

Lojistik Regresyon: Olasılıkları Modelleme Sanatı

İkili sonuçları anlamaktan, bilinçli kararlar almaya giden yolculuk.

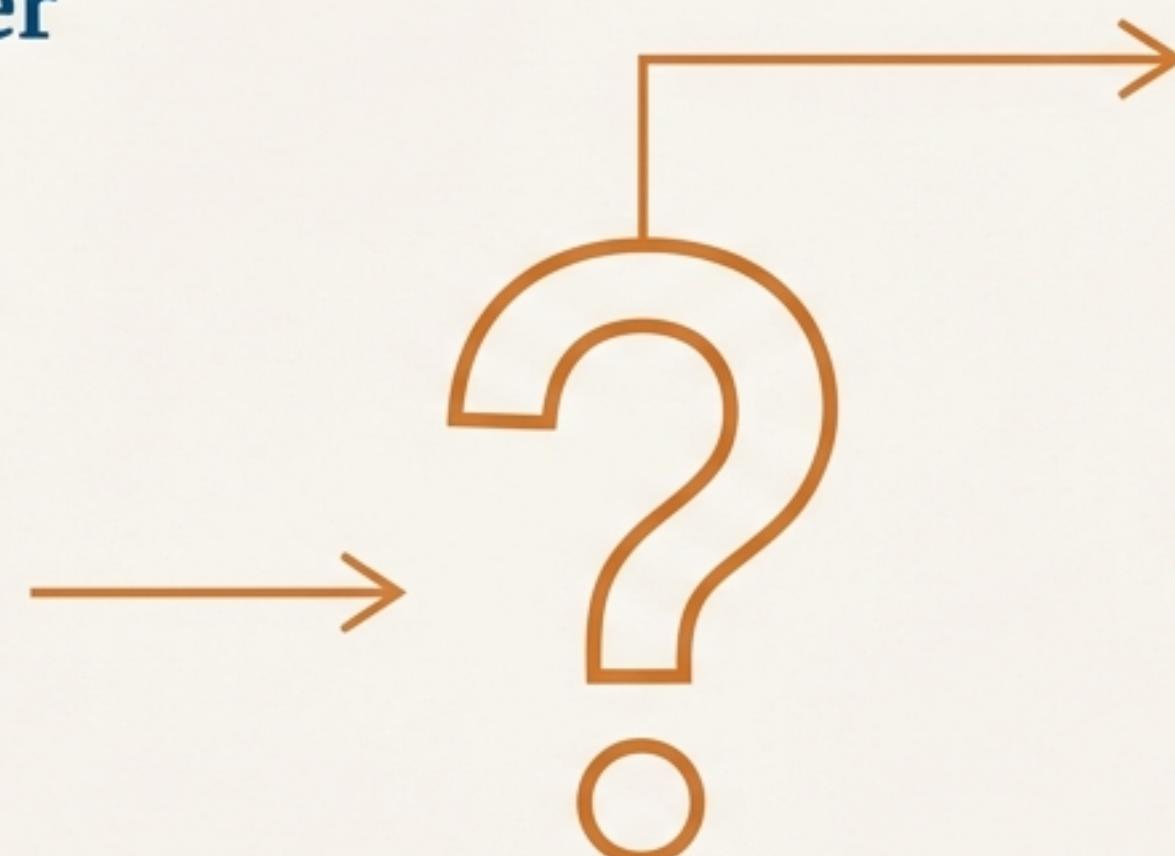


Meydan Okuma: Bir 'Evet/Hayır' Sorusunu Tahmin Etmek

Merkezi bir soru: "Yaşam tarzı özellikleri, koroner kalp hastalığı (KKH) için bir risk faktörü müdür?"

Bağımsız Değişkenler (Tahmin Ediciler)

-  Sigara kullanımı
-  Diyet
-  Egzersiz
-  Alkol tüketimi

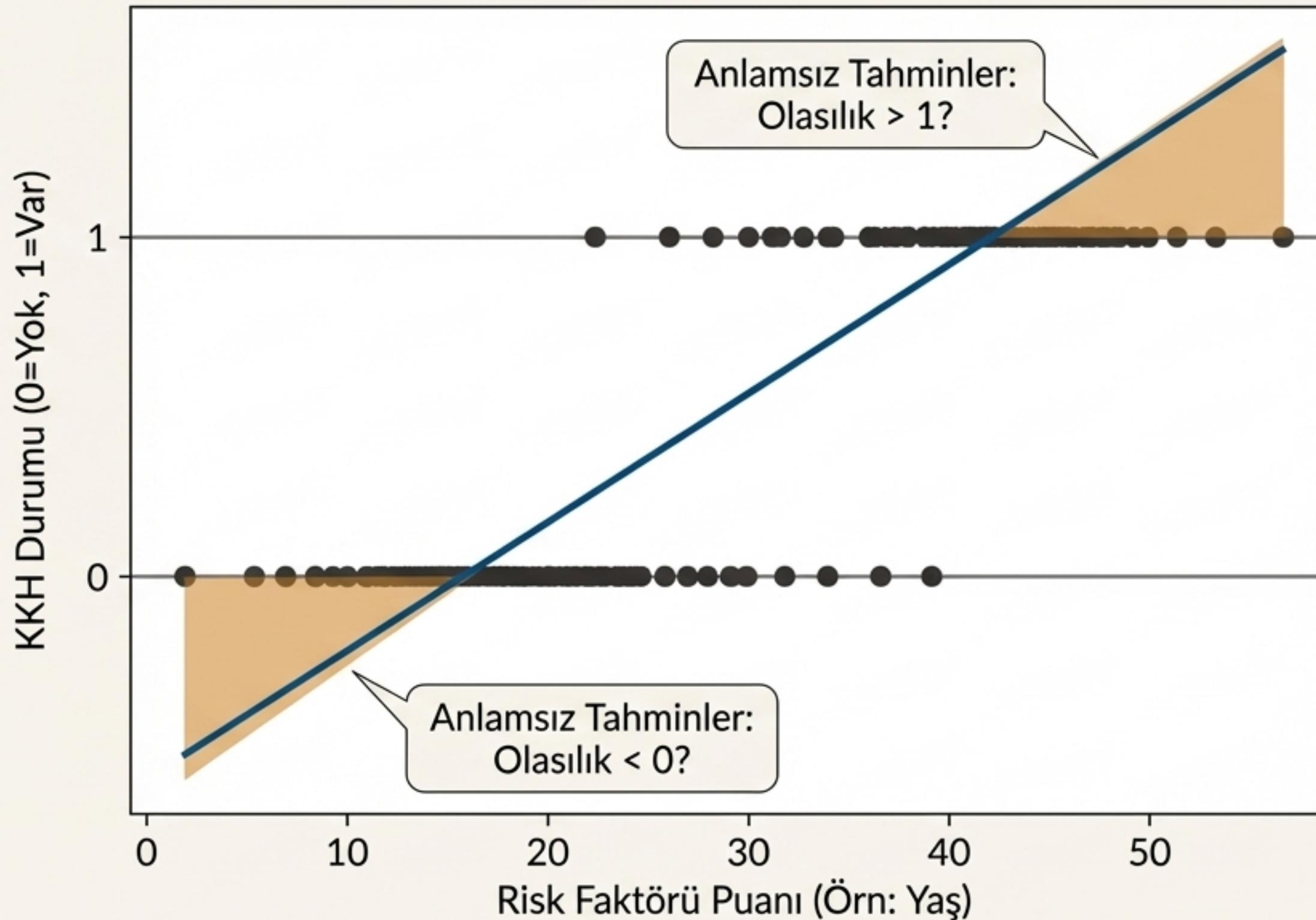


Bağımlı Değişken (Sonuç)



Bu değişkenler ile KKH'nin varlığı veya yokluğu arasındaki ilişkiyi nasıl güvenilir bir şekilde modelleyebiliriz?

İlk Deneme: Lineer Regresyon Neden Yetersiz Kalıyor?



Temel Başarısızlıklar

- **Anlamsız Tahminler**:** Model, olasılık olarak yorumlanamayacak tahminler üretir (örn. $P < 0$ veya $P > 1$). Bu mantıksal olarak imkansızdır.
- **Varsayımlı İhlali**:** Lineer Regresyon, hataların normal dağıldığını varsayar. İki bir değişkende iki bir değişkende bu varsayımlı geçerli değildir.

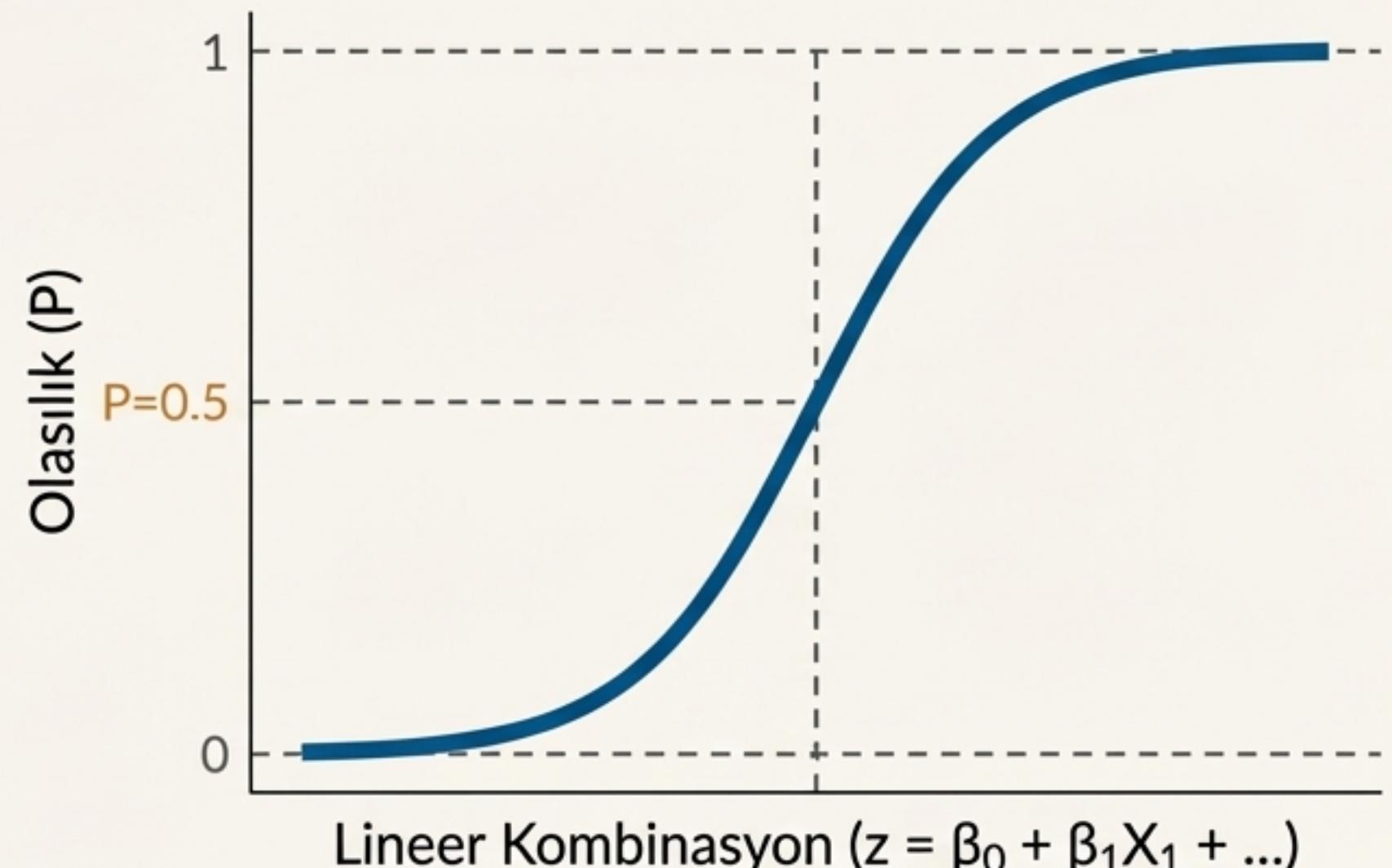
Kıvılcım: Sonucu Değil, Olasılığı Modellersek?

Yeni Fikir

Tahmin hedefimizi 0/1 sonucundan, sonucun 1 olma olasılığına (P) kaydıralım. Bu, tahminlerimizin her zaman mantıksal $[0, 1]$ aralığında kalmasını sağlar.

İhtiyaç

Tahmin edicilerimizin lineer birleşimini (z), yani $(-\infty$ ile $+\infty$ arasında herhangi bir değer alabilen bir sayıyı) alıp, onu $[0, 1]$ aralığına “sıkın” zarif bir fonksiyona ihtiyacımız var.

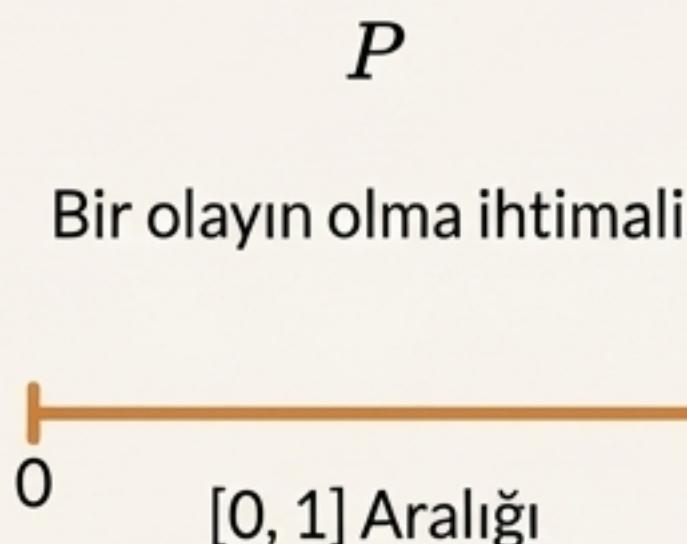


$$P = \frac{1}{1 + e^{-z}}$$

Modeli İnşa Etmek: Doğrusallığa Zekice Bir Geri Dönüş

Hedef: Olasılık (P) ile tahmin edicilerimiz (X) arasında doğrusal bir ilişki kurmak.

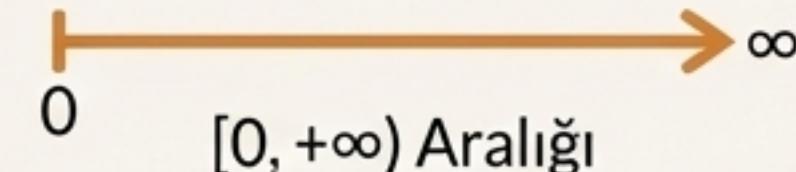
Başlangıç Noktası:
Olasılık



Adım 1: Odds'a Geçiş

$$\text{Odds} = \frac{P}{1 - P}$$

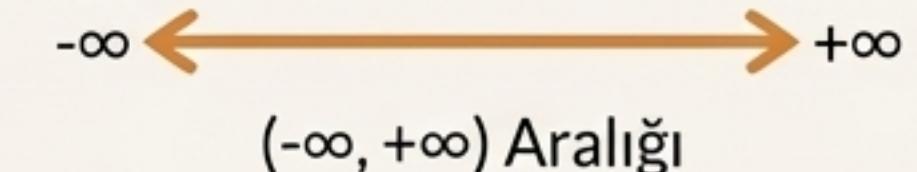
Bir olayın olma olasılığının,
olmama olasılığına oranı.



Adım 2: Log-Odds'a
(Logit) Geçiş

$$\text{Logit}(P) = \ln(\text{Odds})$$

Odds'un doğal logaritması.



Nihai Model: Lojistik Regresyon Denklemi

$$\ln\left(\frac{P}{1 - P}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots$$

Sol taraf (Logit), sağ taraftaki lineer denklemle artık doğrudan ilişkilidir.

En İyiyi Bulmak: Maksimum Olabilirlik Kestirimi

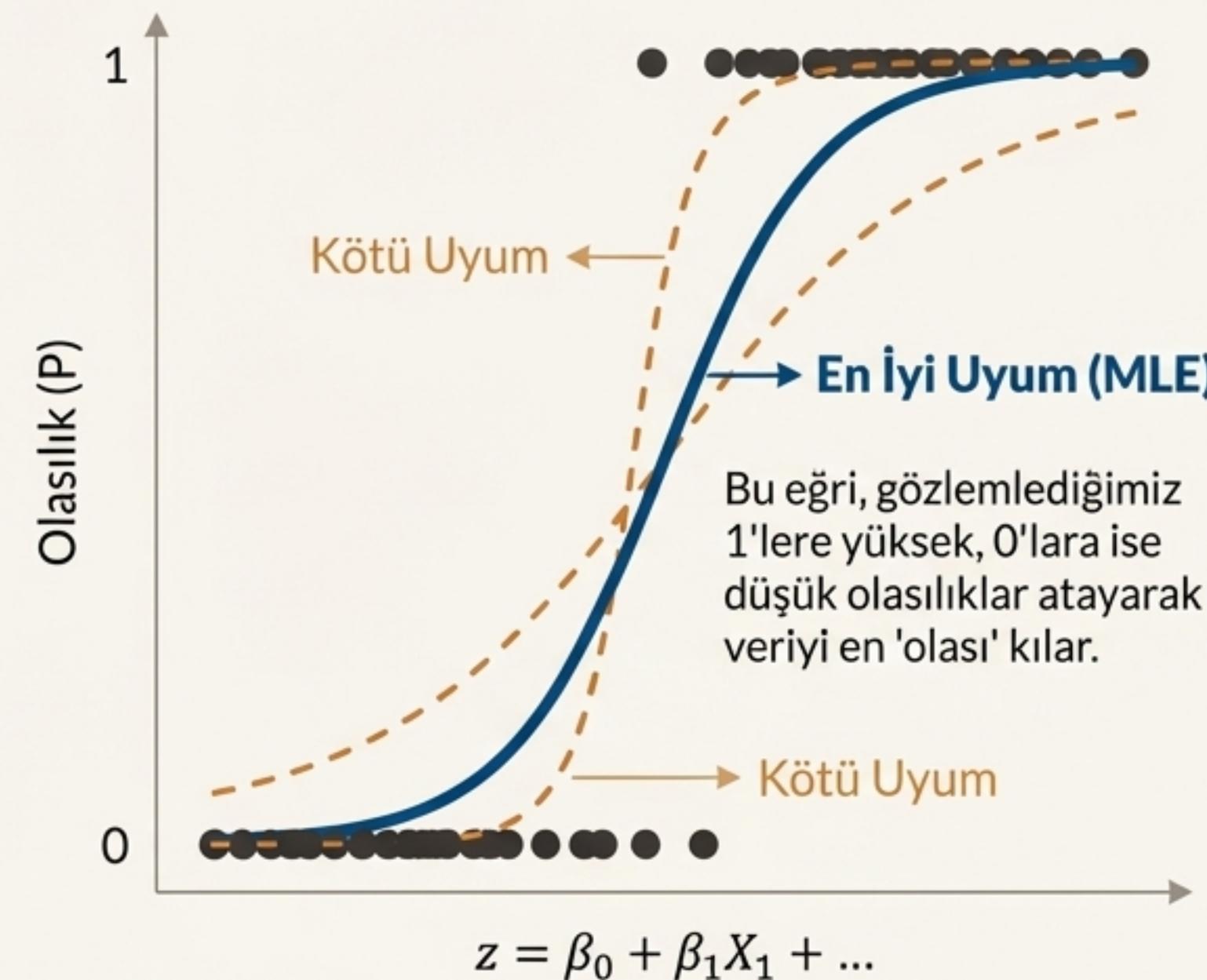
Modelimiz için en uygun β_0, β_1, \dots katsayılarını nasıl belirleriz?

Lineer Regresyon

Hataların kareleri toplamını minimize eder (En Küçük Kareler Yöntemi).

Lojistik Regresyon

İkili veriler için farklı bir mantık kullanır:
Gözlemlenen veriyi en olası kıyan modeli arar.



Maksimum Olabilirlik Kestirimi (MLE)

Sezgisel olarak, "Veri setimizdeki gerçek sonuçları (gözlemlediğimiz 1'ler ve 0'lar dizisini) ortaya çıkarma olasılığını en üst düzeye çıkarılan model katsayılarını (β 'ları) bulma" prensibidir.

Motor Kaputunun Altı: Optimizasyon Algoritmaları

Pratik Uygulama

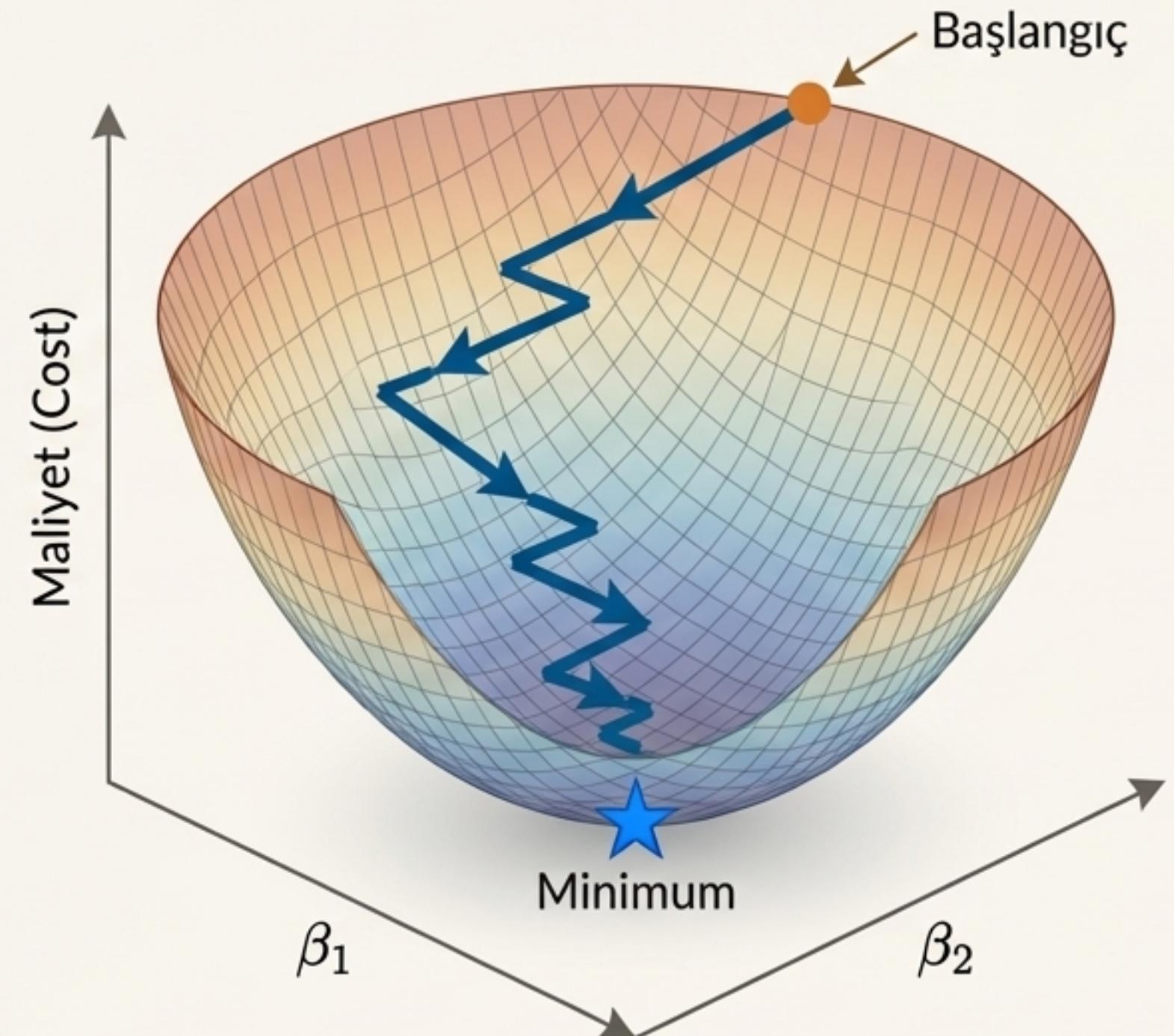
Maksimum Olabilirlik Kestirimi bir hedef tanımlar. Bu hedefe ulaşmak için optimizasyon algoritmaları kullanılır.

Maliyet Fonksiyonu (Log Kaybı)

Olasılığı maksimize etmek, genellikle **Negatif Log-Olasılık** adı verilen bir maliyet fonksiyonunu minimize etmeye denktir. Modelin tahminleri gerçek değerlerden ne kadar uzaksa, bu maliyet o kadar artar.

Çözüm Yöntemi: Gradyan İnişi

Maliyet fonksiyonunun en düşük noktasına ulaşmak için kullanılan iteratif bir yöntemdir. Her adımda, maliyeti en hızlı azaltan yönde küçük bir adım atılır.



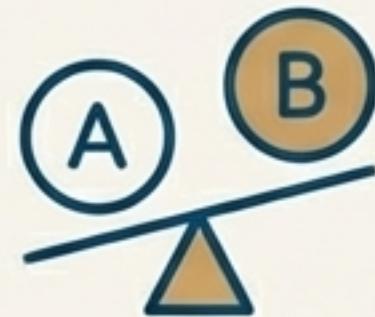
Anlamı Keşfetmek: Üç Yorumlama Merceği

Lojistik regresyon katsayıları (β), doğrusal regresyondakinden farklı olarak birden çok şekilde yorumlanabilir. Her yorum, farklı bir içgörü seviyesi sunar ve farklı bir kitleye hitap eder.

$$f(x) \rightarrow \beta_1$$

1. Matematikçinin Merceği: Log-Odds

Doğrusal ve doğrudan, ama sezgisel değil.



2. İstatistikçinin Merceği: Odds Oranları ($\text{Exp}(\beta)$)

Grupları karşılaştırmak için en güçlü ve en yaygın kullanılan yöntem.



3. Paydaşın Merceği: Olasılıklar

En sezgisel, iş kararlarına en yakın, ancak doğrusal olmayan yorum.

Mercek 1: Log-Odds (Doğrusal ama Sezgisel Değil)

Temel Denklem

$$\ln\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 X_1 + \dots$$

Yorumlama Kuralı

“Bir X değişkenindeki bir birimlik artış, diğer değişkenler sabitken, olayın gerçekleşme **log-odds**'unu β katsayısı kadar değiştirir.”

Örnek (Sigara İçme Modeli)

Eğitimdeki bir yıllık artış, sigara içme log-odds'unu **-0.183** birim azaltır.

Değerlendirme

- ⊕ İlişki doğrusal ve katkısaldır.
- ⊖ ‘Log-odds’ birimi, çoğu paydaş için sezgisel bir anlam taşımaz ve pratik kararlara dönüştürülmesi zordur.

Mercek 2: Odds Oranları (Karşılaştırmanın Gücü)

Dönüşüm

Odds Oranı = $\exp(\beta)$

Yorumlama Kuralı “Bir X değişkenindeki bir birimlik artış, olayın gerçekleşme **odds’unu $\exp(\beta)$ katına çıkarır** (veya $\exp(\beta)$ ile çarpar).”

Örnek (Sigara İçme Modeli)

Cinsiyet (Erkek=1) katsayısı (β): **0.254**

Calculation: Odds Oranı: $\exp(0.254) \approx \textbf{1.289}$

Yorum: “Erkeklerin sigara içme odds’u, kadınların odds’undan **1.289 kat daha fazladır.**”

Alternatif Yorum: “Erkeklerin sigara içme odds’u, kadınlara göre **%28.9 daha yüksektir.**”



$\exp(B) < 1$: Azalan risk/olasılık

$\exp(B) = 1$: Etki yok

$\exp(B) > 1$: Artan risk/olasılık

Mercek 3: Olasılıklar (En Sezgisel Ama En Karmaşık Yorum)

Temel Gerçek: Etki Sabit Değildir

Bir değişkenin olasılık üzerindeki etkisi sabit değildir.

Doğrusal ve katkışal değildir. Bu etki, diğer tüm değişkenlerin mevcut değerlerine bağlı olarak değişir.



Çözüm: Marjinal Etkiler (Marginal Effects)

Bir bağımsız değişkendeki bir birimlik değişimin, diğer değişkenler belirli değerlerde sabit tutulduğunda, tahmin edilen olasılık üzerindeki etkisidir.

Ortalamadaki Marjinal Etki (MEM)

“Ortalama” bir birey için etkiyi hesaplar (tüm değişkenler ortalama değerlerindeyken).

Ortalama Marjinal Etki (AME)

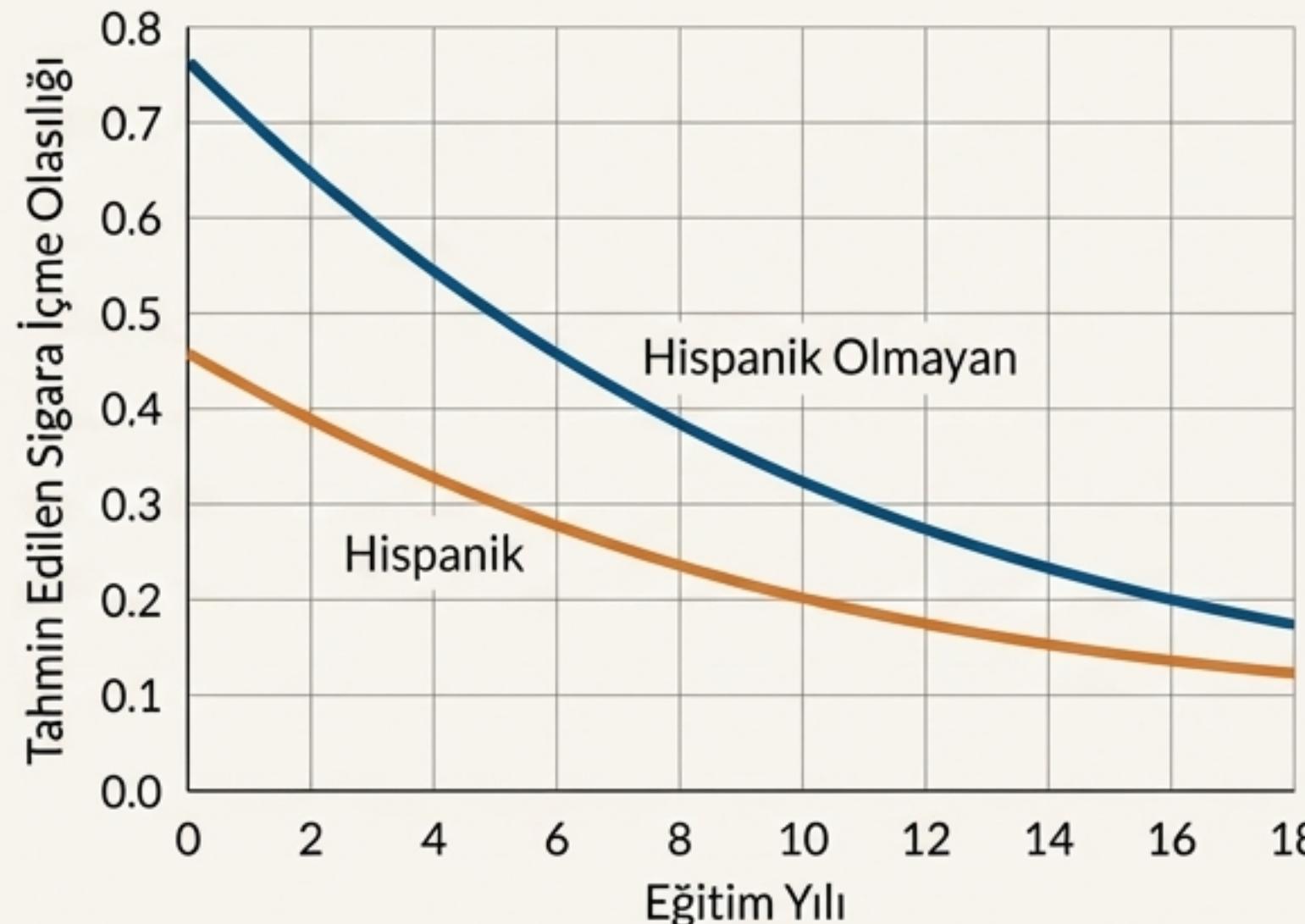
Her birey için etkiyi ayrı ayrı hesaplar ve sonra bu etkilerin ortalamasını alır. Genellikle **daha sağlam ve tercih edilen** bir yöntemdir.

Örnek (AME): “Ortalama olarak, eğitimdeki fazladan bir yıl, sigara içme olasılığını **0.023 puan** (veya **%2.3**) düşürür.”

Olasılıkları Görselleştirmek: Tek Bir Rakam Yeterli Olmadığında

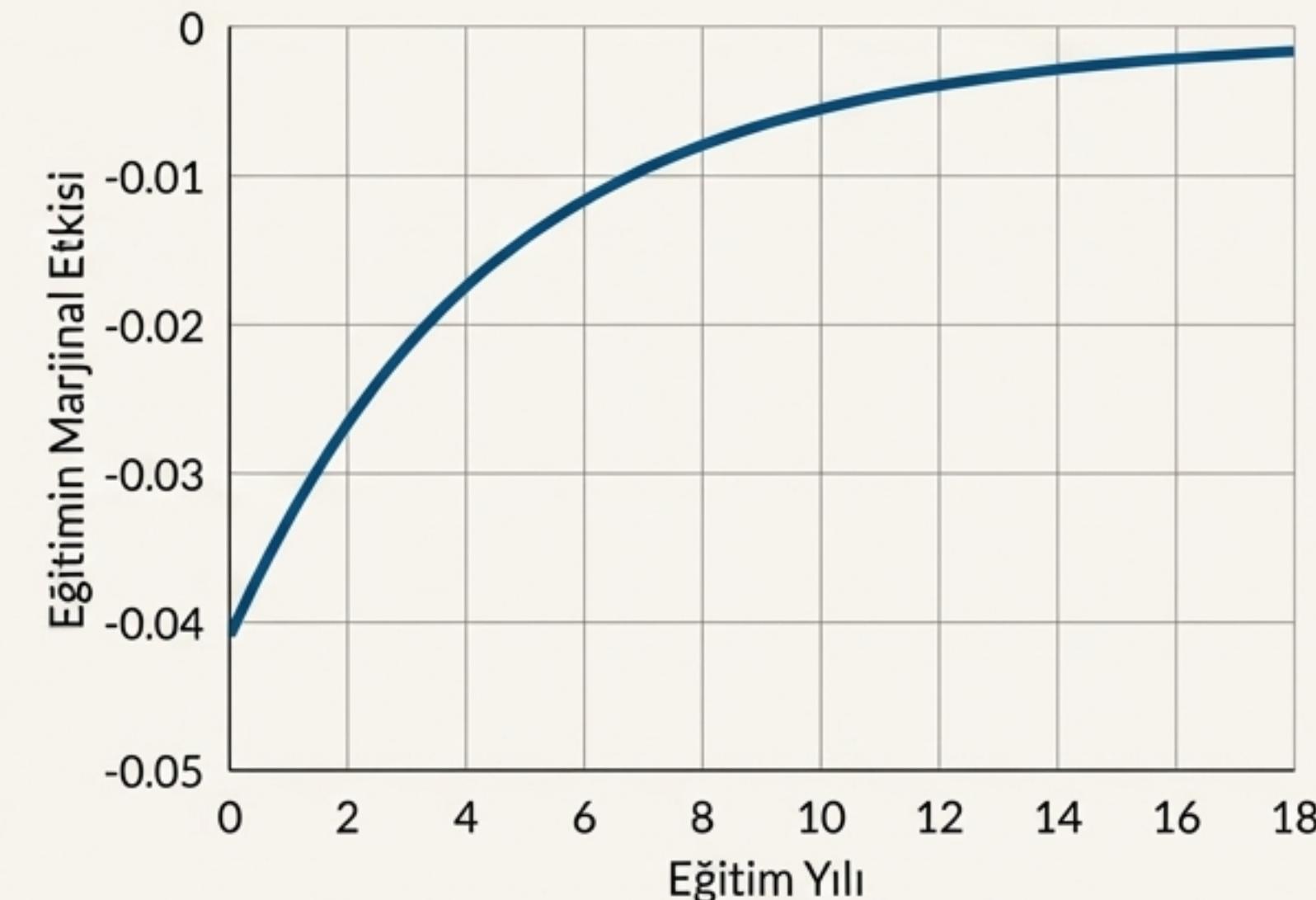
Grafik 1: Tahmin Edilen Olasılıklar (Etkileşim Örneği)

Bu grafik, Hispanik olan ve olmayan bireyler için eğitim seviyesine göre sigara içme olasılığını gösterir. İki grubun yolları doğrusal değildir ve aralarındaki fark eğitim arttıkça daralır.



Grafik 2: Marjinal Etkiler (Doğrusal Olmayan Etki)

Bu grafik, eğitimin marjinal etkisinin (olasılık üzerindeki bir yıllık etkinin) eğitim seviyesine göre nasıl değiştiğini gösterir. Etki, düşük eğitim seviyelerinde en güçlüyken, eğitim arttıkça zayıflamaktadır.



Model Ne Kadar İyi? Performans Değerlendirme Metrikleri

Metrikler

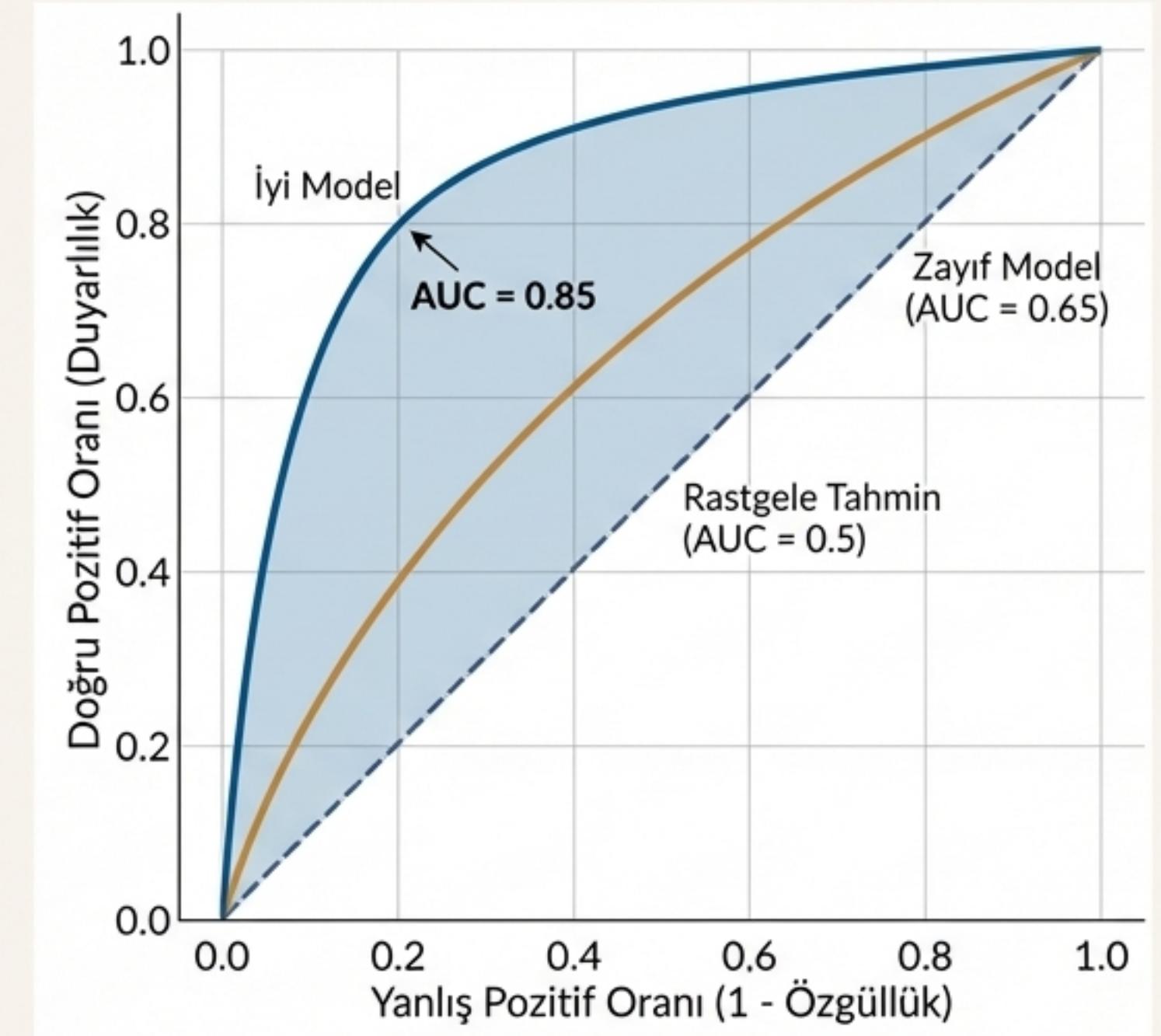
1. Genel Model Uyumu (Goodness-of-Fit)

- **-2 Log-Likelihood:** Model tarafından açıklanamayan varyansın ölçüsü. Düşük değer daha iyi uyum demektir.
- **Hosmer-Lemeshow Testi:** Gözlemlenen ve beklenen frekansları karşılaştırır. Anlamlı olmayan bir p-değeri (> 0.05) iyi uyum gösterir.

2. Tahmin Gücü (Predictive Power)

- **Sınıflandırma Tablosu (Confusion Matrix):** Modelin doğru ve yanlış tahminlerini özetler (Doğruluk, Kesinlik, Duyarlılık).
- **AUC - ROC Eğrisi:** Modelin iki sınıfı ayırt etme yeteneğinin görsel ve sayısal ölçüsüdür. 1'e ne kadar yakınsa o kadar iyidir.

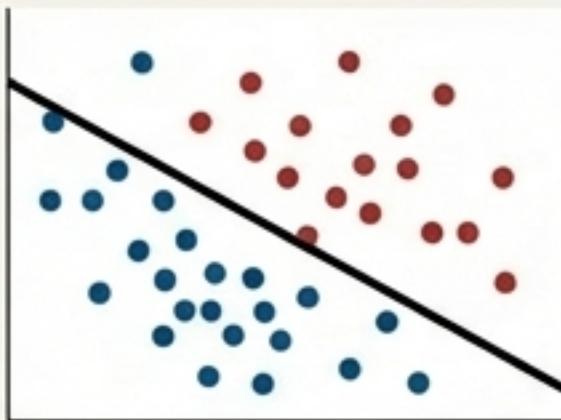
AUC - ROC Eğrisi



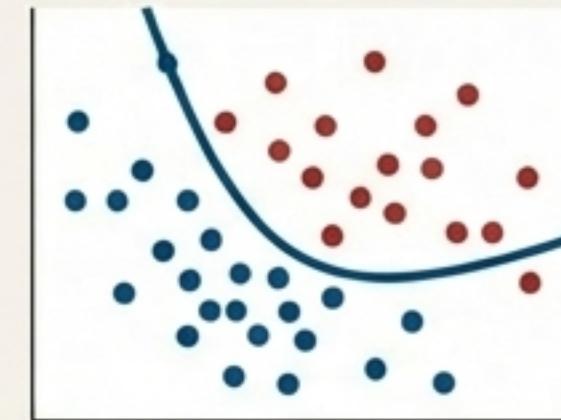
Modeli İyileştirmek: Aşırı Öğrenmeyi Düzenlileştirme ile Önlemek

Problem: Aşırı Öğrenme (Overfitting)

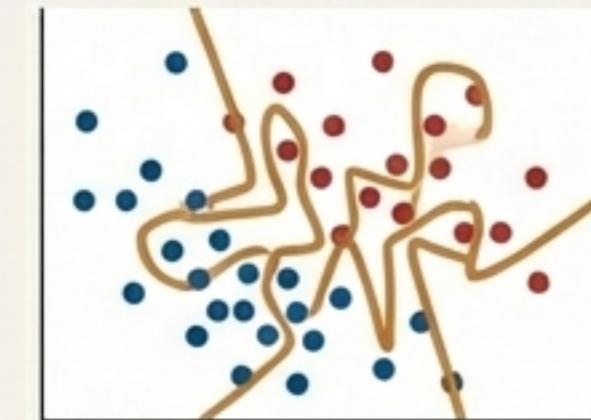
Yetersiz Öğrenme



İyi Uyum



Aşırı Öğrenme



Model, eğitim verisindeki gürültüyü “ezberler” ve yeni verilere genelleme yapamaz.

Çözüm: Düzenlileştirme (Regularization)

Modelin karmaşıklığını cezalandırmak için maliyet fonksiyonuna bir ceza terimi eklenir. Bu, katsayı değerlerini küçülterek modeli daha basit hale getirir.

L2 (Ridge)

Katsayıların karelerinin toplamını cezalandırır. Tüm katsayıları sıfıra yaklaşır ama genellikle tam sıfır yapmaz.

L1 (LASSO)

Katsayıların mutlak değerlerinin toplamını cezalandırır. Önemsiz değişkenlerin katsayılarını tam olarak sıfır indirgerek otomatik değişken seçimi yapabilir.

Lojistik Regresyon Yol Haritası: Problemden Ustalığa



Lojistik regresyon, yalnızca bir sınıflandırma algoritması değil, aynı zamanda olasılıkların arkasındaki '**neden'i anlamak ve anlatmak** için güçlü bir araçtır.