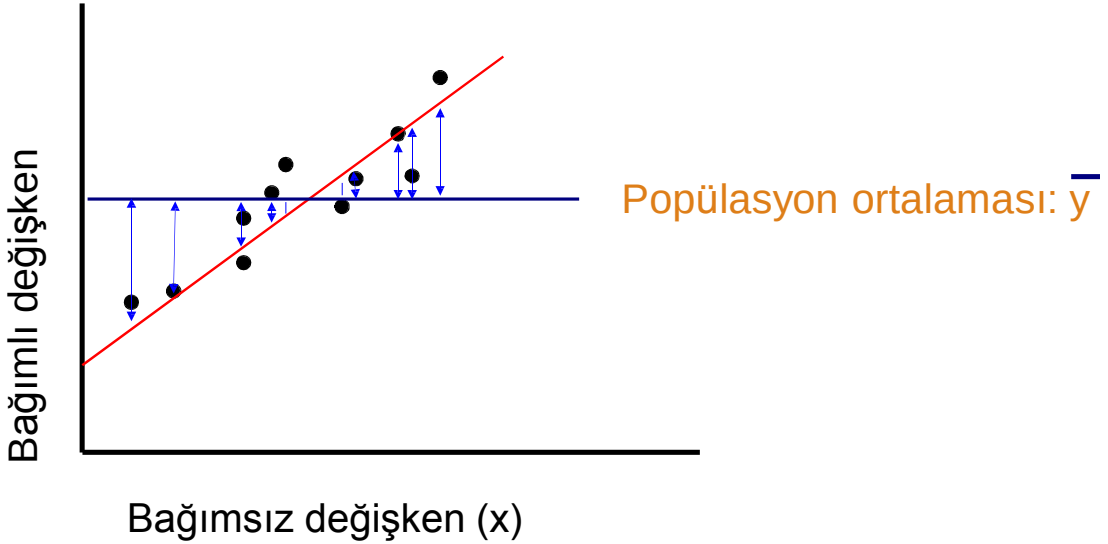
Hataların kareleri toplamı (SSE)



En küçük kareler yöntemi; hataların kareleri toplamından en küçük toplamı veren modeli seçer. Hataların kareleri toplamı (Sum of Squares of Error) veya SSE olarak isimlendirilir.

Regresyon kareleri toplamı (SSR)

- En küçük kareler toplamından başka bir de regresyon kareleri toplamı hesap edilir.
- Regresyon kareleri toplamı tahmin değerleri ile popülasyon ortalaması arasındaki farkların kareleri toplamını elde etmektir.



Regresyon kareleri toplamı (SSR); popülasyon ortalaması ile gözlemler için tahmin değerleri arasındaki farkların kareleri toplamıdır.

Regresyon formülleri

- Regresyon kareleri toplamı

$$SSR = \sum_i (\hat{y}_i - \bar{y})^2$$

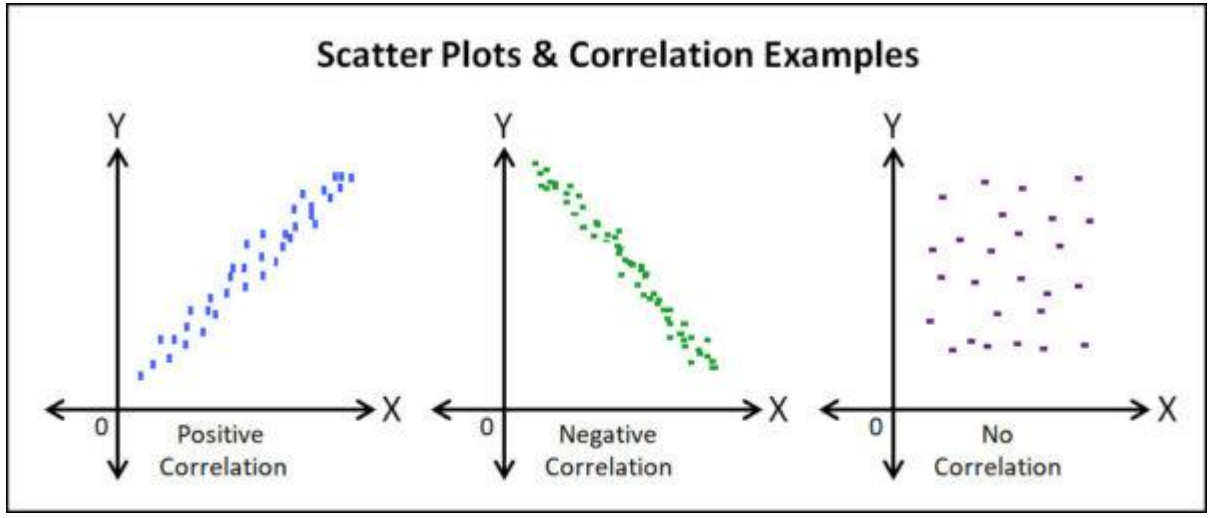
- En küçük kareler toplamı

$$SSE = \sum_i (y_i - \hat{y}_i)^2$$

- SST=SSR + SSE

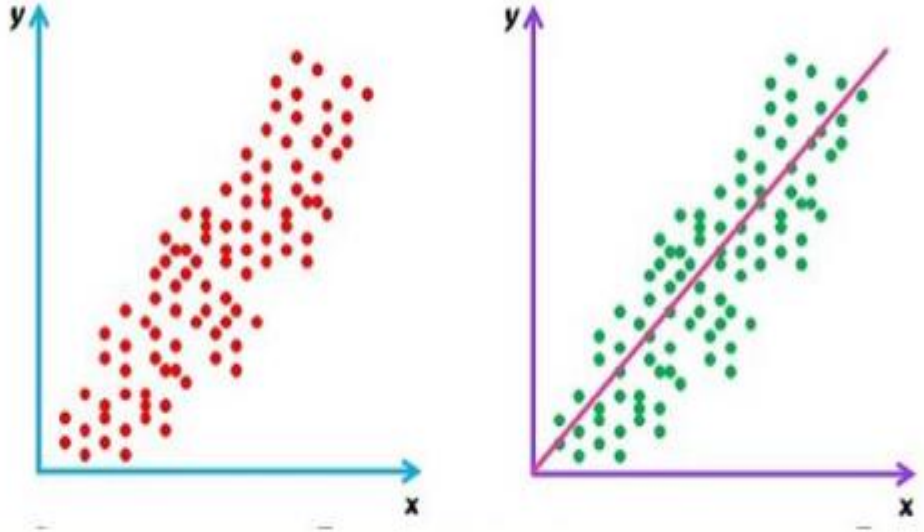
Saçılma grafiği (Scatter plot)

- Regresyon analizi görsel araçlardan saçılma grafikleri ile görülebilir.
- Bağımsız değişken ile bağımlı değişken arasındaki ilişkinin gücüne göre noktalar köşegen boyunca yerleşirler.



https://www.cqeacademy.com/cqe-body-of-knowledge/continuous-improvement/quality-control-tools/the-scatter-plot-linear-regression/

Korelasyon vs Regresyon



Karşılaştırma Yönü	Korelasyon	Regresyon
Anlam	Korelasyon, iki değişken arasındaki ilişkiyi belirleyen istatistiksel bir ölçüdür.	Regresyon, bağımsız bir değişkenin, bağımlı değişkenle sayısal olarak nasıl ilişkili olduğunu açıklar.
Kullanım	İki değişken arasındaki doğrusal ilişkiyi göstermek	En iyi satıra sığdırmak ve bir değişkeni başka bir değişken temelinde tahmin etmek.
Bağımlı ve Bağımsız Değişken	Hangisinin bağımlı hangisinin bağımsız değişken olduğu fark etmez.	İki değişken de farklıdır.
Gösterge	Korelasyon katsayısı, iki değişkenin birlikte hareket etme derecesini gösterir.	Regresyon, yordayıcı değişkendeki (x) bir birim değişikliğinin yordanan değişken (y) üzerindeki etkisini gösterir.
Amaç	Değişkenler arasındaki ilişkiyi ifade eden sayısal bir değer bulmak.	Sabit değişkenli değerleri baz alarak rasgele değişkenin değerlerini tahmin etme

https://acikders.ankara.edu.tr/pluginfile.php/169660/mod_resource/content/0/9_REGRESYON.pdf

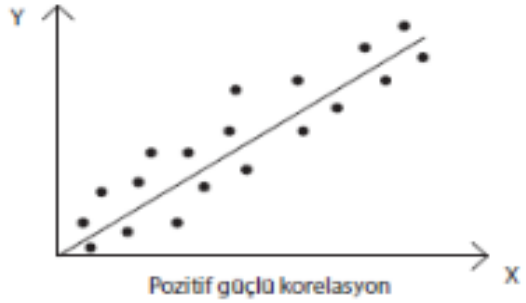
Korelasyon

- Korelasyon analizi iki değişkenin gözlem değerlerinin birlikte nasıl bir değişim içinde olduğunu araştırır. Dolayısıyla, değişkenlerin aldıkları değerlerin seyrine bakar.
- Bu sebeple, eğer serilerin gözlem değerlerinin seyri benzerlik gösteriyorsa, gerçekte ilişkisiz iki seri arasında da güçlü korelasyon çıkabilir. Bu durumu bir örnek ile açıklamaya çalışalım.

- İstanbul'da son 10 yıl içinde şehir hatları vapurlarının yolcu sayıları ile İstanbul'daki evlenme sayısı arasında bir ilişki olup olmadığını korelasyon analiziyle araştırdığımızı ve yüksek pozitif ilişki bulduğumuzu varsayalım. İki değişken arasında güçlü pozitif ilişki olduğunu gösteren korelasyon katsayısına bakarak, vapura binen insan sayısı arttıkça evlenme sayısı artıyor, gibi bir sonuca varırız ki bu tebessüme sebep olacak türden bir yorum anlamı taşır. Vapura binen insan sayısı ile yapılan evlilik sayısı arasında bir ilişki olduğunu düşünmemizi sağlayacak ne bir teori ne de bir mantıklı kurgu bulunmamaktadır.
- Nitekim, bu iki değişken arasında güçlü bir ilişki olduğu sonucunu doğuran asıl sebep, İstanbul'un nüfus artışıdır. Nüfus arttıkça daha çok insan vapura binmekte, nüfus arttıkça daha çok insan evlenmektedir.
- Dolayısıyla burada vapura binen insan sayısı ile evlenen insan sayısı üzerinde ortak bir nüfus etkisi bulunmaktadır ve bu etki göz ardı edildiği taktirde sanki bu iki değişken birbirini etkiliyormuş gibi bir sonuç ortaya çıkmaktadır.
- Şüphesiz bu tamamen yanlış bir değerlendirme ve yorum olacaktır.
- Dolayısıyla, iki değişken arasında mantıksal bir ilişki bulunması korelasyon analizi için vazgeçilmez bir gerekliliktir.

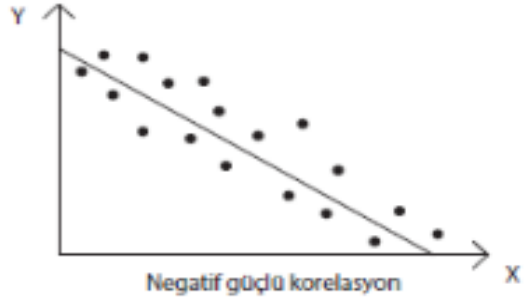
http://auzefkitap.istanbul.edu.tr/kitap/kok/istanaau216.pdf

Korelasyon – Scatter Plot

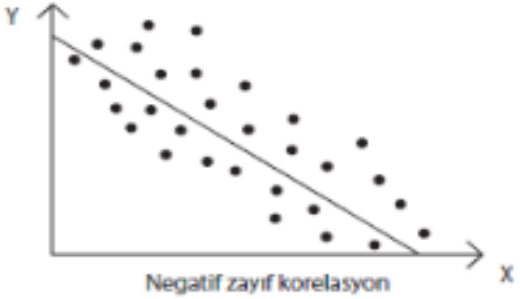


Pozitif güçlü korelasyon

Pozitif zayıf korelasyon



Negatif güçlü korelasyon

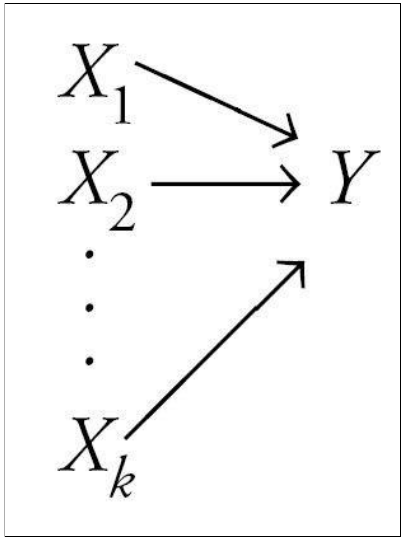


Negatif zayıf korelasyon

Doğrusal korelasyon yok
r=0

ÇOKLU REGRESYON

- Çoklu regresyon birden çok giriş ile tek bir cevap değişkeni değişkeni arasındaki ilişkiye odaklanır.



Amaç; bir taraftan değişkenlerin toplu etkisini görmek diğer taraftan her birinin bireysel katkıları görebilmektir.

İşlem?

- Basit doğrusal regresyonda olduğu gibi model seçiminde en küçük kareler yöntemini kullanır.
- Gerçek çizgi ile tahmin çizgisi arasındaki kareler toplamını minimizi eden farkları bulmaya çalışır.
- Aşağıdaki denklem, hataların karesi toplamını minimize eden fonksiyonu sunar.
- $y_{pred} = a + b_1x_1 + b_2x_2 \ldots + b_nx_n$

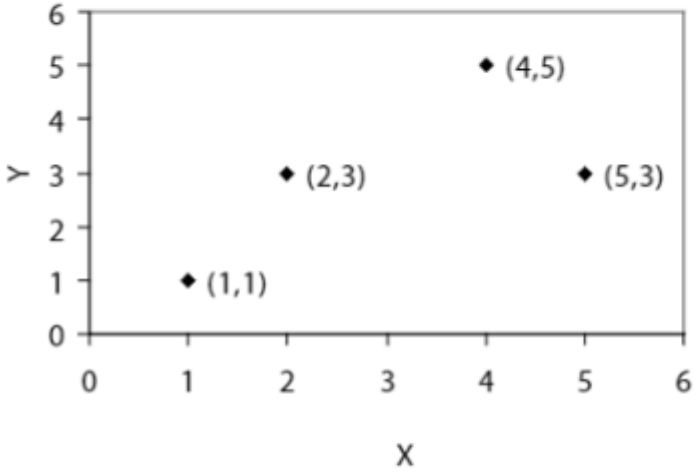
- Çoklu regresyon bağımlı ve bağımsız değişkenler arasındaki en iyi eşleşmeyi sunan bir model üretir.

$$y_{pred} = a + b_1x_1 + b_2x_2 \ldots + b_nx_n$$

- Çoklu regresyon, varsayılan tahmincilerin bazısının diğer bazılarından daha önemli olduğunu varsayar.
- Ayrıca, çoklu regresyon bağımlı değişkendeki değişim için bir araya getirilmiş değişkenlerin katkısını değerlendirir.

ÖRNEK I: Bir grup öğrencinin çalışma saati ve başarı puanı değerleri verilmiştir.

X:	2	4	1	5
Y:	3	5	1	3



- Çalışma saati (X) bağımsız değişken ve başarı puanı (Y) olarak alındığında regresyon eşitliğini hesaplayalım.

(Büyüköztürk ve diğerleri, 2018)

http://www.alcula.com/calculators/statistics/linear-regression

https://acikders.ankara.edu.tr/pluginfile.php/169660/mod_resource/content/0/9_REGRESYON.pdf

Örnek 2

- Bir kasabanın son yedi yılına ait veriler aşağıdaki gibidir. Buna göre bundan 10 yıl sonra nüfusun kaç olması beklenebilir.

yıllar	nüfus
1	3000
2	3110
3	3120
4	3000
5	2900
6	2750
7	2500