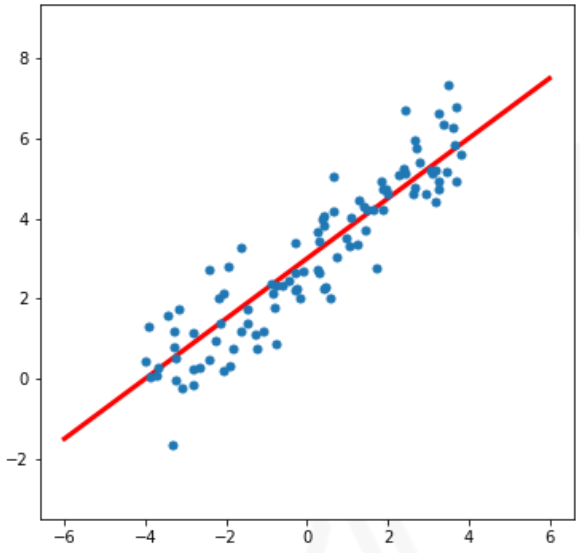
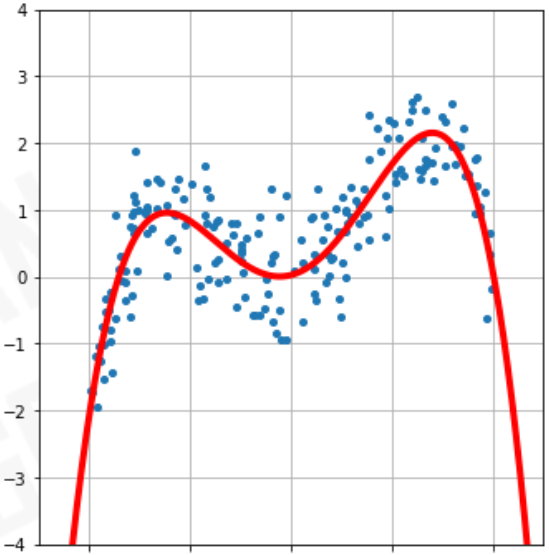
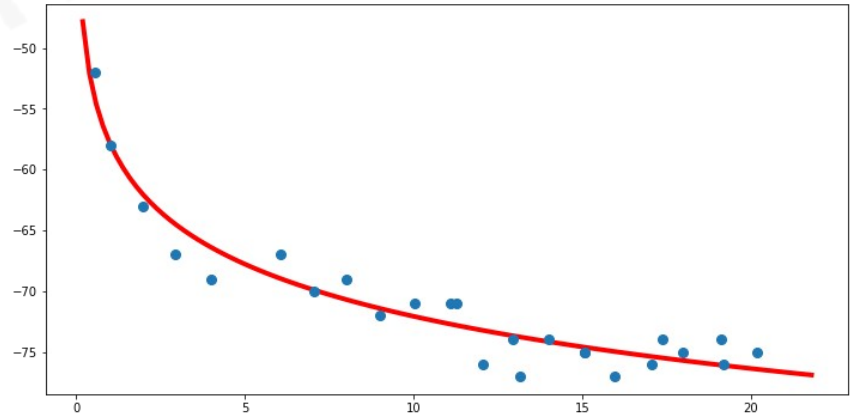
Regresyon
Bölüm I

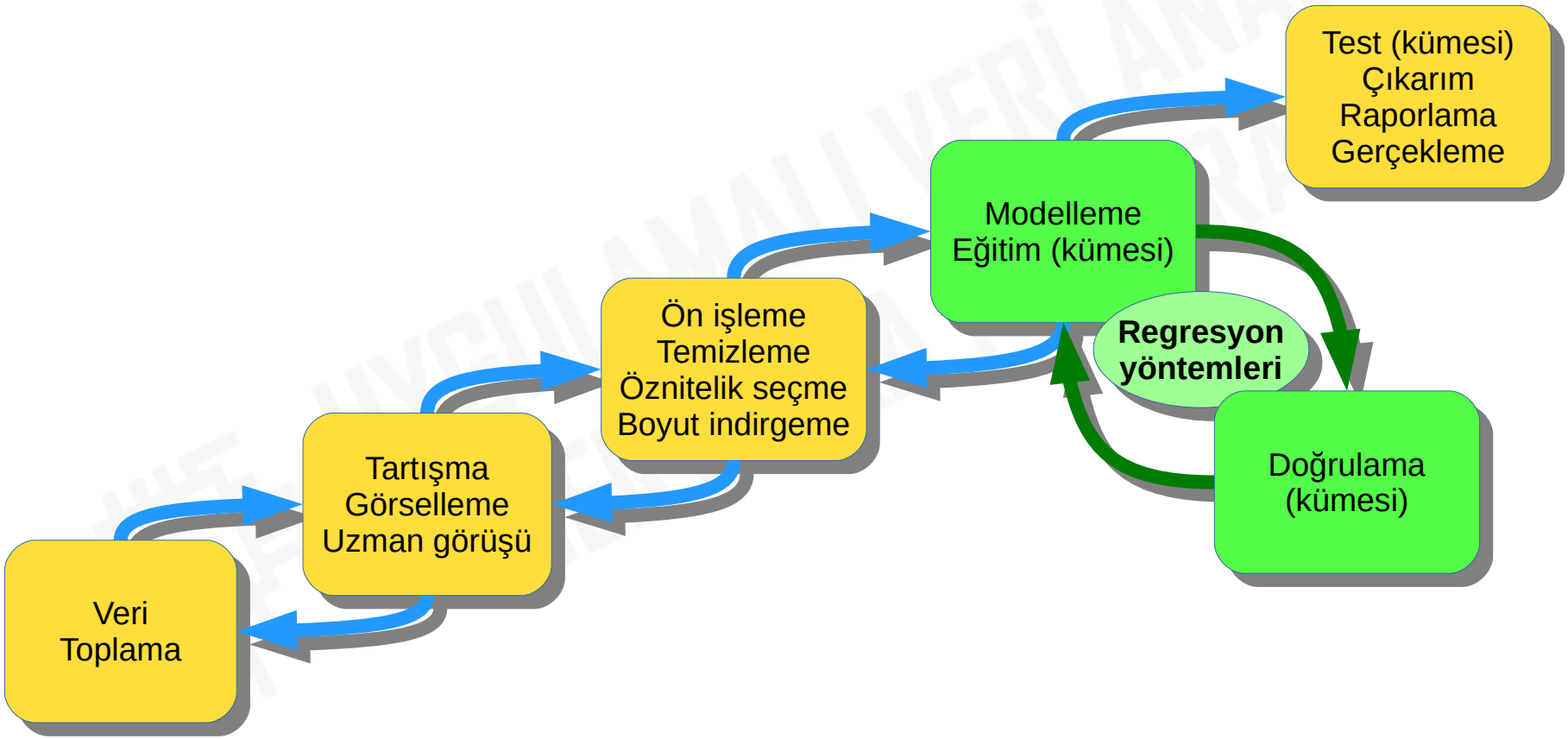
Serhan Daniş



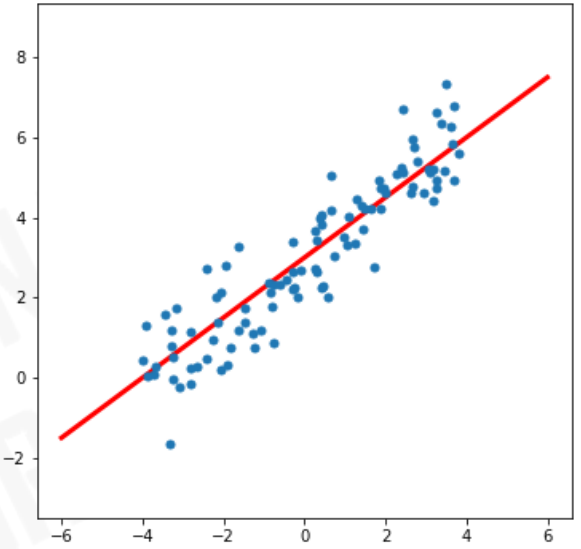




Yapay Öğrenme



Bölüm I

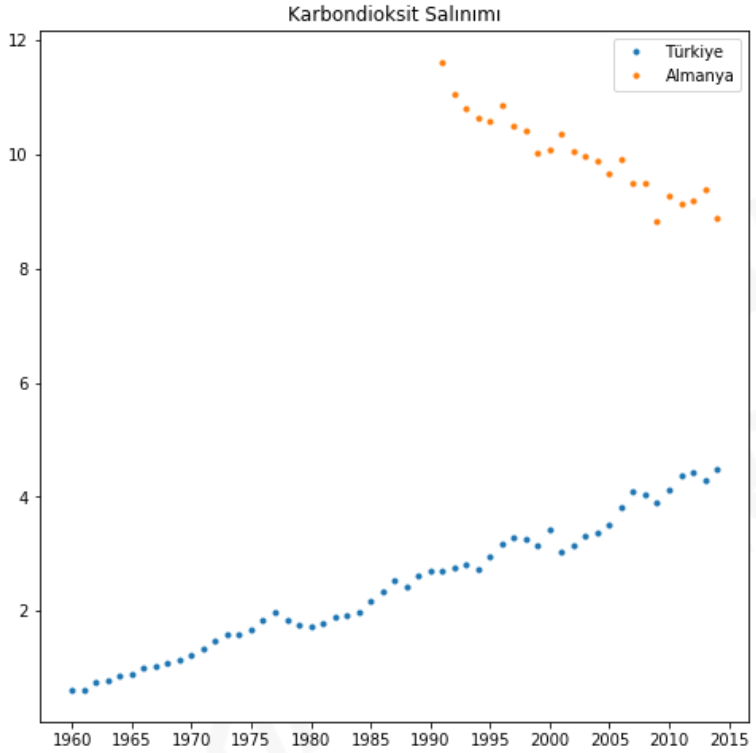


Doğrusal Regresyon

jupyter logosu yansının jupyter-notebook karşılığı olduğunu göstermektedir.

Örnek Senaryo: CO2 Salınımı

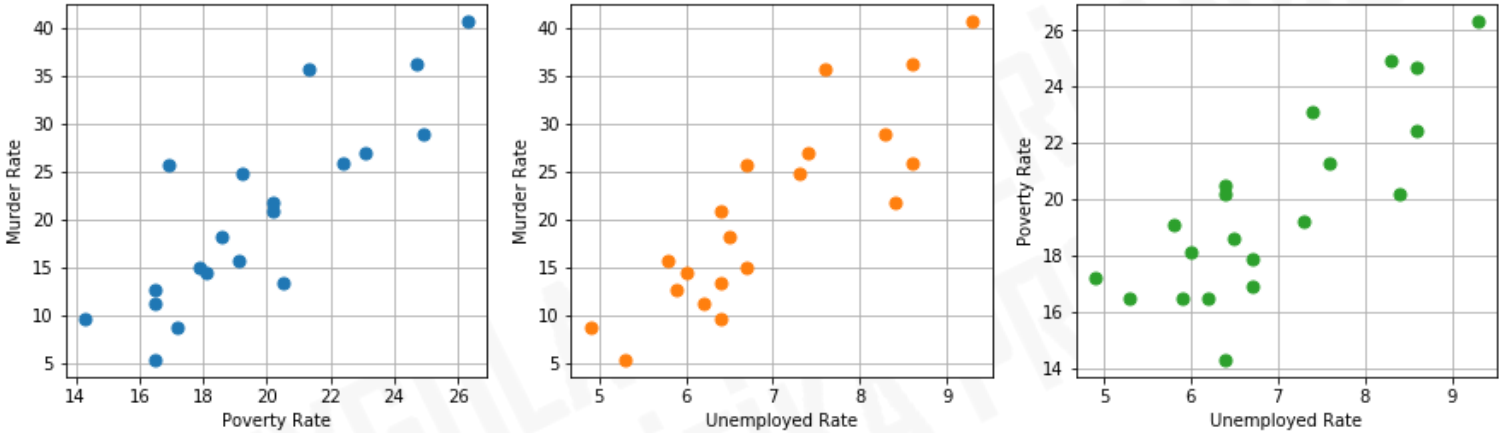
Zamana bağlı veriler:



- Tahminleme:
  - 2022'de Almanya'ın karbondioksit salınımı ne olacak?
- İleri analiz:
  - Türkiye ile Almanya'nın karbondioksit salınımları ne zaman eşitlenir?

Örnek Senaryo: Ekonomi ve Asayiş

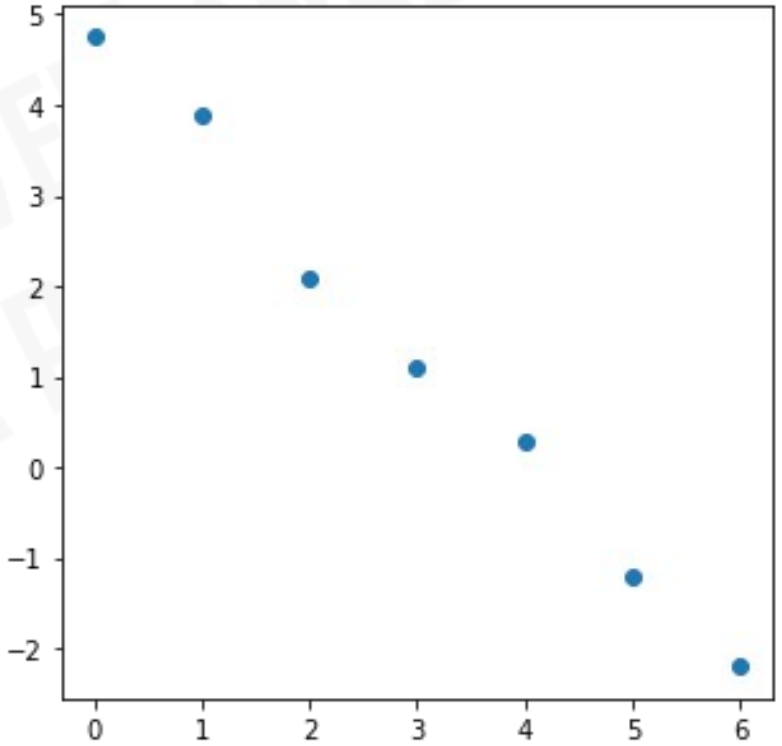
Zamandan bağımsız veriler:



- Modelleme: Cinayet oranlarıyla fakirlik veya işsizlik oranı arasında bir ilişki kurulabilir mi?
- Tahmin: Bir yerdeki halkın %20'si 5000 dolar altında kazanıyorsa, buradaki cinayet oranı nasıl tahminlenir?
- Modelleme: Fakirlikle işsizlik arasında bir bağlantı kurulabilir mi?

Doğrusal Regresyon (Linear regression)

- Regresyon bir **denetimli öğrenme (supervised learning)** yöntemidir.
- Regresyonda amaç **iki farklı veri arasındaki ilişkiyi (model)** tanımlayabilmektir.
- Doğrusal regresyonda ise iki (veya daha çok) veri arasındaki ilişkinin bir **doğru (veya düzlem) ile ifade** edilebileceği farzedilir.
- Sonucunda,
  - Verilerin ilişkisi hakkında **yorum** yapabilir.
  - Hem de doğru bir model kurulabilirse, **görülmeyen veriler** hakkında yorum ve tahmin yapabiliriz.

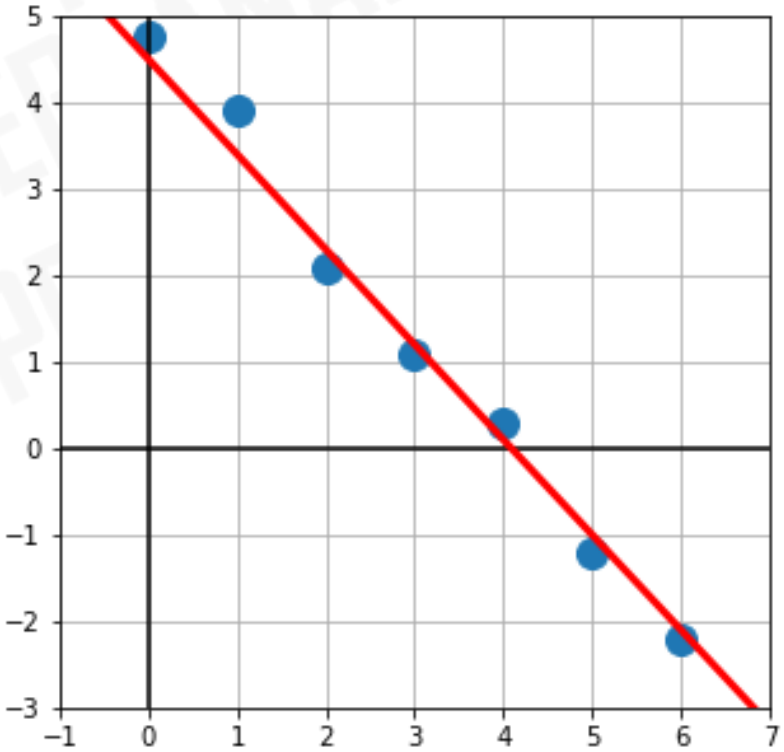


Doğru nedir?

- Çözülmesi gereken problem ise bu **doğrunun denklemini bulmak**:
- Doğru denklemi:

$$y = b_1 x + b_0$$

- x : Girdi verileri (bağımsız)
- y : Çıktı verileri (bağımlı)
- b_1 : Doğru eğimi
- b_0 : Kesim noktası



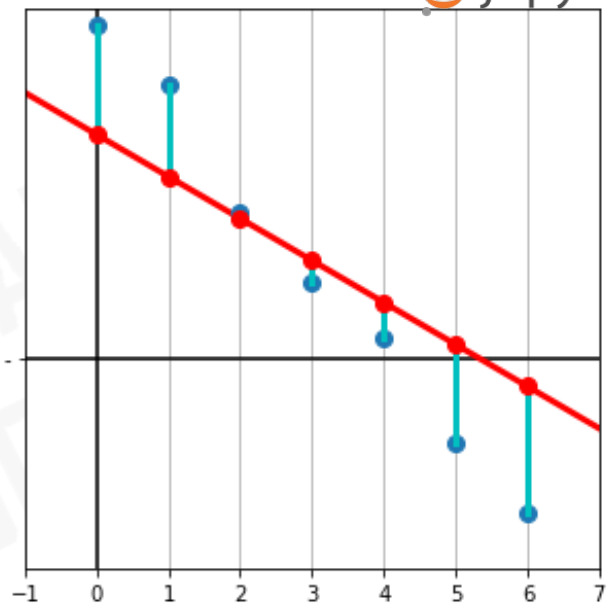
Modelleme:
Doğrusal Regresyon Modeli

- Alternatif gösterim: Her bir veri noktası bir indeks ile gösterilir.

$$y_i = b_1 x_i + b_0 + \epsilon_i$$

- Elimizde verilerin olduğunu farzediyoruz ve **genel bir doğru denklemini arıyoruz**. Daha doğrusu b_1 ve b_0 katsayılarını arıyoruz. ϵ_i, her veri noktası üzerindeki olası **hata**yı ifade eder.
- Elimizde bu katsayılar varsa, bu durumda da **gelecek veya bilinmeyen** çıktı değerlerini **tahmin**leyebiliriz. Bu tahminler genellikle şapka (^) ile gösterilir. $\hat{y}_i = b_1 x_i + b_0$

Modeli ölçme



- Tahmin edilen bir veri ile gerçek veri arasındaki hatayı temel alırız:

$$\epsilon_i = \hat{y}_i - y_i$$

- Bu hatayı bütün verileri kullanarak ölçeriz. Sonuç olarak kurduğumuz modelin **bütün verilere** uygun olup olmadığını ölçmemiz gerekiyor.

- Temelde iki tane hata ölçüm yöntemi vardır:
  - **Ortalama kare hata kökü** (Root Mean Squared Error):

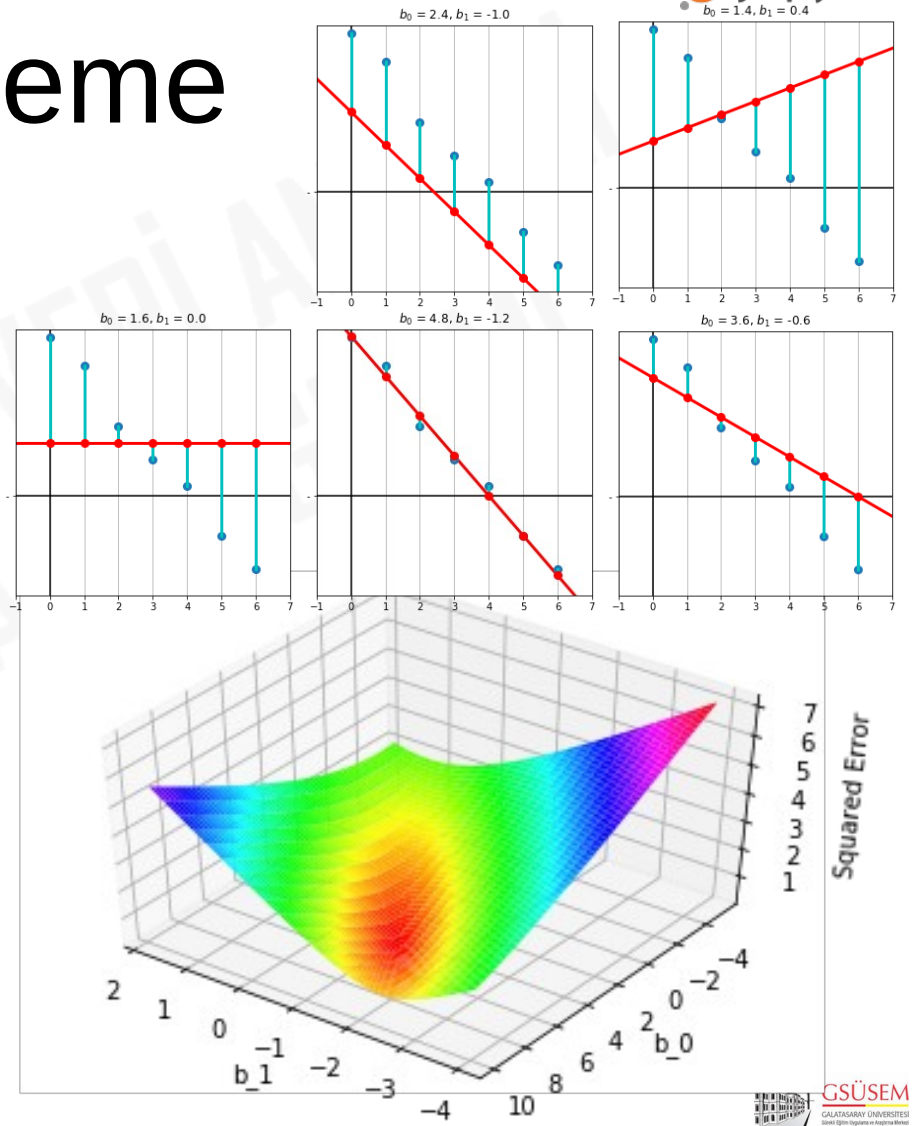
$$RMSE = \sqrt{\frac{1}{N}\sum_{i=0}^{N}(\hat{y}_i - y_i)^2}$$

  - **Ortalama mutlak hata** (Mean Absolute Error):

$$MAE = \frac{1}{N}\sum_{i=0}^{N}|\hat{y}_i - y_i|$$

Hata inceleme

- RMSE hata fonksiyonunu detaylıca inceyelim. Bunun için b_1 ve b_0 katsayılarının farklı değerleri için birer doğru denklemi oluşturuyoruz.
- Bilinen x_i değerlerine göre $\hat{y}_i$ tahminlerini yapıp, bunların da bilinen y_i değerleri ile arasındaki hatayı RMSE ile ölçelim.



En iyileme (Optimization)

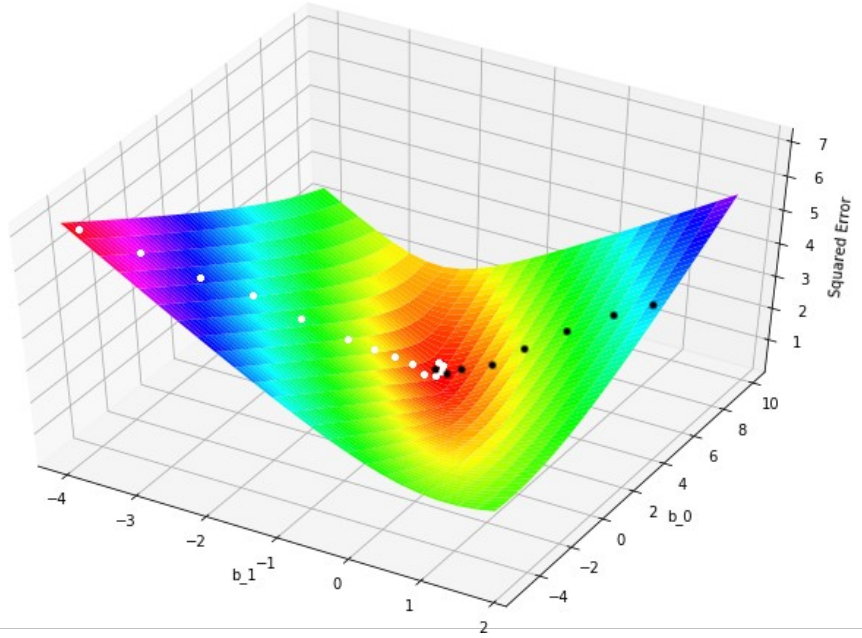
*Doğru katsayıları nasıl bulacağız veya **NASIL EĞİTECEĞİZ**?*

Bayır İnişi (Gradient descent)

- Yinelemeli (iterative)
- Bir katsayı değerinden başlayıp, hatayı azaltacak yönde ilerleriz.
- Hata değişmemeye başladığı noktadaki katsayılar hatayı en aza indirecek katsayılar olacaktır.

En Küçük Kareler Yöntemi (Least Squares Method)

- Doğrudan yöntem
- Hata fonksiyonunun yapısı gereği tek bir çukur nokta vardır. Hata fonksiyonunun türevi alınıp sıfıra eşitlendiğinde elde edilen parametreler en iyi parametrelerdir



Bu çalışmada yöntem olarak En Küçük Kareler Yöntemini kullanacağız.

En Küçük Kareler Yöntemi ve `polyfit()`

- Temelde hata fonksiyonunun her katsayıya göre parçalı **türevinin alınıp sıfıra eşitlenmesi** ile katsayıların en uygun değerleri bulunuyor.

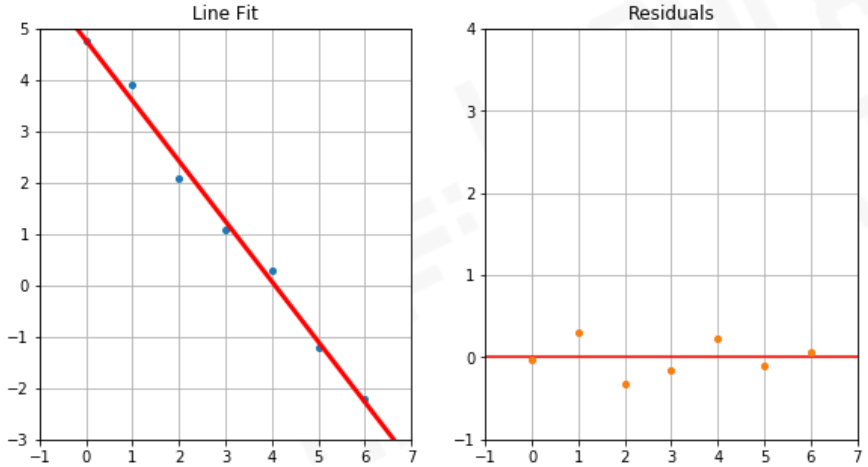
$$\boldsymbol{b} = \left(X^T X \right)^{-1} X^T \boldsymbol{y}$$

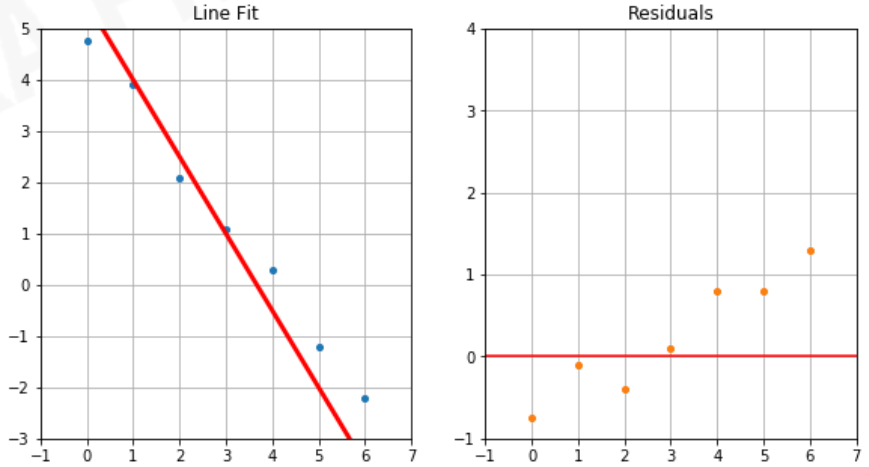
- Yukarıdaki karmaşık problemin çözümü yandaki **matris operasyonları**na indirgeniyor.
- X matrisi ve $\boldsymbol{y}$ vektörü düzgün olarak kurulduğunda ve gerekli matematiksel operasyonlar uygulandığında $\boldsymbol{b}$ vektörü elde edilecektir.
- Ama bunu da kullanmayacağız, bunun yerine `polyfit()` fonksiyonunu kullanacağız.
- $\boldsymbol{b}$ vektörünün elemanları, x_i ve y_i verilerinin bir doğru oluşturduğunu farzettiğimizde çizelecek **doğrunun katsayılarını** oluşturur.

$$\boldsymbol{b} = \begin{bmatrix} b_0 \\ b_1 \end{bmatrix}$$

Residual (hata) incelemesi

- Elde ettiğimiz doğru modelinin ne kadar gerçekçi olduğunu görmek için **veri noktalarının tahminleriyle gerçekleri** arasındaki “residual”lara bakabiliriz.
- Regresyon ile hatayı minimize etmemize rağmen **seçilen regresyon modelinin uygunluğu**na ayrıca karar vermek için başka incelemelerde bulunmalıyız.
- Residualları çizdirdiğimiz zaman x ekseninin etrafında ve her yerde eşit (**homojen**) olarak dağılmış **bir tüneli** andırması öncelikli istenen bir görseldir.





Residual Analizi: R-Kare

- Modelimizin genel **isabet oranı**nı ölçmek için R-Kare (R-Squared) yöntemi kullanılır. Aşağıdaki formülle hesaplanır:

$$R^2 = \frac{TSS - RSS}{TSS} \qquad TSS = \sum_{i=1}^{N} (y_i - \bar{y})^2 \qquad RSS = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$$

- *RSS,* residual kareleri toplamı*; TSS,* toplam kareler toplamı olmak üzere*, R-kare* 0 ile 1 arasında bir değer alır. Bu değer ne kadar **1'e yakınsa** regresyon modelimiz o kadar **isabetli**dir.
- Bu değere ayrıca "**tanımlanan çeşitliliğin oranı**" (fraction of the explained variance) da denir.
- Regresyonu değerlendirebilmek için hem **R-kare istatistiği**ne hem de **residual grafikleri**ne bakmak gerekir.

Örnekler ve Çalışma